EVALUATION OF Teacher Preparation Programs

Purposes, Methods, and Policy Options

Michael J. Feuer, The George Washington University
Robert E. Floden, Michigan State University
Naomi Chudowsky, Caldera Research, and
Judie Ahn, National Academy of Education

National Academy of Education
Washington, DC

NATIONAL ACADEMY OF EDUCATION   500 Fifth Street, NW Washington, DC 20001

NOTICE: This report was approved by the National Academy of Education Board of Directors. This study was supported by a grant from the National Science Foundation, Award No. 1153848. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the National Academy of Education and the authors and do not necessarily reflect the views of the National Science Foundation.

International Standard Book Number   978-0-615-90081-0

Additional copies of this report are available from the National Academy of Education, 500 Fifth Street, NW, Washington, DC 20001; Internet, http://www.naeducation.org.

Suggested citation: Feuer, M. J., Floden, R. E., Chudowsky, N., and Ahn, J. (2013). *Evaluation of teacher preparation programs: Purposes, methods, and policy options.* Washington, DC: National Academy of Education.

NATIONAL
ACADEMY
of
EDUCATION

The **National Academy of Education** advances high quality education research and its use in policy formation and practice. Founded in 1965, the Academy consists of U.S. members and foreign associates who are elected on the basis of outstanding scholarship related to education. Since its establishment, the Academy has undertaken research studies that address pressing issues in education, which are typically conducted by members and other scholars with relevant expertise. In addition, the Academy sponsors professional development fellowship programs that contribute to the preparation of the next generation of scholars.

EVALUATION OF TEACHER EDUCATION PROGRAMS: TOWARD A FRAMEWORK FOR INNOVATION[1]

Steering Committee

Michael J. Feuer (*Chair*), Graduate School of Education and Human Development, The George Washington University
Deborah Loewenberg Ball, School of Education, University of Michigan
Jeanne M. Burns, Louisiana Board of Regents
Robert E. Floden, College of Education, Michigan State University
Susan H. Fuhrman (*Ex Officio*), Teachers College, Columbia University
Lionel C. Howard, Graduate School of Education and Human Development, The George Washington University
Brian Rowan, Institute for Social Research and School of Education, University of Michigan

Staff

Judie Ahn, Senior Program Officer
Naomi Chudowsky, Consultant
Nancy Kober, Editor
Gregory White, Executive Director

[1] It is noted that the title of this project: *Evaluation of Teacher Education Programs: Toward a Framework for Innovation*, as awarded by the National Science Foundation, differs from the title of this report.

Acknowledgments

Public concern for the effectiveness of teacher preparation programs has sparked renewed interest in the attributes of evaluation systems used to gauge their quality. There are many such systems currently in place, with different purposes and consequences, and a growing need to clarify their advantages and drawbacks as the basis for developing new and innovative approaches. That need is the principal impetus for this report, which addresses a number of questions: What are the relative strengths, limitations, and consequences of existing approaches to evaluating teacher preparation programs? How well do evaluation methods align with multiple intended uses of their results? What principles should guide the design, implementation, and interpretation of evaluation systems?

The work leading to this report was made possible by a generous grant from the National Science Foundation. We are especially grateful to Dr. Janice Earle, Senior Program Director in the Division of Research on Learning, for her interest and wise and careful guidance.

The National Academy of Education steering committee formed to undertake this project included Deborah Ball, Jeanne Burns, Robert Floden, Susan Fuhrman (ex-officio), Lionel Howard, and Brian Rowan. I thank them all for their willingness to take time from very busy schedules to help shape the study, contribute their knowledge and experience to our deliberations, and fine-tune key sections of the report. Gregory White, executive director of the National Academy of Education, and Maxine Freund, Associate Dean for Research and External Relations at the Graduate School of Education and Human Development at the George Washington University, provided much needed intellectual and logistical support.

I am indebted to the three principal co-authors for their superb contributions. Bob Floden drafted key sections and read every word of the whole report (several times), added subtlety and nuance to complex explanations, and never let us stray from our main mission. Naomi Chudowsky produced elegant and eloquent text and made sure our discussions were evidence-based and methodologically defensible. Judie Ahn took responsibility for managing and then integrating the work of our outside authors, organized our workshops and meetings, and steadfastly and efficiently kept the project on course. In addition, I want to say a special thanks to Nancy Kober, whose fine editorial hand is evident throughout.

The Committee commissioned four papers to enrich our deliberations with information regarding the special issues surrounding evaluation of future STEM educators; lessons from evaluation systems in other countries; and lessons from pre-service evaluation in the nursing profession. We are grateful to Suzanne Wilson; John Furlong; Maria Teresa Tatto, Joseph Krajcik, and James Pippin; and Jean Johnson and Christine Pintz, for their excellent papers and their interest and willingness to participate in the project.

The Committee convened two workshops in which policy makers, researchers, and evaluation practitioners contributed their expertise. We are grateful to Chad Aldeman, James Cibulka, Susan Fuhrman, Heather Harding, George Noell, Charles Peck, Robert Rickenbrode, Kate Walsh, Nancy Cartwright, Carl Cohn, James Kelly, Richard Shavelson, Lee Shulman, William Trent, Brenda Turnbull, Suzanne Wilson, Joe Conaty, and Patricia Tate. I thank Emerson Elliott for help with fact-checking, and Lorraine McDonnell and Lorrie Shepard for wise counsel early on.

Peer review is a linchpin of the National Academy of Education's process, designed to ensure that its reports are accurate, objective, and responsive to the study charge. I thank the reviewers for their careful reading of the draft and their thoughtful insights and suggestions: Christopher Edley, Rick Ginsberg, Daniel Goldhaber, Shirley Malcom, and Brenda Turnbull were selected for their diverse perspectives and technical expertise and served as anonymous, independent reviewers. They were not asked to endorse the contents of the report or review the final draft of the report, responsibility for which rests with the National Academy of Education. The review process was seamlessly overseen by Academy member Judith Warren Little, who was appointed to make sure that reviewers represented a range of relevant perspectives and expertise and that their comments were carefully considered.

Michael J. Feuer
Chair, Steering Committee

Contents

Appendixes

Summary

Teacher preparation programs (TPPs) are where prospective teachers gain a foundation of knowledge about pedagogy and subject matter, as well as early exposure to practical classroom experience. Although competence in teaching, as in all professions, is shaped significantly by on-the-job experiences and continuous learning, the programs that prepare teachers to work in K-12 classrooms can be early and important contributors to the quality of instruction. Evaluating the quality and effectiveness of TPPs is a necessary ingredient to improved teaching and learning.

Many aspects of the relationship between teacher preparation and instructional quality are not fully understood, and existing approaches to TPP evaluation are complex, varied, and fragmented. Designers and consumers of TPP evaluations could benefit from clear information about the purposes, effects, strengths, and limitations of current evaluation approaches and from guidance for designing and using future evaluations. This report, the product of an analysis by a committee of the National Academy of Education, aims to fill that need.

THE CURRENT LANDSCAPE

Systems for evaluating TPPs use various types of evidence—each with its particular strengths and limitations—to make inferences

about the quality of the preparation experience and its role in producing employable, high-quality teachers.

Evaluations use a variety of evidence to approximate the aspects of teacher preparation that are not all easily observable or quantifiable. "Inputs," such as selectivity in admissions, faculty qualifications, the quality and substance of teacher preparation course instruction, and the quality of student teaching experiences, are typically measured using grade point averages and SAT, ACT, or GRE scores of incoming students; academic credentials, experience, and full-time, adjunct, or part-time status of faculty in TPPs; syllabi, lectures, course offerings, and required hours of coursework; and fieldwork policies and records of observations of student teaching experiences. Evaluations also rely on a variety of "output" measures that typically include teacher licensure test results, surveys of program graduates and their employers, and so-called "value-added" estimates of graduates' impact on the learning of students in their classrooms.

A combination of input and output measures forms the basis for a variety of inferences—findings and interpretations—about the quality of TPP programs. For instance, some users of data about pass rates on licensure tests collected for the federal evaluation system may draw inferences about the degree to which TPPs prepare their students to pass the tests; other users may infer that these rates are more a reflection of the general ability of students in the program.

The sources of evidence used to develop inferences about program quality each have strengths and limitations. For example, the average SAT or ACT scores of an incoming class of TPP participants require relatively little effort to collect, are familiar to the public, and are standardized to enable comparisons across institutions. But they have also been criticized for being more a function of socioeconomic status than individual instructional capability, and in any event do not readily support inferences about the quality of training provided by the TPP. Scanning course syllabi is a less costly and less obtrusive means of determining program content than observing courses. But syllabi are apt to reflect the "intended" rather than the actual curriculum: some material in a formal written syllabus might not get taught, while material *not* included in the syllabus might actually be taught.

On the output side, data from teacher licensure tests can be easily obtained, but the wide variety in test content and passing scores makes it difficult to compare results, especially across states. Moreover, passing these tests is intended primarily to signal that candidates have a *minimum* level of knowledge and competency, rather than to predict their future effectiveness in the classroom.

And although an increasingly popular form of output evidence—value-added models—aims to estimate what some educators believe matters most for teacher preparation—the teacher's impact on student academic achievement—they also pose significant challenges. Problems arise in determining the extent to which differences in teachers' measured impact are due to training rather than to the institution's admission requirements, and in dealing with the fact that many graduates are omitted from the analysis because they teach untested subjects or grades or have left the school system.

The entities that evaluate teacher preparation programs in the United States have developed evaluation systems with different purposes, consequences, advantages, and disadvantages.

The ***federal government***, primarily through Title II of the Higher Education Act, seeks to hold TPPs accountable for performance by requiring them to report large amounts of data, and by requiring states to compile this information into publicly available "report cards" and to identify low-performing TPPs. The evidence employed for this purpose ranges from the performance of teacher candidates on licensure tests to student teaching requirements, among others. The Race to the Top initiative took this federal accountability system a step further by encouraging states to link information on student achievement with specific TPPs, publicly report these data on program impact for each TPP in their state, and expand those TPPs that seem to produce teachers who are effective in promoting student growth.

National, nongovernmental bodies make judgments about the quality of TPPs to determine whether they merit *accreditation*. Toward this end, the main accrediting organization, the Council for the Accreditation of Educator Preparation (CAEP), has issued a revised set of accrediting standards that includes evidence from course syllabi, observations of student teaching experiences, estimates of student achievement growth, results of surveys of program graduates and their employers, and other information. CAEP's new standards are intended to make the accreditation process more rigorous and outcome-focused by setting minimum criteria for program admissions and requiring programs to demonstrate their graduates' impact on student achievement.

State governments evaluate TPPs as part of their responsibility to approve programs—an important designation because graduates of approved programs can automatically be recommended for state teacher certification. Some states rely on the CAEP process for approval decisions, and others conduct their own program reviews using evidence similar to that used by CAEP. Several states, including Race to the Top grantees, are developing or implementing innovative evaluation systems that include value-added measures.

Media and independent organizations have long played a role in rating and ranking educational institutions. One relatively new partnership has focused on TPPs specifically: the project of the National Council on Teacher Quality and *U.S. News and World Report* aims to rate TPPs in colleges and universities against a set of standards that pertain to how well the program seems to cover particular aspects of teaching (as evidenced by review of available syllabi) and includes indicators of admissions selectivity and the quality of student teaching experiences.

Some ***teacher preparation programs*** also engage in self-evaluation to spur and inform program self-improvement. This can be done by a single institution or through a voluntary network of TPPs. These networks encourage members to base their self-studies on evidence and use a "clinical" model of teacher preparation that connects research and practice and involves collaboration with arts and sciences faculty and with K-12 schools.

TPP evaluations serve three basic purposes—holding programs accountable, providing consumer information to prospective TPP students and their potential future employers, and supporting program self-improvement.

Program evaluation has many plausible goals and the policy challenge is to select the system or approach that is best suited for a defined purpose. For example, although an evaluation alone may not provide all the information needed to hold a TPP accountable for producing well-trained and effective educators, it can provide relevant facts to the general public and to education policy makers. Evaluations with more of a consumer information purpose can give prospective teachers data to help them choose from among the broad array of preparation programs and can provide future employers of TPP graduates with information to assist them in hiring decisions. Evaluations for program self-improvement can yield information about an existing program's strengths and weaknesses, which program faculty and leaders can use to guide innovation and positive change.

These purposes are not always clear-cut, and there is nearly always some "drift": evaluation data will almost surely be used or interpreted in ways for which they may not have been planned and validated. Still, developers and users of evaluations should take a lesson from the history of standardized testing and explicitly consider the intended purposes of an evaluation—and take steps to reinforce its integrity by attempting to mitigate the misuse of data for unintended purposes.

Designers of evaluation systems should take account of the incentives those systems create and the consequences of their uses.

The uses and consequences attached to evaluation results, including direct consequences for TPPs and indirect consequences for faculty, students, and the education system as a whole, create incentives for individual and organizational behavior. In many cases, people and organizations will take actions in response to those incentives that will lead to genuine improvements.

But they may also seek better results by "gaming" the system. When evaluations have potentially high-stake consequences—when they are used to make important decisions about such issues as program recognition, accreditation, and closure or resource allocation—there is a danger that they will create incentives for people to manipulate the measure or focus undue attention on attributes that are measured at the expense of those that are not. For example, if course syllabi are perceived to be the basis for important decisions by prospective students, then faculty and program leaders might be tempted to produce syllabi that exaggerate what is taught in the course—thus corrupting the measure and undermining its usefulness.

A FRAMEWORK FOR MAKING DECISIONS ABOUT TPP EVALUATION

A set of core principles can serve as a starting point for thinking about TPP evaluation.

Chief among these principles is *validity*, i.e., the requirement that an evaluation system's success in conveying defensible conclusions about a TPP should be the primary criterion for assessing its quality. Validity refers both to the quality of evidence and theory that supports the interpretation of evaluation results and to the effects of using the evaluation results; the *consequences* of evaluation matter.

Other core principles include these reminders, cautions, and caveats:

- Although program evaluation is important, it is not sufficient in itself to bring about improvements in teacher preparation, teaching quality, and student learning.
- TPP evaluations in the U.S., with its fragmented education system, will always be likely to involve multiple players with different purposes and interests.
- The limitations of any evaluation system should be weighed against its potential benefits.

- Evaluation systems may have differential and potentially unfair effects on diverse populations of prospective teachers and communities.
- Evaluation systems should be designed to be adaptable to changes in education goals and standards.
- Designers and users of evaluation systems should communicate their intended purposes and be held accountable for their quality.

Addressing a sequence of questions can help evaluators decide on the approaches that are best suited to their main purposes.

A rational approach to designing TPP evaluations is to consider their likely direct and indirect, and positive and negative, impacts on teacher education and the broader educational system. Asking and attempting to answer the following questions in the early stages of evaluation design can increase the likelihood that an evaluation system will be coherent, serve its intended purposes, and lead to valid inferences about TPP quality.

Question 1: What is the primary purpose of the TPP evaluation system?

The TPP evaluation design process should begin with a clear understanding of what the system is intended to accomplish. Although evaluation systems often serve more than one purpose, designers should be able to articulate the *primary* purpose, then perhaps one or more secondary purposes.

Question 2: Which aspects of teacher preparation matter most?

Given the reality of limited resources, no single evaluation can measure every aspect of teacher preparation. Choices will have to be made about which aspects are of greatest interest, based on the purpose of the evaluation and the values of the organization doing the evaluating.

Question 3: What sources of evidence will provide the most accurate and useful information about the aspects of teacher preparation that are of primary interest?

Designers and users of TPP evaluations should examine the advantages and disadvantages of each type of available evidence and decide which *combination* of measures will yield the most useful information about program aspects of interest. With the education policy emphasis currently being on outcomes, one might be tempted to favor output over input measures, but both types of evidence should be considered. A key question is whether, on balance, the types of evidence included will lead to the desired inferences about TPP quality.

Question 4: How will the measures be analyzed and combined to make a judgment about program quality?

When a system collects multiple sources of evidence, decisions must be made about how they will be combined, particularly when they seem to conflict. Translating data into evaluation results entails decisions about how the data will be analyzed and interpreted. Applying the best available methods to ensure that the resulting interpretations are valid and fair is a key requirement.

Question 5: What are the intended and potentially unintended consequences of the evaluation system for TPPs and education more broadly?

Decisions will need to be made about the actions that will be taken based on evaluation results. The overall goal should be improving programs rather than punishing or embarrassing the low-performing ones. Initial evaluation results should ideally be followed up by gathering in-depth information to avoid wrongly identifying a program as low-performing or wrongly concluding that a relatively low-performing program should be denied resources that would enable it to improve. No evaluation system should, in itself, be the trigger of decisions to close or eliminate programs without careful attention to direct and indirect long-term effects.

Question 6: How will transparency be achieved? What steps will be taken to help users understand how to interpret the results and use them appropriately?

Those who design and implement TPP evaluations should clearly communicate their purposes and methods and propose the appropriate interpretation of results. All of the information about the evaluation should be easily accessible (through the Internet or other means) and communicated in a way that is understandable to users and the public.

Question 7: How will the evaluation system be monitored?

Once an evaluation system is underway, the consequences of the system, both intended and unintended, should be monitored, and the accuracy, reliability, and validity of the measures and methods should be studied. Designers and implementers should consider whether the system could adapt to new expectations for teacher training and recruitment.

PRIORITIES FOR RESEARCH TO PROMOTE CONTINUOUS IMPROVEMENT OF TPP EVALUATION

There are many complexities involved in TPP evaluation, but the education research community is well poised to take on the challenges. As evaluation systems evolve there will be ample opportunity to review their respective strengths and weaknesses. The credibility of results from TPP evaluations will hinge largely on the extent to which their implementation is monitored and their key features are revised and refined based on independent and objective research. Many issues will require continuous study, and the committee has identified these priorities for continued research:

- the effects of differences in teacher preparation on graduates' effectiveness in the classroom;
- the impact of different TPP evaluation systems on teacher preparation;
- ways to integrate comprehensive measures of teacher effectiveness, including non-cognitive student output measures, into evaluation systems; and
- ways to improve transparency, communication, and trust in evaluation systems.

1

Introduction: Purposes, Context, and Principles

There is widespread agreement about the importance of evaluating the many ways in which prospective teachers are recruited, selected, prepared, and licensed to work in K-12 classrooms. The most recent comprehensive review of teacher preparation in the United States underscored this need (National Research Council, 2010). And there is no shortage of evaluation systems: not surprisingly, in a society steeped in traditions of democratic accountability, reliant on data and formal inquiry to support improved decision making, and accustomed to a complex and fragmented educational landscape, many systems using many different types of data and metrics have evolved to evaluate teacher preparation program (TPP) quality.

This study by the National Academy of Education was motivated by the need to clarify how those systems vary in the evidence they collect, the purposes they are expected to serve, and the effects they have on a multiplicity of valued educational outcomes; and by the need to provide guidance in the development of new and better systems. Although our focus here is on the evaluation of programs, rather than the measurement of the quality of teachers and teaching, the underlying connections are obvious. Teachers, like all professionals, learn a lot about their complex role on the job; but how they are prepared before entering the classroom is assumed to make a difference—which is why the evaluation of preparation programs is so important.

The logic that links teacher preparation to teaching quality and, ultimately, to student learning may seem simple, but anyone who has tried to study these relationships knows they are fraught with methodological complexity and incomplete data. The good news is that empirical evidence now increasingly validates the intuition that teachers matter and that the quality of classroom instruction makes a substantial difference in the performance of students (Shulman, 1986; National Academy of Education, 2005; Hanushek, 2010; National Research Council, 2010; Bill & Melinda Gates Foundation, 2013).

On the other hand, less is known about exactly *how* teachers matter. Social science is still far from reaching a conclusive judgment about how to measure pedagogical skills, content knowledge, temperament, interpersonal styles, empathy, and understanding of the learning needs of children, and how those attributes, however they might be measured, combine to make the most difference in teachers' effectiveness. For some subject areas, educators and researchers may be converging on a set of baseline requirements for entry-level teaching (e.g., Hill, Ball, and Schilling, 2008; Ball and Forzani, 2011). But by and large the knowledge base about essential qualities of teaching is still at a rudimentary stage, a reality that necessarily places limits on the design of valid measures for assessing TPPs.

The methodological challenges are rendered even more complicated by the difficulties of defining educational goals in a system with diffused governance (Vinovskis, 1999, 2009; Fuhrman and Elmore, 2004); assessing student and teacher performance with tests of variable quality and, especially, using the results for high-stakes decisions (Linn, 2000; Shepard, 2003; National Research Council, 2011a); and curbing the potential negative effects of rigorous public accountability on the morale and motivation of teachers (Feuer, 2012b). Perhaps most important, although there is abundant evidence that poverty and inequality are strong correlates of variation in student achievement (OECD, 2010; Duncan and Murnane, 2011), more work needs to be done to untangle those effects and, more specifically, to understand how they relate to the attributes of TPPs.

Given these difficulties, it may be tempting to leave the whole messy problem of assessing the quality of teacher preparation to some combination of intuition, faith, and tradition. But this is a temptation to be resisted at all costs. Complexity is not an excuse for complacency. The state of science may be wanting, but the need remains for the research and policy communities to develop defensible measures of teacher preparation quality. This need is especially urgent in an era of rapid technological and demographic transition and in a culture of public accountability. As a note of encouragement, the education research community has an honorable track record for taking on the most pressing problems in policy and practice and has been responsible for some of the most important methodological advances in the behavioral and social sciences (see National Research Council, 2002). It is on solid footing to meet the current challenges.

PURPOSES OF THIS REPORT

The most powerful rationale for focusing attention on the strengths and weaknesses of alternative approaches to evaluating TPPs and on criteria for designing new and improved approaches comes down to two interconnected realities. First, the landscape of TPPs has become substantially denser in recent years. Between 70 and 80 percent of the roughly 200,000 new teachers entering the profession each year are prepared in traditional programs housed in postsecondary institutions, with the rest entering through 1 of approximately 130 alternate routes (National Research Council, 2010). Amid this hodgepodge of professional preparation routes and systems, demand for evidence of their quality has naturally grown (e.g., Crowe, 2010; Anderson, 2013; National Council on Teacher Quality, 2013). Second, with states, districts, the federal government, teacher education associations, and various independent accrediting and ratings organizations all experimenting with new evaluation tools and techniques (such as value-added modeling), attention increasingly turns to their intended and unintended consequences. The fear is palpable that flaws in these new methods and their applications might perversely undermine rather than enhance teacher quality and effectiveness (e.g., Darling-Hammond, Amrein-Beardsley, Haertel, and Rothstein, 2012).

Finding imperfections in evaluation methods is easier than fixing them; "perfecting" them is probably out of the question altogether. The principal purposes of this report are therefore more modest: to clarify complexities inherent in teacher preparation evaluation and to propose a decision framework for the design of improved evaluations in the future.

In carrying out our charge, as defined by the grant from the National Science Foundation that supported this work, the National Academy of Education convened an expert committee that held two workshops, reviewed relevant literature in both the academic and popular media, commissioned papers and presentations from outside experts, and met in person and electronically to discuss and debate findings, conclusions, and recommendations.

These are the main questions we addressed:

- How are federal, state, and local agencies and other organizations reacting to the public demand for evidence of the quality of teacher preparation?
- How are institutions that prepare future teachers—universities, teacher colleges, private non-university organizations, and

others—handling the challenges of providing better information about the quality of their programs?

- What is known about the relative effectiveness of different approaches to evaluating TPPs?
- How well do different existing or potential methods align with the multiple intended uses of evaluation results?
- What are the most important principles and considerations to guide the design and implementation of new evaluation systems?

This report synthesizes relevant research on existing approaches to evaluating TPPs and analyzes issues relevant to the design and implementation of new or improved approaches. Although general in its overall scope, the report also suggests issues pertaining more narrowly to the evaluation of programs in which future science, technology, engineering, and mathematics (STEM) educators are prepared.

Assumptions, Context, and Core Principles

This report is built on three basic assumptions:

1. The quality of instruction plays a central role in student learning.
2. Teacher preparation programs contribute to the quality of instruction.
3. The evaluation of teacher preparation programs can provide useful information for the improvement of teacher preparation policy and practice.

It is worth underscoring that although we are cognizant of the inherent connections between teacher preparation and teaching, this report focuses on systems of *program evaluation* rather than on the assessment of the effectiveness of individual teachers.

Three attributes of American educational culture and politics define the context of education reform generally and the improvement of TPP evaluation specifically:

- The historical and ongoing struggle for equity *and* excellence in the public education system (Cremin, 1990)
- The fragmented and decentralized nature of schooling, which makes it adaptable to change (Goldin and Katz, 2008) but not easily amenable to the design of common standards for content and performance (e.g., Zehr, 2009)
- The continuing emphasis on outcomes of education (e.g., student achievement), rather than inputs to education (e.g., per pupil

expenditures) as the core elements of accountability (Fuhrman and Elmore, 2004).

Seven core principles have guided our work:

Principle 1: Because we assume there is a basic linkage among teacher preparation, teaching quality, and student learning, the main goal of TPP evaluation is the continuous improvement of teaching quality and student learning (we use "learning" as shorthand for academic, behavioral, and social outcomes of education). However, *although program evaluation is important, it is not sufficient in itself to bring about improvements in teacher preparation, teaching quality, and student learning.*

Principle 2: Because authority for education in the United States is, by design, diffused, *the evaluation of TPPs will always include multiple systems operated by different groups with different purposes and interests.* Unless the decentralized system of governance over American education changes, we assume that there will always be different evaluation methods that rely on different data with results intended for different audiences. No single method or mechanism is likely to completely satisfy multiple, legitimate, and potentially incompatible demands for valid, reliable, and usable information.

Principle 3: *Validity should be the principal criterion for assessing the quality of program evaluation measures and systems.* The word "validity" is shorthand for the extent to which evaluation data support specific inferences about individual or organizational performance. In this report we define validity broadly to include (1) the quality of evidence and theory that supports the interpretation of evaluation results; and (2) the *consequences* of the use of evaluations for individuals, organizations, or the general public. (See Box 1-1.)

Principle 4: We assume that *any measure—or, for that matter, any TPP evaluation system that uses multiple measures—has limitations that should be weighed against potential benefits.* The fact that there are imperfections in evaluation systems is not a sufficient reason to stop evaluating, but rather an argument for investing in the development of improved methods. But trying to find the "perfect" set of measures is a fool's errand; a more rational approach is to explore the relative benefits and costs of alternative approaches (see also Feuer, 2006, 2008) and to consider whether, on balance, there is evidence that the benefits outweigh the costs—not just the costs in dollars and cents but the costs defined more generally in terms of unintended negative consequences.

Principle 5: We assume that *differential effects of TPP evaluation systems—for diverse populations of prospective teachers and the communities in which they may work—matter, and should be incorporated as a component* *of validity analysis and as a design criterion.* It is especially important to consider potential inequities that may arise from the interpretation of evaluation results and their application. Special attention should be paid to unintended impacts on the morale, capacity, and reputation of TPPs that cater to different pools of potential teacher candidates or that are

BOX 1-1
Validity

Validity is defined in the literature of measurement and testing as "the extent to which evidence and theory support the interpretations of test scores" (Messick, 1989; American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999). There is a vast literature about the concept of test validity that goes back many decades (in addition to Messick, 1989, see, for example, Cronbach and Meehl, 1955; Shepard, 1993).

Evaluations typically make use of multiple measures rather than a single test, but key questions about validity, including the following, apply to TPP evaluation:

- To what extent does the evaluation measure what it claims to measure? (This is sometimes referred to as *construct validity.*)
- Are the right attributes being measured in the right balance? (This is sometimes referred to as *content validity.*)
- Is there evidence that teachers graduating from highly rated TPPs prove more effective in the classroom? (This is sometimes referred to as *predictive validity.*)
- Is a measure subjectively viewed as being important and relevant to assessing TPPs? (This is sometimes referred to as *face validity.*)

The committee takes the view that *consequences* are central to judging the soundness of a TPP evaluation system. Questions about consequential validity—an aspect of validity that addresses the intended and unintended consequences of test interpretation and use (Messick, 1989)—include the following:

- To what extent does the evaluation affect the behavior of teacher educators in the ways intended?
- To what extent does the evaluation create perverse incentives such as "gaming" of the system on the part of teacher educators, lead to policy decisions with unknown or unwanted long-term effects, or create other unintended consequences?

Although debate continues among education and measurement researchers about whether consequences should be included in the formal definition of validity (Messick, 1989; Linn, 1997; Popham, 1997; Shepard, 1997; Feuer, 2013a), there is widespread agreement that monitoring consequences of an assessment system is crucial in determining the system's soundness and value. For discussion of a particularly important aspect of consequential validity, see Principle 5.

intended to serve communities struggling to overcome socioeconomic disadvantage. We are not suggesting differential standards for the evaluation of program quality, but rather we are flagging the importance of studying how those standards may, for example, lead to the reduction in the supply of prospective future teachers and/or an interruption in the flow of potentially excellent and dedicated teachers into poor neighborhoods.

Principle 6: *TPP evaluation systems should themselves be held accountable*. Private and commercial organizations and government agencies that produce, promulgate, or mandate evaluations must be clear about their intents and uses and, to the extent possible, provide evidence that intended and unintended consequences have been considered. Evaluators and users of evaluation data must be open to critique as the basis for continuous improvement and be willing and able to explore policies aimed at reinforcing appropriate uses of evaluation information.

Principle 7: *TPP evaluation systems should be adaptable to changing educational standards, curricula, assessment, and modes of instruction.* As we prepare this report, expectations for teaching and standards of student learning (and other valued outcomes) are again changing. The implementation of the Common Core State Standards, for example, and associated new assessment technologies will necessarily shape the context for evaluating TPPs. The need for flexibility in adapting to these types of changes must be balanced against the legitimate desire for evaluations designed to provide reliable *trend* information. Achieving a workable balance requires an appreciation of tradeoffs and an acceptance of compromises in designing systems with diverse purposes.

THE POLICY CONTEXT

The following observation will undoubtedly resonate with anyone who keeps up with the current education reform debate and recurring attacks on teacher preparation institutions:

> A professional educator would have had to restrict his reading almost entirely to children's literature in order to escape notice of the recurrent criticisms of American teacher education appearing in popular and professional publications in recent years. . . . In the midst of today's heated discussions of the adequacy of teacher education programs, one might conclude that, in addition to their intensity, the number of teacher education criticisms has been great during the past decade.

Indeed, the critiques are frequent and fierce. But this quotation is not from a recent issue of *Education Week* or *Educational Researcher.* It is from an article published in the December 1958 edition of the *Phi Delta Kappan* (Popham and Greenberg, 1958). The remainder of the article summarizes results of a survey of the types of criticisms that had been leveled against teacher education in the prior decade, ranging from "overemphasis on pedagogy" and "inadequate philosophical bases" to "anti-intellectualism" and "educationists' reluctance to accept criticism." The authors admonish educators to "not content [themselves] with merely throwing up a handful of 'defenses' in professional journals . . . [but also to not] resign [themselves] to cower helplessly before the blistering attacks." Clearly the 2013 zeitgeist of angst and anger, recrimination and rebuttal, and claim and counterclaim about the teaching profession is not unprecedented.

It is worth noting in this context that other professions—and not just teaching—are targets of heightened public critique for the ways their preparation programs are evaluated. Business schools and law schools, for example, are under a great deal of pressure to change. The *Harvard Business Review* (2009) featured a discussion about MBA training that included criticisms of the focus on academic as opposed to professional skill training and inattention to needed areas like ethics. Law school preparation is also being widely critiqued for not preparing lawyers for the type of work they will actually do once they graduate (see, e.g., Spencer, 2012). In both cases, much of the criticism laments the push in preparation programs to focus on scholarly achievements and funding of faculty rather than other arguably more practical needs. Regarding medical preparation, a recent RAND study found no link in the research literature between health care training and quality of care (Nolte, Fry, Winpenny, and Brereton, 2011). Clearly we live in an era in which many professional fields and their preparation activities are subjected to intense scrutiny.

Still, even against this backdrop of a generally more heated environment of accountability in many professions, one cannot escape the impression that in education the elbows have really gotten sharper. The attitude about public schooling generally, and teacher education specifically, has grown more dismal, and the rhetoric has become not only more strident but also more prevalent. Figure 1-1 provides a graphical representation

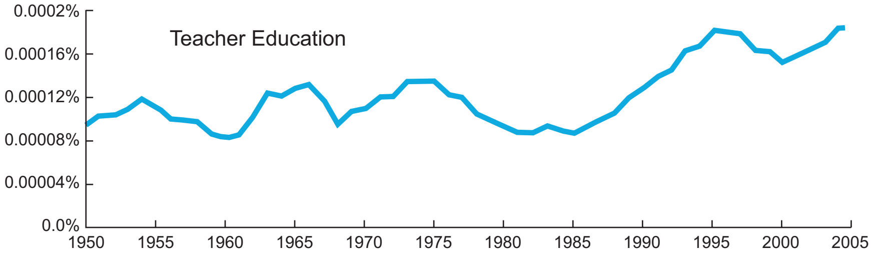

FIGURE 1-1 Mentions of "teacher education" in published books, 1950-2005.
SOURCE: Figure based on Google Books Ngram Viewer, http://books.google.com/ngrams.

of the increased appearance of the phrase "teacher education" in books published in English between 1950 and 2005. Perhaps not surprisingly, the slope of the trend line became steeper in the 1980s, an era most well known in education for the publication of *A Nation at Risk* (National Commission on Excellence in Education, 1983), one of the most influential treatises on the condition of education with some of the most memorable rhetoric ever seen in official policy literature (for a critique see, e.g., Stedman and Smith, 1983).[1]

Two overlapping perspectives discernible in the public and professional rhetoric of education reform help explain why the debate over teacher preparation and its evaluation has become so intense. What might be called the "excellence" perspective emphasizes changing global competition and its implications for the American education system as a whole (e.g., National Academy of Sciences, National Academy of Engineering, and Institute of Medicine, 2007; Hanushek, Peterson, and Woessmann, 2012; Schmidt, Burroughs, and Cogan, 2013). This "macro" view starts with the perhaps obvious assumption that education enhances economic capacity and competitiveness, and builds the case for higher standards of performance in our public schools principally as a means of reinforcing (or regaining) America's prominence as a world economic leader. According to the underlying logic of this perspective, aggregate academic outcomes as measured by test scores affect long-term economic capacity; educational performance, in turn, hinges to a significant extent on the quality of classroom instruction; and the quality of instruction hinges to a significant extent on the quality of pre-service teacher preparation. Holding TPPs accountable as a means of effecting positive change in the condition of American schooling is viewed as an essential ingredient of educational improvement needed to sustain and grow America's global stature.

A somewhat different perspective challenges the broad characterization of the American education system as essentially failing (e.g., Xie and Killewald, 2012; Breunig, 2013; Salzman, 2013) and emphasizes instead its persistent and gnawing inequities. According to this line of reasoning, even if American education *on the whole* is not performing as poorly as is often claimed, in comparison with earlier times or with our global competitors (Loveless, 2011; Carnoy and Rothstein, 2013; Feuer, 2012a, 2013a), the main problem is evidence of rising inequality in both the allocation of resources—including the allocation of quality teachers to students who

[1] Numerous factors could explain the apparent increase in attention to teacher education suggested by Figure 1-1, including the rapid expansion of the Internet and electronic media, along with growth in education advocacy groups and think tanks. We include this graph without wishing to overstate its empirical significance.

need them most—and the distribution of educational outcomes (Duncan and Murnane, 2011; Malcom-Piqueux and Malcom, 2013; Rothstein, 2004). In other words, regardless of whether one considers the *average* level of educational performance to be adequate to current and projected needs, the pervasive and growing *variance* in educational opportunity and outcomes is simply unacceptable. It is assumed that how teachers are prepared for work makes a difference to their classroom performance, but the *main emphasis is on how differences in prior preparation affect the quality of instruction for the schools and children with the greatest educational needs*. A fundamental goal of evaluation, therefore, should be to remedy imbalances in teaching quality that perpetuate the achievement gap and, by extension, socioeconomic and educational inequalities that threaten to rip the fabric of American society.

It may be tempting to view these perspectives as irreconcilable. Indeed, in much of the debate over education reform policies there is an increasingly noticeable (and unproductive) tug between those who caution that school reform without eradication of poverty is futile and that it is unfair to hold teachers accountable given their students' circumstances (e.g., Berliner, 2012); and those who argue that economic disadvantage cannot be an excuse for a status quo of low-performing teachers and schools and that to remedy economic inequality we need to raise the productive efficiency of teachers (e.g., Hanushek, 2010). Both sides in this debate share the conviction that schools and teachers matter, and that children's life circumstances affect their educational chances; it is a matter of emphasis, rather, and an argument about what policy levers are most likely to effect positive change. The debate is a variation on a familiar tension in the rhetoric of education reform, especially in the U.S., with its long history of struggling to expand access and opportunity and simultaneously raise standards (see, e.g., Tyack, 1974; Cremin, 1990; Office of Technology Assessment, 1992).

But as the pre-eminent education historian Lawrence Cremin so eloquently argued in his last book, this false dichotomy of standards versus access has been and should continue to be resisted (Cremin, 1990). We agree. First of all, at a general level both perspectives have merit, and no good will come from framing education policy as a simple either-or proposition. Second, and perhaps more to the point, there is a simple reason why today, again, the equity and excellence perspectives converge: in the light of incontrovertible evidence of demographic changes affecting the American population, attention to access and equity is a necessary condition for sustaining and growing our aggregate economic and social performance. This joint perspective should be at the center of discussions about the design of improved TPP evaluation systems.

THE SPECIAL CASE OF STEM

Science, technology, engineering, and mathematics (STEM) education has an especially prominent place in the debate about excellence and equity. There is a great deal of concern about the condition of science generally and the state of STEM education specifically, in this country and in others. As the President's Council of Advisors on Science and Technology (PCAST) stated (2010), "STEM education will determine whether the United States will remain a leader among nations and whether we will be able to solve immense challenges in such areas as energy, health, environmental protection, and national security" (p. vii).

Recent developments in STEM education intersect with issues concerning the development of appropriate systems of TPP evaluation. Here we briefly describe three sets of STEM efforts with implications for TPP evaluation: (1) national initiatives to improve STEM education broadly; (2) the development of STEM standards, curriculum, and assessments; and (3) initiatives specifically aimed at STEM teacher preparation.[2]

The first set of national improvement efforts grows out of widespread concern about the state of STEM education in the United States. President Obama mentioned STEM education and the preparation of STEM teachers in myriad presidential campaign speeches. PCAST issued two reports, the first on STEM in K-12 schools (2010) and the second on STEM in higher education (2012). In addition, Congress asked the National Science Foundation to identify highly successful STEM schools, which resulted in two National Research Council (NRC) reports (2011b, 2013). These reports identify the school conditions that shape effective STEM instruction and more than a dozen indicators that might be used to track progress toward improved STEM education. Wilson (2013) concludes that if TPPs responded to the messages of these and other reports, they would focus on demonstrating the following outcomes and characteristics:

- Graduates of TPPs would have extensive STEM content knowledge for teaching and have sufficient pedagogical knowledge and skills.
- Teachers would be prepared through a genuine partnership among disciplinary departments and with K-12 schools.
- TPPs would use data for program improvement, including tracking graduates into their early careers.
- TPPs would use carefully crafted, aggressive, and innovative recruitment strategies.

[2] This section draws from Wilson (2013).

- Graduates of TPPs would be aware of and prepared to work in the varied school structures that support STEM learning.
- TPPs would provide new teachers with clinical experiences in STEM schools that have the appropriate conditions for high-quality instruction.

A second set of STEM efforts has focused on clarifying what should be taught in K-12 schools. Currently, the mathematics standards developed through the Common Core State Standards Initiative (n.d.) and the Next Generation Science Standards (n.d.) have garnered much attention. While there is considerable overlap between these documents and previous visions for STEM curricula, there are a few significant differences. In both cases, the developers of the standards attempted to constrain the list of topics to be covered so that teachers and students could explore significant STEM concepts in depth. Both emphasize "practices," perhaps best understood as an effort to capture how scientists, engineers, and mathematicians work: posing questions, gathering evidence, offering arguments, presenting ideas, and the like. In addition both the Common Core State Standards in Mathematics (CCSSM) and the Next Generation Science Standards (NGSS) emphasize the "progression" of student learning—how the concepts studied by students evolve, expand, and increase in sophistication and how students' understanding broadens and deepens over time.

Along with these new standards, groups of states and other organizations have put considerable effort into developing assessments that can be used to track student mastery of the standards. Two state consortia—the Smarter Balanced Assessment Consortium and the Partnership for Assessment of Readiness for College and Careers—have received funding from the U.S. Department of Education to develop "next-generation" assessments that will gauge students' "college and career readiness." These new K-12 standards have substantial ramifications for teacher preparation and TPP evaluation. TPPs will need to demonstrate that the relevant disciplinary and pedagogical courses have been altered to provide teachers with the knowledge and skills embodied in the new standards and tested by the new assessments.

Wilson (2013, p. 7) raises several concerns about this process:

> The default will be to highlight the similarities between prior reforms and to simply package the "old wine" in new bottles. Old textbooks will be stamped with official announcements that they are "aligned," teachers will declare their efforts to teach inquiry identical to the calls to teach practices, and the like. For teacher preparation programs, the challenge of finding high quality clinical placements is already considerable, and

> just finding enough teachers who know about CCSSM and/or NGSS and are adjusting their practice in light of those changes seems considerable. To now locate clinical settings in which STEM teaching and learning is experiencing radical overhaul might be even more challenging.

A third set of efforts focuses on the role of teacher preparation in addressing the supply and quality of STEM teachers. The PCAST (2010) report recommended recruiting 100,000 teachers in the next 10 years; in 2011, the Carnegie Corporation of New York convened a group of diverse stakeholders to rise to this challenge. The resulting “100K-in-10” network consists of more than 150 partners; some of them are funders, but most are teacher preparation programs, ranging from university-based programs to residencies to Teach For America, that have pledged to educate a certain number of teachers in the coming ten years (100K-in-10, n.d.). Another prominent effort is the Science and Mathematics Teacher Imperative (SMTI) started by the Association of Public and Land-Grant Universities (APLU, n.d.). University presidents pledged to increase their enrollments and graduation rates of STEM teacher candidates, and SMTI began working with a smaller group of universities to lay the foundations and develop tools and resources that would be useful to universities involved in this effort. SMTI has produced an analytic framework to help teacher preparation institutions evaluate what they do against a set of best practices.

In sum, issues of teacher preparation and TPP evaluation are magnified when it comes to STEM. While there may appear to be considerable agreement about what needs to be done, it is agreement about general ideas rather than about the details of what teacher education should look like. Consider the issue of content knowledge. Efforts focused on the preparation of STEM teachers (such as those cited above) all claim that STEM teachers need to have substantially more content knowledge; but teacher preparation efforts vary widely in how they conceptualize, assess, and deliver STEM content knowledge. Educators have yet to take up the issue of the differences between content knowledge of a science discipline, as represented by an undergraduate science degree, and content knowledge for science *teaching*. Preparation efforts also vary in how they define “clinical experience” or “a culture of evidence” or “recruiting high-quality candidates.” Wilson concludes that the landscape of STEM teacher preparation is messier than ever and that any evaluation system will need to be adaptable to programs that vary widely in clientele, program purposes, program substance, and the like.

TOWARD A DECISION FRAMEWORK

Our work and the resulting decision framework were organized around an intuitively appealing concept, namely a "mapping" of the purposes or intended uses of program evaluation against the various ways it can be designed and organized. Because TPP evaluation has many plausible goals and can be designed and conducted in many ways, the policy challenge is to align means and ends—to select the methods or approaches that are at least reasonably well suited to defined purposes.

Implicit in this mapping framework is the suggestion that in an ideal world, evaluation methods would be used only to accomplish purposes for which they have been designed and validated. Indeed, this ambitious goal is articulated in the professional measurement community's standards for professional practice (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999). But abundant evidence about the uses of assessment and evaluation and of social science data generally suggests the need for a more nuanced expectation. We assume there will always be some "drift"—in other words, that evaluation data will almost surely be used or interpreted in ways for which they may not have been validated. A central rationale for our framework is the concern, based on substantial evidence and experience, about the problems that arise when the drift from validated uses of evaluations to more casual interpretations becomes too great.

With this caveat about the inevitability of drift, the goal of the proposed decision support framework is not necessarily the *optimal* alignment of means and ends, but rather the design of systems with the best chances of reducing, if not totally eliminating, unintended negative effects. Expressed more optimistically, we hope that using this decision framework will increase the odds of success in applying various evaluation approaches toward various defined goals. Individuals and organizations using the framework will need to apply their own levels of "risk-aversion" to reach judgments about the benefits and costs of alternative evaluation systems.

CONTENT OF THE REPORT

Chapter 2 reviews the existing landscape of teacher preparation program evaluation in the United States. Not surprisingly, the intentionally fragmented "system" of authority and governance for public education in the United States has spawned a remarkably complex array of mechanisms and institutions to evaluate TPPs. We have grouped current approaches, or systems, to TPP evaluation into five basic categories based on the entities responsible for conducting the evaluation: the federal government, national accreditation bodies, state governments, media outlets and independent organizations, and the TPPs themselves. While these systems are separate in terms of their origins and details, in practice they are frequently implemented simultaneously. Moreover, most states and institutions use multiple approaches and data methodologies. Still, these diverse approaches actually rely on similar sources of data and on similar methods for converting data into judgments of program quality.

The largely descriptive analysis in Chapter 2 lays the foundation for Chapter 3, which presents the concept of mapping—linking characteristics of evaluation systems with their various purposes and intended uses. Although the reality is more complicated, we collapse the multiple purposes or uses of evaluation into three main categories: ensuring accountability and monitoring, providing consumer information, and improving teacher preparation programs. These purposes overlap, but the three categories offer a useful framework to support decisions about program evaluation.

Chapter 4 then builds on the descriptions in Chapter 2 and the mapping concept of Chapter 3. We review basic assumptions and underlying principles and propose a decision framework consisting of a sequence of questions that designers should address in the early stages of constructing or revising a TPP evaluation system. We include in the chapter a short discussion of priority areas for further study.

Throughout the report, we intersperse information about how selected other professions and other countries evaluate their pre-service education and training programs. Those discussions are intended for illumination more than emulation. For example, the approaches used for pre-service education and training in the nursing profession may hold useful lessons for the teaching profession, but differences in labor market conditions, required skills and knowledge, and other characteristics of the professions limit the extent to which those approaches are directly applicable to teacher preparation.[3]

[3] For a particularly illuminating review and analysis of how nursing preparation programs are evaluated, see the paper by Johnson and Pintz (2013) commissioned for this study. For more detailed discussion of teacher preparation evaluation in other countries, see Furlong (2013) and Tatto, Krajcik, and Pippin (2013).

2

The Landscape of Teacher Preparation Program Evaluation

Systems for evaluating teacher preparation programs in the United States rely on a complicated array of mechanisms and institutions. At various points, the same TPP may undergo different types of evaluation, each with its own purpose, data sources, methods, and consequences. A useful way to make sense of the complex landscape of TPP evaluation is to consider both the sources of evidence used to assess program quality and the entities doing the evaluation. We begin this chapter by describing the different but often overlapping sources of evidence used in TPP evaluation. We then describe five types of TPP evaluation systems categorized by the entities responsible for conducting the evaluation[1]:

1. The federal government
2. National nongovernmental accrediting bodies
3. State governments
4. Media outlets and other independent organizations
5. The TPPs themselves

[1] One school district has also conducted a TPP evaluation. Recently, the New York City Department of Education (2013) conducted its first evaluation of a dozen public and private TPPs in the city and released score cards. Since this is the first high-profile instance of a school district conducting a formal evaluation of TPPs, we do not address school districts as a separate category of evaluators in this report. However, it is worth noting that district-level TPP evaluations may increase in the future.

For each of these types of evaluations, we discuss its origins, its particular areas of focus, and the processes used to evaluate TPPs. At the end of the chapter we present a matrix that combines the five types of systems with the main sources of evidence used by each system.

SOURCES OF EVIDENCE

Several aspects or attributes of teacher preparation may not be directly observable but are often of interest to TPP evaluators. These include the quality and substance of instruction, faculty qualifications, effectiveness in preparing new and employable teachers, and success in preparing high-quality teachers.

Evaluations use a variety of evidence, or *measures*, to estimate the attributes of interest. For example, to gauge how well TPPs prepare high-quality teachers, evaluators could use several different types of evidence, such as performance assessments of teacher candidates, "value-added" estimates of the impact of a particular teacher on the achievement of his or her students, and surveys of employers. Table 2-1 lists the most common attributes of interest in TPP evaluations and, for each, the various types of evidence that are typically used to measure them.

Ideally, the evidence used to assess TPP quality would reflect the characteristics of teacher education that empirical studies have shown to be most important in preparing effective teachers. Although research provides some information to guide TPP evaluation, there is still much more to be learned about effective teacher preparation practices (Wilson, Floden, and Ferrini-Mundy, 2001; Cochran-Smith and Zeichner, 2005; National Research Council and National Academy of Education, 2010). Most measures of TPP quality in use today seem to have been chosen based on their face validity—in other words, they appear to address important characteristics of teachers and teaching—and on the feasibility of collecting the data, rather than on empirical correlations or "predictive validity" evidence linking qualities of teacher preparation with student outcomes. (See Box 1-1 in Chapter 1 for more about different aspects of validity.) For this and other reasons the professional measurement and evaluation communities continue to advocate strongly for the use of multiple measures (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999; American Federation of Teachers, n.d.).

A distinction is often made between input and output measures. *Input* measures reflect the substance and processes of teacher preparation

TABLE 2-1 Attributes Related to TPP Quality and Evidence Used to Measure Them

Attribute	Measures
Admissions and recruitment criteria	• GPA of incoming class • Average entrance exam scores (e.g., SAT, ACT) • Percentage of minority students in incoming class • Number of candidates in high-need subject areas and specialties
Quality and substance of instruction	• Course syllabi • Lectures and assignments • Textbooks • Course offerings and required hours • Required content courses
Quality of student teaching experience	• Fieldwork policies, including required hours • Qualifications of fieldwork mentors • Surveys of candidates • Records from observations of student teaching
Faculty qualifications	• Percentage of faculty with advanced degrees • Percentage of faculty that are full-time, part-time, adjunct
Effectiveness in preparing new teachers who are employable and stay in the field	• Pass rates on licensure tests • Hiring and retention data
Success in preparing high-quality teachers	• Teacher performance assessments administered near end of program • Ratings of graduates by principals/employers • Value-added estimates

programs and the attributes of students who enter those programs; *output* measures gauge the performance of graduates after completing a TPP.[2]

Input Measures

Input measures seek to assess the quality of the preparation experience and include such evidence as entrance requirements, course syllabi, and fieldwork policies.

[2] Input and output measures are similar to what some evaluators refer to as process and outcome measures, respectively (National Center for Justice Planning, n.d.).

Admissions and Recruitment Criteria

Program admissions criteria, such as average SAT, ACT, or GRE scores, or GPAs of incoming students, are often used as measures of TPP quality, even though most TPPs do not have rigorous admissions requirements. Walsh and Jacobs (2007) found that 40 percent of TPPs set a minimum GPA for admissions. Two-thirds of the TPPs accepted more than half of their applicants, while one quarter accepted nearly all applicants.

Presumably, the rationale for collecting data on the selectivity of TPP admissions criteria is that a TPP with an academically gifted student body is more likely to have a rigorous curriculum and a highly qualified faculty. TPPs with relatively high admissions criteria continue to attract more academically gifted teacher candidates, who may want to enroll in a program with similarly talented peers. Selectivity also affects a program's reputation and prestige, which probably bears some relationship to its underlying quality. But it is not necessarily true that more selective programs produce better-prepared and more effective teachers. A TPP with an open admissions policy may well have an excellent curriculum and faculty and produce good teachers. Taken alone, selectivity data say nothing about the actual quality of the program as gauged by the knowledge and skills acquired by the students who complete it.

There is some evidence of a relationship between the general aptitude of candidates who enter a TPP, as measured by relatively curriculum-neutral standardized tests such as the SAT 1, and their eventual teaching effectiveness. More broadly, teacher candidates with high ACT or SAT scores, GPAs, and class rank have been shown to perform better on Praxis certification tests (Gitomer and Latham, 1999). Some studies indicate that candidates with these academic qualifications also end up being more effective teachers, as measured by growth in their students' math test scores and teacher retention rates (Ferguson and Ladd, 1996; Henry, Bastian, and Smith, 2012). Still, the question of *causation* remains unresolved: Are these results due to the teacher candidates' high level of preparedness before they enter the TPP or to the effect of their completing a selective program? Absent a research design that eliminates or reduces this type of selection bias, the most one can infer is a modest correlation between measured ability on entrance exams and subsequent classroom performance.

Measures like admissions test scores and GPAs have the advantages of being easy to collect, quantify, and compare across TPPs. Leaving aside lingering debates about whether variations in SAT scores, for example, are explained primarily by socioeconomic status rather than by differences in the knowledge or skills the test purports to measure (Jaschik, 2012), the fact remains that the SAT and other college entrance exams were designed to estimate an individual's academic preparedness for college, not the quality of instruction in a higher education program. Furthermore, commonly used entrance exams have never been shown to predict success at work *after* college, whether in classroom teaching or any other field. The use of college entrance exams as an indicator of TPP quality is viewed with substantial skepticism by many professionals, especially measurement experts who know the most about the exams' content and the purposes for which they have been validated.

Quality and Substance of Instruction

Course offerings and required hours in key subjects are often used as components in TPP evaluations. Some evaluations delve more deeply by analyzing course syllabi, lectures, textbooks, and assignments. These sources of evidence are assumed to indicate the extent to which important subject matter and pedagogical content are delivered and, if so, whether the content meets accepted standards. When reviewed systematically and coded consistently, these types of evidence can provide more insight into instruction than the number of course hours or a listing of offerings (Coggshall, Bivona, and Reschly, 2012).

However, syllabi alone provide limited estimates of program quality, given that some material in the formal written syllabus may not get taught and material *not* included in the syllabus may actually be taught. Textbook content is also examined, but again, just because certain material appears in a textbook does not mean it will be covered or emphasized by the instructor. A concept that originated in international comparative studies of curriculum is the distinction between *intended* and *enacted* curricula (see McKnight et al., 1987; Schmidt, McKnight, Cogan, Jakwerth, and Houang, 1999). The intended curriculum is what students are supposed to learn, as laid out in syllabi and textbooks, whereas the enacted curriculum refers to the content actually delivered during instruction and how it is taught. This distinction is germane to discussions about the validity and reliability of using evidence from syllabi and other course materials in TPP evaluations. If syllabi are used as part of a high-stakes evaluation that has serious consequences for TPPs, it is conceivable that TPPs might produce impressive syllabi that give an inflated picture of what is actually taught in courses.

Boyd, Grossman, Landford, Loeb, and Wyckoff (2008) reported some initial findings from New York City about the relationship between the academic content of TPPs and the subsequent impact of program graduates on their students' achievement. In particular, the study found that teachers who had greater opportunities in their preparation to engage in actual teaching practices—for example, listening to a child read aloud for assessment purposes, planning a guided reading lesson, or analyzing student math work—showed greater student test score gains. The same was true for teachers who had the opportunity to learn and review curriculum used in New York City schools prior to actually teaching it. The study uncovered some evidence of improved outcomes for students whose teachers had preparation in math content but found no effects for many other academic factors, such as opportunities to learn about how students learn. The authors urged caution in interpreting these results since "research analyzing such relationships is still in its infancy" (Boyd, Grossman, Lanford, Loeb, and Wyckoff, 2008, p. 28).

The subject area most heavily studied, in terms of the effects of coursework, has been secondary school mathematics teaching. In their review of the literature, Floden and Meniketti (2005) found that most studies report a positive correlation between teachers' study of math and measures of student achievement, but the results are not entirely consistent. Monk and King (1994) found a positive overall association between subject matter study and student achievement but with differential effects for different types of students. Students in advanced math classes, for example, benefited from teachers with more math preparation, but students in remedial math classes did not perform better when they had teachers who had taken more math. Teachers' courses in undergraduate math *pedagogy* contributed more to explaining student performance gains than did undergraduate math courses.

Quality of Student Teaching Experiences

Sources of evidence about TPP quality often include the minimum number of required hours of fieldwork, which refers to student teaching in schools, as well as simulations, case studies, observations, and analyses of teaching, curricula and student work. Other, less frequently used sources of evidence include surveys of teacher candidates about their student teaching experiences, reviews of fieldwork policies and other documents, and records of student teaching observations. These latter sources provide greater insight into the quality of fieldwork experiences than do simple numbers of hours. However, as Coggshall, Bivona, and Reschly (2012) point out, surveys rely heavily on perception rather than reality. Document reviews may provide information about the structure, format, requirements, and expectations of student teaching, but like syllabi reviews, they may reveal more about intentions rather than about actual practices. Finally, some evaluations collect information about the qualifications of fieldwork mentors.

Evidence about the quality of student teaching experiences is central to TPP evaluation. Practicing teachers often view clinical experiences as the most powerful component of teacher preparation, but more research is needed to identify the specific qualities of the field experience that enhance teacher preparation (Wilson, Floden, and Ferrini-Mundy, 2001). Boyd, Grossman, Lanford, Loeb, and Wyckoff (2008) present initial evidence that high-quality student teaching experiences, with strong oversight by the TPP program rather than the host school, have positive effects on program graduates' impact on student achievement. More specifically, new teachers had higher student achievement gains in their first year of teaching if they graduated from programs that (1) were actively involved in selecting field placements, (2) had minimum experience thresholds for cooperating teachers, and (3) required supervisors to observe student teachers at least five times. Ronfeldt (2010) found that teacher retention rates and student achievement were higher among teachers who had preservice field placements in easier-to-staff schools (those that served relatively privileged student populations), even if those teachers ended up working in the most difficult-to-staff schools (those with high proportions of poor, minority, and low-achieving students). These kinds of empirical findings point to specific aspects of the student teaching experience that may be important to assess when conducting TPP evaluations.

Faculty Qualifications

Measures of faculty qualifications included in TPP evaluations usually consist of data on the numbers of instructors that are full-time faculty, adjunct faculty, graduate teaching assistants, and the like. Data are often collected on the numbers of practitioners in K-12 schools who provide instruction and supervision for candidates during fieldwork. Information is also sometimes collected on the proportion of faculty with doctorates or with exceptional expertise and contemporary professional experiences relevant to their assignments and roles. Evaluation systems often place value on faculty qualifications, which intuitively seem important, but empirical evidence of correlations with outcomes for aspiring teachers is lacking.

One problem is that TPP evaluations tend to take into account only the qualifications of the faculty of the school of education, when in fact prospective teachers take most of their content courses (e.g., biology or math) in other departments. When judging faculty, a TPP evaluation will ideally pay attention to non-education faculty too, and will gather evidence of their qualifications to prepare prospective K-12 teachers.

Output Measures

A variety of output measures are used to gauge how well TPPs are preparing new teachers who are employable and effective in the classroom. Sources of evidence typically include teacher licensure test results, surveys of program graduates and their employers, and measures of graduates' impact on the learning of students in their classrooms.

Teacher Licensure Tests

Most TPP evaluations take into account the results of program participants on teacher licensure or certification tests (which should not be confused with the certification tests of the National Board for Professional Teaching Standards; see National Research Council, 2008). These licensure exams may take the form of paper-and-pencil or computerized tests, may use multiple-choice as well as essay questions, and may cover basic skills, subject matter knowledge, and/or pedagogical knowledge. Different tests are used to evaluate candidates' knowledge in more than 25 credential areas, such as elementary education, chemistry, art, and special education. More than 600 teacher tests are currently in use (National Research Council, 2010). Variations in how these tests are developed and used make it difficult to generalize about them or compare results across states. For example, even states that use the same test often set different cut scores for passing. In addition, candidates take the tests at different points in their preparation program and thus have completed varying amounts of coursework and student teaching experience at the time of testing.

Teacher licensure tests have been the subject of public concern. Some critics have complained that the tests are too easy and the cut scores too low, so that the percentage of teacher candidates passing is higher than it should be (e.g., Fowler, 2001; Crowe, Allen, and Coble, 2013; Sawchuk, 2013a). However, it is not possible to determine from high pass rates alone whether the test is too easy, whether TPPs are doing a good job preparing teacher candidates, or whether people entering TPPs tend to have the minimum competencies to begin with. On medical board exams, for instance, the percentages of first-time test-takers who pass are typically in the mid-90s in many specialties (American Board of Internal Medicine, n.d.), but one seldom hears concern about whether this is an indication of the relatively low quality of doctors. Rather, it tends to be assumed that medical schools are highly selective in the applicants they admit and that medical preparation programs are of high quality.

Licensure tests are designed to protect the public from those who have not mastered the *minimum* competencies necessary to perform their job (Mehrens, 1990); they are used to make dichotomous (pass/fail) decisions. Licensure tests are not designed or intended to predict the level of a person's future performance. Still, researchers have explored such relationships in the teaching field. Some studies have shown that teachers with higher scores on licensure tests do have a positive impact on student test scores, especially in math (Clotfelter, Ladd, and Vigdor, 2007), but this relationship should not mask the important fact that some teachers who do not perform well on licensure tests still have a positive impact in the classroom (Goldhaber, 2006). Misclassification errors may cause teachers who would be excellent in the classroom to be categorized as not ready or relatively incompetent based on low test scores, while less effective teachers may be categorized as competent by dint of their high exam scores. The National Research Council (2001a) conducted a review of the widely used Praxis tests and found them to be technically sound and "content-valid"—i.e., the tests did assess candidates on skills that committees of practitioners and others deemed useful in the classroom. What the NRC committee found lacking was evidence that "test results relate to other relevant measures of candidates' knowledge, skills, and abilities" (p. 114). Presumably, this includes the relationship between Praxis scores and the difference a teacher makes in actual student learning.

Hiring and Retention

In an era of heightened concern about the economic returns on investments in teacher preparation (or, for that matter, in other higher education programs), data on hiring and placement of teachers is a legitimate component of a broader evaluation of TPP quality. What are the job prospects of TPP graduates? To what extent do TPPs prepare candidates to teach in schools with large proportions of low-income and minority students? Detailed information about supply and demand can inform TPPs about the labor market need for particular kinds of teachers, such as the need for teachers in certain subject areas or grade levels. But while hiring data can be informative, it is important to remember that larger economic and social forces affect the number of teacher vacancies in a particular region and that TPPs have limited control over hiring and placement (Coggshall, Bivona, and Reschly, 2012). For example, Kukla-Acevedo, Streams, and Toma (2009) found that many TPPs are geographically isolated and that most of their graduates end up teaching at the same few schools. Schools in a certain geographic area may suffer from high teacher turnover for reasons unrelated to TPP quality.

Retention data, which tracks whether new teachers stay in the teaching profession, can also be informative. If a particular TPP produces an unusually high percentage of graduates who leave after their first few years of teaching, this may be a sign that something is amiss in either the TPP's selection process or preparation program. However, reviews of research on why teachers leave the profession do not indicate that teachers' preparation has much to do with these decisions (Ingersoll and Smith, 2003; Johnson, Berg, and Donaldson, 2005; Berry, Smylie, and Fuller, 2008).

Many leave for family or other personal reasons. The most likely in-school factors affecting decisions to leave the profession include poor interactions with principals and staff, lack of a sense of efficacy in the classroom, poor physical conditions of schools, lack of time, heavy teaching loads, large class sizes, and pay. Over 50 percent of the former California teachers surveyed by Futernick (2007) mentioned "bureaucratic impediments" and "poor staff support" as reasons for leaving, but only about 13 percent mentioned "poor teacher prep coursework."

Teacher Performance Assessments

Unlike traditional multiple-choice teacher tests, teacher performance assessments typically include observations of actual teaching and portfolios of lessons and student work and are designed to capture how teacher candidates apply what they have learned in the classroom. Some experts have advocated the use of performance assessments not only to measure the skills of individual teacher candidates, but also to evaluate the quality of TPPs when results of their enrollees are aggregated (Pecheone and Chung, 2006; Darling-Hammond, 2010). There is some evidence that candidates' scores on performance assessments can predict their subsequent effectiveness in the classroom (Darling-Hammond, Newton, and Wei, 2013).

Currently, the most widely used performance assessment for teacher candidates during their final year in preparation is the edTPA, developed at Stanford University. The assessment is administered at or near the end of a candidate's pre-service experience. Candidates are videotaped while teaching three to five lessons from a unit of instruction to one class of students. Evidence of teacher competence includes video clips of instruction, lesson plans, student work samples, analyses of student learning, and reflective commentaries by the candidate. The results of this assessment are reported back to the TPP.

Despite the appeal of performance assessments, they have drawbacks as tools for TPP evaluation. For one thing, they are costly to administer and score on a large scale. Further, as Greenberg and Walsh (2012) have noted, what may be a very good culminating exercise for a TPP to administer is not necessarily a sufficiently valid and reliable measure of either the skills of an individual teacher or the quality of a program. For example, the edTPA allows candidates to choose the lessons they will deliver, rehearse as many times as they wish, and edit the videotape of their teaching. This type of editing raises questions about whether the resulting project is a valid assessment of a candidate's teaching skills.

Surveys of Graduates and Employers

Another way of collecting information about the quality of TPPs is to survey the graduates themselves, after they have secured a teaching job, and their principals. Graduates are typically asked questions about how well-prepared they feel to handle certain demands of their job (or how much opportunity they had to learn how to handle those demands), such as meeting the instructional needs of English language learners, conducting student assessments, and effectively teaching the subjects for which they are responsible. Graduates are also often asked about their student teaching experience—for example, whether they received useful feedback from their mentors after teaching a lesson. Principals might be asked about the extent to which a new teacher was prepared to effectively implement discipline and class management procedures, meet the learning needs of students with disabilities, and fulfill other responsibilities of teaching (Coggshall, Bivona, and Reschly, 2012).

A general concern about using surveys for evaluation is that they rely on individuals' perceptions, which do not necessarily comport accurately with reality. Still, Jacob and Lefgren (2008) found that principals can effectively identify which teachers produce the largest and smallest student test score gains in their schools, although they are far less able to distinguish between teachers in the middle of the distribution. Harris and Sass (2009) found a positive correlation between principal ratings and teachers' impact on growth in student test scores as measured by value-added models (explained below), although the correlations are weak (in the .15 to .30 range). While teachers' past value-added results predict their future ones, principals' subjective ratings can provide additional information and substantially increase this predictive power.

Value-Added Models

The most recent innovation in TPP evaluation is the use of value-added models (VAMs) that purport to measure the impact on student achievement of new teachers prepared by a particular program. VAMs are statistical techniques that measure student achievement gains on standardized tests while controlling for differences in students' prior achievement and other background factors such as family income that are not under teachers' control. In this way, VAMs are intended to help "level the playing field" among teachers with different class compositions and enable valid comparisons of teachers' effectiveness (National Research Council and National Academy of Education, 2010). Typically, the value-added estimate for a teacher is the difference between the actual improvement and the statistically expected improvement of his or her students. When VAMs are used to evaluate TPPs, the process goes a step further by analyzing the extent to which graduates of particular TPPs have raised their students' scores.

Research on using VAMs for evaluating TPPs is still in the early stages, and there are still many more questions than answers. One important question is whether the differences in VAM estimates among TPPs are more a function of the training the graduates receive or the population of teacher candidates they attract (Goldhaber and Liddle, 2012). Given the relatively small differences between programs and the variations and tradeoffs associated with different statistical models, state-level decision makers and K-12 administrators should avoid placing too much weight on VAM scores when making critical decisions about program accountability or hiring (Koedel, Parsons, Podgursky, and Ehlert, 2012). Box 2-1 lays out some of the strengths and weaknesses of VAMs and considers the most appropriate and valid ways that they might be used in TPP evaluation. For an in-depth review of the many issues involved in the use of VAMs for TPP evaluation, see Meyer, Pyatigorsky, Rice, and Winter (2013).

BOX 2-1
The Potential Value and Risks of Using VAMs for TPP Evaluation

Value-added models (VAMs) hold promise for moving TPP evaluation forward. They are an important development because they represent the only approach to TPP evaluation that actually judges TPP quality based on the effectiveness of their graduates in producing growth in student achievement, while controlling for out-of-school factors that are not subject to teachers' influence. The results can help determine which TPPs produce the most effective teachers and can spur weaker providers to emulate those programs' practices. VAMs allow for repeated measurement of a relevant, meaningful outcome of interest, and if results are stable or show clear trends over time, they offer the potential to improve programs by providing feedback in a domain in which data have not been available in the past (Reusser, Butler, Symonds, Vetter, and Wall, 2007; Gansle, Noell, and Burns, 2013).

Critics argue that the value-added approach is fraught with methodological difficulties, which render the results untrustworthy. Many of the difficulties relate to the general use of VAMs for measuring teacher effectiveness. A joint report of the National Research Council and National Academy of Education (2010) details some of the problems, including concerns about the standardized tests that provide the raw data for value-added analyses and technical problems related to bias, imprecision, and instability. There are also issues of transparency and public understanding of the results.

Most of the research on the use of VAMs specifically for TPP evaluation has focused on how well these models differentiate between different TPPs. Findings have been mixed. Several studies have found significant variation across

Other Potential Sources of Evidence

Other sources of evidence are not currently being used for TPP evaluation but might be in the future. For example, the Measures of Effective Teaching (MET) Project, funded by the Bill & Melinda Gates Foundation, emphasizes the importance of using well-designed classroom observations as one component of a teacher evaluation system. There are a number of classroom observation protocols based on years of research, such as the Classroom Assessment Scoring System (Pianta and Hamre, 2009), the Framework for Teaching (Danielson Group, 2013), the Mathematical Quality of Instruction (Hill, Ball, and Schilling, 2008), and the Protocol for Language Arts Teaching Observations (Grossman et al., 2010). The MET Project has demonstrated that it is possible to identify effective teachers by combining teacher observation data with student surveys (using the Tripod surveys developed by Ferguson and Ramsdell, 2011) along with VAM results (Bill & Melinda Gates Foundation, 2013). This type of system that uses multiple measures to evaluate individual teachers raises intrigu-

TPPs in the average effectiveness of the teachers they produce (Boyd, Grossman, Landford, Loeb, and Wyckoff, 2008; Noell and Gleason, 2011; Goldhaber and Liddle, 2012; Henry, Bastian, and Smith, 2012; Plecki, Elfers, and Nakamura, 2012), but a few other studies have found only very small differences between programs (Mason, 2010; Koedel, Parsons, Podgursky, and Ehlert 2012). Other problems include incomplete data and the fact that methodological variations in statistical models can produce different judgments about TPP effectiveness (Mihaly, McCaffrey, Sass, and Lockwood, 2012). It is difficult to separate TPP effects from school-level factors (e.g., the culture at a school, the effectiveness of principals). The fact that some schools tend to hire teachers from particular TPPs makes this especially challenging (Mihaly, McCaffrey, Sass, and Lockwood, 2012). Another complexity is whether the VAM accounts for the possibility that training program effects decay or potentially grow over time; while it makes sense to evaluate TPPs based only on the most recent three cohorts of program graduates, limiting analyses to a few cohorts creates significant sample size problems if the programs are small (Goldhaber and Liddle, 2012).

As Harris (2011) explains, many of the most serious criticisms about VAMs assume they will be used as the basis for high-stakes decisions about individual teachers, such as decisions on hiring, firing, and pay. TPP evaluations avoid this problem by aggregating results from many teachers to make judgments about programs rather than individuals (Bryk, 2012). The odds of making valid decisions using VAMs can be further increased if the results are based on two or more years of data and if the VAM is just one of the multiple measures in an evaluation system (Harris, 2011; Meyer, Pyatigorsky, Rice, and Winter, 2013). Evaluation systems could use a VAM as an initial filter or trigger to identify the very lowest-performing TPPs that need further examination using additional methods.

ing possibilities for TPP evaluation; if such a system were implemented widely, then the resulting data could be fed back into new teacher preparation programs.

Other potential sources of evidence for TPP evaluation include new types of teacher assessments now being developed. For instance, the Mathematical Knowledge for Teaching (MKT) assessment developed by Hill, Rowan, and Ball (2005) assesses teachers' ability to *teach* math rather than just their mastery of elementary or middle school math content. In short, MKT tests the kind of math knowledge that supports teaching math to students, including the ability to explain how mathematical procedures work, to effectively define a mathematical term, and to recognize mistakes students are likely to make. Hill and colleagues found a correlation between teachers' MKT scores and their students' achievement. The test was designed for professional development but may also hold potential for TPP evaluation.

EXISTING SYSTEMS FOR EVALUATING TPPS

The first part of this chapter reviewed various types of evidence used to measure the quality of TPPs in the United States. We now turn to describing some of the larger systems for evaluation, each of which uses multiple sources of evidence. The committee has categorized existing TPP evaluation systems based on which entity is doing the evaluating: (1) the federal government; (2) national nongovernmental accrediting bodies; (3) state governments; (4) media outlets and other independent organizations; and (5) the TPPs themselves.

Federal Approaches

The federal government has become increasingly involved in evaluating TPPs, with the U.S. Department of Education (ED) moving from virtually no direct involvement in teacher preparation, to the introduction of accountability systems that require TPPs and states to report vast amounts of data, and finally to offering monetary incentives for states to develop innovative systems that measure how program graduates perform on the job.

The Higher Education Act

The legislative cornerstone of federal policy on the evaluation of TPPs is the 1965 Higher Education Act (HEA), part of President Johnson's Great Society efforts. The original legislation contained no accountability or reporting requirements for TPPs. Then, in the 1990s, a number of well-publicized news reports highlighted the poor reading and writing skills of some teachers and the fact that many teachers were teaching without having passed certification tests. Report cards and other forms of public accountability became increasingly popular tools for rating K-12 schools, and soon the idea was picked up by higher education policy makers (Russo and Subotnik, 2005). The 1998 reauthorization of the HEA added a new Title II, which created an accountability system that called on TPPs to collect data on a wide range of indicators (about 400 data points altogether) and required states to compile the results into report cards. At the program level, the state report cards included indicators such as pass rates on teacher licensure tests and admission requirements, and identified which TPPs were low-performing. The report cards also included state-level teacher education statistics and policies, such as the number of educators teaching on licensure waivers, information on alternative routes into teaching, and procedures for identifying and helping low-performing TPPs.

The next reauthorization of the HEA in 2008 tweaked Title II by adding a few more categories of measures. For example, in addition to reporting the percentage of teachers passing certification tests, states had to report average scores on these tests, which would allow for more nuanced comparisons of programs within states or across states using the same test. States also had to report on indicators related to student teaching requirements, training on the use of technology in the classroom, and progress in preparing teachers in high-need subjects (Sawchuk, 2011).

Critics of the HEA Title II system pointed to large amounts of missing data and noted that "much of the reporting is inconsistent, incomplete, and incomprehensible" (Huang, Yi, and Haycock, 2002, p. 4). Michelli and Earley (2011) asserted that the promised results of the 1998 reauthorization were never realized. They noted that because of differences in cut scores and difficulty of the certification tests, "a pass rate in one state could not be directly compared with a pass rate in another state and furthermore there were different tests in different states, making comparison across states impossible" (p. 9). Crowe (2010) concluded that HEA Title II does not represent a real system of accountability, based on the high rates of teachers passing the tests and the small percentage of TPPs identified as low-performing (only about 2 percent).

Race to the Top

In the Obama Administration, Secretary of Education Arne Duncan took the next big step in federal involvement in the evaluation of TPPs. After a few high-profile speeches in which he said teacher education was "doing a mediocre job," Secretary Duncan announced plans for a revolutionary change in TPP evaluation (Duncan, 2009). The Race to the Top (RTTT) program would "reward states that publicly report and link student achievement data to the programs where teachers and principals were credentialed." He called for a shift in focus from program inputs to program outputs—measuring program graduates' effectiveness in the classroom during their first few years of teaching by looking at the learning gains of their students. Typically this would be done using VAMs, and he held up the example of Louisiana, the first state to have put such a program in place.

The Obama Administration used economic stimulus spending as the means to pursue this policy. The American Recovery and Reinvestment Act set aside $4.35 billion for competitive awards under RTTT to states that could demonstrate progress on four goals:

1. Adopting standards and assessments that prepare students to succeed in college and the workplace and to compete in the global economy
2. Building data systems that measure student growth and success and inform teachers and principals about how they can improve instruction
3. Recruiting, developing, rewarding, and retaining effective teachers and principals, especially where they are needed most
4. Turning around the lowest-achieving schools

As of late 2012, 16 school districts and 22 states had won RTTT grants (U.S. Department of Education, 2011a, 2012). According to early analyses, implementation of the reforms at the state level has been uneven and characterized by numerous delays, primarily due to lack of capacity to implement reforms, continued funding problems, or resistance from educational organizations to new methods of teacher evaluation (Boser, 2012; Rentner and Usher, 2012).

The Obama Administration is also pursuing its reform agenda through proposed changes to HEA Title II. In the fall of 2011, the Administration laid out its plan for evaluating the effectiveness of TPPs (U.S. Department of Education, 2011b). Most importantly, this document asserts that it is possible to differentiate between TPPs using VAMs and that such a process would accomplish the following goals:

> [T]he federal government can shine a spotlight on exemplary models for replication and scaling. It can and should address challenges that for too long have been neglected by supporting state-level policies that reward the best programs, improve the mid-performing programs, and transform or ultimately shut down the lowest performers (p. 8).

The Administration's plan embraces three primary measures of TPP program quality:

1. Achievement growth of elementary and secondary school students taught by program graduates, through VAMs
2. Job placement and retention rates
3. Surveys of program graduates and their principals

The Administration's 2012 budget request for HEA Title II introduced the Presidential Teaching Fellows program, which would provide funds to states for scholarships in return for states instituting the aforementioned accountability measures. The Administration also sought to ease some of the onerous data requirements under the previous two HEA reauthorizations in favor of a more streamlined approach that focuses on outcomes, as in the three measures of quality listed above.

Various stakeholders discussed these measures in the Department of Education negotiated rulemaking sessions for several months in 2012, but the negotiations ended in a deadlock about which specific measures to use. The Department of Education declined to extend the rulemaking process beyond April 2012, so it will now craft its own rules.

Government Evaluation of Teacher Preparation in Other Countries

A common question about the evaluation of TPPs is how other countries do it and whether their experiences can be relevant and informative in the U.S. context. We do not assume that such comparisons are easy to make or that methods used elsewhere can be readily imported to the specific context of teacher preparation here; but we do think that comparisons provide useful insights.

Tatto, Krajcik, and Pippin (2013) argue that national-level efforts to strengthen accountability and quality assurance in higher education appear to be increasing and intensifying across the globe. Assessments of teacher preparation programs tend to be encompassed within broader evaluations of higher education institutions. In Finland, for example, there are no systems aimed specifically at external evaluation of TPPs; that task is left to the programs themselves. Singapore has a highly centralized structure for teacher preparation, which enables the country to manage the quantity and quality of its teaching workforce and carefully monitor problems or areas needing improvement as they arise. England has a strong national TPP evaluation system, as described by Furlong (2013) and summarized in Box 2-2.

BOX 2-2
Evaluation of Teacher Education in England[1]

Government inspection of public services, including teaching, has a long history in England. From the earliest days, the government in England has insisted on maintaining tight control over the numbers of teachers in training and the quality of that training. In the 1990s a new body—the Office for Standards in Education, Children's Services and Skills (Ofsted)—was formed, and its role was "collecting objective evidence about schools and reporting on their failings" (Lawlor, 1990). During the 1990s, another notable development was the government's insistence that student teachers spend a substantial proportion of their training in schools, and that schools have a leading role in designing, overseeing, and assessing initial teacher education in partnership with colleges.

How Ofsted inspections work. England's inspection system is evolving and has become more rigorous with time. Under the current system, there are two inspection processes: a majority system of "routine" inspections (conducted about every six years), and a minority system of "no notice" inspections.

For a "routine" inspection, the TPP to be inspected is contacted about one week prior to an inspection visit, and then a team of inspectors conducts a site visit. It is their task to inspect the program wherever it is delivered; this means that in university-led courses, schools that work in partnership with the program are subject to the same inspection process as the university itself. Following the site visit, inspectors will assign the program with a single grade for its overall effectiveness: Grade 1 (outstanding), Grade 2 (good), Grade 3 (requires improvement), and Grade 4 (inadequate). Inspectors' evaluation of a program's effectiveness must take into account three key indictors:

1. **Outcomes for trainees** (assessing the quality of teaching of a sample of current and former trainees, and considering trainees' program completion rates and employment rates)
2. **Quality of training across the school–TPP partnership** (assessing the consistency, coherence, and quality of all aspects of the training, e.g., the quality of placements and quality of mentoring, through direct classroom observation)
3. **Leadership and management of the school–TPP partnership** (assessing how well leaders and managers are focused on improving or sustaining outcomes for trainees, e.g., whether strong partnerships between the school and TPP exist, whether effective monitoring and evaluation

Accreditation by Nongovernmental Bodies

In the U.S., the federal government does not accredit or approve TPPs. This is done by national, nongovernmental accrediting bodies and state governments. Both types of review are meant to assure prospective teachers, employers, policy makers, and the public that TPPs meet a certain processes are in place, and whether the partnership meets current statutory teacher training requirements).

A minority of inspections are "no notice" or "focused monitoring" inspections, a concept introduced in 2013. Currently, these inspections focus solely on the quality of phonics training in TPPs (i.e., trainees' skills in teaching early reading using phonics).

Those TPPs that are judged to "require improvement" undergo a reinspection within 12 months, and if upon reinspection the program is still judged to require improvement, it will be considered "inadequate." Programs that remain "inadequate" after one further reinspection may be withdrawn by their universities.

Inspection results are used in several other major ways. Individual program reports are published on government websites, so they influence the "market choices" of prospective students, and individual institutions use these results as an indicator of departmental and faculty quality. Also, the Ofsted inspectorate issues more general reports on "the state of the nation," particularly the *Chief Inspector's Annual Report*, that is presented to Parliament and is widely publicized. Data from inspections are also used to construct league (ranking) tables in the annual *Good Teacher Training Guide* (Smithers and Robinson, 2011). These league tables, though somewhat questionable in terms of statistical robustness, are widely publicized and read by prospective students, TPPs, and the government.

After reviewing the limited evidence on the impacts of the inspection system, Furlong (2013) concludes that teacher education in England has improved in many ways over the last 20 years—certainly in terms of its consistency and coherence. For one, the inspection framework ensures that all programs conform to national standards. But, according to Furlong, important questions remain. Has the inspection system actually improved program quality? (Many would argue that teacher education in England has become too narrow and too technical.) Are the positive changes over time in fact attributable to Ofsted, or to the fact that student teachers have been spending a substantial proportion of their training in clinical settings since the 1990s? Finally, to what extent is the Ofsted experience transferrable to other countries? Compared with many other countries, and particularly the United States, England's system of teacher preparation is a highly centrally managed system, which makes this type of inspection system possible.

[1] Box 2-2 draws from a comprehensive description of England's teacher preparation evaluation system prepared by Furlong (2013).

standard of quality. The processes of accreditation and state approval are similar and often overlap in ways that can be confusing.

Like other programs that prepare candidates for various professions, TPPs have their own accrediting bodies, which make judgments about the quality of programs to prepare teachers going into preK-12 settings. The general TPP accreditation process is summarized in Box 2-3.

BOX 2-3
The TPP Accreditation Process

Accreditation of TPPs typically works as follows (U.S. Department of Education, n.d.):

1. ***Standards.*** The accrediting agency, in collaboration with teacher education programs, establishes standards.
2. ***Self-study.*** The institution or program seeking accreditation prepares an in-depth self-evaluation study that measures its performance against the standards established by the accrediting agency.
3. ***On-site evaluation.*** A team selected by the accrediting agency visits the institution or program to determine first-hand if the applicant meets the established standards.
4. ***Publication.*** Upon being satisfied that the applicant meets its standards, the accrediting agency grants accreditation or pre-accreditation status and lists the institution or program in an official publication with other similarly accredited or pre-accredited institutions or programs.
5. ***Monitoring.*** The accrediting agency monitors each accredited institution or program throughout the period of accreditation to verify that it continues to meet the agency's standards.
6. ***Reevaluation.*** The accrediting agency periodically reevaluates each listed institution or program to ascertain whether continuation of its accredited or pre-accredited status is warranted.

Changes in Accreditation of Teacher Preparation Programs

Until recently, the main accrediting body for TPPs was the National Council for the Accreditation of Teacher Education (NCATE). The organization was launched in 1954 by a coalition of professional organizations from across the education community. NCATE aimed to professionalize teaching by establishing national standards for accreditation, similar to the process used in medicine and law (Vergari and Hess, 2002).

The NCATE standards were developed by representatives of its 33 member organizations. Over time, NCATE changed its standards in response to criticisms that they were not rigorous enough or not supported by evidence that NCATE-accredited programs produce greater student learning (Vergari and Hess, 2002). For instance, NCATE released updated standards in 2000 that put greater emphasis on teacher candidates' demonstrated knowledge of their subjects and their skills in teaching subject matter content to children (Bradley, 2000). TPPs seeking accreditation had to assess their students' performance regularly by gathering evidence from projects, journals, videotapes and other work, and to share the results with accreditors. NCATE accreditation also started taking into account outcome measures, including prospective teachers' pass rates on licensure tests, evaluations conducted during their induction periods, and reports from employers.

In 1997, the Teacher Education Accreditation Council (TEAC) was founded with the support of presidents of small, independent colleges who thought that NCATE's prescriptive standards favored larger schools and neglected program outputs. The distinguishing feature of TEAC's approach was that TPPs could set their own standards, within TEAC guidelines. The organization's audits focused on three quality principles: evidence of teacher candidate learning, evidence that the assessment of such learning is valid, and evidence of the program's own continuous improvement and quality control (Teacher Education Accreditation Council, n.d.).

By 2010, just over half of the teacher education programs in the United States were accredited by NCATE or TEAC, with the bulk of these accredited through NCATE (National Research Council, 2010). In October of that year, the boards of NCATE and TEAC unanimously agreed to merge. The resulting Council for the Accreditation of Educator Preparation (CAEP) was led by a design team with equal representation from the two organizations. One of its goals was to enable the education profession to speak with a single, unified voice about the preparation of teachers. Another was to boost the status of the profession by raising standards for the evidence the field relies on to support its claims of quality (NCATE, n.d.).

In summer 2013, the CAEP Commission on Standards and Performance Reporting submitted a set of recommended standards for TPP accreditation to the CAEP Board of Directors, which the Board approved at the end of August. These standards, more specific and more outcome-focused than the previous ones (Council for the Accreditation of Educator Preparation, 2013a,b), fall into five broad categories: (1) equipping candidates with content knowledge and appropriate pedagogical tools; (2) working in partnership with districts to provide strong student-teaching experience and feedback; (3) recruiting a diverse and academically strong group of candidates, and developing them through all phases of preparation from admission to program completion; (4) demonstrating a program's impact, using measures of students' academic achievement, indicators of teaching effectiveness in the classroom, satisfaction of employers of program graduates, and satisfaction of program graduates themselves; and (5) maintaining a quality-assurance and continuous improvement system. The CAEP standards also suggest the types of evidence that a program can submit to demonstrate that it has met each standard. CAEP will offer TPPs a choice of accreditation processes to follow, which encompass the variety of pathways offered previously by NCATE and TEAC.

Some of the new CAEP requirements are controversial. A diverse group of organizations ranging from the American Federation of Teachers to the Council of Chief State School Officers are in agreement about the need for more selectivity in teacher preparation, but others have expressed various concerns, including the possibility that selectivity stipulations may harm the diversity of the teaching force (see, e.g., Sawchuk, 2013d). Note that neither of the two accrediting bodies that merged to form CAEP previously had a standard on selectivity. CAEP accreditation will now incorporate admissions requirements, including an average GPA of its accepted cohort of program candidates that meets or exceeds the CAEP minimum of 3.0, and the cohort's average performance on nationally normed achievement tests.

CAEP is aiming for a nuanced approach to standards on recruitment and selectivity. The new admissions criteria are being introduced gradually: the top 50 percent of the distribution by 2016-2017; the top 40 percent of the distribution by 2018-2019; and the top 33 percent of the distribution by 2020. In addition, programs that do not conform to CAEP's standard (such as those that have open enrollment) can offer evidence, such as the admitted cohort's positive impact on preK-12 student learning, to meet the standard for "candidate quality, recruitment, and selectivity."

Research on accreditation has found little empirical evidence of its impacts (e.g., Tamir and Wilson, 2005). Goodlad (1990) concluded that accreditation produced a "stifling conformity" and lack of innovation. National accreditation was seen as important by regional institutions, Goodlad observed, but less so by more prestigious flagship and major public and private universities. Findings are mixed about whether accredited programs produce higher-quality teachers than non-accredited ones. Gitomer and Latham (1999) found that NCATE accredited schools were more successful than non-NCATE accredited schools in getting their students to pass the Praxis tests—even when students from the non-NCATE schools had higher scores on the SAT and ACT. But Ballou and Podgursky (1999) uncovered no appreciable difference in licensure exam results from NCATE and non-NCATE accredited schools of education. They also made the following point:

> [T]here is little evidence that teachers trained in NCATE-accredited schools conduct themselves more professionally, are more likely to continue teaching, or experience more satisfaction with their career choice. Perhaps more revealing, there is no evidence that those hiring new teachers think so either. The percentage of non-NCATE applicants who found a teaching job was as high as among NCATE applicants. The jobs they obtained paid as well (p. 47).

CAEP intends to have a stronger impact on teacher education through its more rigorous and evidence-based approach to accreditation. The

CAEP standards and system call for constant review of their impact, including research on the effects on the field, which will be discussed in a series of annual reports (for the first baseline annual report, see CAEP, 2013). The annual reports will be a source of information on trends and conditions in teacher preparation and will include findings on strengths, weaknesses, and areas for improvement in the CAEP evaluation system itself.

Comparison with Nursing Preparation

Box 2-4 describes the accreditation process used for nursing preparation programs, which has some commonalities with accreditation for teacher preparation but also some notable differences.

State Review Systems

States have the primary responsibility for establishing teacher policies, including standards for teacher education and requirements for certification. States exercise authority over TPPs through program approval processes that allow for graduates who meet state criteria to be automatically recommended for certification at the program's discretion. (An individual teacher can still apply directly to the state department of education for certification.) Every state but Arizona requires its public TPPs to undergo some sort of approval process, and some extend those requirements to private institutions (Education Commission of the States, 2013). But program approval processes vary widely across states, and there is currently no systematic information or objective analysis of how each state carries out its process.

Many states accomplish TPP approval under formal partnerships with the national accrediting body (CAEP). The purpose of those partnerships is to save states and institutions time and expense by eliminating duplication of effort and paperwork in conducting state approval and national accreditation (NCATE, n.d.). Other states conduct reviews and make program decisions on their own. An informal scan of state websites reveals that these states tend to use a process that is similar to national accreditation but is based on the state's own standards for TPPs. The state approval process usually begins with the TPP producing a self-study report, followed by an on-site evaluation by state reviewers, an approval decision by the state, monitoring, and periodic reevaluation.

Several states are on the forefront of change in TPP evaluation. Six states that won Race to the Top grants—Florida, Louisiana, North Carolina, Ohio, Tennessee, and Texas—are recognized for having innovative systems in place or in development (Coggshall, Bivona, and Reschly, 2012). These states are all using VAMs, along with other measures, to

BOX 2-4
Accreditation of Nursing Schools[1]

On the surface, there are striking similarities between nursing and teaching. Both fields are female-dominated, experience fluctuations in supply and demand, must deal with new challenges of serving populations with changing demographic characteristics, and are grappling with the integration of new technologies. Perhaps most significantly, good nurses, like good teachers, must have a blend of technical (content) knowledge and adaptability to perform in complex and dynamic situations; in fact, both professions require a high concentration of tacit skills, or techniques that cannot be easily explained or written down, along with codified or explicit knowledge (Polanyi, 1966; Murnane and Nelson, 1984).

In nursing, two main national associations act as accrediting bodies: the Accreditation Commission for Education in Nursing (ACEN) and the Commission on Collegiate Nursing Education (CCNE) of the American Association of Colleges of Nursing. There are also separate, smaller accrediting organizations for programs in some specific nursing areas, such as midwifery. The accrediting bodies set standards that guide curriculum, program implementation, and evaluation of nursing preparation programs. Although accreditation is voluntary, 96 percent of preparation programs that offer a baccalaureate degree in nursing are accredited by one of the major accrediting organizations. Of the programs that offer an associate's degree, however, only 52 percent are accredited. Usually, these are vocational school or community college programs that are accredited by the state board of nursing but not by one of the national accrediting organizations. The field is moving toward getting state boards of nursing to require the accreditation of all nurse preparation programs (National Council of State Boards of Nursing, 2012).

The accreditation process for nursing schools is similar to that followed by TPP accreditation agencies. It consists of an initial period of program self-study, peer review, site visits, and a rating and/or acceptance decision by the accrediting agency, followed by periodic monitoring and oversight. Nursing schools are reevaluated every five to ten years, and in intervening years they submit annual reports to the accrediting agencies (Heydman and Sargent, 2011).

Separate from the accreditation agencies are state boards of nursing and licensure. Similar to the process for licensing individual teachers, state-level governmental entities regulate the practice of nursing within a state, including nurse education, and oversee license approval and renewal for registered nurses. The primary purpose of these boards is to protect the public; the boards grant permission to an individual to engage in the practice of nursing under protected titles (e.g., registered nurse, nurse practitioner, clinical nurse specialist, nurse anesthetist, or nurse midwife).

The National Council of State Boards of Nursing, the umbrella organization for all the boards, administers the National Council Licensure Examination for RNs (NCLEX-RN), a national assessment that graduates must pass to be eligible to practice nursing. In addition, nursing schools must meet minimum pass rates on this exam, which are set by individual state boards; a school's failure to meet a minimum pass rate could result in a review of the program by the state board of nursing. Therefore, nursing programs have a responsibility to prepare their students to pass the national licensure exam. The pass rate for first-time test takers is around 90 percent (National Council of State Boards of Nursing, 2012). Programs commonly administer practice exams to gauge how well their nursing candidates are performing and to identify areas where curriculum may need to be addressed. There are also specific licensure tests for nursing subfields and for advanced practice registered nurses (APRNs), who have graduate-level training.

Despite the many apparent similarities, there are some notable differences between the evaluation of nursing and teaching programs. The nursing school evaluation system is more coherent and tighter across states and institutions, and the standards are widely adhered to and accepted by nursing candidates, educators, and professionals in the medical fields. All registered nurse candidates nationwide take the same licensure test, and all candidates within a particular specialty, or who wish to be certified as APRNs, take the same licensure test. This is in contrast to teacher candidates, who take different tests in different states.

Unlike the movement in the teaching field to examine the link between student performance and teacher preparation programs, nursing has not tried to link patient outcomes to nurse preparation programs. This is because of the many confounding variables in patient care. On any given day, two to three nurses may treat a single patient. Nurses care for hundreds of patients in a year, and they work in teams with many other health professionals who also have an impact on patient outcomes. Under such a system, linking specific patient outcomes to specific nurses is not feasible.

Part of the differences in evaluation and accreditation between teacher and nursing education may have to do with the extent of agreement about the particular knowledge and skills that nurses and teachers need to be effective on the job. There is a good deal of agreement over the competencies that need to be taught to produce competent nurses. That may be because nurses' work involves relatively more discrete, concrete, and technical skills (for example, giving an injection or physical assessment or assisting in surgery). Also, perhaps nursing must adhere to a stricter set of standards because inadequate training could be life threatening. This stands in contrast to teacher education, where there is debate over standards and a questioning of the value of accreditation and other review systems, as well as the rigor of licensure tests. As Wilson (2011) states, "Three major reviews of research on teacher preparation in general have all drawn the same unsatisfying conclusion that we know very little about effective teacher preparation based on empirical research" (p. 3).

[1] Box 2-4 draws in part from a paper by Johnson and Pintz (2013) commissioned for this report.

evaluate TPPs. All plan to provide more and better information to TPPs and the public based on their multiple-measure systems.

There has been little research on the effects of state approval systems on teacher education and other aspects of the education system. Levine (2006) found that mediocre TPPs easily receive state approval because the process in most states is procedural rather than substantive. He found that states tend to examine TPPs in a cursory way, without looking at quality. In a presentation to the committee, Aldeman (2012) characterized state accountability as "weak." He reported that out of the 1,400 institutions preparing teachers, only 37 TPPs were identified by states as low-performing in 2011, and 27 states had never identified a single low-performing TPP. Whether this will continue to be the case under the new state approval systems being developed is a question for further research.

Media and Other Independent Organizations

Media outlets, advocacy groups, and other independent organizations develop rankings and ratings of higher education programs to inform potential students, their parents, and other consumers about the programs' quality. By *rankings*, we mean instances whereby organizations collect information about higher education programs, assemble it into an index, and rank the programs in order (first, second, third, and so on). *Ratings* are similar, in that organizations collect various types of information from higher education programs, but rather than being listed in rank order, the institutions are placed into performance categories based on how well they meet certain criteria.

Since the 1980s, rankings have become a prominent part of the U.S. higher education landscape. The *U.S. News and World Report* rankings of colleges and universities are the best known; this media company also produces annual rankings of schools in business, education, engineering, law, and medicine, among others. But the specific programs for teacher preparation within schools of education have only recently been subject to this type of evaluation. Currently, only one prominent evaluation system falls into this category—the review of TPPs conducted by the National Council on Teacher Quality (NCTQ) in collaboration with *U.S. News and World Report* (2013). Below we describe their approach, while realizing that in the future other media and independent organizations may start rating TPPs using a variety of different approaches.

NCTQ/U.S. News Ratings

In 2011, *U.S. News* announced that it would partner with the National Council on Teacher Quality to develop a methodology to rate TPPs

(Morse, 2011). Together they developed a 5-point scale on 18 standards, basing the methodology on deliberations of an expert panel and pilot studies (National Council on Teacher Quality, 2013). The standards focus on how well programs cover aspects of teaching such as early reading instruction; the ability to work with English language learners, students with disabilities, and struggling readers; understanding of the Common Core State Standards; classroom management skills; lesson planning; and knowledge of assessment practices. Other standards relate to selectivity in admissions and the quality of student teaching experiences.

To determine how well TPPs are meeting these standards, NCTQ analysts gathered information on admissions criteria, course syllabi, textbooks, student teaching policies, and program outcome data where available. They decided not to use some of the more controversial and criticized indicators used in *U.S. News* ratings of other types of professional schools, most notably data on research expenditures, student/faculty ratios, and peer ratings of prestige. They did not place particularly heavy emphasis on VAM results, but gave credit to TPPs that demonstrate an interest in using outcome data, where available, for their own self-improvement.

Many TPPs declined to participate in the NCTQ/*U.S. News* ratings process, due to concerns about the methodology and amount of time and trouble required to supply the requested documentation. Of the more than 2,400 TPPs that NCTQ sought to examine, it was able to issue an overall rating to only 1,200; among these, complete data were provided from roughly 10 percent. Many TPPs cooperated only after receiving open-records requests. In particular, private colleges are underrepresented in the ratings because their documents generally do not fall under states' open-records statutes.

The results were released by NCTQ/*U.S. News* in June 2013 and concluded that only a small number of TPPs across the nation adequately prepare new teachers. Just four programs, all in secondary teacher preparation, earned a four-star overall rating (the highest possible score); about 160 programs were deemed so weak that they were put on a "consumer alert" list by the council. The report generated a great deal of discussion, positive and negative. The Fordham Foundation (which founded NCTQ in 2000) applauded the study and the wealth of data that was collected, claiming that it will have lasting impact on policy and practice (Tatz, 2013). Some educators commented that the report bolsters other studies that highlighted the wide variation in what teacher candidates are expected to learn (Sawchuk, 2013c). But many education scholars criticized the heavy reliance on document review and failure to check for inaccuracies in the data, and claimed that the measures bore little or no relationship to the quality of training (Sawchuk, 2013c; Strauss, 2013). For example, graduate level TPPs at highly selective universities like Harvard, Columbia, and Stanford earned low ratings for selectivity because they do not require a minimum grade point average or GRE score, although their students in fact rank far above national averages on these measures (Strauss, 2013).

Rankings in Other Professions

Since this sort of system for rating TPPs is a new enterprise, empirical studies of its consequences have yet to be done. But other professional schools have long been ranked, and a great deal has been written about how ranking has affected programs in law, medicine, business, and engineering. Although the rating and ranking systems used for other professional programs differ from each other and can be expected to have different consequences, it is worth examining effects of rankings that are similar across programs. A review of the literature identifies three main problems with ratings and rankings: perverse incentives, poor indicators, and the "rich get richer" phenomenon known as the Matthew Effect. Each of these issues can be said to apply more generally to any potential high-stakes uses of evaluation data, but raise particularly important problems for rankings and ratings. The effects described below have been found from research on ratings and rankings of professional schools.

Perverse Incentives and Gaming

Some evidence suggests that when academic programs are rated or ranked, they will respond to incentives and take action to increase or maintain their status. Some of these actions may not serve to improve the program and may even be harmful in some ways. Negative effects uncovered by various studies include the stratification of higher education systems and incentives to increase selectivity at the expense of more inclusive access. For example, a program may seek to become more selective by stiffening admissions requirements or using financial aid to attract the most talented students regardless of need—an action that could harm underrepresented groups of students (Zell, 2001; Stake, 2006; Institute for Higher Education Policy, 2007). In addition, to attract more students, higher education institutions often make "image-enhancing" investments that bear little relation to academic quality, such as building impressive sports and recreation facilities (Institute for Higher Education Policy, 2007).

Incomplete Indicators

Some researchers argue that the indicators used in rating and ranking methodologies inadequately capture the quality of preparation programs or focus narrowly on selected aspects of a program's mission. Mullan, Chen, Patterson, Kolsky, and Spagnola (2010) took issue with a *U.S. News* indicator of the quality of some medical colleges—namely, the amount of research money the institution received from the National Institutes of Health (NIH). Neglected were any indicators that placed a value on a medical school's social mission, such as training doctors to work in poor communities or with underserved populations. The research team re-ranked medical schools on this basis, and three historically black medical schools came out on top. In fact, the authors found an inverse relationship between NIH funding and social mission and concluded that *U.S. News* rankings miss out completely on the social mission and thus ignore a vital component of U.S. health care delivery.

The reverse seems to be true for MBA programs. For business schools, *U.S. News* does not use research funds from external sources as an indicator of quality. Zell (2001) asserts that as a result, research programs at one major business school have been devalued. This school has also shifted toward using adjunct faculty with "real-world" business experience.

The Matthew Effect

The *U.S. News* rankings for both undergraduate institutions and graduate programs are based in part on rankings of their prestige or reputation, as determined from responses to questionnaires by faculty and staff at surveyed institutions. The *Matthew Effect* is a term coined by Columbia University social scientist Robert Merton, which draws from a passage in the Gospel of Matthew: "[F]or to everyone who has will more be given, and he will have abundance; but from him who has not, even what he has will be taken away." In other words, the rich get richer. In the case of prestige rankings of higher education institutions, the term refers to the phenomenon whereby an institution's ranking in one year reinforces or strengthens its rankings in later years. This creates a self-referential, self-reinforcing cycle that continues to reward highly ranked institutions (McGaghie and Thompson, 2001). Several studies have suggested that the Matthew Effect is a reality when peer reviews are used as a measure for rankings.

Positive Effects of Rankings and the Need for a Balanced Approach

While the research reviewed above raises important cautions about negative aspects of rankings and ratings, it is important to note that some analyses have identified positive effects as well. Zell (2001) reports that business school professors and administrators believe that rankings have forced improvements in the relevance of course offerings and the quality of teaching. They provide prospective students with some sort of benchmark or measure of quality that is unavailable from other sources (Institute for Higher Education Policy, 2007). A 2005 policy report by the Association to Advance Collegiate Schools of Business, while somewhat critical of media rankings, still states that "as a whole, media rankings have raised the overall visibility of MBA programs and business schools" (p. 6).

The available evidence suggests that rankings and ratings, which appear to have substantial intuitive appeal, can have beneficial as well as undesirable attributes, which underscores the importance of looking more closely at the details of any given rating or ranking project. The NCTQ/*U.S. News* system for rating TPPs, for example, has deliberately avoided some pitfalls by eschewing indicators based on peer ratings of prestige, research expenditures, and student/faculty ratios. At the very least, by doing away with faculty and superintendent prestige rankings, NCTQ has attempted to address the Matthew Effect problem and some of the problems of incomplete indicators. NCTQ has also attempted to focus on the content that is actually taught. It does use selectivity as one indicator, citing research on the relationship between measures such as SAT scores and teacher effectiveness (National Council on Teacher Quality, 2013).

Program Self-Study

Thus far, we have described how TPPs in the United States must deal with a variety of accountability policies at the federal and state levels, as well as those adopted by outside organizations. According to Peck, Gallucci, and Sloan (2010), these contexts may create serious dilemmas for teacher educators: on one hand, compliance with prescriptive government mandates is often interpreted by faculty as a demoralizing loss of program autonomy and integrity; on the other hand, noncompliance may result in a loss of program accreditation.

Another type of evaluation occurring in some TPPs originates in the TPP itself rather than from external forces. Across the country, some TPPs are voluntarily undertaking efforts to evaluate their programs for the purposes of program improvement and inquiry.

The term "self-study" is used in different ways in the context of teacher education. Here we are not referring to the initial self-study reports that TPPs are often required to submit for national accreditation or state approval. Although these reports are supposed to be used for institutional self-improvement, more often they are completed mainly to fulfill reporting requirements. We are also not referring to the literature on teacher education self-study practices, which deals with reports that individuals make about their own teacher preparation practices rather than reports that institutions make to study the effectiveness and operations of their programs. Rather, we focus on program self-studies for which the audience is the TPP itself—that is, a TPP gathers information about its own program to decide how to improve.

Creating Cultures of Evidence

Peck and colleagues at the University of Washington are engaged in a program of research that examines the uses of evidence for TPP improvement (Peck, Gallucci, and Sloan, 2010; Peck and McDonald, 2013). More specifically, they have explored how TPPs can benefit from state-mandated performance assessments of prospective teachers by using the resulting evidence for program improvement and inquiry. As described by Peck, Gallucci, and Sloan (2010), inquiry is a process that involves faculty in making data-based decisions about such areas as organizational change, institutional policies, collective values, curriculum, and assessment. Their research draws on sociocultural theory to analyze processes of learning and change within TPPs.

For example, Peck, Gallucci, and Sloan (2010) conducted a case study of one TPP within the University of California system; the researchers were faculty of the program at the time. Over an 18-month period, they studied the process by which the faculty implemented a new California requirement that required TPPs to use standardized performance assessments in making teacher-credentialing decisions. At first, the faculty perceived the new state policy mandates as demanding and intruding strongly on local program values and practices. In a strategic effort to negotiate the tension between these perceptions and the institution's need to implement the new policies, the researchers helped to develop an approach that shifted the discourse from a focus on compliance to a focus on inquiry.

In the end, as a result of the performance assessment, faculty and staff made a number of changes in program structure, working practices, and ways of thinking about their program. One such change was the emergence of new modes of direct and indirect interaction among program faculty, staff, and students. Another category of changes included the development of new types of program-wide meetings in which faculty examined samples of candidate work on the new performance assessments. The process also led to more clearly articulated connections across the coursework and fieldwork dimensions of the program. Conversations within the organization moved from a focus on reacting to the imposition of external standards to articulating valued outcomes for the whole program. Subsequent studies by Peck and McDonald (2013) note that performance assessment data has been used successfully to foster positive changes in organizational practice in a diverse set of TPPs operating in different contexts.

Voluntary Networks of TPPs

In another type of self-study, TPPs have come together to create voluntary networks that work cooperatively to evaluate and improve their own programs. One of the most visible examples of this approach was Teachers for a New Era (TNE), which was started in 2001 by the Carnegie Corporation and planned as a five-year investment. This initiative, which included 11 TPPs, aimed to stimulate development of excellent TPPs that were guided by respect for evidence-based decision making (Kirby, McCombs, Barney, and Naftel, 2006). The strategy began with a small, select group of TPPs (among them, Boston College, Michigan State University, Stanford University, University of Connecticut, and University of Virginia) that agreed to create exemplary models for teacher preparation that could be replicated elsewhere.

TNE was guided by three principles established by Carnegie:

1. TPPs should be driven by evidence. A culture of research, inquiry, and data analysis should permeate the program. Gains in student learning in classes taught by program graduates should be measured with standardized tests.
2. TPPs should have greater engagement with arts and sciences faculty in order to strengthen content knowledge and ensure that teacher candidates possess integrative knowledge of the nature, premise, modes of inquiry, and limits of various disciplines.
3. Teaching should be seen as an academically taught, clinical practice profession. There should be close cooperation between colleges of education and actual K-12 schools. Master teachers should be appointed as clinical faculty, and graduates should undergo a two–year residency induction period (Kirby, McCombs, Barney, and Naftel, 2006, p. xvi).

TNE strongly emphasized a “clinical” model of teacher preparation that connects research and practice and involves collaboration among the TPP, arts and sciences faculty, and K-12 schools. The clinical model gives greater emphasis to actual classroom experience through student teaching and related fieldwork, interwoven with academic content and coursework. The TNE initiative also emphasized program evaluation; in fact, the project's funders hired RAND to conduct studies of how TNE was being implemented at participating institutions. These evaluations indicated that implementation was proceeding very slowly:

> Thus far, the actual changes in the teacher education programs at the TNE sites appear to be small and incremental. This is not surprising, given that these institutions were selected because they were among the best in their "class" of institutions. However, the process by which these incremental changes to a program will result in highly qualified, competent teachers who will be markedly "better" than the graduates before them is not well defined. (Kirby, McCombs, Barney, and Naftel, 2006, p. xxi).

There are other networks of TPPs aimed at self-improvement, and all of them urge member TPPs to adopt a more clinical model of teacher preparation that creates strong connections between the TPP and schools in their geographic area. TNE and The Renaissance Group (TRG) both insist that member TPPs apply some sort of measure of TPP effectiveness, whether measures of student learning or other measures of graduates' effectiveness on the job. The National Network for Educational Renewal is built around the educational reform ideas of John Goodlad. Box 2-5 describes the Science and Math Teacher Initiative (SMTI), which aims to increase the supply of highly effective STEM teachers and includes a strong program self-evaluation component.

SUMMARY MATRIX

In this chapter we have provided an overview of the TPP evaluation landscape by describing the primary sources of evidence used for program evaluation and the five main types of evaluation systems in the United States. Table 2-2 brings together these two strands by showing the types of evidence used by each of the five main systems.

While the systems use some similar types of evidence, they also diverge in important ways. For instance, the federal HEA evaluation system relies heavily on easily quantifiable data such as admissions criteria and results of teacher licensure tests. In contrast, the media rating system implemented by NCTQ/*U.S. News* leans more heavily toward evidence of what teacher candidates are being taught. Accreditation and state government reviews tend to use a wide variety of indicators. All of the TPP evaluation systems are exploring the use of outcome measures that gauge graduates' effectiveness in raising student achievement. Whether outcome measures will prove to be better for determining and advancing teacher quality is an open question that will require careful monitoring and research.

BOX 2-5
Framework for Evaluating the Preparation of Science and Mathematics Teachers

The Science and Mathematics Teacher Imperative (SMTI) is aimed at increasing the supply of talented K-12 science and math teachers. It is a network of 132 universities in 45 states, graduating over 8,000 science and math teachers each year; a goal is to increase this figure to 10,000 teachers (Association of Public and Land-Grant Universities, n.d.). Administration of the project is housed at the Association of Public and Land-Grant Universities. With additional funding from the Carnegie Corporation and the National Science Foundation, APLU created its “Analytic Framework,” an evaluation system that specifically rates the status and progress of a higher education institution in preparing math and science teachers. The evaluation framework is referred to as “a common framing tool for use in analyzing, designing and implementing more coherent, engaging and effective science and mathematics teacher education programs” (Coble, DeStefano, Shapiro, Frank, and Allen, n.d.; Coble, 2012).

The evaluation framework was designed around a group of concepts or ideas meant to increase communication and efforts across disciplines at universities. Goals include bolstering science teacher preparation, making maximum use of clinical practice opportunities by partnering with school districts, and preparing teachers in such a way as to demonstrate impacts on student achievement. These concepts were drawn from and informed by the National Research Council report *Educating Teachers of Science, Mathematics and Technology: New Practices for a New Millennium* (National Research Council, 2000), as well as Goodlad’s work on partnerships between TPPs and school districts.

Table 2-2 is a simplified representation and may give the sense of more uniformity across TPP evaluation systems than is actually the case. Even though several groups may use the same general type of data, they often analyze and operationalize the data in different ways. For example, selectivity is measured in a number of ways by the various TPP evaluation systems. NCTQ judges programs on whether or not they require a 3.0 high school GPA for admission, and whether program applicants have an SAT score above 1120, an ACT score above 24, or a score in the upper half of the distribution on another norm-referenced test (2013). The federal government, under Title II of the HEA, not only requires TPPs to provide information on average GPAs and SAT/ACT scores of incoming classes, but also to report whether programs require fingerprint and background checks, whether candidates have experience working in a classroom, and whether the TPP requires an essay, personal statement, interview, or personality test.

The framework consists of five core components with associated goals:

1. ***Institutional commitment.*** The higher education institution promotes and sustains the program across university departments as well as with partner school districts.
2. ***Recruitment, selection and admission.*** The program is highly selective, recruits interested students, and ensures diversity.
3. ***Content, pedagogy and clinical practice.*** The program ensures that teachers have good content knowledge and the skills to impart this knowledge to students, includes quality clinical practice, and incorporates relevant state and national standards.
4. ***Support for beginning teachers.*** The program provides mentors and other types of support for recent graduates and tracks their effectiveness in the classroom.
5. ***Professional development.*** The program provides continuous and advanced learning opportunities for in-service math and science teachers.

The Framework is to be used for program self-study purposes; no team of auditors from APLU or other organizations will conduct site visits or review material. Instead, institutional leaders assess their own programs. A five-point scale is used to rate the extent to which the program places a high value on the components and goals and has taken steps to implement them. The Framework is a relatively parsimonious document, just 10 pages, and is estimated to take one hour to complete. The job of conducting the assessment does not fall on one person; APLU suggests that numerous persons within the program and its partners conduct the assessment and compare and discuss the outcomes. In addition to identifying shortcomings, the Framework can also be used to identify promising practices that can be documented and shared with other participating higher education institutions.

TABLE 2-2 Main Types of Evidence Used by Different TPP Evaluation Systems

	Federal HEA-II	National Accreditation[a]	State Program Approval[b]	Media and Independent Ratings[c]	TPP Evaluations for Program Improvement
Input Measures					
Selectivity (e.g., average SATs or GPA)	X	X	X	X	
Faculty qualifications		X	X	X	
Substance of instruction (e.g., syllabi, lectures, textbooks)		X	X	X	X
Student teaching experience (e.g., minimum number of hours, records of observations)	X	X	X	X	X
Surveys of program graduates		X	X		X
Output Measures					
Teacher certification tests (pass rates, average scale scores)	X	X	X		
Hiring and retention		X	X		X
Candidate performance assessments		X	X		X
Surveys of principals and employers		X	X		X
Impact on student learning (VAMs)		X	X	X	X

[a] This reflects CAEP, which is currently the single national accreditation program.

[b] State systems vary in the evidence they use; these types of measures are used by at least some states.

[c] The only existing evaluation of this type is done by NCTQ/*U.S. News*, so this column reflects their approach; other organizations may produce ratings in the future using different types of evidence.

3

Program Evaluation: Mapping Approaches to Purposes

Program evaluation has many plausible goals and can be designed and conducted in various ways. The policy challenge is to select the system or approach that is best suited for a defined purpose. To help policy makers and practitioners make informed decisions, in this chapter we map the various purposes or intended uses of TPP evaluations against the different ways in which these evaluations can be designed.

This chapter begins with a discussion of three main purposes for evaluating TPPs: ensuring accountability, providing consumer information, and enabling self-improvement of teacher preparation programs. We then analyze the five types of existing TPP evaluation systems presented in Chapter 2 in terms of four elements: (1) the evidence they use; (2) the inferences they are intended to support; (3) the incentives they offer for TPPs to participate; and (4) the perceived and real consequences they bring—including direct consequences for TPPs and indirect consequences for teachers, students, and the education system as a whole.

There are several reasons why it is important for evaluation designers to pay attention to these elements of TPP evaluation systems. First, analyzing these features can help determine the purposes best served by existing systems. Second, thinking through these elements early in the design of a new system can increase the chances that the evaluations will be used coherently and effectively. Third, carefully considering the alignment between evaluation methods and purposes can focus attention on the benefits and potential risks (or unintended negative consequences) of using various evaluation approaches.

For this analysis we draw, in part, on lessons learned from the history of standardized testing. Standardized, or large-scale, tests have long been used to assess student achievement and, by extension, the effectiveness of K-12 education. Testing has been the subject of a great deal of theory and research, as well as controversy (e.g., Cronbach, 1975; Office of Technology Assessment, 1992; Kaestle, 2013). Many of the lessons learned from testing, both about the potential of tests to inform education and the dangers of their misuse (National Research Council, 1999, 2001a, 2011), are germane to the evaluation of TPPs.

PURPOSES FOR TPP EVALUATION

The purpose of evaluation is to judge the worth or merit of something (Scriven, 1967). That very broad definition encompasses many reasons for evaluation. We condense the more complex and comprehensive list of purposes for TPP evaluation into three main categories:

1. Ensuring accountability, which involves monitoring program quality and providing reliable information to the general public and policy makers
2. Providing information for consumers, which includes giving prospective teachers data that can help them make good choices from among the broad array of preparation programs, and giving future employers of TPP graduates information to help with hiring decisions
3. Enabling self-improvement by teacher preparation programs, which entails providing institutions with information to help them understand the strengths and weaknesses of their existing programs and using this information to spur innovation and continuous improvement.

The first and third purposes are well-established functions of program evaluation in general and are often referred to, respectively, as "summative" and "formative" strategies (Scriven, 1967; Worthen, Sanders, and Fitzpatrick, 1997). The second purpose, providing consumer information, could also be considered a type of summative evaluation, but is more specific to systems in which consumers use evaluation results and other information to choose among competing options (see also Scriven, 1983, for discussion of the importance of "consumer-oriented evaluation").

Accountability

Accountability means holding TPPs responsible for accomplishing, or at least pursuing, their goals. Assuming that the principal goal of TPPs is to prepare future teachers for effective work in classrooms that will lead to increased student learning and other valued educational outcomes, it follows that holding TPPs accountable requires a focus both on the general criteria for effective teaching (the "excellence perspective" for reform described in Chapter 1) and on the specific requirements for teaching in diverse and economically disadvantaged communities (the "equity perspective").

Evaluation should not be misconstrued as the setting of a simple or single criterion for judging programs, but rather as a process for providing information relevant to making complex human judgments. As an accountability tool, evaluation is oriented ultimately to the general public—taxpayers and voters who are legitimately curious about whether their resources are being spent wisely. The audiences for accountability-driven evaluations tend to be policy makers at the federal, state, and institutional levels who are responsible for allocating resources and need information about the relative quality of programs. In the American system of government, people seldom expect a single metric, no matter how it is defined or derived, to lead automatically to particular policies or practical actions. The data are meant to inform, not to trigger, decisions.

In addition, accountability as a linchpin of modern democratic governance is inseparable from notions of trust: citizens trust their leaders with stewardship of resources, and leaders trust that they will be judged fairly (Feuer, 2012b). From both vantage points, however, trust comes with and is enhanced by the notion of validation, or verification. Data for accountability need to be collected scrupulously and interpreted rigorously, according to defined and accepted professional standards.

Accountability-driven evaluations also have a more subtle but compelling indirect purpose—to spur reform or improvement. In other words, they are frequently intended not only to measure individuals and institutions but also to *influence* their behavior. An underlying assumption of accountability systems is that publicizing the results of evaluations of program quality creates incentives for program managers to do better; to avoid being tagged for not meeting standards, TPPs that are weak in certain areas will be motivated to take steps to improve on their own. Setting standards in accountability systems is no trivial matter: they must be developed fairly and realistically and should reflect the most valued goals for programs to strive toward.

Consumer Information

The second purpose for evaluating TPPs is to provide information to guide prospective teachers and their parents in choosing a preparation program and, more indirectly, to guide future employers from K-12 school districts in hiring graduates of these programs. Here, the evaluation aims to identify high-performing programs on the assumption that some consumers want to find the "best" TPPs and that information about key attributes of programs will help shape their decisions. (There is little empirical evidence about the actual ways in which decision makers frame their choices in these situations—for example, whether or how they seek the "best" schools—or about the assumptions institutions make regarding consumers' behavior. Most likely, many students choose a TPP based on practical considerations such as proximity to their home and cost of tuition as opposed to rigor and other qualities of the program. But the general idea that consumers are looking for "best bets" is entirely plausible.) "High performance" for a TPP may be defined as either providing the best education for a teacher candidate or producing the best teachers. In addition, competition between TPPs can serve as an impetus for reform and improvement if the indicators on which they are evaluated reflect the most valued goals for teacher education.

Providing consumer information is an inexact science. To be useful the data must distill key indicators from the vastly complex domains of qualities and characteristics of programs and must be suitable for making meaningful, though incomplete, comparisons. The challenge in providing useful information about TPPs to prospective teachers and/or future employers of new teachers is to identify a set of program attributes that are as succinct and relevant as, say, the qualities of refrigerators described and rated in *Consumer Reports*. Indeed, it is a vexing and hotly debated question as to whether the complexities and subtleties of teaching and learning and, by extension, the preparation of future teachers can or should be reduced to a set of relatively superficial proxies, especially if the resulting array of comparative data is deemed unreliable, unfair, or misleading.

It may be intuitively appealing to reduce complex qualitative variables to a set of minimal or baseline requirements, such as the skills and knowledge that all beginning teachers should have on the first day of their new jobs. But to make this idea operational requires agreement on what those skills and knowledge are, which itself is an exercise in approximation and compromise. And there is always the fear that setting minimal standards will lead to minimal performance, a criticism often raised about the minimum competency testing movement (see, e.g., Office of Technology Assessment, 1992). These challenges notwithstanding, there is substantial public demand for at least some relevant comparative information. In light of this demand, the measurement and evaluation profession has an obligation to use reasonably good methods to supply that kind of information while indicating whether and how the information is limited or misleading.

Program Improvement

The third purpose for evaluating TPPs is perhaps the most easily understood and the most readily embraced: to help institutions make evidence-informed decisions to improve their programs. Evaluation in this context and for this purpose is aimed at identifying specific program strengths and weaknesses and is most often initiated by the program itself, principally for the use of its faculty, staff, and administration.

Unlike evaluations done for other purposes, the results of this type of evaluation are not necessarily used to prove or even to suggest that the program meets an *external* standard. In its pure form, this approach to evaluation would not be considered "high-stakes" in the same way that accountability or consumer-oriented evaluations are. This is because the information from program improvement evaluations does not typically have to be made public and is not intended to be used by external authorities or prospective "buyers" to make decisions with potentially dramatic consequences for individuals or institutions.

It may seem ironic that evaluation systems intended for what may be the most important use, diagnosing program deficiencies and developing innovative remedies, should be labeled "low-stakes." Indeed, even evaluations that are internally driven and organized can have significant consequences for individual faculty and staff—for example, if programs are deemed ineffective and then slated for reduction or elimination. Deans of teacher education schools have considerable experience walking these tightropes. Moreover, if institutions are either required or inclined to release their internal reviews, this may bring high-stakes consequences of the more familiar sort: public institutions, for example, are subject to Freedom of Information Act rules that may result in disclosures of information originally intended only for internal diagnostic and program improvement uses.

It is clear that a fundamental tension affects all program evaluation. On one hand, evaluation is meant to influence positive change, whether by providing decision makers with reliable information or by creating more indirect incentives for changes in behavior and performance. Evaluations that are momentous enough to induce change have potentially high-stakes consequences for individuals or institutions. On the other hand, as the consequences become increasingly significant, threats to the validity of the underlying measures also become more powerful, as some faculty and administrators may seek ways to show results that do not necessarily reflect real improvements in student learning.

Setting the dial at the right level—one that avoids the extremes of indifference on one end and incentives for opportunistic tampering on the other—is a major challenge in designing and using evaluation or any system of performance measurement (see Koretz, 2009; National Research Council, 2011a; Feuer, 2013c; Linn, 2013).

The real picture is even more complicated because just about any TPP evaluation effort may be expected to serve more than one purpose. *In general, the more purposes a single evaluation mechanism aims to serve, the more likely it is that each specific goal or purpose will be compromised and that problems of misuse will arise.* This is another hard-learned lesson from the history of standardized testing in the United States.

Considering Purpose in Designing Evaluations

As the preceding discussion suggests, a federal approach to evaluating TPPs for accountability will, and should, be designed differently than one conducted by a TPP itself for program improvement. Specifically, a federally mandated evaluation will aim to standardize the types of data provided by states and TPPs—perhaps to assert greater centralized authority, but also to produce a reliable and fair picture of the quality of teacher education that is comparable across states. Such an approach will have to rely on data, such as teacher test results, that are widely available across states and programs and are not tied to a particular set of standards or approach to teacher education. In contrast, an evaluation aimed at providing useful feedback to TPPs for self-improvement needs to be well-aligned with specific program goals. Peck and McDonald (2013) presented evidence that when faculty generally perceive measures as being aligned with a program's guiding values and beliefs, they may be more receptive to evaluation results and better able to use them constructively for program improvement. Standardization, which is useful and necessary for reliable and fair comparisons across states and localities, is not as crucial in this situation.

Questions remain about whether using an evaluation system for purposes other than those for which it was originally designed and validated is always bad, and whether the risks associated with this kind of "drift" are justified by potential or actual benefits of using evaluation data for varied purposes simultaneously. The literature on unintended negative consequences of such drift is compelling (e.g., RAND's studies of the Vermont portfolio assessment program, in Koretz, Stecher, Klein and McCaffrey, 1994). Research has given less attention to the notion of weighing costs and risks against the intended (and perhaps unintended) benefits in judging the quality and use of any measurement system (see Feuer, 2008, 2010).

For an evaluation system to serve more than one purpose is not necessarily wrong, and in reality TPP evaluation results will almost inevitably be used for multiple purposes. Such is the nature of information, especially in a society that believes in increasingly free and open access. But those who design or mandate evaluations of TPPs should be explicit about the primary purpose of the evaluation, cautious about the possibility of its results being used in unintended ways, and transparent about the design of mechanisms to reduce (if not eliminate) the hazards of misuse. What is most important is to avoid the temptation to oppose evaluation solely because of its imperfections.

ELEMENTS OF TPP EVALUATION SYSTEMS

How can policy makers and practitioners make informed decisions about selecting or designing a TPP evaluation system that is well suited to its intended purpose? How well do different existing systems align with various purposes for TPP evaluation? What considerations can help guide the design of improved evaluation systems?

Several key elements or characteristics of TPP evaluation systems are important to consider when choosing from existing approaches or designing new ones. The first set of elements includes the systems' evidence base and types of inferences that are supported by particular evidence. The second set includes the intended or unintended incentives created by evaluation systems and their positive and negative consequences on TPPs directly and on the education system indirectly.

An overarching concept that guides this discussion is *validity*. In the theory and practice of educational and psychological testing, validity is a central idea and refers to the extent to which interpretations of test scores are defensible or trustworthy. Although the concept of validity is emphasized to a lesser extent in the literature of evaluation, the American Evaluation Association (n.d.) does include this idea among its standards: *"Evaluation information should serve the intended purposes and support valid interpretations."* The committee believes that validity should be the primary criterion for assessing the value of various TPP evaluation approaches for different purposes. More information about validity appears in Box 1-1 in Chapter 1.

Evidence and Inferences

Evaluation is a process of reasoning from evidence. Just as a test is designed to "observe students' behavior and produce data that can be used to draw reasonable inferences about what students know" (National Research Council, 2001b, p. 42), an evaluation is intended to produce information that can be used to draw reasonable inferences about the quality of programs. *Evidence* refers to the measures or data collected. Chapter 2 describes the many sources of evidence used by existing TPP evaluation systems. For instance, the current federal TPP evaluation system emphasizes results on teacher certification tests, while the consumer information system emphasizes selectivity and academic content.

By *inferences*, we mean interpretations or findings based on the evidence. For example, in evaluations conducted to meet federal requirements, users of data on certification test pass rates may draw inferences about the degree to which TPPs prepare teacher candidates to pass the tests. Others may infer that pass rates are more of a reflection of the general ability of the students who entered the program. In media ratings meant to support consumer decision making, users of data on academic content are likely to draw inferences about the quality and rigor of TPP course offerings. Users of data from both federal and media evaluation systems may be tempted to make inferences about the general quality of TPP programs, but (as discussed below) these inferences are not necessarily valid. Sources of evidence used by a system will dictate the types of valid inferences that can be drawn. Designers of a new evaluation system should identify the types of inferences intended up front because that will point to the types of evidence that need to be collected.

Identifying which types of evidence will lead to valid inferences is never simple. What seems like a good measure or piece of evidence may, in the end, not support the type of inference required or sought. For example, selectivity data, such as average SAT scores of those admitted to a TPP, may not be a valid indicator of TPP quality. Whether or not one takes into account current controversies over the SAT as a surrogate for socioeconomic status (Jaschik, 2012), a TPP may *not* be highly selective but will still do an excellent job of preparing teacher candidates. Similarly, the federal approach puts a lot of emphasis on the percentage of teachers who pass licensure exams, but since the vast majority of candidates pass these it is not clear what one can meaningfully infer. Are TPPs doing a good job of preparing candidates for these exams, or is it simply that the exams are easy to pass? Is the high passing rate partly due to selection of qualified candidates in the first place? A simple but often overlooked point is that TPP "effects," as gauged by output measures such as pass rates on certification tests, VAM estimates, and employer satisfaction surveys, are actually a combination of selection and training effects that are difficult to separate. One danger to be avoided is misclassification—that is, inferring from output measures that a TPP is doing a good job of preparing teachers for the classroom when it is not, or inferring that it is not doing a good job when it is.

Because the connections between evidence and inference are often complex and because no single piece of evidence captures every important attribute of a program, validity is best served by using multiple types of evidence. Indeed, as shown in Chapter 2, most TPP evaluation approaches do rely on multiple measures to make judgments about program quality.

Incentives and Consequences

When evaluations have potentially high-stakes consequences—in other words, when they are used to make important decisions about individuals or institutions— incentives are created. We use the term *incentive* here to mean a tangible or intangible reward or sanction tied to the results of an evaluation (National Research Council, 2011a). Some incentives work in the desired direction: TPP faculty and administrators may be motivated to initiate actions that will lead to genuine improvements. If a TPP is judged on the basis of a certain piece of evidence, indicator, or attribute, then faculty and administrators at a TPP are expected to take appropriate action. Research in economics has shown that people tend to choose the quickest and easiest action to improve measured performance (Prendergast, 1999; Rothstein, 2008; National Research Council, 2011a). At best, such actions will actually improve the TPP. Any evaluation method provides feedback about performance, and incentives are attached in the form of recognition, embarrassment, possible loss of accreditation, reorganization forced by state authorities, or loss of federal or state funding. Since these incentives are designed to change the behavior of TPP faculty and administrators, an important question is whether the changes in behavior created by these incentives serve the goal of improving TPP quality.

Indeed, incentives can sometimes distort the intent of the evaluation, if, for example, faculty and administrators seek better results by "gaming" the system or following other opportunistic instincts that ultimately confound the meaning of the data and undermine progress toward the goal of program improvement. Faculty and other staff may manipulate the measures or devote attention and resources to attributes that are measured at the expense of other important attributes or goals that are not measured. This idea stems from the assertion by Donald T. Campbell (1976), which has come to be known as "Campbell's Law:"

> The more any quantitative social indicator is used for social decision making, the more subject it will be to corruption pressures and the more

> apt it will be to distort and corrupt the social processes it is intended to monitor (p. 49).

At the risk of overemphasizing the potential negative effects of incentives[1] there is a legitimate concern that Campbell's Law affects TPP evaluation and accountability programs by creating incentives to engage in opportunistic practices that are counterproductive for reform (see also Springer, 2009). In the case of HEA Title II, for example, Aldeman, Carey, Dillon, Miller, and Silva (2011) reported that measures based on the percentage of graduates passing required licensure and certification exams created an incentive for TPP organizations and states to require students to pass these exams to graduate, thus ensuring very high pass rates:

> The initial response to the 1998 HEA accountability requirements illustrates the level of intransigence and bad faith among state policymakers when it comes to improving teacher preparation. Some states rated programs based on the number of program participants who passed the program's entrance test. Thus, by definition, all programs in those states reported 100 percent pass rates. Other states rated programs based on the licensure exam pass rate of "program completers"—and then defined "program completer" as "a person who has passed the licensure exam (p. 3).

Thus, pass rates became so high that they were no longer a meaningful measure of TPP quality. In short, incentives created by the HEA corrupted the validity of inferences derived from the measure.

Fleener and Exner (2011) speculate that use of value-added measures may incentivize TPPs to focus on training candidates to deliver the content demanded by K-12 standards and assessments, at the expense of aspects of teaching that are less tangible and measurable but still valuable. The VAM measure directs resources toward what will be measured—the delivery of standards-derived content in tested subjects. Similarly, state accreditation programs, with their numerous and sometimes onerous demands, may incentivize TPPs toward a "stifling conformity," in the words of Goodlad (1990), rather than innovation. This criticism has been directed in the past toward NCATE accreditation, which some researchers assert is a matter of clearing bureaucratic hurdles.

These cases highlight the importance of identifying the right measures and understanding how they may create unintended incentives that distort the goals for improving teacher education. Faulty indicators and incentives, for example, may lead to the misallocation of resources. We

[1] See, e.g., Hirschman, 1991 for discussion of the risks of exaggerating the potential negative consequences of policy without adequate attention to benefits.

live in an era in which states, districts, the federal government, teacher education associations, and various independent accrediting and ratings organizations are increasingly experimenting with new evaluation tools and techniques. These systems, along with their consequences, need to be monitored to ensure they do not unintentionally undermine teacher preparation and effectiveness.

ANALYSIS OF EXISTING TPP EVALUATION SYSTEMS

If there is one simple lesson to be taken from this discussion, it is that no evaluation system is perfect in its design or immune from misuse. But it would be naïve to assume or require that evaluation data could or should be completely embargoed or quarantined from unintended uses. The policy challenge is to understand the strengths and limits of evaluation systems, whether used alone for specific purposes or in conjunction with other systems for multiple purposes, and to apply that understanding to set reasonable and rational expectations for the effectiveness of evaluation in improving teacher preparation and student learning.

Table 3-1 summarizes the five types of existing TPP evaluation systems in terms of the elements outlined above: evidence, inferences, incentives, and consequences. Unpacking these elements can help one better understand these systems and determine the most appropriate uses of each approach.

As an example, the second column of Table 3-1 shows a breakdown of the elements of the CAEP national accreditation system. The accreditation process uses a wide variety of input and output data so that inferences can be drawn about whether a TPP meets a defined set of national accrediting standards. Accreditation typically involves a self-study component and cycles of review and feedback, so that inferences can be made about program strengths and weaknesses relative to teacher education standards. The system creates an incentive for TPPs to earn a national "stamp of approval," which can help with recruitment of highly qualified candidates and faculty. The intended and likely consequences are that TPPs will be adequately aligned with standards for teacher education and that most TPPs will work to address areas of weakness. One potential unintended consequence is that such a system will lead to more conformity and less innovation by programs. Table 3-1 also lays out the elements of evaluation systems conducted by the federal government, state governments, media and other independent organizations, and TPPs themselves.

TABLE 3-1 Attributes of Existing TPP Evaluation Systems

Attribute	Type of Evaluation System				
	Federal HEA-II	National Accreditation[a]	State Program Approval	Media and Independent Ratings[b]	TPP Evaluations for Program Improvement[c]
Evidence	Inputs: admission requirements, student teaching requirements Outputs: teacher certification test results	Inputs: admission requirements, program documents and policies, faculty qualifications, case studies, student teaching observations Outputs: candidate performance assessments, teacher certification test results, hiring and retention, VAM if available, employer satisfaction surveys	Varies by state; some use national accreditation process Increasing number of states use VAMs	Inputs: admission requirements, course offerings, syllabi, textbooks, student teaching observation instruments	Varies by study. Some use- Inputs: course syllabi, student teaching observations, candidate work samples Outputs: candidate performance assessments, hiring and retention, VAM
Inferences	Proportion of students passing certification tests; average scale score for the program	Program meets or does not meet national teacher education standards Level of satisfaction of employers of graduates	Program meets or does not meet state teacher education standards Other inferences vary by state	Rating of programs according to performance categories (more nuanced information than pass/fail)	Program strengths and weaknesses, aspects of program that need improvement

	Whether the program has a minimum GPA or SAT/ACT admission requirement How much student teaching experience candidates are gaining Whether a TPP has been identified by the state as low-performing	Program strengths and weaknesses, aspects of program that need improvement		Extent to which program requirements and course materials align with standards held by the organization	
Incentives for TPPs	Must comply in order to receive any federal funds	National "stamp of approval" can help with recruitment of highly qualified candidates and faculty Often also serves as state approval If detailed data are made public, incentive to improve in areas of identified weakness	Maintain ability to recommend teachers for state certification by having "stamp of approval" from state If detailed data are made public, incentive to improve in areas of identified weakness	Prestige associated with being highly rated, which may draw stronger faculty and students, as well as support from state policy makers	Strengthen the program

continued

TABLE 3-1 Continued

	Type of Evaluation System				
Attribute	Federal HEA-II	National Accreditation[a]	State Program Approval	Media and Independent Ratings[b]	TPP Evaluations for Program Improvement[c]
Likely consequences + Intended - Unintended	+ Very low-performing programs identified, encouraging states to take remedial action or move toward closure - Gaming pass rates	+ TPPs will be aligned with standards for teacher education + Programs may work to address areas of weakness identified - Possibly more conformity and less innovation by programs	+ TPPs will be aligned with state standards for teacher education + Programs may work to address areas of weakness identified - Possibly more conformity and less innovation by programs	+ Improvement due to increased competition and desire to improve ratings - Overreliance on syllabi may create perverse incentives (e.g., to create impressive syllabi that do not reflect actual instruction)	+ Program improvement + Culture of evidence within the program

[a] This reflects CAEP, which is currently the single national TPP accreditation program.

[b] The only existing evaluation of this type is done by NCTQ/*U.S. News*, so this column reflects their approach; other organizations may produce ratings in the future using different types of evidence.

[c] TPP faculty and staff also produce self-studies that are often an integral part of the national accreditation and state program review processes, but here we are referring to internal studies conducted by TPPs for the purpose of program improvement.

PURPOSES BEST SERVED BY EACH EVALUATION APPROACH

Each type of TPP evaluation system relies on somewhat different evidence that can be used to draw different inferences. Each system also creates different incentives and consequences for TPPs. Thus, instead of asking which TPP evaluation approach is *best*, the more appropriate and important question is, *how well does each approach serve a particular purpose?*

Table 3-2 summarizes the committee's conclusions about the strengths and weaknesses of existing systems for particular purposes. For example, the federal approach is not useful for providing TPPs with feedback for improvement; the other four types of systems are better suited to that purpose. On the other hand, the federal approach does fill a unique niche in terms of accountability and monitoring, in that it is the only system that provides *some* data that can be compared across states and can be used to gauge the overall status of teacher education in the nation. Accreditation and state reviews are similar to reviews conducted by the media and other independent organizations in that they are useful for determining whether TPPs have met a certain set of standards for teacher education. But there is an important difference related to who is doing the evaluating. Whenever a review is conducted, there is a concern about the values and self-interests of the evaluator unduly influencing the reports. Accreditation and state reviews may be perceived by some as lacking objectivity, because of the heavy involvement of teacher preparation practitioners in setting the standards and serving as peer reviewers. On the other hand, independent media ratings might not be trusted if the teacher preparation community views them as inadequately sensitive to the needs of the profession and conducted without sufficient attention to technical and procedural practicalities.

In addition to mapping current approaches to the purposes they best serve, Table 3-2 is also intended to provide guidance on the development of new or innovative TPP evaluation systems. These new systems might combine the attributes from existing systems with new types of evidence and other new attributes.

TABLE 3-2 Mapping TPP Evaluation Approaches to Purposes

	Advantages and Disadvantages of Existing Systems				
If the goal is:	Federal HEA-II	National Accreditation[a]	State Program Approval	Media and Independent Ratings[b]	TPP Evaluations for Program Improvement[c]
Accountability and monitoring	+ Gives national picture and enables some cross-state comparisons + Perceived as a legitimate role of the federal government, i.e., providing objective and reliable data (analogous to other federally supported data systems) – Most indicators are at the state level; little information at the program level for evaluating TPPs	+ Indicates that TPPs have met widely agreed upon standards for teacher education ~ Influenced by the teacher preparation community, given its involvement in peer review of TPPs	+ Indicates that TPPs have met state standards for teacher education ~ Influenced by the teacher preparation community, given its involvement in peer review of TPPs	+ Ratings enable comparisons across participating TPPs on a useful but limited set of attributes ~ Reviewers are often from outside the teacher preparation community – Methodological constraints in transforming substantive information from comparative ratings into unidimensional rank orderings	Lacks independence/ objectivity so not used for this purpose

Consumer information	+ Provides a "one-stop" shopping venue for an overview of data (on limited set of attributes) covering all states – Provides inadequate program-specific information to be of use to prospective students or other "consumers"	+ Provides equivalent of "stamp of approval" from credible body of professionals with deep content and pedagogical knowledge – Simple pass/fail designation does not provide readily accessible information to prospective students, who will not be able to sift through large quantities of dense program-specific reports to make informed consumer decisions	+ Indicates that TPPs have met state standards for teacher education + Some states provide public reports that include more detailed information about each TPP than simply approved/not approved	+ Enables comparisons across programs on multiple criteria – Is limited to a subset of relevant indicators of quality that may bias overall comparative impressions	+ Faculty and staff are in the best position to provide detailed information (e.g., via website) about program content, etc., that can be useful to prospective students or other "consumers"

continued

TABLE 3-2 Continued

	Advantages and Disadvantages of Existing Systems				
If the goal is:	Federal HEA-II	National Accreditation[a]	State Program Approval	Media and Independent Ratings[b]	TPP Evaluations for Program Improvement[c]
Program improvement	– Inadequate data, not intended or useful for this purpose	+ Provides in-depth feedback from reputable and neutral expert body; cyclical process fosters continuous improvement – May be perceived as product of "entrenched" system without adequate attention to public (external) preferences or changing norms	+ Some states provide in-depth feedback; cyclical process fosters continuous improvement + VAM can provide useful information about impact on student achievement – Results may be too large-grained to guide program improvement	+ Suggests areas of strength and weakness relative to other programs – Ratings lack credibility, trustworthiness, especially if they are transformed (tacitly or via statistical formulas) into rankings	+ Faculty and staff know the most about their TPP and can identify and discuss areas where improvement is needed – Does not provide a basis for performance to be benchmarked to a defined set of objective standards

NOTE + Advantage; – Disadvantage; ~ May be perceived as an advantage to some, a disadvantage to others

[a] This reflects CAEP, which is currently the single national accreditation program.

[b] The only existing evaluation of this type is done by NCTQ/*U.S. News*, so this column reflects their approach; other organizations may produce ratings in the future using different types of evidence.

[c] TPP faculty and staff also produce self-studies that are often an integral part of the national accreditation and state program review processes, but here we are referring to internal studies conducted by TPPs for the purpose of program improvement.

4

Designing, Using, and Interpreting Teacher Preparation Program Evaluations: Toward A Decision Framework

The federal government, state departments of education, national accreditation bodies, teacher preparation institutions, and media and other organizations that provide information to prospective teachers and their employers face different challenges in designing TPP evaluations. But they share a common need to assess the relative strengths and limitations of different types of evidence for defined evaluation goals. Our goal in this chapter is to provide a practical tool for decision makers working in diverse environments.

We begin with an abridged review of the core assumptions and principles of our report, and then turn to a set of questions that designers and users of TPP evaluation systems might refer to as they consider their specific needs and contexts. The chapter ends with a short list of priority topics for further study.

ASSUMPTIONS AND CORE PRINCIPLES

This report is built on three assumptions:

- The quality of instruction plays a central role in student learning.
- Teacher preparation programs contribute to the quality of instruction.
- Evaluation of teacher preparation programs can provide useful information for improving teacher preparation policy and practice.

As intuitively and logically compelling as these assumptions are, *the details of exactly how differences in instructional method and style affect student learning and how differences in teacher preparation affect instructional quality are not fully understood*. Still, there is little disagreement among professional teacher educators, policy makers, and the general public about the importance of assessing the many ways in which prospective teachers are recruited, selected, prepared, and licensed (or certified) for classroom work. The key challenge faced by designers and users of evaluation systems, therefore, is to understand how those systems vary in the evidence they collect, the purposes they are expected to serve, and their likely effects on valued educational outcomes.

Building off these assumptions, our core principles (see Chapter 1) are the prelude to questions intended to guide a rational and coherent approach to the design and use of TPP evaluations.

1. Although program evaluation is important, it is not sufficient in itself to bring about improvements in teacher preparation, teaching quality, and student learning.
2. Because authority for education in the United States is, by design, diffused, the evaluation of TPPs will likely always include multiple systems operated by different groups with different purposes and interests.
3. Validity—the extent to which evaluation data support specific inferences about individual or organizational performance—should be the principal criterion for assessing the quality of program evaluation measures and systems.
4. Any measure—or, for that matter, any TPP evaluation system that uses multiple measures—has limitations that should be weighed against its potential benefits.
5. Differential effects of TPP evaluation systems—for diverse populations of prospective teachers and the communities in which they may work—matter and should be incorporated as a component of validity analysis and as a design criterion.
6. TPP evaluation systems should themselves be held accountable.
7. TPP evaluation systems should be able to adapt to changing educational standards, curricula, assessments, and modes of instruction.

FROM THEORY TO ACTION

We turn now to a sequence of questions to guide the thinking and work of TPP evaluation designers and users. We believe that focusing attention on these questions will increase the likelihood of creating a coherent evaluation system that serves its intended purposes and leads to valid inferences about TPP quality.

Question 1: What is the primary purpose of the TPP evaluation system?

The TPP evaluation design process should begin with a clear statement about intent: what does the system aim to accomplish? Clearly, evaluation systems will often serve more than one purpose. But designers should be able to articulate the *primary* purpose, and then perhaps one or two secondary purposes. Is the primary goal accountability? (If so, to whom or what authority is the TPP being held accountable?) Is the system intended primarily to provide consumers with information about the quality of individual TPPs? Or is the primary purpose to provide the program itself with information for self-improvement?

Once the central purpose is determined, can a more specific statement be made about what the system is intended to accomplish? For instance, a federal evaluation for accountability may aim specifically to accomplish its main purpose through public reporting of a large variety of state and national data about the quality of teacher preparation. A national accreditation system, also generally aimed at accountability, may more specifically be intended to spur reform in teacher education by implementing rigorous standards that TPPs must meet to earn accreditation.

Being explicit about the purpose of the evaluation is important for at least two reasons. First, the purpose will guide many of the subsequent design decisions. Second, it will be important to communicate the purpose of the evaluation to end users in order to guard against the tendency to use the evaluation results for purposes for which they were not intended and which may not be appropriate.

Question 2: Which aspects of teacher preparation matter the most?

Attributes of teacher preparation that may not be directly observable could be of interest to TPP evaluators. These attributes, which are not necessarily amenable to direct measurement, might include the following:

- Qualifications of students admitted
- Quality and substance of the postsecondary instruction provided to TPP students by all faculty in the program
- Quality of student teaching experience
- Expertise of faculty
- Effectiveness in preparing new teachers who are employable and stay in the field
- Success in preparing teachers who are effective in the classroom

This is not an exhaustive list, and its elements may shift as curricular and instructional reforms are implemented. The point is that *no single evaluation, given the reality of limited resources, will be sufficient to measure every aspect of teacher preparation*. Choices will have to be made about which attributes are of most interest, based on the anticipated uses of the evaluation and the values of the organization doing the evaluating. Evaluators interested in using accountability to spur reform might focus on rigor in the substance of instruction, while those wanting to hold TPPs to a certain minimum standard might focus on course hour requirements and pass rates on licensure tests. Evaluators who are interested in accountability but do not want to prescribe the elements of a high-quality TPP may choose to focus on the extent to which TPPs produce graduates who are employable and demonstrate effectiveness in the classroom.

It is important to maintain a flexible and adaptive approach to evaluation, especially in an era of reforms motivated by changing conceptions about the most valued outcomes of education. Evaluators will face a familiar dilemma: while changing measures to align with new definitions of teaching quality is logical, it reduces the validity of the results as estimates of program improvement over time.

Question 3: What sources of evidence will provide the most accurate and useful information about the aspects of teacher preparation that are of primary interest?

TPP evaluation designers should examine the types of evidence available and decide which will give them the most useful information about the attributes of interest. Because any single type of evidence will give an incomplete picture of TPP quality and because each type of evidence has limitations, a good evaluation system will carefully select the combination of measures that makes the best use of available resources. Evaluators should carefully consider *each type of evidence* that might be used and what can be learned from that data, while attending to its advantages and disadvantages.

These are the most important considerations in addressing this question:

- How much effort is involved in collecting the data?
- Will double-checking for accuracy be feasible?
- How prone is the measure to gaming and corruption?
- Is there empirical evidence tying the measure to future teacher performance? Or can a rationale for using the measure be based on its face validity—that it subjectively appears to measure an important construct?

Given limited resources, investing in one measure will likely mean giving less attention to TPP attributes that are not included. The question then becomes whether, on balance, the types of evidence included will lead to the desired inferences about overall TPP quality. Table 4-1 summarizes the strengths and limitations of the most commonly used measures of TPP quality.

Inputs, Outputs, and Outcomes: Context and Clarification

In the past quarter century, education accountability has shifted markedly from a heavy reliance on input measures toward a greater emphasis on measures of educational outcomes. This shift has influenced the design and purposes of TPP evaluation (Crowe, Allen, and Coble, 2013). As explained in Chapter 2, TPP evaluation data are commonly categorized as input and output measures. Examples of input measures include average SAT scores of incoming students, course syllabi, and required hours for student teaching. Examples of output measures are pass rates on licensure tests, graduates' VAM estimates, and surveys of employers. The major national accrediting organization, CAEP, favors a shift toward using more output measures, in part because there is not a solid empirical base that links input characteristics of TPPs to measures of student achievement (Sawchuk, 2013a). However, focusing on these outputs raises a problem of selectivity bias: are the effects of teacher preparation courses confounded with prior attributes of the admitted candidates?

Our recommendation is to avoid framing this issue as an either-or proposition. It is not a question of input versus output measures, as if that distinction were a sufficient design criterion. Instead of assuming that either input or output measures are generally superior, we suggest that each type of evidence be considered in terms of its relative strengths and limitations and what can be learned from it.

Consider, for example, a widely used input measure: course syllabi. Although syllabi may appear to offer a more relevant proxy for the substantive content and quality of teacher preparation than the scores of teacher candidates on admissions tests, it is important to recall the distinctions between the "intended" and "enacted" curricula. In other words, what appears in printed syllabi may bear little resemblance to what is actually taught in TPP courses. In a high-stakes environment—that is, if syllabi are perceived as the basis for important decisions by prospective students—then the use of such a measure might create an incentive for faculty or program administrators to produce impressive syllabi that exaggerate what is taught in the course, thereby corrupting the measure and undermining its usefulness.

TABLE 4-1 Strengths and Limitations of Commonly Used Measures of TPP Quality

Measure	Strengths	Limitations
Admissions and Recruitment Criteria		
Average GPA of incoming class	Single number representing academic ability of the student body Easy to collect Easily understood by the general public as an approximation of overall level of incoming students	Grading is not uniform across educational institutions Grades are weak indicators of the quality of training provided by TPP Average GPA may be less important than the minimum required
Average entrance exam scores	Single number representing academic ability of the student body Some research shows positive link between candidates' performance on entrance exams and the achievement of candidates' eventual students Easy to collect Standardized measure that makes for easy point of comparison Familiar to the public	Criticized for simply being a measure of socioeconomic status Average entrance exam scores are weak indicators of the quality of training provided by TPP
Percentage of minority students in incoming class	Encourages TPPs to recruit minority candidates Easy to collect Easy to make comparisons across programs Easily understood by the public	Minority participation rate is a weak indicator of the quality of training provided by TPP May provide incentive for program to admit students who are academically unprepared and end up dropping out

TABLE 4-1 Continued

Measure	Strengths	Limitations
Number of candidates admitted in high-need areas (e.g., teachers of STEM, special education, English language acquisition)	Encourages TPPs to recruit candidates to teach in high-need areas Easy to collect Easy to make comparisons across programs	Distribution of admitted candidates by content area concentration is a weak indicator of the quality of training provided by TPP
Quality and Substance of Instruction		
Course syllabi	Contract or agreement that a course will cover certain material Less costly than actually observing courses	Syllabi may reflect intended curriculum vs. enacted curriculum (what is actually taught) Process must be developed and implemented to enable reliable coding of syllabi; can be labor intensive Syllabi may not reflect instruction—many syllabi are terse, faculty may alter courses mid-stream, and using results for high-stakes decisions may corrupt validity
Lectures and assignments	May be a more accurate reflection than syllabi of what is actually taught	Process must be developed and implemented for reliably coding documents; can be labor intensive Quantity of documents that needs to be collected and coded makes this costly Reflect content of instruction, but not quality of instruction

continued

TABLE 4-1 Continued

Measure	Strengths	Limitations
Textbooks	Can give additional information about course coverage	Not all material in the textbook may be covered in the course; other material may be added Process must be developed and implemented for analyzing textbook content; can be labor intensive
Course offerings and required hours	Easy to collect Easy to make comparisons across programs	Does not indicate actual quality of instruction in TPP courses
Number of required content courses	Evidence of positive effect on student achievement, especially for secondary mathematics teachers	Courses may not cover content most important for effective K-12 teaching
Quality of Student Teaching Experience		
Fieldwork policies including required hours	Easy to collect Easy to make comparisons across programs	Does not indicate actual quality of fieldwork experience
Qualifications of fieldwork mentors	One aspect of quality of fieldwork experience	Little empirical evidence links characteristics of mentors to their success in teacher preparation
Surveys of candidates	TPP students can report on actual experience in the field, e.g., frequency of observations, specificity of feedback	Requires development of survey and analysis of responses, which may be time-consuming Based on individual perceptions; may be biased
Records from observations of student teaching	Can gauge quality of feedback from mentor Can assess whether candidates are applying what they have learned in the TPP	Requires developing and implementing a method to analyze observation records for TPP evaluation purposes; can be labor intensive

TABLE 4-1 Continued

Measure	Strengths	Limitations
Faculty Qualifications		
Percentage of faculty with advanced degrees, part-time, adjunct, etc.	Easy to collect East to make comparisons across programs Face validity—TPP faculty should have appropriate expertise and credentials	Many instructors of teacher candidates are in departments other than education and tend not to be included in the evaluation Little empirical evidence to support connection to effective teacher preparation
Effectiveness in Preparing Candidates Who Are Employable and Stay in the Field		
Pass rates and/or average scores on licensure tests	Easy to collect	Wide variety in tests and cut scores makes comparisons difficult, especially across states Controversy over rigor and relevance of current exams Often misinterpreted: indicates that candidates have minimum competencies to enter teaching profession but does not predict future effectiveness in the classroom May be corrupted (e.g., requiring TPP students to pass a test in order to graduate to ensure 100% pass rates)
Hiring and retention data	Important to potential candidates; face validity	Influenced by numerous geographic and non-TPP factors May be inaacurate and/or difficult to collect; have to track graduates post-TPP

continued

TABLE 4-1 Continued

Measure	Strengths	Limitations
Success in Preparing High-Quality Teachers		
Teacher performance or portfolio assessments administered near end of program	Detailed and comprehensive measure of candidates' skills, results of which can be aggregated to make judgments about TPP outputs Some evidence shows that these can predict future classroom performance	Costly to administer and score Validity issues arise when candidates can choose what to include in their portfolios
Ratings of graduates by principals/employers	High face validity Some research shows that principals can accurately identify teachers with low VAM scores	May be costly or time-consuming to gather Subjective; may be biased
Value-added models	Measures teacher impact on student achievement, while attempting to take into account out-of-school factors that affect achievement	Requires state to have VAM system in place (not currently the case in most states) Numerous methodological issues related to reliability and validity still need to be addressed Incomplete data Difficult to explain and understand

Similarly, consider the increasingly popular output measures that come under the general heading of "value-added models," or VAM. These measures purport to link TPP attributes to the subsequent performance of teachers in their classrooms and hold special appeal because they aim to isolate the effects of classroom teachers from the many other factors that affect K-12 student achievement. But VAMs also pose a number of challenges, such as determining the extent to which observed differences are due to training rather than selection effects or addressing the problem posed by the many graduates omitted from the analysis because they teach untested subjects or grades or have left the state. But just because these measures have limitations does not mean they should be put aside; rather, decision makers must weigh the pros and cons of each type of measure and decide whether, on balance, it is worth using.

If important goals for TPPs are to bring more minorities and individuals from disadvantaged backgrounds into the teaching profession and to supply schools in impoverished communities with effective teachers, then it is especially important to focus attention on how different measures will honor and reinforce—or undermine—those values. Regardless of whether the evaluation system is being planned at the federal, state, or institutional level, a key consideration should be whether the measures selected might produce results that diminish the supply of teacher candidates who are willing and able to work in poor or rural areas. If an evaluation system were to produce this unintended effect, then it would undermine the goal of enhanced education for all children, regardless of where they live.

We should underscore that we are *not* talking about different standards for different TPPs. Rather, our concern is that the design of TPP evaluations should not inadvertently cause or exacerbate inequities in resource allocation, especially in communities with the greatest need for high-quality teachers. An overreliance on selectivity (using input measures, such as performance on standardized admissions tests) might unintentionally lead not only to the unfair misclassification of programs but ultimately to the perpetuation of disadvantage in the communities that most need effective teachers.[1] Including measures that "give credit" for diversity in admissions should be considered a safeguard against this undesirable outcome.

Question 4: How will the measures be analyzed and combined to make a judgment about program quality?

Evidence or data do not automatically translate into evaluation results. Decisions must be made about how the data will be analyzed and interpreted. If admissions or licensure tests are used, evaluation designers will need to decide whether to employ average scores and/or pass rates. The latter implies the need to determine cut scores (passing scores), which is a somewhat complex technique in itself. (There is a body of literature from the field of testing about cut scores. See, e.g., Cizek and Bunch, 2007.)

[1] The possibility that error in classification based on quantitative measures might disproportionately and systematically affect certain individuals or groups has been a central issue in measurement for at least a half century. See, for example, National Research Council, 1982, 1989.

For other types of evidence, scoring rubrics (guidelines) will need to be developed. For instance, if course syllabi are collected to assess the substance of instruction, rubrics will be needed to specify the characteristics that must be evident in the syllabus to demonstrate that it meets a certain standard. Here, too, designers need to be aware of the subtleties and complications of establishing the rubrics. In any event, raters will need to be trained and monitored to ensure that they code documents reliably.

If the goal is to come up with a single indicator of TPP quality, as is often the case with evaluations for accountability or consumer information purposes, evaluation designers must make additional decisions about how to weight and combine the various sources of evidence. The single indicator of quality may be a pass/fail decision, a ranking of programs from highest to lowest quality, or some other sort of summary rating. Several questions should be considered. For example, will TPPs be required to meet a certain level on each measure (referred to as a *conjunctive* model)? Or will a high score on one measure be allowed to compensate for a low score on another (a *compensatory* model)? Does each piece of evidence carry the same weight, or is one measure more important than another? Or will the measures each be reported on separately, leaving it to users to make their own summary judgments?

In order to earn CAEP accreditation, for example, a TPP must demonstrate to the review team that it meets each of five major standards. Based on documentation and site visits, review teams rate the TPP in one of three levels on each standard: *unacceptable, acceptable* or *target* (Council for the Accreditation of Educator Preparation, 2013b). TPPs must meet at least the acceptable level on each standard to earn accreditation. With their consumer-oriented rankings, NCTQ/*U.S. News* gives each TPP a score on each standard, while weighting some standards more heavily than others in computing the overall ratings (National Council on Teacher Quality, 2013). The overall score consists of a weighted sum of the component ratings; this is a compensatory model because a high score on one standard can help make up for a low score on another. In contrast, Ohio produces a set of "performance reports" on each of the state's TPPs. The reports seek to give the public information about how well the states' TPPs are operating. They report on a number of variables separately and intentionally avoid the assignment of an overall score or grade (Bloom, 2013).

Some flexibility may need to be built into the analysis of data for the sake of equity. Ideally, evidence will be interpreted within the context of program participants, resources, and communities served by the TPPs. This may include, but not be limited to, demographics, ecological/environmental context, and policy climate. To yield an overall judgment about TPP quality, a compensatory model might give TPPs credit for seeking diversity in their candidate population or being located in a disadvantaged community; these can make up for lower scores on other indicators.

Question 5: What are the intended and potentially unintended consequences of the evaluation system for TPPs and education more broadly?

Consequences of evaluation should be determined with the overall goal of improving teacher preparation, rather than punishing or embarrassing low-performing programs. The results of a TPP evaluation aimed at program improvement might be shared and discussed only among internal users to enable them to identify steps for improvement. Systems aimed at producing consumer information will publicize the results more broadly. Evaluations for accountability may be publicized and may also carry higher stakes that could include directives, mandates, or even program closures. If the results trigger serious consequences, then ideally the initial evaluation should be followed up by a more in-depth one to ensure that the TPP was not wrongly identified as low performing. This is especially important when relying on measures like VAMs, which are a few steps removed from the actual training taking place in a TPP and have problems of measurement error.

Decision makers should also try to anticipate unintended negative consequences of the system. Is the evaluation likely to identify a disproportionate number of TPPs in disadvantaged communities as failing? If those TPPs are closed or sanctioned, what impact will that have on the production of minority teachers? And how will this closure affect the supply of teachers in the community where the TPP is located? Can decision makers avoid these negative consequences by thinking early in the process about how the results of an evaluation will be used? If, as we assume, the overarching goal is to improve the quality of teacher preparation, a first step could involve anticipating the likely need to allocate extra resources to TPPs that need them to make improvements.

Question 6: How will transparency be achieved? What steps will be taken to help users understand how to interpret the results and use them appropriately?

Transparency, or open communication, is crucial if users are to trust the results of an evaluation. Those who design and implement TPP evaluations have the responsibility to clearly communicate the purpose for the evaluation and the methods used to collect and analyze the data. It is also important to communicate appropriate interpretations of the results, along with the limitations in what one can infer from the data. One caution, for example, is that while an evaluation system may be adequate for approximating the general quality of an entire program, the result may not pertain to the quality of specific individual graduates. This is one example of what is known as classification error in measurement: good teachers may come from programs that are labeled as poor or substandard, and inferior teachers may come from programs that received an overall high rating. All of the information about the evaluation should be easily accessible on the Internet or otherwise and communicated in a way that is easily understood by users and the public.

Transparency is especially important for technically sophisticated measures like VAMs. Research in the neurosciences and mathematics suggests that people tend to believe data that they do not understand very well (Weisberg, Keil, Goodstein, Rawson, and Gray, 2008; Eriksson, 2012) because of the way the data are presented. Sperber (2010) calls this the "guru effect," which occurs when readers judge technical information as profound without really understanding how it was derived. VAMs, like many contemporary measurement systems, rely on complex statistical models that lead to a heightened perception of their scientific accuracy. Admonitions from psychometricians, who know the most about the potential for error in these systems and who caution against their overuse, are often ignored or dismissed by policy and education communities eager to treat the quantitative data as scientifically valid and therefore trustworthy. Thus, evaluation designers must make special efforts to convey the limitations of VAM results in terms of the validity of interpretations that can be drawn from them.[2]

But transparency is important with all types of measures, quantitative and qualitative, even those that seem more intuitively understandable. Users should be reminded, for instance, that syllabi may not reflect the actual content of instruction as delivered, that licensure tests are not designed to predict future teacher performance, and that hiring and placement results are something TPPs generally have little control over. Developing innovative and effective ways to promote transparency should become a research priority, as discussed below.

Question 7: How will the evaluation system be monitored?

One should not assume that an evaluation system is functioning as envisioned and producing the intended impacts on teacher preparation. Consequences of the system, both intended and unintended, should be studied. For the program improvement purpose of evaluation, for example, key issues are whether the evaluation promotes increased com-

[2] The idea of requiring labels on score reports as an antidote to overinterpretation of their accuracy was suggested in National Research Council, 1999.

munication among faculty about how they can improve teacher training at their institution; whether evaluation results encourage specific actions to improve the program; the extent to which the evaluation creates incentives for opportunistic behavior that distort the meaning of the results; and whether different groups of teacher educators are affected differently and perhaps unfairly by the application of evaluation results.

In addition to monitoring consequences of the system, evaluation leaders should arrange for ongoing studies of the accuracy and reliability of the measures and analytic methods being used in the system. If documents are being coded, auditing studies can be conducted to check on rater agreement. To the extent possible, validity studies should be conducted to see if the ratings that result from the evaluation correlate with other, independent indicators of TPP quality. Are the results of the evaluation corroborated by other evidence not used in the evaluation? States that rely heavily on VAM results, for instance, might conduct surveys of graduates to see if their perceptions support the conclusions drawn from the VAMs about highest- and lowest-performing TPPs.

Evaluation systems should be flexible and adaptable. Earlier we noted that changing standards for K-12 STEM education will require changes in TPPs, as they align their programs with the new expectations for teacher training and recruitment. Likewise, evaluations of TPPs will need to adapt to measure how well programs are meeting the new STEM goals, according to an appropriate timeline that allows TPPs adequate time to adjust.

Of course, there is a tension between adaptability and stability in an evaluation system. Keeping measures and analytic methods stable is important to allow results to be compared from one year to the next for purposes of tracking trends in teacher preparation. Thus, decisions will have to be made about whether a certain change to the system will have enough positive impact on teacher preparation to counterbalance some loss of comparability in the data.

Holding evaluation systems accountable is necessary for building trust in the communities most likely to use and be affected by their results (Feuer, 2012b). Ultimately, a major purpose of evaluation is to contribute to the improvement of student learning and other valued educational outcomes. For this goal to be advanced, designers and operators of teacher preparation program evaluations need to consider the extent to which they build or erode trust among the professionals who prepare future educators and among the participants in those programs.

As a recap, Box 4-1 briefly summarizes the main questions to be addressed in the development of TPP evaluation systems.

BOX 4-1
Decision Framework for Constructing or Revising a TPP Evaluation System

Below are the key questions that designers and users of TPP evaluation systems should address. Referring to these questions early and often will increase the likelihood of creating a coherent evaluation system that serves its intended purposes and leads to valid interpretations about TPP quality.

Question 1: What is the primary purpose of the TPP evaluation system?
Question 2: Which aspects of teacher preparation matter the most?
Question 3: What sources of evidence will provide the most accurate and useful information about the aspects of teacher preparation that are of primary interest?
Question 4: How will the measures be analyzed and combined to make a judgment about program quality?
Question 5: What are the intended and potentially unintended consequences of the evaluation system for TPPs and education more broadly?
Question 6: How will transparency be achieved? What steps will be taken to help users understand how to interpret the results and use them appropriately?
Question 7: How will the evaluation system be monitored?

A FOCUS ON PURPOSE

When designing an evaluation for the purpose of *program self-improvement*, issues related to the motivation of faculty and staff will need to be considered. An advantage of conducting an evaluation for program improvement is that faculty and staff are generally more willing to be honest about program weaknesses because it is an internal rather than a public conversation. At the same time, people directly connected to a TPP may be somewhat complacent about or unaware of ongoing problems that would have been identified by an independent review. Since the evaluation has not been mandated by a government agency or other authority, the leader of the evaluation will have to think about how to get the faculty engaged in the process. People will have to be guided to think about the program as a whole, rather than their own little piece of it. They will have to be encouraged to think outside of the box, be open to major changes that might be indicated, and not limit themselves to tinkering with minor details of the program as it currently exists.

A particular concern when designing an evaluation for *accountability* is the corruption of measures. Attaching serious consequences to evaluation results can create incentives for people to increase performance in the easiest ways possible and can lead to gaming of the system. Even evaluation systems that do not take punitive action against low-performing TPPs can be considered high-stakes; simply publicizing results that identify the best and worst programs can put serious pressure on TPP faculty and staff. When choosing the measures that will be used in an evaluation aimed at accountability, decision makers should consider the extent to which a measure is prone to corruption. For example, syllabi are more prone to being "faked" than licensure test or VAM results. Once an evaluation system is implemented, it should be monitored on an ongoing basis to make sure that measures are not being corrupted.

Evaluations designed by the media or other independent organizations to provide consumers with information also carry high stakes in the form of good or bad publicity, and therefore issues of corruptibility of measures also apply. In this case, designers of these evaluations need to pay special attention to issues of trust and how the evaluation results are likely to be received by the education community. Whenever evaluative judgments are made, there is concern about the values and biases of the evaluator unduly influencing the results. If teacher preparation practitioners are not centrally involved in designing the evaluation, independent media ratings may not be trusted and may be viewed as insensitive to the needs and practicalities of the profession. At the same time, evaluators who are external to the profession may be more objective. Trust in the system will probably be best served by forming an advisory group that includes a balance of individuals from inside and outside the field of teacher preparation.

PRIORITIES FOR FUTURE RESEARCH AND DEVELOPMENT

A number of areas warrants further analysis to improve TPP evaluation. The committee has identified the following priorities for research and development:

How do differences in teacher preparation affect graduates' effectiveness in the classroom? The details of how differences in teacher preparation affect teachers' later instructional quality are only partially understood. A strong consensus exists about the need for ongoing, high-quality research on these issues. General sources of guidance for what teacher preparation should look like and entail are plentiful: substantial content knowledge and pedagogical knowledge, extensive clinical experiences, and the like. The history of teacher education comes back to these features again and again. Yet, as Wilson (2013) point outs, while there may appear to be considerable agreement over basic principles, the details can be devilish. In STEM in particular, the current rhetoric involves a level of abstraction that masks the considerable variation on the ground in how programs implement commitments to clinical experience or content knowledge, or collaborations with STEM disciplinary faculty or preK-12 school professionals. More needs to be done to move from general recommendations to detailed descriptions of what effective teacher preparation entails.

How might comprehensive measures of teacher effectiveness, including non-cognitive student output measures, be integrated into evaluation systems? There is a need for more innovative and comprehensive output measures that move beyond defining teacher effectiveness as simply growth in student achievement test scores. Heckman and Kautz (2012), for example, admonish that achievement test scores are not as predictive of student success in school, career, or health as are other factors. Their research in economics makes a case for paying greater attention in schools to "soft skills," such as conscientiousness, motivation, and curiosity that have been shown to predict success in life. In other words, research is showing the importance of other student outputs beyond achievement test scores in math and reading and needs to be integrated into the study of teacher preparation and its evaluation.

How do different TPP evaluation systems affect teacher preparation? TPP evaluations are being implemented with very little knowledge of their impacts on teacher preparation. We need to know more about what happens as a result of these evaluations and use that knowledge to improve the systems. What is the impact of using VAMs for state TPP evaluations? How will the new CAEP accreditation standards affect teacher preparation? How will standards calling for greater selectivity shape the pool of new teachers, particularly the supply of minority teachers and placements in hard-to-staff schools?

How could different requirements for explaining the strengths and limitations of evaluation systems improve transparency, communication, and trust? The NRC report *High Stakes* (1999) includes an important suggestion about how *labeling* might be used to promote appropriate test use in K-12 schools. The report proposes that high-stakes testing programs might be required to supply certain types of information to users (including educators and parents), such as the purpose of the test, how individual test results will be used, whether these results will be the sole basis for a particular decision or whether other indicators will be used, evidence of the validity of the results, and so on. Parallels are drawn to various "right-to-know" policies that provide information to the public about the health risks and benefits associated with various drugs, food products, and toxins. The assumption behind these policies is that disclosures will correct the information imbalance between produces and consumers, enabling people to make informed purchases and participate more equitably in public decisions. Would such an approach, i.e., the requirement for greater transparency about the intricacies of evaluation systems, provide users (policy makers, TPPs, prospective students and employers) with useful and relevant information about appropriate interpretations of results and their limitations? It would be worth investing in studies to ascertain how to best communicate key information in an understandable way and whether doing so actually leads to more appropriate use of evaluation results and better consequences for teacher education.

References

100K-in-10. (n.d.). Providing America's classrooms with 100,000 excellent science, technology, engineering, and math (STEM) teachers by 2021 [Web page]. Retrieved from http://www.100kin10.org/.

Aldeman, C. (2012). *Teacher preparation strategy.* Presentation to the National Academy of Education Steering Committee on the Evaluation of Teacher Education Programs: Toward a Framework for Innovation, June 25.

Aldeman, C., Carey, K., Dillon, E., Miller, B., and Silva, E. (2011). *A measured approach to improving teacher preparation.* Retrieved from the Education Sector website: http://www.educationsector.org/sites/default/files/publications/TeacherPrep_Brief_RELEASE.pdf.

American Board of Internal Medicine. (n.d.). Exam pass rates [Web page]. Retrieved from http://www.abim.org/about/examInfo/data-pass-rates.aspx.

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.

American Evaluation Association. (n.d.). The program evaluation standards. Retrieved from http://www.eval.org/evaluationdocuments/progeval.html.

American Federation of Teachers. (n.d.). *A guide for developing multiple measures for teacher development and evaluation.* Retrieved from http://www.aft.org/pdfs/teachers/dev multiplemeasures.pdf.

Anderson, J. (2013, March 30). Curious grade for teachers: Nearly all pass. *New York Times*. Retrieved from http://www.nytimes.com/2013/03/31/education/curious-grade-for-teachers-nearly-all-pass.html?pagewanted=all&_r=0.

Association of Public and Land-Grant Universities. (n.d.). Science and mathematics teacher imperative [Web page]. Retrieved from http://www.aplu.org/page.aspx?pid=2182.

Association to Advance Collegiate Schools of Business. (2005). *The business schools rankings dilemma*. Tampa, FL: Author. Retrieved from http://www.aacsb.edu/publications/researchreports/archives/rankings.pdf.

Ball, D. L., and Forzani, F. M. (2011). Building a common core for learning to teach, and connecting professional learning to practice. *American Educator, 35*(2), 17-21, 38-39.

Ballou, D., and Podgursky, M. (1999). Teacher training and licensure: A layman's guide. In M. Kanstoroom and C. E. Finn, Jr. (Eds.), *Better teachers, better schools*. Washington, DC: Thomas B. Fordham Foundation.

Berliner, D. C. (2012). Effects of inequality and poverty vs. teachers and schooling on America's youth. *Teachers College Record, 116*(1). Retrieved from http://www.tcrecord.org/Content.asp?ContentID=16889.

Berry, B., Smylie, M., and Fuller, E. (2008). *Understanding teacher working conditions: A review and look to the future.* Hillsborough, NC: Center for Teaching Quality.

Bill & Melinda Gates Foundation. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET Project's three-year study* (Final research report). Retrieved from http://www.metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf.

Bloom, M. (2013). *Five things to learn from Ohio's new teacher preparation program evaluations.* Retrieved from http://stateimpact.npr.org/ohio/2013/01/22/five-things-to-learn-from-ohios-new-teacher-preparation-program-evaluations/.

Boser, U. (2012). *Race to the Top: What have we learned from states so far?* Washington, DC: Center for American Progress. Retrieved from http://www.americanprogress.org/wp-content/uploads/issues/2012/03/pdf/rtt_states.pdf.

Boyd, D., Grossman, P., Landford, H., Loeb, S., and Wyckoff, J. (2008). *Teacher preparation and student achievement.* National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w14314.pdf.

Bradley, A. (2000, May 24). NCATE unveils standards based on performance. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2000/05/24/37ncate.h19.html.

Bruenig, M. (2013, April 25). The STEM-shortage myth. *The American Prospect.* Retrieved from http://prospect.org/article/stem-shortage-myth.

Bryk, A. (2012, September 20). *Setting the context for teacher evaluations.* Introductory remarks at the Revisiting Teacher Evaluation Forum, Washington, DC. Retrieved from http://vimeo.com/53124661.

Campbell, D. T. (1976). *Assessing the impact of planned social change.* Hanover, NH: The Public Affairs Center, Dartmouth College.

Carnoy, M., and Rothstein, R. (2013). What do international tests really show about U.S. student performance? Retrieved from Economic Policy Institute website: http://www.epi.org/publication/us-student-performance-testing/.

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

Clotfelter, C. T., Ladd, H. F., and Vigdor, J. L. (2007). Teacher credentials and student achievement in high school: A cross-subject analysis with student fixed effects. *The Journal of Human Resources, 45*(3), 655-681.

Coble, C. R. (2012). *Developing the Analytic Framework: Assessing innovation and quality design in science and mathematics teacher preparation.* Washington, DC: Association of Public and Land-Grant Universities. Retrieved from http://www.aplu.org/document.doc?id=3652.

Coble, C. R, DeStefano, L., Shapiro, N., Frank, J., and Allen, M. (n.d.). *The Analytic Framework: A tool for assessing innovation and quality design in science and mathematics teacher preparation*. Washington, DC: Association of Public and Land-Grant Universities. Retrieved from http://www.aplu.org/document.doc?id=4106.

Cochran-Smith, M., and Zeichner, K. M. (2005). *Studying teacher education: The report of the AERA Panel on Research and Teacher Education.* Retrieved from http://books.google.com/books?id=JYywbNudIoMC&pg=PA8&dq=teacher's+characteristics+research+on+the+indicators+of+quality&hl=en&sa=X&ei=j0eBUbSwBuSUiAKynoEY&ved=0CDkQ6AEwAA#v=onepage&q=teacher's%20characteristics%20research%20on%20the%20indicators%20of%20quality&f=false.

Coggshall, J. G., Bivona, L., and Reschly, D. J. (2012). *Evaluating the effectiveness of teacher preparation programs for support and accountability*. Washington, DC: National Comprehensive Center for Teacher Quality.

Common Core State Standards Initiative. (n.d.). *Standards for mathematical practice.* Retrieved from http://www.corestandards.org/Math.

Council for the Accreditation of Educator Preparation. (2013a). *Annual report.* Retrieved from http://caepnet.files.wordpress.com/2013/05/annualreport_final.pdf.

Council for the Accreditation of Educator Preparation. (2013b). *CAEP accreditation standards and evidence: Aspirations for educator preparation.* Retrieved from http://caepnet.files.wordpress.com/2013/02/commrpt.pdf.

Cremin, L. (1990). *Popular education and its discontents*. New York: Harper & Row.

Cronbach, L. J. (1975). Five decades of public controversy over mental testing. *American Psychologist, 30*, 1-14.

Cronbach, L. J., and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

Crowe, E. (2010). *Measuring what matters: A stronger accountability model for teacher education*. Washington, DC: Center for American Progress.

Crowe, E., Allen, M., and Coble, C. (2013, June 11). Time for progress in teacher prep. *Education Week.* Retrieved from http://www.edweek.org/ew/articles/2013/06/12/35crowe.h32.html?qs=crowe+teacher+licensure+tests.

Danielson Group. (2013). *The framework for teaching evaluation instrument.* Retrieved from http://www.danielsongroup.org/userfiles/files/downloads/2013EvaluationInstrument.pdf.

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

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., and Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan, 93*(6), 8-15.

Darling-Hammond, L., Newton, S., and Wei, R. C. (2013). *Developing and assessing beginning teacher effectiveness: The potential of performance assessments.* Retrieved from Stanford Center for Opportunity Policy in Education website: http://edpolicy.stanford.edu/sites/default/files/publications/developing-and-assessing-beginning-teacher-effectiveness-potential-performance-assessments.pdf.

Duncan, A. (2009, October 22). *Teacher preparation: Reforming the uncertain profession.* Remarks at Teachers College, Columbia University. Retrieved from http://www2.ed.gov/news/speeches/2009/10/10222009.html.

Duncan, G., and Murnane, R. (Eds.). (2011). *Whither opportunity: Rising inequality, schools, and children's life chances*. New York: Russell Sage and Spencer Foundations.

Education Commission of the States. (2013). *Eight questions on teacher preparation: What does the research say?* Retrieved from http://www.ecs.org/html/educationissues/teachingquality/tpreport/home/summary.pdf.

Eriksson, K. (2012). The nonsense math effect. *Judgment and Decision Making, 7*(6), 746.

Ferguson, R. F., and Ladd, H. F. (1996). How and why money matters: An analysis of Alabama schools. In H. F. Ladd (Ed.), *Holding schools accountable: Performance-based reform in education* (pp. 265-298). Washington, DC: The Brookings Institution.

Ferguson, R., and Ramsdell, R. (2011). *Tripod classroom-level student perceptions as measures of teaching effectiveness.* Retrieved from http://www.gse.harvard.edu/ncte/news/NCTE_Conference_Tripod.pdf.

Feuer, M. (2006). *Moderating the debate: Rationality and the promise of American education*. Cambridge, MA: Harvard Education Press.

Feuer, M. (2008). Future directions for educational accountability: Notes for a political economy of measurement. In L. A. Shepard and K. Ryan (Eds.), *The future of test-based educational accountability*. New York: Routledge.

Feuer, M. J. (2010). Externalities of testing: Lessons from the blizzard of 2010, *Measurement: Interdisciplinary Research and Perspectives, 8,* 59–69.

Feuer, M. (2012a, August). *No country left behind: Notes on the rhetoric of international comparisons of education.* William Angoff Invited Lecture, Educational Testing Service, Princeton, NJ.

Feuer, M. (2012b, September 24). Measuring accountability when trust is conditional. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2012/09/24/05feuer_ep.h32.html.

Feuer, M. (2013a). International large-scale assessments: Validity in presentation and use. In M. Chadhabi (Ed.), *Validity issues in international large scale assessment*. New York: Routledge (forthcoming).

Feuer, M. (2013b). STEM education: Progress and prospects. *The Bridge*, *43*(1), 3-6.

Feuer, M. J. (2013c, April 9). It's not the test that made them cheat. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2013/04/09/28feuer.h32.html.

Fleener, M. J., and Exner, P. D. (2011). Dimensions of teacher education accountability: A Louisiana perspective on value-added in teacher education policy in the United States. In P. Earley, D. Imig, and N. Michelli (Eds.), *Issues and tensions in an era of evolving expectations*, pp. 26-43. New York: Routledge.

Floden, R., and Meniketti, M. (2005). Research on the effects of coursework in the arts and sciences and in the foundations of education. In M. Cochran-Smith and K. Zeichner (Eds.) *Studying teacher education: The report of the AERA panel on research and teacher education,* pp. 261-308. Washington, DC: American Educational Research Association.

Fowler, R. C. (2001). What did the Massachusetts teacher tests say about American education? *The Phi Delta Kappan*, *83*(10).

Fuhrman, S., and Elmore, R. (Eds.). (2004). *Redesigning accountability systems for education*. New York: Teachers College Press.

Furlong, J. (2013). *Inspecting initial teacher education in England—the work of Ofsted.* Paper commissioned for the National Academy of Education Steering Committee on the Evaluation of Teacher Education Programs: Toward a Framework for Innovation.

Futernick, K. (2007). *A possible dream. Retaining California teachers so that all students learn.* Retrieved from http://www.calstate.edu/teacherquality/documents/possible_dream_exec.pdf.

Gansle, K. A., Noell, G., and Burns, J. M. (2013). Do student achievement outcomes differ across teacher preparation programs? An analysis of teacher education in Louisiana. *Journal of Teacher Education*, *63*(5), 304-317.

Gitomer, D. H., and Latham, A. S. (1999). *The academic quality of prospective teachers: The impact of admissions and licensure testing.* Retrieved from the Educational Testing Service website: http://www.ets.org/Media/Research/pdf/RR-03-35.pdf.

Goldhaber, D. (2006). National Board teachers are more effective, but are they in the classrooms where they're needed the most? *Education Finance and Policy*, *1*(3), 372-382.

Goldhaber, D., and Liddle, S. (2012). *The gateway to the profession: Assessing teacher preparation programs based on student achievement* (CALDER Working Paper No. 65). Retrieved from http://www.caldercenter.org/upload/Goldhaber-et-al.pdf.

Goldin, C., and Katz, L. (2008). *The race between education and technology*. Cambridge, MA: Belknap Press.

Goodlad, J. (1990). *Teachers for our nation's schools.* San Francisco, CA: Jossey-Bass.

Greenberg, J., and Walsh, K. (2012, August 27). EdTPA: Slow this train down [Web log post]. Retrieved from http://www.nctq.org/p/tqb/viewStory.jsp?id=32495.

Grossman, P., Loeb, S., Cohen, J., Hammerness, K., Wyckoff, J., Boyd, D., and Lankford, H. (2010, May). *Measure for measure: The relationship between measures of instructional practice in middle school English Language Arts and teachers' value-added scores* (NBER Working Paper No. 16015). Retrieved from http://www.nber.org/papers/w16015.pdf.

Hanushek, E. (2010). The economic value of higher teacher quality (CALDER Working Paper No. 56). Retrieved from http://www.caldercenter.org/UploadedPDF/1001507-Higher-Teacher-Quality.pdf.

Hanushek, E., Peterson, P. E., and Woessmann, L. (2012). Achievement growth: International and U.S. state trends in student performance. Retrieved from http://hanushek.stanford.edu/sites/default/files/publications/Hanushek%2BPeterson%2BWoessmann%20 2012%20PEPG.pdf.

Harris, D. N. (2011). *Value-added measures in education: What every educator needs to know*. Cambridge, MA: Harvard Education Press.

Harris, D. N., and Sass, T. R. (2009). *What makes for a good teacher and who can tell?* (Working Paper 30, National Center for the Analysis of Longitudinal Data in Education Research). Retrieved from http://www.urban.org/uploadedpdf/1001431-what-makes-for-a-good-teacher.pdf.

Harvard Business Review. (2009, March-May). The HBR debate: How to fix business schools [Web log posts]. Retrieved from http://blogs.hbr.org/how-to-fix-business-schools/.

Heckman, J. J., and Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics, 19*, 451-464.

Henry, G. T., Bastian, K. C., and Smith, A. A. (2012). Scholarships to recruit the "best and brightest" into teaching. *Educational Researcher*, *41*(3), 83-92.

Heydman, A. M., and Sargent, A. (2011). Planning for accreditation: Evaluating the curriculum. In Keating, S. B. (Ed.), *Curriculum development and evaluation in nursing* (pp. 311-332). New York: Springer Publishing.

Hill, H., Ball, D. L., and Schilling, S. (2008). Unpacking "pedagogical content knowledge": Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education, 39*(4), 372-400.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, L. S., and Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction, 26*(4), 430-511.

Hill, H. C., Rowan, B., and Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42,* 371-406.

Hirschman, A. O. (1991). *The rhetoric of reaction.* Cambridge, MA: Belknap Press.

Huang, S., Yi, Y., and Haycock, K. (2002). *Interpret with caution: First state Title II reports on the quality of teacher preparation.* Retrieved from the Education Trust website: http://www.edtrust.org/sites/edtrust.org/files/publications/files/titleII.pdf.

Ingersoll, R. M., and Smith, T. M. (2003). The wrong solution to the teacher shortage. *Educational Leadership*, *60*(8), 30-33.

Institute for Higher Education Policy. (2007). *College and university ranking systems: Global perspectives and American challenges.* Retrieved from http://www.ihep.org/assets/files/publications/a-f/CollegeRankingSystems.pdf.

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

Jaschik, S. (2012, September 14). Renewed debate on SAT and wealth. *Inside Higher Ed.* Retrieved from http://www.insidehighered.com/news/2012/09/14/new-research-finds-sat-equally-predictive-those-high-and-low-socioeconomic-status.

Johnson, J., and Pintz, C. (2013). *Protecting the public: Ensuring nursing education quality.* Paper commissioned for the National Academy of Education Steering Committee on the Evaluation of Teacher Education Programs: Toward a Framework for Innovation.

Johnson, S. M., Berg, J. H., and Donaldson, M. L. (2005). *Who stays in teaching and why: A review of the literature on teacher retention* [Harvard Graduate School of Education report]. http://assets.aarp.org/www.aarp.org_/articles/NRTA/Harvard_report.pdf.

Kaestle, C. (2013). *Testing policy in the U.S.: A historical perspective*. Paper prepared for the Gordon Commission. Retrieved from www.gordoncommission.org/pdf/kaestle_testing_policy_us_historical_perspective.pdf.

Kirby, S. N., McCombs, J. S., Barney, H., and Naftel, S. (2006). *Reforming teacher education: Something old, something new.* Retrieved from the RAND website: http://www.rand.org/content/dam/rand/pubs/monographs/2006/RAND_MG506.pdf.

Koedel, C., Parsons, E., Podgursky, M., and Ehlert, M. (2012). *Teacher preparation programs and teacher quality: Are there real differences across programs?* (CALDER Working Paper No. 79). Retrieved from http://www.caldercenter.org/publications/upload/WP-79.pdf.

Koretz, D. (2009). *Measuring up: What educational testing really tells us.* Cambridge, MA: Harvard University Press.

Koretz, D., Stecher, B., Klein, S. and McCaffrey, D. (1994). The Vermont portfolio assessment program. *Education Measurement: Issues and Practice, 13*(3), 5-16.

Kukla-Acevedo, S., Streams, M., and Toma, E. F. (2009). *Evaluation of teacher preparation programs: A reality show in Kentucky* [Working paper 2009-09, University of Kentucky Institute for Federalism and Intergovernmental Relations]. Retrieved from http://www.ifigr.org/publication/ifir_working_papers/IFIR-WP-2009-09.pdf.

Lawlor, S. (1990). *Teachers mistaught: Training in theories or education in subjects?* London: Centre for Policy Studies.

Levine, A. (2006). *Educating school teachers.* Retrieved from http://www.edschools.org/pdf/Educating_Teachers_Report.pdf.

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice, 16*(2), 14-16.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher 29*(2), 4-16.

Linn, R. L. (2013). Test-based accountability. Paper prepared for the Gordon Commission. Retrieved May 27, 2013 at http://www.gordoncommission.org/rsc/pdf/linn_test_based_accountability.pdf.

Loveless, T. (2011). *The 2010 Brown Center report on American education*. Washington, DC: Brookings Institution. Retrieved from http://www.brookings.edu/about/centers/brown/brown-center-reports .

Malcom-Piqueux, L., and Malcom, S. M. (2013). Engineering diversity: Fixing the education system to promote equity. *The Bridge, 94*(1), 24-34.

Mason, P. (2010). *Assessing difference: Examining Florida's initial teacher preparation programs and exploring alternative specifications of value-added models* (MPRA Paper #27903). Retrieved from http://mpra.ub.unimuenchen.de/27903/1/MPRA_paper_27903.pdf.

McGaghie, W. C., and Thompson, J. (2001). America's best medical schools: A critique of *U.S. News and World Report* rankings. *Academic Medicine*, (*76*)10. Retrieved from http://journals.lww.com/academicmedicine/Fulltext/2001/10000/America_s_Best_Medical_Schools__A_Critique_of_the.5.aspx.

McKnight, C. C., Crosswhite, F. J., Dossey, J. A., Kifer, E., Swafford, J. O., Travers, K. J., and Cooney, T. J. (1987). *The underachieving curriculum: Assessing U.S. school mathematics from an international perspective.* Retrieved from http://www.metproject.org/faq.php.

Mehrens, W. A. (1990). *Assessing the quality of teacher assessment tests. Assessment of teaching: Purposes, practices, and implications for the profession* (Paper 6). Retrieved from http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1005&context=burosassess teaching.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5-11.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8.

Meyer, M., Pyatigorsky, M., Rice, A., and Winter, J. (2013) *Student growth and value-added information as evidence of educator preparation program effectiveness: A review.* Wisconsin Value-added Research Center.

Michelli, N. M., and Earley, P. M. (2011). Teacher education policy in context. In P. M. Earley, D. G. Imig, and N. M. Michelli (Eds.), *Teacher education policy in the United States: Issues and tensions in an era of evolving expectations.* New York: Routledge.

Mihaly, K., McCaffrey, D., Sass, T. R., and Lockwood, J. R. (2012). *Where you come from or where you go? Distinguishing between school quality and the effectiveness of teacher preparation program graduates* (CALDER Working Paper No. 63). Retrieved from http://www.caldercenter.org/upload/Mihaly_TeacherPrep.pdf.

Monk, D. H., and King, J. K. (1994). Multilevel teacher resource effects on pupil performance in secondary mathematics and science: The case of teacher subject-matter preparation. In R. Ehrenberg (Ed.) *Choices and consequences: Contemporary policy issues in education*, pp. 29-58. Ithaca, NY: ILR Press.

Morse, R. (2011). FAQs on new teacher preparation program rankings. *U.S. News and World Report.* Retrieved from http://www.usnews.com/education/blogs/college-rankings-blog/2011/02/10/faqs-on-new-teacher-preparation-program-rankings.

Mullan, F., Chen, C., Patterson, S., Kolsky, G., and Spagnola, M. (2010). The social mission of education: Ranking the schools. *Annals of Internal Medicine*, *152*(12), 804-811. Retrieved from http://annals.org/article.aspx?volume=152&issue=12&page=804.

Murnane, R. J., and Nelson, R. R. (1984). Production and innovation when techniques are tacit: The case of education. *Journal of Economic Behavior and Organization*, *5*(3-4), 353-373.

National Academy of Education. (2005). *A good teacher in every classroom: Preparing the highly qualified teachers our children deserve*. Committee on Teacher Education. L. Darling-Hammond and J. Baratz-Snowden (Eds.). San Francisco, CA: Jossey-Bass.

National Academy of Sciences, National Academy of Engineering, and Institute of Medicine. (2007). *Rising above the gathering storm: Energizing and employing America for a brighter economic future*. Committee on Prospering in the Global Economy of the 21st Century: An Agenda for American Science and Technology. Washington, DC: The National Academies Press.

National Center for Justice Planning. (n.d.). Process and outcome measures [Web page]. Retrieved from http://www.ncjp.org/process-outcome-measures.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: U.S. Department of Education.

National Council for Accreditation of Teacher Education. (n.d.). Council for the Accreditation of Educator Preparation [Web page]. Retrieved May 27, 2013 http://www.ncate.org/Default.aspx.

National Council on Teacher Quality. (2013). *Teacher prep review 2013 report.* http://www.nctq.org/dmsStage/Teacher_Prep_Review_2013_Report.

National Council of State Boards of Nursing. (2012). *Nurse licensee volume and NCLEX® Examination* [NCSBN Research Brief 52]. Chicago, IL: Author.

National Research Council. (1982). *Ability testing: Uses, consequences, and controversies, parts I and II.* A. K. Wigdor and W. R. Garner (Eds.). Committee on Ability Testing. Washington, DC: National Academy Press.

National Research Council. (1989). *Fairness in employment testing*. J. A. Hartigan and A. K. Wigdor. Committee on the General Aptitude Test Battery. Washington, DC: National Academy Press.

National Research Council. (1999). *High stakes: Testing for tracking, promotion, and graduation*. J. P. Heubert and R. M. Hauser (Eds.). Committee on Appropriate Test Use. Washington, DC: National Academy Press.

National Research Council. (2000). *Educating teachers of science, mathematics and technology: New practices for a new millennium.* Committee on Science and Mathematics Teacher Preparation. Washington, DC: National Academy Press.

National Research Council. (2001a). *Testing teacher candidates: The role of licensure tests in improving teacher quality.* K. J. Mitchell, D. Z. Robinson, B. S. Plake, and K. T. Knowles (Eds.). Committee on Assessment and Teacher Quality. Washington, DC: National Academy Press.

National Research Council. (2001b). *Knowing what students know: The science and design of educational assessment*. J. Pellegrino, N. Chudowsky, and R. Glaser (Eds.). Committee on the Foundations of Assessment. Washington, DC: National Academy Press.

National Research Council. (2002). *Scientific research in education*. R. J. Shavelson and L. Towne (Eds.). Committee on Scientific Principles for Education Research. Washington, DC: National Academy Press.

National Research Council. (2008). *Assessing accomplished teaching: Advanced-level certification programs.* M. Hakel, J. Koenig, and S. Elliott (Eds.). Committee on Evaluation of Teacher Certification by the National Board for Professional Teaching Standards. Washington, DC: The National Academies Press.

National Research Council. (2010). *Preparing teachers: Building evidence for sound policy*. Committee on the Study of Teacher Preparation Programs in the United States. Washington, DC: The National Academies Press.

National Research Council. (2011a). *Incentives in test-based accountability and education*. M. Hout and S. Elliott (Eds.). Committee on Incentives and Test-Based Accountability in Public Education. Washington, DC: The National Academies Press.

National Research Council. (2011b). *Successful K-12 STEM education: Identifying effective approaches in Science, Technology, Engineering, and Mathematics*. Committee on Highly Successful Schools or Programs for K-12 STEM Education. Washington, DC: The National Academies Press.

National Research Council. (2013). *Monitoring progress toward successful K-12 STEM education: A nation advancing*? Committee on the Evaluation Framework for Successful K-12 STEM Education. Washington, DC: The National Academies Press.

National Research Council and National Academy of Education. (2010). *Getting value out of value-added*. H. Braun, N. Chudowsky, and J. Koenig (Eds.). Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability. Washington, DC: The National Academies Press.

New York City Department of Education. (2013). Teacher preparation program reports. Retrieved from http://schools.nyc.gov/NR/rdonlyres/D9840D7D-7A36-4C66-817C-C48CFE5C017C/0/NYCDOETeacherPreparationProgramPresentation_August_2013.pdf.

Next Generation Science Standards. (n.d.). About the standards development process [Web page]. Retrieved from http://www.nextgenscience.org/about-standards-development-process.

Noell, G. H., and Gleason, B. (2011). *The status of the development of the value added assessment model as specified in Act 54.* Retrieved from Louisiana Board of Regents website: http://www.regents.doa.louisiana.gov/assets/docs/TeacherPreparation/LegilsativeValueAddedReportFeb2011FINAL.pdf.

Nolte, E., Fry, C. V., Winpenny, E., and Brereton, L. (2011). *Use of outcome metrics to measure quality in education and training of healthcare professionals.* Cambridge, UK: Rand, Europe.

OECD. (2010). *PISA 2009 results: Overcoming social background—Equity in learning opportunities and outcomes (Vol. II).* Retrieved from http://dx.doi.org/10.1787/9789264091504-en.

Office of Technology Assessment. (1992). *Testing in American schools: Asking the right questions*. Washington, DC: Government Printing Office.

Pecheone, R. L., and Chung, R. R. (2006). Evidence in teacher education: The performance assessment for California teachers (PACT). *Journal of Teacher Education, 57*(1), 22-36.

Peck, C. A., Gallucci, C., and Sloan, T. (2010). Negotiating implementation of high-stakes performance assessment policies in teacher education: From compliance to inquiry. *Journal of Teacher Education, 61*(5), 451-463.

Peck, C. A., and McDonald, M. (2013). Creating "cultures of evidence" in teacher education: Context, policy, and practice in three high-data-use programs. *The New Educator, 9*(1), 12-28.

Pianta, R. C., and Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38,* 109-119.

Plecki, M. L., Elfers, A. M., and Nakamura, Y. (2012). Using evidence for teacher education program improvement and accountability: An illustrative case of the role of value-added measures. *Journal of Teacher Education*, *63*(5), 318-334.

Polanyi, M. (1966). *The tacit dimension*. London: Routledge.

Popham, W. J. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practice, 16*(2).

Popham, W. J., and Greenberg, S. (1958). Teacher education: A decade of criticism. *Phi Delta Kappan, 40*(3), 118-120.

Prendergast, C. (1999). The provision of incentives in firms. *Journal of Economic Literature, 37,* 7-63.

President's Council of Advisors on Science and Technology. (2010). *Prepare and inspire: K-12 science, technology, engineering, and math (STEM) education for America's future.* Washington, DC: Author.

President's Council of Advisors on Science and Technology. (2012). *Engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics*. Washington, DC: Author.

Rentner, D., and Usher, A. (2012). *What impact did education stimulus funds have on states and school districts?* Retrieved from Center on Education Policy website: http://www.cep-dc.org/displayDocument.cfm?DocumentID=407.

Reusser, J., Butler, L., Symonds, M., Vetter, R., and Wall, T. J. (2007). An assessment system for teacher education program quality improvement. *International Journal of Educational Management*, *21,* 105-113.

Ronfeldt, M. (2010). Where should student teachers learn to teach? Effects of field placement school characteristics on teacher retention and effectiveness. *Educational Evaluation and Policy Analysis, 34*(1), 3-26.

Rothstein, R. (2004). *Class and schools: Using social, economic, and educational reform to close the achievement gap*. Washington, DC: Economic Policy Institute.

Rothstein, R. (2008). Holding accountability to account: How scholarship and experience in other fields inform exploration of performance incentives in education (Working Paper 2008-04). National Center on Performance Incentives.

Russo, A., and Subotnik, R. (2005). The teacher education report card: Title II of HEA. In S. Cimburek (Ed.), *Leading a profession: Defining moments in the AACTE Agenda, 1980 to 2005.* Washington, DC: American Association for Colleges of Teacher Education. Retrieved from http://www.apa.org/ed/schools/cpse/publications/report-card.pdf.

Salzman, H. (2013, summer). What shortages? The real evidence about the STEM workforce. *Issues in Science and Technology*. Retrieved from http://www.issues.org/29.4/hal.html.

Sawchuk, S. (2011, March 8). Administration pushes teacher-prep accountability. *Education Week.* Retrieved from http://www.edweek.org/ew/articles/2011/03/09/23hea_ep.h30.html.

Sawchuk, S. (2013a, February 15). Overhaul of teacher prep standards targets recruitment, performance. *Education Week.* Retrieved from http://blogs.edweek.org/edweek/teacher beat/2013/02/teacher_prep_accreditation_ove.html.

Sawchuk, S. (2013b, May 14). States' teacher-exam bar set low, federal data show. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2013/05/15/31tests.h32.html.

Sawchuk, S. (2013c, June 18). Disputed review finds disparities in teacher prep. *Education Week.* Retrieved from http://www.edweek.org/ew/articles/2013/06/18/36nctq.h32.html.

Sawchuk, S. (2013d, July 9). Tougher requirements ahead for teacher prep. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2013/07/10/36caep.h32.html?qs=caep+commission.

Schmidt, W. H., Burroughs, N. A., and Cogan, L. S. (2013). On the road to reform: K–12 science education in the United States. *The Bridge, 43*(1), 7-14.

Schmidt, W. H., McKnight, C., Cogan, L., Jakwerth, P., and Houang, R. T. (1999). *Facing the consequences: Using TIMSS for a closer look at United States mathematics and science education.* Boston: Kluwer Academic Publishers.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, and M. Scriven (Eds.), *Perspectives of curriculum evaluation*, 39-83. Chicago, IL: Rand McNally.

Scriven, M. (1983). Evaluation ideologies. *Evaluation in Education and Human Services*, *6*, 229-260.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, *19*, 405-450.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, *16(*2).

Shepard, L. A. (2003). The hazards of high-stakes testing. *Issues in Science and Technology*. Retrieved from http://www.issues.org/19.2/shepard.htm#.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4-14.

Smithers, A., and Robinson, P. (2011). *The good teacher training guide*. Buckingham: University of Buckingham.

Spencer, A. B. (2012). The law school critique in historical perspective. *Washington and Lee Review, 69,* 1949-2066.

Sperber, D. (2010). The guru effect. *Review of Philosophy and Psychology, 1*(4), 583-592.

Springer. M. G. (2009). *Performance incentives: Their growing impact on American K-12 education.* Washington, DC: Brookings Institution.

Stake, J. E. (2006). The interplay between law school rankings, reputation, and resource allocation: Ways rankings mislead. *Indiana Law Journal*, *81*, 229-269. Retrieved from http://www.law.indiana.edu/ilj/volumes/v81/no1/11_Stake.pdf.

Stedman, L., and Smith, M. (1983). Recent reform proposals for American education. *Contemporary Education Review, 2*(2), 85-104.

Strauss, V. (2013, June 18). Why the NCTQ teacher prep ratings are nonsense. *Washington Post.* Retrieved from http://www.washingtonpost.com/blogs/answer-sheet/wp/2013/06/18/why-the-nctq-teacher-prep-ratings-are-nonsense/.

Tamir, E., and Wilson, S. (2005). Who should guard the gates: Evidentiary and professional warrants for claiming jurisdiction. *Journal of Teacher Education 56*(4), 332-342.

Tatto, M. T., Krajcik, J., and Pippin, J. (2013). *Variations in teacher preparation evaluation systems: International perspectives.* Paper commissioned for the National Academy of Education Steering Committee on the Evaluation of Teacher Education Programs: Toward a Framework for Innovation.

Tatz, P. (2013*). Teacher prep review: A review of the nation's teacher preparation programs.* Retrieved from http://www.edexcellence.net/commentary/education-gadfly-weekly/2013/june-20/teacher-prep-review-a-review-of-the-nations-teacher-preparation-programs-2013.html.

Teacher Education Accreditation Council. (n.d.). TEAC accreditation [Web page]. Retrieved from http://www.teac.org/accreditation/.

Tyack, D. (1974). *The one best system: A history of American urban education.* Cambridge, MA: Harvard University Press.

U.S. Department of Education (2011a, December 23). Department of Education awards $200 million to seven states to advance K-12 reform [Press release]. Retrieved from http://www.ed.gov/news/press-releases/department-education-awards-200-million-seven-states-advance-k-12-reform.

U.S. Department of Education. (2011b). *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. (2012, December 11). Education Department announces 16 winners of Race to the Top—District competition [Press release]. Retrieved from http://www.ed.gov/news/press-releases/education-department-announces-16-winners-race-top-district-competition.

U.S. Department of Education. (n.d.). Accreditation in the United States [Web page]. Retrieved from http://www2.ed.gov/admins/finaid/accred/accreditation_pg2.html.

Vergari, S., and Hess, F. M. (2002, fall). The accreditation game. *Education Next*. Retrieved from http://educationnext.org/files/ednext20023_48.pdf.

Vinovskis, M. (1999). *The road to Charlottesville: The 1989 Education Summit*. Washington, DC: The National Education Goals Panel.

Vinovskis, M. (2009). *From a Nation at Risk to No Child Left Behind: National Education Goals and the creation of federal education policy*. New York: Teachers College Press.

Walsh, K., and Jacobs, S. (2007). *Alternative certification isn't alternative.* Thomas B. Fordham Institute and National Council on Teacher Quality. Washington, DC: Thomas B. Fordham Institute.

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., and Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience, 20*(3), 470-477.

Wilson, S. (2011). *Effective STEM teacher preparation, induction, and professional development*. Paper presented at the National Research Council Workshop on Successful STEM Education in K-12 Schools.

Wilson, S. (2013). Recent developments in STEM education relevant to the qualities of teacher preparation programs. Paper commissioned for the National Academy of Education Steering Committee on the Evaluation of Teacher Education Programs: Toward a Framework for Innovation.

Wilson, S. M., Floden, R. F., and Ferrini-Mundy, J. (2001). *Teacher preparation research: Current knowledge, recommendations, and priorities for the future.* Center for the Study of Teaching Policy, University of Washington, Seattle, WA.

Worthen, B. R., Sanders, J. R., and Fitzpatrick, J. L. (1997). *Program evaluation: Alternative approaches and practical guidelines.* New York: Addison Wesley Longman.

Xie, Y., and Killewald, A. (2012). *Is American science in decline?* Cambridge, MA: Harvard University Press.

Zehr, M. (2009, March 15). The Ed Department's Mike Smith talks about 'common standards.' *Education Week.* Retrieved from http://blogs.edweek.org/edweek/curriculum/2009/03/the_ed_departments_mike_smith.html.

Zell, D. (2001). The market-driven business school: Has the pendulum swung too far? *Journal of Management Inquiry, 10*(4), 324-393. Retrieved from http://www.csun.edu/~dmz51283//pdfs/MarketDriven.pdf.

A

Workshop Agenda and Participants

FIRST WORKSHOP
JUNE 25, 2012

AGENDA

9:30 am – 10:00 am	**Breakfast**
10:00 am – 10:15 am	**Welcome** *Michael Feuer*, George Washington University Steering Committee Chair
10:15 am – 11:45 am	***Current Mechanisms for Evaluation*** **10-minute presentations / 5-minute Q&A** *Presenters:* National accreditation *James Cibulka*, NCATE Federal approaches *Chad Aldeman*, Department of Education Internal program reviews/State level approaches *Charles Peck*, University of Washington

Value-added models
George Noell, Louisiana State University

Internal program review
Heather Harding, Teach For America

Ratings & rankings
Robert Rickenbrode, National Council on Teacher Quality

Guiding questions for presentations:

1. What is the underlying "theory of action" for your institutional evaluation mechanism?
2. What is the nature of evidence used and produced by these methods?
3. What is known about the strengths and possible areas of improvement of these methods in implementation?
4. Who are the intended audiences, and what are the goals and purposes of your institutional mechanism?

11:45 am – 12:15 pm **Working lunch**

12:15 pm – 2:15 pm ***Analytic Comments on Program Evaluation Approaches***

Suzanne Wilson, Michigan State University
William Trent, University of Illinois, Urbana-Champaign
Richard Shavelson, Stanford University
Nancy Cartwright, London School of Economics
Brenda Turnbull, Policy Studies Associates, Inc.
Carl Cohn, Claremont Graduate University
James Kelly, University of Michigan
Lee Shulman, Stanford University

Guiding questions for analytic comments:

1. In what ways do the various mechanisms contribute to the general improvement of teacher education?

2. Which aspects of the various mechanisms should be incorporated into the design of future approaches? What new methods or features should be added?
3. What are the most important criteria in evaluating teacher education programs for the various evaluation purposes?

2:15 pm – 2:30 pm **Break**

2:30 pm – 3:15 pm **Open discussion and next steps: Essential insights, topics of commissioned papers, possible authors**

3:15 pm **Workshop adjourns**

3:30 pm – 4:30 pm {Steering committee closed meeting}

PARTICIPANTS

Judie Ahn, National Academy of Education
Chad Aldeman, U.S. Department of Education
Deborah L. Ball, University of Michigan
Jeanne Burns, Louisiana Board of Regents
Nancy Cartwright, London School of Economics
James Cibulka, National Council for Accreditation of Teacher Education
Carl Cohn, Claremont Graduate University
Joe Conaty, U.S. Department of Education
Janice Earle, National Science Foundation
Emerson Elliott, National Council for Accreditation of Teacher Education
Michael Feuer, The George Washington University
Robert Floden, Michigan State University
Susan Fuhrman, Teachers College, Columbia University
Heather Harding, Teach For America
Lionel Howard, The George Washington University
James Kelly, Kelly Advisors
George Noell, Louisiana State University
Charles (Cap) Peck, University of Washington
Robert Rickenbrode, National Council on Teacher Quality
Brian Rowan, University of Michigan
Richard Shavelson, Stanford University (by phone)
Lee Shulman, Stanford University

Patricia Tate, The George Washington University
William Trent, University of Illinois, Urbana-Champaign
Brenda Turnbull, Policy Studies Associates
Kate Walsh, National Council on Teacher Quality
Gregory White, National Academy of Education
Suzanne Wilson, Michigan State University

SECOND WORKSHOP
FEBRUARY 25, 2013

AGENDA

8:30 am – 9:15 am	**Breakfast**
9:15 am – 9:30 am	**Welcome and project overview** *Michael Feuer*, George Washington University Steering Committee Chair
9:30 am – 11:45 am	***Presentations and Q&A*** *(30 minutes each)* Evaluating STEM Teacher Preparation: The Implications of New Curricular, Assessment, and Teacher Quality Initiatives *Suzanne Wilson*, Michigan State University Protecting the Public: Ensuring Nurse Education Quality *Jean Johnson* and *Christine Pintz*, George Washington University [15 minute break] Variations in Teacher Preparation Evaluation Systems: International Perspectives *Maria Teresa Tatto*, Michigan State University Inspecting Teacher Education in England–the Work of Ofsted *John Furlong*, University of Oxford
11:45 am – 12:30 pm	**Lunch**

12:30 pm – 1:15 pm	*Presentation and Q&A (con't)* *Naomi Chudowsky*, Education Research Consultant • Evaluation of Teacher Education: What We've Learned from K-12 Test-based Accountability • Federal Approaches to Evaluating Teacher Education Programs
1:15 pm – 2:15 pm	**Open group discussion: essential insights and next steps**
2:15 pm	**Workshop adjourns**
2:30 pm – 5:00 pm	{Steering committee closed meeting}

PARTICIPANTS

Judie Ahn, National Academy of Education
Deborah L. Ball, University of Michigan
Jeanne Burns, Louisiana Board of Regents
Naomi Chudowsky, Caldera Research
Michael Feuer, The George Washington University
Robert Floden, Michigan State University
Susan Fuhrman, Teachers College, Columbia University
John Furlong, University of Oxford
Lionel Howard, The George Washington University
Jean Johnson, The George Washington University
Christine Pintz, The George Washington University
Maria Teresa Tatto, Michigan State University
Gregory White, National Academy of Education
Suzanne Wilson, Michigan State University

B

Biographical Sketches of Steering Committee Members

Michael J. Feuer (*Chair*) is dean and professor of education at The George Washington University Graduate School of Education and Human Development. Previously he served as the executive director of the Division of Behavioral and Social Sciences and Education at the National Academy of Sciences, where he had also been the first director of the Board on Testing and Assessment and the Center for Education. He received a B.A. from Queens College (CUNY), and an M.A. and a Ph.D. from the University of Pennsylvania. He has published in education, economics, philosophy, and public policy journals. Most recently he was the guest editor of "The Bridge," the flagship journal of the National Academy of Engineering. He is President-elect of the National Academy of Education, fellow of the American Association for the Advancement of Science, and fellow of the American Educational Research Association. He is chair-elect of the AERA Organization of Institutional Affiliates executive committee and member of the AERA government relations committee.

Deborah Loewenberg Ball is the William H. Payne collegiate professor in education at the University of Michigan, and an Arthur F. Thurnau professor. She currently serves as dean of the School of Education and as director of TeachingWorks. She taught elementary school for more than 15 years, and continues to teach mathematics to elementary students every summer. Her research focuses on the practice of mathematics instruction, and on the improvement of teacher training and development. She is an expert on teacher education, with a particular interest in how professional training and experience combine to equip beginning teachers with the skills and knowledge needed for responsible practice. She has authored or co-authored more than 150 publications and has lectured and made numerous major presentations around the world. Her research has been recognized with several awards and honors, and she has served on several national and international commissions and panels focused on policy initiatives and the improvement of education, including the National Mathematics Advisory Panel, the National Science Board, and the Michigan Council for Educator Effectiveness. She is a fellow of the American Mathematics Society and of the American Educational Research Association, and an elected member of the National Academy of Education.

Jeanne M. Burns is the associate commissioner for teacher and leadership initiatives for the Louisiana Board of Regents. She previously taught and served in district administrative roles in Florida and Louisiana. After receiving an M.Ed. and a Ph.D. from Louisiana State University and A&M College, she taught at Stetson University and Southeastern Louisiana University in the areas of leadership for change, gifted education, psychometrics, and reading assessment. She is currently on loan to the State to work full-time for the Louisiana Board of Regents. She has published in professional journals and presented papers at over 150 international, national, regional, and state conferences. During the last twenty years, she has helped the State obtain external grant funds to support the development of a state plan for K-12 education, create a K-12 technology initiative, develop the K-12 school accountability system, coordinate efforts to redesign all public and private teacher education and educational leadership programs within the state, implement a new teacher preparation accountability system, support the implementation of a value-added teacher preparation assessment model, and support campuses as they have provided input into the development of the PARCC assessments and integrated the Common Core State Standards into the teacher preparation curriculum.

Robert Floden is university distinguished professor of teacher education, measurement & quantitative methods, educational psychology, educational policy, and mathematics education at Michigan State University. He received an A.B. with honors in philosophy from Princeton University and an M.S. in statistics and Ph.D. in philosophy of education from Stanford University. He has studied teacher education and other influences on teaching and learning, including work on the cultures of teaching, on teacher development, on the character and effects of teacher education, and on how policy is linked to classroom practice. He is currently working on the development of tools for studying classroom processes that help students develop robust mathematical understanding for use in solving algebra word problems. He has been president of the Philosophy of Education Society, a member of the NRC Committee on Education Research, an Alexander von Humboldt fellow at the University of Tuebingen, and Fulbright Specialist at Pontificia Universidad Católica, Santiago, Chile. He received the Margaret B. Lindsey Award for Distinguished Research in Teacher Education from the American Association of Colleges for Teacher Education. His work has been published in the *Handbook of Research on Teaching*, the *Handbook of Research on Teacher Education*, the *Handbook of Research on Mathematics Teaching and Learning*, and in many journals and books.

Susan H. Fuhrman (*ex-officio*) is the president of Teachers College, Columbia University, founding director and chair of the Management Committee of the Consortium for Policy Research in Education (CPRE), and president of the National Academy of Education. Her substantial leadership record includes her term as dean of the University of Pennsylvania's Graduate School of Education from 1995-2006, where she was also the school's George and Diane Weiss professor of education. She is a former vice president of the American Educational Research Association as well as a former trustee board member of the Carnegie Foundation for the Advancement of Teaching and a former non-executive director of Pearson plc, the international education and publishing company. She received bachelor's and master's degrees in history from Northwestern University and a Ph.D. in political science and education from Teachers College and Columbia University. Her research interests include accountability in education, intergovernmental relationships, and standards-based reform, and she has written widely on education policy and finance.

Lionel C. Howard is an assistant professor of educational research at The George Washington University. His research interests include, broadly, gender identity development and socialization, motivation and academic achievement, and quantitative and qualitative research methodology. He has worked on several local and national research projects focused on improving the educational trajectory and schooling experiences of African American and Latino students. He has also served as a consultant on education policy and evaluation studies. He has published in *Thymus: Journal of Boyhood, Journal of Black Psychology*, *International Journal of Inclusive Education, Journal of Orthopsychiatry,* and *Harvard Educational Review*, and is co-editor of *Facing Racism in Education (3rd Ed)*, published by Harvard University Press. He received a B.A. in applied mathematics and statistics from William Paterson University of New Jersey; an M.A. in measurement, statistics, and evaluation from the University of Maryland, College Park; and an Ed.D. in human development and psychology from Harvard University, Graduate School of Education. He completed an NICHD postdoctoral fellowship at the University of North Carolina at Chapel Hill in the Department of Psychology and the Frank Porter Graham Child Development Institute.

Brian Rowan is the Burke A. Hinsdale collegiate professor in education at the University of Michigan, where he also is a research professor at the Institute for Social Research and a professor of sociology. A sociologist by training, he is a member of the National Academy of Education and a recipient of the William J. Davis Award for outstanding scholarship in the field of education administration. Over the years, he has conducted pioneering studies of schools as organizations as well as important research on school and teaching effectiveness. Currently, he is principal investigator of two efficacy trails examining the effects of educational interventions on teaching and learning in elementary and secondary schools. He also is a contributing researcher on the Measures of Effective Teaching extension project, where he is leading various efforts to collect, archive, disseminate, and analyze video and quantitative data on effective teaching practices. In 2011, he was appointed director of pilot research for the Michigan Council for Educator Effectiveness, in which role he conducted a study of more than 100 Michigan schools as they implemented new teacher evaluation practices in response to changes in Michigan's teacher tenure laws. Born in New Jersey, he received a B.A. from Rutgers University and a Ph.D. at Stanford University. He has been at the University of Michigan since 1991.