



CAEP Evidence Guide

February 2014

FOREWORD

The *CAEP Evidence Guide* is one of several papers linked with the implementation of CAEP accreditation standards adopted in August 2013. This *Evidence Guide* contains information for Educator Preparation Providers (EPPs) on two large topics.

The first, Part A, describes CAEP's perspectives on a "culture of evidence" and the use of data in educator preparation and accreditation. This section describes roles of EPPs, especially with regard to their responsibility for validity of evidence offered to demonstrate that standards are met and for continuous improvement efforts. And it describes a role that CAEP has established for itself as the new sole accreditor for educator preparation. CAEP is committed to strengthen the quality and utility of preparation and accreditation data. Through its own direct actions and through collaborative activities with states and other organizations, CAEP seeks data that are more descriptive, more comparable, and more useful to inform decision making. Part A also describes the role of "annual reporting measures" that are integral to the CAEP standards and accreditation reviews. An additional section will be written in the coming months to describe how accreditation evidence will be accumulated to inform judgments that standards are met, and then how accreditation decisions are reached across all standards.

The second portion of the *Evidence Guide*, Part B, provides protocols and instructions for EPPs on the design for data collections, gathering data, and data analysis. Three situations are addressed in this part of the *Evidence Guide*: principles of "good evidence," especially emphasizing validity; guidelines on "case studies," a frequent source of accreditation data; and suggestions for data on the impact of candidates and completers on P-12 student learning. Two additional sections will be written in the coming months. One of these will suggest guidelines for assessments that EPPs create, administer, and score, and another will examine the information that licensure tests can provide.

CAEP recognizes that the 2013 standards require, in some place, evidence that has not been required or collected in the past. Accordingly, CAEP has established developmental expectations for EPPs with visits during the transition period when the new standards are being phased in (2014 and 2015) and for EPPs with visits in the first two years in which the standards are required for all visits (those with visits in 2016 and 2017).

- ***EPPs with visits in 2014 and 2015*** may present plans in the self study for collecting the required evidence and, once approved by the CAEP Accreditation Council, will present in their annual reports their progress in implementing these plans along the approved timeline.
- ***EPPs with visits in 2016 and 2017*** may also present plans in their self study in lieu of unavailable data and in addition will be expected to provide evidence of implementation in their self study.

EPPs which do not have access to state P-12 student learning data and EPPs that are supplementing state or district data with data on subjects or grades not covered should refer to the *CAEP Evidence Guide*.

In each case, site visitors will investigate the EPP's capacity to carry out and implement the plans with progress to-date.

CAEP Evidence Guide

Table of Contents	
Outline	Page
Part A: PERSPECTIVES ON DATA FOR EDUCATOR PREPARATION	4
1. A culture of evidence	4
2. Role of Educator Preparation Providers (EPPs)	5
a. Responsibility for Quality Assurance Systems and Evidence Used for Accreditation	5
b. Continuous Improvement	6
3. Role of the Council for the Accreditation of Educator Preparation (CAEP)	8
4. Eight Annual Reporting Measures	10
a. Defining the measures	10
b. The purposes of annual reporting measures	10
c. Use of the measures in CAEP accreditation	11
d. What CAEP means by “consumer information”	12
Part B: GUIDELINES FOR DATA DESIGN, COLLECTION, AND ANALYSIS	12
5. Validity and other Principles of “Good Evidence”	12
6. Case Studies; guidelines for a frequent type of data gathering	18
7. Impact of Candidates and Completers on P-12 Student Learning	23
Appendix I–Applying principles of “good evidence” to typical accreditation measures	28
Appendix II–Developing “Ideal” annual reporting measures	32
Appendix III–NAE table on strengths and weakness of typical preparation measures	35

Part A: Perspectives on Data for Educator Preparation

1. **A CULTURE OF EVIDENCE**—Educator Preparation Providers (EPPs) gather data on all aspects of their preparation programs and use them for continuous improvement. Data are not an end in themselves, but the basis for beginning a conversation.

New CAEP standards adopted by the Board of Directors on August 29, 2013 anticipate that educator preparation accreditation will be characterized as a “culture of evidence.” One regional accreditor defines that term this way:

Culture of evidence – a habit of using evidence in assessment, decision making, planning, resource allocation, and other institutional processes that is embedded in and characteristic of an institution’s actions and practices^{1,2}.

This idea of using evidence, “embedded in and characteristic of an institution’s actions and practices,” is the essence of the Baldrige *Education Criteria for Performance Excellence*, updated each year by the Baldrige Performance Excellence Program at the National Institute of Standards and Technology, U.S. Department of Commerce. These criteria are intended to:

empower your organization—no matter its size or the type of education programs and services offered—to reach your goals, improve student learning and other results, and become more competitive by aligning your plans, processes, decisions, people, actions, and results. Using the Criteria gives you a holistic assessment of where your organization is and where it needs to be.³

The CAEP Commission on Standards and Performance Reporting adapted these culture-of-evidence concepts to educator preparation and accreditation in its June 11, 2013 recommendations to the CAEP Board of Directors. The perspective was focused particularly in the Commission’s Standard 5, *Provider Quality Assurance and Continuous Improvement*, but was a strong influence throughout the comments and advice contained in the report. This *Evidence Guide* paper has been prepared to make these ideas about evidence explicit and accessible for EPPs and the public. Some providers, of course, will interpret these ideas as descriptions of what they are already doing. Others may find them different from their prior understandings about use of data in preparation and about accreditation procedures and functions.

For CAEP, the culture of evidence is summed up by the language of the Commission’s Standard 5⁴:

The provider maintains a quality assurance system comprised of valid data from multiple measures, including evidence of candidates’ and completers’ positive impact on P-12 student learning and development. The provider supports continuous improvement that is sustained and evidence-based and that evaluates the effectiveness of its completers. The provider uses the results of inquiry and data collection to establish priorities, enhance program elements and capacity, and test innovations to improve completers’ impact on P-12 student learning and development.

¹ The Western Association of Schools and Colleges. (2013). *2013 Handbook of Accreditation* (p. 47). Retrieved from <http://www.wascenior.org/resources/handbook-accreditation-2013>

² The Western Association of Schools and Colleges (WASC). (2002). *A Guide to Using Evidence in the Accreditation Process: A Resource to Support Institutions and Evaluation Teams*, A Working Draft. Alameda, CA: WASC. Retrieved from <http://www.wascenior.org/search/site/guide%20to%20using%20evidence>

³ Baldrige Performance Excellence Program. (2011). *2011-2012 Education criteria for performance excellence* (p. 2). Gaithersburg, MD: National Institute of Standards and Technology.

⁴ Council for the Accreditation of Educator Preparation (CAEP). (2013). *CAEP Accreditation Standards* (p. 14). Washington, DC: CAEP. Retrieved from http://caepnet.files.wordpress.com/2013/09/final_board_approved1.pdf (The CAEP Accreditation Standards were adopted by the CAEP Board of Directors on August 29, 2013.)

This excerpt outlines critical aspects of the CAEP perspective on evidence. Evidence is not something that an EPP “does for the accreditor.” It is not a “compliance” mechanism. The data are not an end in themselves or “the answer” for accreditation. Instead, data are the basis to begin a conversation. An accredited EPP will have a functioning *quality assurance system*. That system includes data from *multiple measures* and those measures address *evidence of candidates’ and completers’ positive impact on P-12 student learning and development*. Evidence is marshaled to demonstrate that the *data are valid*. But notice, especially, the data are *used* by the EPP for purposes of *continuous improvement*.

In addition to this description of culture-of-evidence ideas that are embedded in the 2013 CAEP standards, the perspectives in Part A are structured around three large topics:

- EPPs serve in a pivotal position. CAEP looks to them as the responsible parties to demonstrate that evidence offered for accreditation is valid and reliable. But EPPs have a unique role, as well, to combine information about program impacts and outcomes together with detailed knowledge of the recruitment, selection, courses, and experiences that comprise preparation. EPPs maintain quality assurance systems with capacity to compile and analyze data, and make use of the results for continuous improvement of the preparation programs.
- The qualities of educator preparation data fall far short of an “ideal.” CAEP, as the EPP-accreditation organization, must play a prominent role to advance evidence-informed accreditation as one of its professional responsibilities. With strong interest across states, and a heightened awareness among policymakers, CAEP’s first years should be an ideal time to define, reach consensus on, and put strong assessments and statistical measures into place.
- The CAEP Board of Directors has adopted the Commission’s recommendations for gathering and publishing data on eight “annual reporting measures.” These collectively describe important aspects of the impact of preparation completers on the job, and preparation outcomes and information of significance to stakeholders and consumers.

2. **ROLE OF EDUCATOR PREPARATION PROVIDERS—EPPs maintain quality assurance systems to support data that can inform continuous improvement, and they take responsibility for the credibility of evidence they use to demonstrate that CAEP standards are met.**

a. Responsibility for Quality Assurance Systems and Evidence

An EPP’s quality assurance system is an essential foundation for any EPP that focuses on results. The Baldrige criteria, for example, focus on common needs to foster understanding, communication, sharing, alignment, and integration while supporting innovative and diverse approaches. And they focus on results—not procedures, tools, or organizational structure.⁵ The CAEP Standards identify some features of an EPP quality assurance system. It:

- Relies on a variety of measures that are relevant to the EPP mission,
- Defines performance benchmarks for its measures (compared with external references where possible),
- Investigates the quality of evidence and the validity of EPP interpretations of that evidence,
- Seeks the views of all relevant stakeholders, and
- Shares evidence widely with both internal and external audiences.⁶

⁵ Baldrige Performance Excellence Program. (2011). *2011-2012 Education criteria for performance excellence* (p. 3). Gaithersburg, MD: National Institute of Standards and Technology.

⁶ Council for the Accreditation of Educator Preparation (CAEP). (2013). *CAEP Accreditation Standards* (p. 15). Washington, DC: CAEP. Retrieved from http://caepnet.files.wordpress.com/2013/09/final_board_approved1.pdf

The new CAEP Standards place direct responsibility on EPPs for the quality of evidence on which they rely for continuous improvement and for accreditation. Providers demonstrate that the data used in decision making are valid, reliable, and fair (free from bias). In keeping with the Commission's perspective that results matter, providers give equal weight to interpretation of the values or results; through benchmarks, comparisons, and other means, accredited providers describe their status and trends in relation to CAEP standards.

For accreditation self studies, providers present empirical evidence of each measure's psychometric and statistical soundness. They describe their processes for testing the validity, reliability, and fairness of measures and instruments used to determine candidates' progress through the preparation program, at completion of the program, and during the first years of practice.

b. Continuous Improvement

CAEP defines continuous improvement as:

An organizational process through which data are collected on all aspects of a provider's activities; analyzed to determine patterns, trends, and progress; and used to define changes for the purpose of improving the quality of programs, faculty, candidates, policies, procedures, and practices of educator preparation⁷.

The purpose of a robust quality assurance system is to *inform* policies and practices in consultation with partners and stakeholders. *Data are to be used.*

Accredited providers develop new models and evaluate and scale up effective solutions. Research and development in the accreditation framework can deepen the knowledge of existing best practices and provide models of emerging innovations to transform educator preparation.

Even the best programs can improve. The quality assurance systems described in Standard 5 are characterized by clearly articulated and effective processes to define and assure quality outcomes and for using data in a process of continuous improvement. A robust EPP quality assurance system provides a continuing stream of relevant data to evaluate "the effectiveness of its completers," "to establish priorities," to "enhance program elements and capacity," and to "test innovations to improve completers' impact on P-12 student learning and development". The provider assures that appropriate stakeholders, including alumni, employers, practitioners, school and community partners, and others defined by the provider, are involved in program evaluation, improvement, and identification of models of excellence. The Commission writes:

The quality of an EPP is measured by the abilities of its completers to have a positive impact on P-12 student learning and development. Program quality and improvement are determined, in part, by characteristics of candidates that the provider recruits to the field; the knowledge, skills, and professional dispositions that candidates bring to and acquire during the program; the relationships between the provider and the P-12 schools in which candidates receive clinical training; and subsequent evidence of completers' impact on P-12 student learning and development in schools where they ultimately teach. To be accredited, a preparation program must meet standards on each of these dimensions and demonstrate success in its own continuous improvement efforts.⁸

⁷ Ibid.

⁸ Ibid.

In essence, continuous improvement is an ongoing learning process driven by evidence that results in intentional organizational changes with the purpose of ultimately improving performance. This idea has served as the principal focus for recent work of the Carnegie Foundation for the Advancement of Teaching, under the headings of “improvement research” and “networked learning.” The following excerpts from a recent paper by Carnegie President, Anthony Bryk, and his colleagues elaborate on continuous improvement in education settings and can be applied to all EPPs.

Bryk labels the evidence for improvement research as:

“*practical measurement*—that is distinct from those commonly used by schools for accountability or by researchers for theory development...”⁹”

He spells out the implications of practical measurement this way:

- First, improvement efforts require *direct measurement of intermediary targets* (i.e., “mediators”) in order to evaluate key change ideas and inform their continued refinement. For example, is a student’s mindset actually improving in places where a change has been introduced, and for whom and under what set of circumstances?
- Second, practical measurement often presses toward *greater specificity* than occurs with measurement for theory development. Educators need data closely linked to specific work processes and change ideas being introduced in a particular context.
- Third, increased sensitivity can be gained when measures are *framed in a language specific to the populations targeted for improvement* (e.g., adult community college students) and *contextualized around experiences common* to these individuals (e.g., classroom routines they are likely to experience).
- Fourth, and most significant from a practical perspective, they need to be *engineered to embed within the constraints of everyday school practice*. For example, a survey routinely given to students during regular classroom time would need to be brief—for instance, no more than 3 minutes.

Bryk concludes by identifying how practical measures are used in (1) assessing changes—whether a change is actually an improvement; (2) in predictive analytics—which individuals or groups are at higher risk for problematic outcomes; and (3) for priority setting—making choices about where best to focus improvement efforts.¹⁰

The Bryk notion of “practical measurement” is expressed again in section 6, below on case studies. Good data should provide specific guidance for action and improvement. Data that are costly, complex, or that entail too much time or political cost are difficult to justify. Simple measures are often the best, even if they are less technical. And measurements should have properties that make disaggregation possible so that underlying patterns can be uncovered for different populations, or different programs.

⁹ Bryk, A. S., Yeager, D. S., Hausman, H., Muhich, J., Dolle, J. R., Grunow, A., ... & Gomez, L. (2013, June). *Improvement Research Carried Out Through Networked Communities: Accelerating Learning about Practices that Support More Productive Student Mindsets* (pp.6-9). Retrieved from

http://www.carnegiefoundation.org/sites/default/files/improvement_research_NICs_bryk-yeager.pdf (In A White Paper prepared for the White House meeting on “Excellence in Education: The Importance of Academic Mindsets.”)

¹⁰ Ibid.

3. ROLE OF THE COUNCIL FOR THE ACCREDITATION OF EDUCATOR PREPARATION—CAEP has a responsibility to collaborate with states and other stakeholders to make the data available for preparation and accreditation more consistent, discriminating, and useful.

The CAEP Commission on Standards and Performance Reporting included an extended description of the current state of educator preparation data, concluding with an outline of a role that CAEP, itself, should take on:

In an ideal world, EPP accreditation would draw its evidentiary data from a wide array of sources that have different qualitative characteristics from many of those currently available. There would be elements of preparation that are quantified with common definitions or characteristics (e.g., different forms or patterns of clinical experiences) that everyone would understand and that providers would use in their own data systems. There would be comparable experiences in preparation that providers as well as employers, state agencies, and policymakers agree are essential. There would be similar requirements across states for courses, experiences, and licensure. There would be a few universally administered examinations that serve as strong anchors for judgments about effective preparation and that are accepted as gateways to preparation programs, employment, or promotion.

The qualities of educator preparation data fall far short of such an ideal system. However, Commission members are optimistic that advances in the quality of evidence are at hand. From many arguments that might be made in defense of that optimism, three stand out:

1. The current policy interest in well-prepared teachers and leaders is probably higher than it has ever been, especially in states,
2. Several research projects promise to increase knowledge about critical relationships between preparation and effective teaching that can inform EPPs and also next-generation accreditation standards (for example, the U.S. Department of Education’s Institute for Education Sciences is supporting randomized controlled trials to examine elements of preparation, including selection and clinical experiences); and
3. The Gates Foundation’s “Measures of Effective Teaching” project has recently concluded a large research study of instruments used to evaluate teacher performances, some or all of which might be adapted to serve as pre-service measures.¹¹

The CAEP Data Task Force recommended that CAEP exercise a role in continuing efforts that refine and validate all measures in the accreditation process. This would include formal construct validation for each of the measures. The Task Force concurs with the concluding paragraphs of the CAEP Commission report on Standards and Performance Reporting: “CAEP should hold itself to the same standard of evidence-based practice that it calls on providers to meet . . . it should monitor new evidence of implementation of existing assessments, the development of new assessments, and improper uses of assessment tools.”¹²

CAEP should collaborate with the Council of Chief State School Officers (CCSSO) and states on these studies, it should seek guidance from all relevant groups about how to improve measures, and undertake systematic experiment and piloting of measures and processes. CAEP should also develop guidelines and a structure for an “ideal” data system for EPPs, together with advice about how to implement and use such a system. It should work with the states to determine the current conditions of state databases, especially how well they are suited to providing EPPs with data that would be useful for accreditation, and should develop a description for an ideal state data system for preparation and accreditation as well.

¹¹ Council for the Accreditation of Educator Preparation (CAEP). (2013). *CAEP Accreditation Standards* (pp. 26-27). Washington, DC: CAEP. Retrieved from http://caepnet.files.wordpress.com/2013/09/final_board_approved1.pdf

¹² Ibid. (p. 32).

CAEP is making numerous efforts to instantiate the Commission’s vision:

- It is collaborating with Pearson PLC on the preparation of a current status report on education data. This report will update and detail the general conclusions of the 2010 National Research Council (NRC) teacher preparation report with current data, a description of what is available, and outline of where the future opportunities for better data lie.
- It has employed a Director of Research and Data Strategy as a focal point and representative of an expanded role for accreditation in this area.
- It is joining with CCSSO and a group of states with which that organization has affiliated in order to advance goals of the 2012 NRC report on teacher preparation—in program approval, in licensure/certification, and in improved data.
- CAEP will participate in an exploratory “workshop” with the National Research Council to determine whether, and if so, how, the various EPP actions, courses, policies, and experiences that comprise “preparation” can be defined in statistical terms so that a national descriptive database on preparation can be constructed.

The annual report of CAEP, initiated in 2013, will provide continuing updates on these efforts.

Gaining rigor in EPP accreditation is dependent on good data. The Commission supported recommendations from the CAEP Data Task Force about the direction that data improvement efforts should tend. Common assessments that can serve as anchor measures are one such direction. Others include data gathered around standard statistical definitions that would permit comparisons across peers, or over geographic areas; and benchmarks—that is, identification of “best in class” levels of accomplishments. Completer and employer surveys would be constructed so that the opinions they record can be associated with elements of preparation rather than represent simple popularity polls. Assessments and surveys would systematically be validated. Observation evaluations would follow design protocols and be judged by trained third parties. Unintended consequences of measures, and the data burden and human resource challenges of CAEP’s move to more rigorous evidence would be evaluated as well.¹³

4. EIGHT ANNUAL REPORTING MEASURES—what they are, their purpose, their role in accreditation, and consumer measures

One feature of the CAEP Standards is annual reporting and CAEP monitoring. When fully developed, indicators of *program outcome* and *program impact* are intended to be a means for EPPs to report to the public, prospective candidates, policymakers, the media, and CAEP on the pathways of its graduates and the results of preparation. CAEP expects these data will be accessible on the EPP website and serve as a crucial source of information for its own continuous improvement efforts. These data will also build a significant accreditation resource to assist CAEP’s monitoring of EPP performance, establish useful comparisons and benchmarks, serve researchers, and undergird CAEP’s national reporting on accredited EPPs.¹⁴ The subsections, below, describe the annual reporting measures, their purposes, their use by EPPs and in accreditation, and, finally, the role of student loan default and other consumer information reporting.

a. Defining the measures

In brief, the CAEP standards describe the measures in two categories. One category indicates the results of preparation—that is, the performance of completers once they are employed. The measures of “program impact” are:

¹³ Ibid. (p. 37).

¹⁴ Ibid. (p. 17).

- Impact that completers' teaching has on P-12 learning and development,
- Indicators of teaching effectiveness,
- Results of employer surveys, and including retention and employment milestones, and
- Results of completer surveys.

The second category indicates outcomes of programs and consumer information, yardsticks that states and policymakers particularly find useful. The measures of "program outcome and consumer information" include:

- Graduation rates from preparation programs,
- Ability of completers to meet licensing (certification) and any additional state requirements (i.e., licensure rates),
- Ability of completers to be hired in education positions for which they were prepared (i.e., hiring rates), and
- Student loan default rates and other consumer information.

b. The purposes of annual reporting measures

The CAEP Commission on Standards and Performance Reporting described multiple purposes for the annual reporting measures. Some of these are directed at the EPP.

- They are incentives for providers to routinely gather, analyze, and report critical data about their programs;
- The data serve as one means for public accountability and transparency;
- The measures encourage more in-depth evaluation, self-interrogation, and reporting on the full breadth of standards and components—they are a resource for continuous improvement; and
- Employers and prospective applicants for admission need this kind of information in user-friendly, transparent forms.

Others fall to CAEP and outline a new and demanding role for the EPP accreditor:

- The data will become the foundation of a national information base that increases in value over time;
- The data can trigger an alert to CAEP that further examination may be warranted (see below, on how the annual reporting measures are used in CAEP accreditation);
- The data will be a source of information for CAEP's annual report, complement descriptive measures for all accredited providers, facilitate monitoring of trends over time, allow analysis of preparation patterns for different subgroups of providers (e.g., state, regional, urban, rural), and be a resource for identifying benchmark performances; and
- The database will enable CAEP to report on the progress of continuous improvement not just for an individual provider but for educator preparation across all accredited providers.¹⁵

The conclusion of this part of the Commission's recommendations, subsequently adopted by the CAEP Board of Directors, emphasized CAEP's own annual reporting, the developing and differing state and provider data systems that will influence the data available to CAEP, and provider responsibilities for data gathering and reporting:

CAEP should be committed to annual reporting of data on the aforementioned measures, while allowing for a degree of flexibility that recognizes some states and providers may need to develop needed data gathering and reporting capacities. CAEP has a responsibility to work with states and the

¹⁵ Ibid.

Council of Chief State School Officers to assist providers with these efforts, but providers also have a responsibility for maintaining a system of ongoing data collection and reporting.¹⁶

Appendix II displays the initial CAEP expectations for the “ideal” directions a mature version of the annual reporting measures might take. CAEP would work with stakeholders on measures that are consistently defined, gathered in systematic ways on agreed-upon timelines, that make use of appropriate standardized assessments with well-established validity as preparation formative or exit measures, that contain a few assessments (as do other professional fields) that can serve as anchoring points for a portion of accreditation judgments, and that include common descriptive statistical indicators.

c. Use of these measures in CAEP accreditation

The CAEP Board of Directors policy is that a CAEP accreditation decision signals that the provider’s efforts and results substantially comply with the rigorous levels recommended by the Commission. That is, accreditation can be achieved if there are some areas where component evidence fails to reach decision guidelines. But there are two exceptions, one regarding the eight annual reporting measures. The policy is “the provider must meet CAEP’s guidelines for evidence for the annual report measures”¹⁷.

The other exception is that EPPs must meet CAEP evidence guidelines for the standards on continuous improvement.

The Commission’s deliberations included consideration of trends in other accreditation bodies. For example, some accreditors rely more heavily on data than EPP accreditation has. Some have accreditation review procedures that shift the balance a little more in favor of off-site review and a little less to the on-site visit. Some use, or are considering, how “experts” might complement the traditional “peers” for accreditation, especially in areas such as assessments.

Among these trends identified by the Commission was the idea that annual reporting data would be monitored by CAEP. That monitoring could have either positive or negative consequences for the EPP and its accreditation review in the following way:

CAEP would identify both “levels” of performance on the annual reporting measures and “significant amounts of change.” When the EPP’s annual report value exceeds those identified markers, a closer examination will be prompted by the CAEP Accreditation Council’s Annual Report & Monitoring Committee. No action would be taken automatically, but the data could initiate (1) a follow up in future years, (2) steps toward adverse action that could include revocation of accreditation status, or (3) steps leading toward eligibility for a higher level of accreditation.¹⁸

d. What CAEP means by “consumer information”

The final annual measure is labeled “student loan default rates and other consumer information.” The Commission’s intent was to allow prospective candidates to assess the cost and potential benefit of a provider’s programs.¹⁹

Note that these rates are not be considered for accreditation decisions. They are not intended as an indicator of EPP quality. Instead, the information would be furnished to prospective applicants as part of a

¹⁶ Ibid. (pp. 17-18).

¹⁷ Ibid. (p. 18).

¹⁸ Ibid. (p. 16).

¹⁹ Ibid. (p. 17).

suite of information that might include cost of attendance for enrolled candidates, typical employment placement sites for completers, and typical first year salaries for completers.

This type of consumer information is often included in accreditation requirements, and is a part of the public accountability standard of CAEP's own accreditor, the Council for Higher Education Accreditation (CHEA). The Commission suggested that EPPs publish these data along with the other seven annual measures.

Part B: Guidelines for Appropriate Data Design, Collection and Analysis

5. VALIDITY AND OTHER PRINCIPLES OF “GOOD EVIDENCE”–Characteristics of good evidence and useful data for improvement.

This section draws together important attributes of “good evidence” found in three sources. The Western Association of Schools and Colleges (WASC) has created many materials to support college and university accreditation in California, Hawaii, Guam, the Pacific Basin and international sites. One that has influenced these pages is their evidence guide²⁰. A second reference is the National Academy of Education report on *Evaluation of Teacher Preparation Programs*, released in the fall of 2013²¹. And the third reference is a report from CAEP's Data Task Force, an advisory group convened by President Cibulka to advise on CAEP's use of evidence in accreditation²².

a. What is evidence?

The WASC description of “what is evidence” in the box below is equally applicable to CAEP.

Exhibit 1 • WASC: A Guide to Using Evidence in the Accreditation Process: A Resource to Support Institutions and Evaluation Teams²³

NOTE: [Bracketed] edits to substitute “CAEP” for “WASC” are inserted

At the most fundamental level, “evidence” is the substance of what is advanced to support a claim that something is true. This makes evidence different from things like “information,” “data,” or “facts” in at least five subtle but important ways:

- First, evidence is *intentional* and *purposeful*; it is advanced to address deliberately posed questions that are important to both institutions and their stakeholders. One implication is that evidence is always implicitly or explicitly located within a dialogue among those who seek to reach agreed upon conclusions about what is true. What counts as evidence, then, is not a given but rather a particular community of judgment.

²⁰ The Western Association of Schools and Colleges (WASC). (2002). *A Guide to Using Evidence in the Accreditation Process: A Resource to Support Institutions and Evaluation Teams*, A Working Draft. Alameda, CA: WASC. Retrieved from <http://www.wascenior.org/search/site/guide%20to%20using%20evidence>

²¹ 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. Retrieved from http://www.naeducation.org/xpedio/groups/naedsite/documents/webpage/naed_085581.pdf

²² Ewell, P. (2013). *Principles for measures used in the CAEP accreditation process*. Washington, DC: CAEP. Retrieved from <http://caepnet.files.wordpress.com/2012/12/caep-measure-principles.pdf>

²³ The Western Association of Schools and Colleges. (2002). *A Guide to Using Evidence in the Accreditation Process: A Resource to Support Institutions and Evaluation Teams*, A Working Draft, (p. 7). Alameda, CA: WASC. Retrieved from <http://www.wascenior.org/search/site/guide%20to%20using%20evidence>

The [CAEP guide to principles of good evidence, below, contain] a set of basic principles designed to help determine what constitutes good evidence in the context of [education] accreditation. In applying these principles, it is important to remember that the essential setting in which evidence is advanced remains a continuing *dialogue* – an exchange in which information is advanced, critiqued, refined, and enhanced.

- Second, evidence always entails *interpretation* and *reflection*; it does not “speak for itself.” This means that sound evidence involves more than simply presenting a body of data or “listing the facts.” Instead, it implies that the party advancing the evidence has thought about what it means and can interpret it appropriately to support a conclusion. Indeed, for purposes of accreditation, as much emphasis should be placed on what an institution *makes* of the information that it advances – and how it is using the conclusions it has drawn to improve itself – as on the information itself.
- Third, good evidence is *integrated* and *holistic*; it does not consist merely of a list of unrelated “facts.” Individual pieces of data are thus never advanced as evidence on their own. Rather, they take on meaning in the overall context in which they are presented. This means that individual pieces of evidence should mutually reinforce one another, based on the fact that information of quite different kinds, drawn from diverse sources, point in a similar direction. It also implies that judgments need to be made about any body of evidence as a whole – on the “weight” of the evidence, in common parlance.
- Fourth, what counts as evidence can be both *quantitative* and *qualitative*; it is not just confined to numbers. Certainly, where available and appropriate, quantitative data will be powerful and it is expected that much of the information an institution advances in support of its claims for capacity and educational effectiveness will be in numeric form. But it is important for institutions to avoid automatic assumptions that “measurement” is what is wanted. Indeed, narrowly confining the body of evidence submitted to things like disembodied test scores or facilities inventories is precisely the opposite of what [CAEP] seeks from institutions.
- Fifth, good evidence can be either *direct* or *indirect*; it does not always require obtrusive data gathering that uses specially designed instruments. Indeed, as emphasized in the [WASC] *2001 Handbook of Accreditation*, the process should “rely heavily on existing institutional evidence and sampling of institutional exhibits and processes²⁴.” While there may be many occasions on which new data will need to be collected, institutions should be certain that they have creatively tapped the wealth of information on their own performance that is already available.

b. Principles of good evidence

In its *Evaluation of Teacher Preparation Programs* volume, the National Academy of Education gives, as does CAEP, a primary status to *validity*. Validity is:

...the requirement that an evaluation system’s success in conveying defensible conclusions about a [Teacher Preparation Program] 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.²⁵

²⁴ The Western Association of Schools and Colleges (WASC). (2001). *2001 Handbook of Accreditation*, (p. 4). Alameda, CA: WASC. Retrieved from http://vpapf.chance.berkeley.edu/accreditation/pdf/WASC_Handbook.pdf

²⁵ Feuer, M. J., et al. (2013). *Evaluation of teacher preparation programs: Purposes, methods, and policy options* (p. 5). Washington, DC: National Academy of Education. Retrieved from http://www.naeducation.org/xpedio/groups/naedsite/documents/webpage/naed_085581.pdf

The principles developed by Peter Ewell, below, are intended as a guide to EPPs in making their own determination of the adequacy of measures proposed for use in the CAEP accreditation process. Appendix I also illustrates how the principles can be applied to typical preparation or accreditation measures.

1. **Validity and Reliability.** All measures are in some way flawed and contain an error term that may be known or unknown. In general, the greater the error, the less precise—and therefore less useful—the measure. But the level of precision needed depends on the circumstances in which the measure is applied. To be used in accreditation decisions, measures need to be founded upon reliable measurement procedures, but they also need to be designed to operate under less-than-ideal measurement conditions. Even the most rigorous measures, moreover, may not embrace the entire range of validities—construct, concurrent, and predictive. The meaning of “validity” has evolved and has come to embrace the appropriateness of the use to which the measure is put (“consequential validity” as in Messick, 1995²⁶). This means, for example, that studies of value added measures (VAM) that explicitly consider their use as program evaluation indicators, rather than as a component of teacher or school evaluation, are more applicable for preparation program review situations.²⁷

In its data analyses to support continuous improvement and accreditation self studies, accredited EPPs meet accepted research standards for validity and reliability of comparable measures and, among other things, rule out alternative explanations or rival interpretations of reported results. Validity can be supported through evidence of:

- Expert validation of the items in an assessment or rating form (for convergent validity),
- A measure’s ability to predict performance on another measure (for predictive validity),
- Expert validation of performance or of artifacts (expert judgment), and
- Agreement among coders or reviewers of narrative evidence.

And reliability in its various forms can be supported through evidence of:

- Agreement among multiple raters of the same event or artifact (or the same candidate at different points in time),
- Stability or consistency of ratings over time, and
- Evidence of internal consistency of measures.²⁸

²⁶ Messick, S. (1995). *Standards of validity and the validity of standards in performance assessment*. Educational Measurement: Issues and Practice, 14(4), 5-8.

²⁷ Ewell, P. (2013). *Principles for measures used in the CAEP accreditation process*. Washington, DC: CAEP. Retrieved from <http://caepnet.files.wordpress.com/2012/12/caep-measure-principles.pdf> (Appendix I provides examples of the application of these principles to three different measures—licensure passage rates, completer employment rates, and case studies.)

²⁸ Council for the Accreditation of Educator Preparation (CAEP). (2013). *CAEP Accreditation Standards* (pp. 33-34). Washington, DC: CAEP. Retrieved from http://caepnet.files.wordpress.com/2013/09/final_board_approved1.pdf

**Exhibit 2 • Excerpt from National Academy of Education report,
*Evaluation of Teacher Preparation Programs*²⁹**

NOTE: “TPP” = Teacher Preparation Program

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 5³⁰.

2. **Relevance.** The measures advanced ought to be demonstrably related to a question of importance that is being investigated. This principle implies validity, but it goes beyond it by also calling for clear explanation of what any information put forward is supposed to be evidence of and why it was chosen. This principle also implies that there is a clear and explicable link between what a particular measure is established to gauge and the substantive content of the Standard under which it is listed.

²⁹ Feuer, M. J., et al. (2013). *Evaluation of teacher preparation programs: Purposes, methods, and policy options* (p. 14). Washington, DC: National Academy of Education. Retrieved from http://www.naeducation.org/xpedio/groups/naedsite/documents/webpage/naed_085581.pdf

³⁰ The reference is to “Principle 5” in the NAE report, which the report summarizes (p. 6): Evaluation systems may have differential and potentially unfair effects on diverse populations of prospective teachers and communities.

3. Verifiability. The validity of any measure advanced should be subject to independent verification. This is partly a matter of whether the process of creating the current value of the measure is replicable, and if repeating the process would yield a similar result. This principle implies reliability, but goes beyond it to require transparency and full documentation—whether sufficient information is available to enable any third party to independently corroborate what was found.
4. Representativeness. Any measure put forward should be typical of an underlying situation or condition, not an isolated case. If statistics are presented based on a sample, therefore, evidence of the extent to which the sample is representative of the overall population ought to be provided, such as the relative characteristics of the sample and the parent population. If the evidence presented is qualitative—for example, case studies or narratives, multiple instances should be given or additional data shown to indicate how typical the examples chosen really are. CAEP holds that sampling is generally useful and desirable in generating measures efficiently. But in both sampling and reporting, care must be taken to ensure that what is claimed is typical and the evidence of representativeness must be subject to audit by a third party.
5. Cumulativeness. Measures gain credibility as additional sources or methods for generating them are employed. The resulting triangulation helps guard against the inevitable flaws associated with any one approach. The same principle applies to qualitative evidence whose “weight” is enhanced as new cases or testimonies are added and when such additions are drawn from different sources. In sum, the entire set of measures used under a given Standard should be mutually reinforcing.³¹

All aspects of a preparation program from recruitment and admissions, through completion and into on-the-job performance should be informed by multiple measures. These measures will:

- Document and monitor effects of EPP admissions selection criteria,
 - Monitor candidate progress;
 - Monitor completer achievements;
 - Monitor provider operational effectiveness;
 - Demonstrate that the provider satisfies all CAEP standards;
 - Trace status and progress of the EPP on measures of program impact—
 - P-12 student learning and development,
 - Indicators of teaching effectiveness,
 - Results of employer surveys and including retention and employment milestones, and
 - Results of completer surveys;
 - Trace status and progress of the EPP measures of program outcomes—
 - Completer or graduation rates,
 - Ability of completers to meet licensing (certification) and any additional state accreditation requirements,
 - Ability of completers to be hired in education positions for which they are prepared; and
 - Other consumer information, including student loan default rates for completers.³²
6. Fairness. Measures should be free of bias and be able to be justly applied by any potential user or observer. Potential sources of bias might be introduced by the values or beliefs of those applying the measure, such as the conviction that a particular result should be observed. Other sources of

³¹ Ewell, P. (2013). *Principles for measures used in the CAEP accreditation process*. Washington, DC: CAEP. Retrieved from <http://caepnet.files.wordpress.com/2012/12/caep-measure-principles.pdf>

³² Council for the Accreditation of Educator Preparation (CAEP). (2013). *CAEP Accreditation Standards* (pp. 14-15). Washington, DC: CAEP. Retrieved from http://caepnet.files.wordpress.com/2013/09/final_board_approved1.pdf

bias are situational, such as the limited perspective of an untrained observer undertaking a classroom observation or applying a rubric. In this sense, fairness is a special case of reliability: a fair measure will return the same result even if applied by different observers under different circumstances or at different points in time.

7. Stakeholder Interest. A sound set of measures should respect a range of client perspectives including the program, the student, the employer, and the state or jurisdiction. Taken as a whole, a set of measures should potentially support the establishment of an informed dialogue among appropriate parties. A statistic on the employment rates of program completers, for example, can be summarized from the student point of view as the probability of being placed, from the program's point of view as a placement rate, and from an employer's point of view as the proportion of needed job openings filled each year. To reflect stakeholder interests, moreover, proposed measures should be neither arcane nor overly academic.
8. Benchmarks. Without clear standards of comparison, the interpretation of any measure is subject to considerable doubt. Measures can be compared across programs, against peers, against established "best practices," against established goals, against national or state norms, or over time. For every measure under each standard, CAEP should be able to indicate an appropriate benchmark against which a given EPP's performance can be judged.
9. Vulnerability to Manipulation. All measures are to some extent vulnerable to manipulation. This is one reason to insist upon triangulation and mutual reinforcement across the measures used under each standard. For example, program graduation and licensure passage rates depend a great deal on which students are included in the denominator. Because the incentives to perform well on such measures are considerable, programs may identify ways to construct these denominators that yield maximum values on these measures regardless of what they are actually doing.
10. Actionability. Good measures, finally, should provide EPPs with specific guidance for action and improvement. Many promising measures fail simply because they are too expensive, too complex, too time consuming, or too politically costly to implement. Often, the simplest are best, even if they seem less technically attractive. Value-Added Measures, for example, are conceptually compelling, but they demand a good deal of investment to create and interpret. This principle also suggests that any measure should be able to be disaggregated to reveal underlying patterns of strength and weakness or to uncover populations who could be served more effectively. Finally, the measures provided should be reflectively analyzed and interpreted to reveal specific implications for the program.³³

6. CASE STUDIES—guidelines for a frequent type of EPP data gathering

a. Overview

The CAEP Standards, adopted by the CAEP Board of Directors on August 29, 2013, were written by the Commission on Standards and Performance Reporting to integrate "standards" for educator preparation accreditation with the "evidence" to demonstrate that those standards have been met.

The Commission's counsel about evidence appears throughout the final report, and includes an appendix with 79 illustrative examples of evidence across the five standards and annual reporting

³³ Ewell, P. (2013). *Principles for measures used in the CAEP accreditation process*. Washington, DC: CAEP. Retrieved from <http://caepnet.files.wordpress.com/2012/12/caep-measure-principles.pdf>

recommendations³⁴. A quarter of those illustrative examples describe exhibits such as case studies, documentation of particular program features, or demonstrations of the consequences of practice. Among them are examples in which the EPP would develop and evaluate new measures, such as these:

- Assess the effects of a change in admissions that define criteria for "grit," persistence and leadership abilities, as an "innovation"—*for Standard 3 on candidate quality and Standard 5 on continuous improvement/quality assurance*;
- Pilot a new assessment constructed to show developing candidate proficiencies for use of an assessment to enhance learning during clinical experiences—*for demonstration of one InTASC standard in CAEP Standard 1 on content and pedagogical knowledge*; or
- Conduct a case study of completers that demonstrates the impacts of preparation on P-12 student learning and development—*for part of the evidence under Standard 4*.

Other examples of case studies are ones in which the EPP would track the results of a change in preparation practices:

- Ability of candidates to design and use a variety of formative assessments with P-12 students—*for Standard 1 on candidate proficiency in use of assessments to enhance learning, and an indicator of completer capability at exit for Standard 3*;
- Implement a video recording of candidate instruction with review and evaluation based on rubrics as a way to demonstrate particular candidate proficiency and growth—*for CAEP Standard 2 on clinical partnerships and practice*;
- Gather information from P-12 student surveys conducted during candidates' pre-service clinical practice and analyze the data on candidate instructional practices—*for Standard 1 as a performance measure of candidate application of knowledge and pedagogical skills, and for Standard 3 as an indicator of candidate progress during preparation*; or
- Observation measures with trained review procedures and/or faculty observation with rubrics to track progress during candidate preparation and trends across cohorts—*for Standard 1 as an indicator of candidate capacity to use instructional practice and InTASC knowledge, and Standard 3 as an indicator of developing candidate abilities, perhaps conducted multiple times*.

Evidence of this kind is generally most useful in generating hypotheses or ideas, and is less useful or applicable in confirmatory analysis. In assembling such evidence, moreover, the standards that apply to research for peer review and publication cannot be implemented rigidly or in all situations. "The use of evidence in such cases more resembles its use in action research, where the objective is less to establish 'truth' than to discern an appropriate course of action directed at program improvement."³⁵

Many EPPs will have their own internal criteria or operational procedures for gathering data to inform practice as a routine part of internal continuous improvement efforts. The CAEP guidelines in this paper would supplement the practices in those EPPs, and perhaps suggest ways that other EPPs could undertake similar efforts if they have not already done so. The ideas have been assembled from several sources and represent a composite perspective that draws from literature on action research, case study research, and the Carnegie Foundation³⁶ for the Advancement of Teaching improvement research, as well as practices that underlie the Baldrige criteria for high-performing education organizations. A report from the National

³⁴ Council for the Accreditation of Educator Preparation (CAEP). (2013). *CAEP Accreditation Standards* (p. 41-57). Washington, DC: CAEP. Retrieved from http://caepnet.files.wordpress.com/2013/09/final_board_approved1.pdf

³⁵ Ewell, P. (2013). "Action Research" *Measures for Program Improvement* (p. 1). Washington, DC: CAEP. Retrieved from <http://caepnet.files.wordpress.com/2012/12/caep-action-research-measures.pdf>

³⁶ Bryk, A. S., et al. (2013, June). *Improvement Research Carried Out Through Networked Communities: Accelerating Learning about Practices that Support More Productive Student Mindsets*. Retrieved from http://www.carnegiefoundation.org/sites/default/files/improvement_research_NICs_bryk-yeager.pdf

Academy of Education, *Evaluation of Teacher Preparation Programs*³⁷, has served as a source of reference as well. The concepts on which this work is founded have been developed principally in the health care field and Don Berwick is a key spokesperson for their application. He has served as Administrator for the Centers for Medicare and Medicaid Services, and as a leader of the not-for-profit Institute for Healthcare Improvement. He writes:

Effective leaders of improvement insist that the status quo should be challenged continuously through the active testing of promising changes on a small scale. Such testing is totally unfamiliar as part of normal work and most organizations resist the concept. The resistance comes in many disguises, such as the demand for perfect measurement, planning tests so large that they never occur, or extending the time frame ("We'll meet again next month") exactly when it should be shorter ("We'll meet again tomorrow")...

When leaders manage to overcome this fear they often run into a second barrier: the search for perfect measurement. The rooting of health care in scientific research has generated some myopia about the preconditions for inference. When we try to improve a system we do not need perfect inference about a pre-existing hypothesis: we do not need randomization, power calculations, and large samples. We need just enough information to take a next step in learning. Often a small series of patients or a few closely observed events contain more than enough information for a specific process change to be evaluated, refined, or discarded, just as my daughter, in learning to ride her bicycle, sometimes must fall down only once to learn not to try that maneuver again. In measurement for improvement the best is often the enemy of the good.³⁸

This perspective applies equally well to educator preparation.

The case study guidelines are founded on four assumptions:

- **Focus on results**—Data used for improvement efforts and accreditation should ultimately aim to enhance preparation performance outputs related to P-12 student learning;
- **Always improve**—Data for accreditation should be some portion of the data that an EPP uses for its own continuous improvement efforts. A successful EPP builds capacity for improvement rather than for compliance;
- **Rely on data**—Collecting valid and reliable data from multiple sources to inform decision making is an essential component of a continuous improvement system; and
- **Engage stakeholders**—EPPs engage stakeholders as an integral part of the on-going effort to improve programs.

b. Guides for Case Studies

In developing and implementing systems that use evidence for continuous improvement, providers may consider questions posed under the following headings: identify the topic; generate ideas for change; define the measurements; test promising solutions; sustain and scale solutions; and share knowledge. Each of these is presented in the six sections, below.

³⁷ Feuer, M. J., et al. (2013). *Evaluation of teacher preparation programs: Purposes, methods, and policy options* (p. 5). Washington, DC: National Academy of Education. Retrieved from http://www.naeducation.org/xpedio/groups/naedsite/documents/webpage/naed_085581.pdf

³⁸ Berwick, D. M. (1996). A primer on leading the improvement of systems. *BMJ: British Medical Journal*, 312(7031), 619. Retrieved from <http://dx.doi.org/10.1136/bmj.312.7031.619>

i. Identify the topic to study—As Peter Ewell emphasizes, a “worthwhile question is the point of entry into any empirical investigation”³⁹. Questions, and the case study designs developed to investigate them, should reflect a solid understanding of relevant prior theoretical, methodological, and empirical work. Tony Bryk asks, “what specifically is the problem we are trying to solve”⁴⁰. And he observes that engaging key participants early and often at this and later stages is enlivening and important. Questions that EPPs can pose include these:

- Is your improvement work focused on Identifying and solving specific problems of practice that are measurable and whose solutions are reasonably attainable?
- What evidence have you used to identify the problem?
- Does your problem statement (question of inquiry) reflect a solid understanding of relevant prior theoretical, methodological, and empirical work on this topic?

ii. Generate ideas for change—Developing ideas to address the identified problem is not just a matter of brainstorming. Bryk cautions that it is hard to improve what you do not fully understand. He advises: “Go and see how local conditions shape work processes. Make your hypotheses for change public and clear.”

And Don Berwick provides more context for this topic as he considers it from the healthcare perspective: At the heart of a scientifically grounded theory for improving healthcare is the premise that quality is a system property, and that, therefore, what primarily determines the level of performance is the design of a healthcare system, not simply the will, native skill, or attitude of the people who work in that system. This is a relatively rare insight in a world strongly biased toward individual accountability and, when things go wrong, toward blame.⁴¹

Generating ideas should be a deliberative process that considers such questions as the following:

- Do you have a disciplined process in place for generating promising ideas for solving the problem?
- Does the process involve key stakeholders and end users?
- Are the ideas based upon a strong theoretical framework?
- Are the ideas clearly and directly aligned with the problem to be addressed?

iii. Define the measurements—What measures can be used to determine whether the change is an improvement? The National Academy of Education report on teacher preparation evaluation constructed a table listing commonly used measures of provider quality, together with brief descriptions of the strengths and limitations of each. This table is provided as Appendix III and EPPs may find it a useful tool as they define metrics for case studies.

Bryk notes that measures should be embedded to gauge key outcomes and processes, tracking change and supporting judgments that the changes are actually improvements. He also reminds EPPs to anticipate unintended consequences and to measure those as well.⁴² Ewell counsels as follows:

- **Why is the evidence important?** The intent of the evidence presented should be clear and the evidence should directly suggest program improvements. For example, the potential results of a

³⁹ Ewell, P. (2013). “Action Research” Measures for Program Improvement (p. 1). Washington, DC: CAEP. Retrieved from <http://caepnet.files.wordpress.com/2012/12/caep-action-research-measures.pdf>

⁴⁰ *Improvement Research*. Carnegie Foundation for the Advancement of Teaching. Retrieved from <http://www.carnegiefoundation.org/improvement-research/approach>

⁴¹ Berwick, D. M. (2003). *Improvement, trust, and the healthcare workforce* (p. i2). Boston, MA: Institute for Healthcare Improvement. Retrieved from <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1765768/pdf/v012p000i2.pdf> (From an article first presented at the Second Asia Pacific Forum on Quality Improvement in Health Care, Singapore, September, 2002.)

⁴² *Improvement Research*. Carnegie Foundation for the Advancement of Teaching. Retrieved from <http://www.carnegiefoundation.org/improvement-research/approach>

given case study should be important or significant enough to trigger actions to modify the program.

- **Is the evidence direct and compelling?** Evidence should be directly related to the underlying condition or phenomenon under investigation. For example, if the effectiveness of candidate preparation is the object, student testimony through surveys indicating that they feel that they have received effective preparation should not be the only form of evidence submitted.
- **Is the evidence theoretically grounded?** Every body of evidence is situated within a larger theoretical or conceptual framework that guides the entire investigation. Every new piece of evidence generated or applied builds upon this framework to create new understanding. For example, case descriptions of candidate teaching in a clinical setting are located within and made sense of through frameworks that describe sound teaching practice.
- **Can the evidence be corroborated?** Because all evidence is of variable or unknown quality and coverage, it should always be backed up or “triangulated” by evidence from other sources that provide results that are consistent with those originally shown. These sources should be as different from one another as feasible, and the more of them that are presented, the better.
- **Is the evidence part of a coherent and explicit chain of reasoning?** Sound evidence requires the development of a logical chain of reasoning from questions to empirical observations that is coherent, transparent, and persuasive to a skeptical outsider. Detailed descriptions of procedures and analyses are crucial to permit others to critique or attempt to replicate a study.
- **Is the evidence drawn from situations that are typical and potentially generalizable?** All evidence should be drawn from situations that are typical. A given case study advanced as evidence should therefore be closely examined to determine if a similar case study in another situation or setting might show something else.
- **Can the evidence be replicated?** Additional confirmation of what any evidence shows can be provided by clear documentation that might enable the study to be repeated to determine the extent to which the same result is obtained.⁴³

iv. Test promising solutions—As noted above, the Institute for Healthcare Improvement has pioneered a systematic approach for testing promising hospital-based solutions which the Carnegie Foundation has adapted to education. Bryk says that the critical issue is not only what works, but rather what works, for whom, and under what set of conditions. Key questions embodied in this process include:

- Do you have a system in place to test ideas in authentic settings, rapidly collect and analyze results, make adjustments, and test interventions in additional contexts? The Institute for Healthcare Improvement calls this a *Plan-Do-Study-Act cycle*⁴⁴. Bryk notes, “That failures may occur is not the problem; that we fail to learn from them is”⁴⁵.
- Is the EPP using the measures set up in section iii to test promising solutions?
- Is the EPP able to determine if the change is an “improvement” based upon the evidence?
- Is the EPP able to determine through evidence what works, for whom, and under what set of conditions?

These steps imply not only the gathering of data, but also its analysis. Is there, in fact, “improvement”? Is there some unpredicted outcome? What can be learned about the consequences of the intervention that is under study?

⁴³ Ewell, P. (2013). “Action Research” *Measures for Program Improvement* (p. 1). Washington, DC: CAEP. Retrieved from <http://caepnet.files.wordpress.com/2012/12/caep-action-research-measures.pdf>

⁴⁴ Langley, G. J., Moen, R., Nolan, K. M., Nolan, T. W., Norman, C. L., & Provost, L. P. (2009). *The improvement guide: A practical approach to enhancing organizational performance* (2nd ed.). San Francisco, CA: Jossey-Bass. Retrieved from <http://www.ihl.org/resources/Pages/HowtoImprove/default.aspx>

⁴⁵ *Improvement Research*. Carnegie Foundation for the Advancement of Teaching. Retrieved from <http://www.carnegiefoundation.org/improvement-research/approach>

v. Sustain and scale solutions– A key goal of improvement work is the effort to transform promising ideas into sustainable solutions that achieve effectiveness reliably at scale. The term “scaling up” is popularly used to indicate moving from a limited effort to one that is much more widely implemented. Within an EPP, the concept might pertain to moving from piloting a “promising solution” with, say, half of the elementary teacher candidates, to all of the elementary preparation program. Or it might mean adapting a successful “promising solution” developed for the elementary preparation program to secondary preparation or preparation of special education teachers.

There are many steps along the way. The issues of sustainability and scaling should be built into the solution’s design from the outset and not be done as an afterthought of the improvement process. Bryk writes: “Accelerate improvements through networked communities. Embrace the wisdom of crowds. We can accomplish more together than even the best of us can accomplish alone.”

Here are questions to consider at the early stages and into the later steps:

- Does the EPP intend to implement the solution in other programs or contexts over time?
- What level of evidence does the EPP need to begin to scale the solution?
- At what point will the EPP need to conduct an impact study?
- Will scaling require changes in the design of the solution? How will these changes affect performance?

vi. Share knowledge–As Bryk emphasizes, building the field’s capacity to “learn in and through practice to improve” is a critical need⁴⁶. Thus sharing new knowledge about both the solution and the improvement process for developing it is a critical element of this improvement work. Here are several questions to consider:

- What conclusions and inferences can be drawn from the solutions generated through the process?
- How will the EPP share the findings?
- What lessons has the EPP learned about the continuous improvement process itself?
- What kinds of adjustments are needed in the EPP’s continuous improvement process?
- What more does the EPP need to know about the solution and continuous improvement?

7. IMPACT OF CANDIDATES AND COMPLETERS ON P-12 STUDENT LEARNING—guide for EPP information on impact that candidates and completers have on P-12 student learning.

a. Context

CAEP Standard 4, on preparation program impact, begins with a call that providers demonstrate “the impact of (their) completers on P-12 student learning and development, classroom instruction, and schools, and the satisfaction of its completers with the relevance and effectiveness of their preparation.” The concept of teacher impact on P-12 student learning measures as a basis for judging preparation occurs throughout the standards, and includes measures at both pre-service and in-service levels. The Commissioners viewed candidate and completer impact on student learning as the “ultimate” measure by which preparation would be judged. The CAEP Data Task Force characterizes P-12 student learning as “the only *direct* measure” of the results of teacher classroom performances.⁴⁷

⁴⁶ Bryk, A. S., et al. (2013, June). *Improvement Research Carried Out Through Networked Communities: Accelerating Learning about Practices that Support More Productive Student Mindsets* (p. 3). Retrieved from http://www.carnegiefoundation.org/sites/default/files/improvement_research_NICs_bryk-veager.pdf

⁴⁷ Ewell, P. (2013). *Report of the data task force to the CAEP Commission on Standards and Performance Reporting* (p. 1). Washington, DC: CAEP.

In-service measures have received much media attention as groups of researchers have advanced their differing perspectives on P-12 student learning data as a factor in teacher evaluations. However the research knowledge base has accumulated so that the debate is now *less* about should we or should we not, and *more* about what are the appropriate ways to apply these data in different situations.

For additional perspectives, readers are referred to papers prepared with CAEP collaboration by the American Psychological Association⁴⁸ and, through a CAEP commission by the Value-Added Research Center at the University of Wisconsin⁴⁹. Both of these are applications of P-12 student learning data in teacher evaluations for the purposes of program evaluation and accreditation rather than for evaluation of individual teacher performance. The Data Task Force recommended that CAEP continue to undertake and foster additional validation studies for application of P-12 impact data to preparation evaluation and accreditation⁵⁰. Among other topics, these validation studies should document whether particular measures employed are appropriately aligned with the curriculum implemented by the teachers for whom results are reported.

A recent report from the National Academy of Education is addressed entirely to evaluation of teacher preparation programs and contains the boxed summary, below, on P-12 learning measures in teacher evaluations. Note that the statement distinguishes use of P-12 student learning data to evaluate preparation from using them “for high-stakes decisions about individual teachers” (see last paragraph).

**Exhibit 3 • Excerpt from National Academy of Education report,
*Evaluation of Teacher Preparation Programs – The Potential Value and Risks of Using
VAMs for TPP Evaluation*⁵¹**

NOTE: “TPP” = Teacher Preparation Program

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.

⁴⁸ Brabeck, M., Dwyer, C., Geisinger, K., Marx, R., Noell, G., Pianta, R., & Worrell, F. (2013). *Assessing and Evaluating Teacher Preparation Programs*, draft paper. American Psychological Association. Washington, DC.

⁴⁹ Meyer, R., Pyatigorsky, M., Rice, A., & Winter, J. (2013). *Student Growth and Value Added Information as Evidence of Educator Preparation Program Effectiveness: A Review*. Madison, WI: University of Wisconsin, Value Added Research Center.

⁵⁰ Ewell, P. (2013). *Report of the data task force to the CAEP Commission on Standards and Performance Reporting* (p. 2). Washington, DC: CAEP.

⁵¹ Feuer, M. J., et al. (2013). *Evaluation of teacher preparation programs: Purposes, methods, and policy options* (pp. 36-37). Washington, DC: National Academy of Education. Retrieved from http://www.naeducation.org/xpedio/groups/naedsite/documents/webpage/naed_085581.pdf

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 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.

b. Guidelines

All EPPs that seek CAEP accreditation are expected to provide evidence of completer impact on P-12 student learning. The Commission suggests five source options for student-impact data:

- **Pre-service progress**— standardized measures where they are available, or periodic measures, designed and conducted by the provider to supplement other measures;
- **Pre-service exit**— for example, edTPA that includes pre-and post-instruction P-12 student data, the ETS pre-service portfolio with similar student data, or state constructed teaching performance measures;
- **State teacher evaluations**—student learning, VAMs linked with teachers (NOTE: see items a-k, appropriate qualitative characteristics for state P-12 student learning data, in point ii, below);
- **“Teachers of record” for alternative preparation**—state student growth and VAMs apply; and
- **Provider studies**—case studies conducted by the EPP.⁵²

Accreditation information on candidate and completer P-12 student impact will frequently come from “case study” evidence. But the issues attending the gathering and use of these data are sufficiently unique that these supplemental guidelines have been written for EPPs. Note that point ii, below, describes situations where EPPs are recipients of data from states that include P-12 student learning information linked with completers. The Data Task Force recommended that CAEP, as a part of its own efforts to improve accreditation data (see section 3, on pp. 8-9, on the CAEP role), take active steps to determine

⁵² Council for the Accreditation of Educator Preparation (CAEP). (2013). *CAEP Accreditation Standards* (p. 33). Washington, DC: CAEP. Retrieved from http://caepnet.files.wordpress.com/2013/09/final_board_approved1.pdf

which states can currently act as productive partners or will soon be in such a position, and work with them to determine how EPPs can access the relevant data.⁵³ CAEP is eager to begin this undertaking.

i. All EPPs provide the following information in their self studies about impact on P-12 student learning:

- **Their case for the validity and reliability of P-12 student learning impact information as they use it for preparation and accreditation purposes.** Each EPP interprets the meaning and significance of the pre-service and in-service data, and describes how the data have been used for program- or continuous-improvement purposes.
- Information taken from **pre-service assessments** of candidate impact on P-12 student learning.
 - **All providers administer assessments that monitor candidate proficiencies, including impact on P-12 student learning,** at various points during their developmental preparation experiences.
 - **All providers administer capstone assessments that sample multiple aspects of teaching.** These routinely include measures of impact on P-12 student learning and development as well as lesson plans, teaching artifacts, examples of student work and observations or videos judged through rubric-based reviews by trained external reviewers.

ii. EPPs that have access to data from states about completer impact on P-12 student learning:

- **Demonstrate that they are familiar with the sources of the P-12 student learning impact data and the state’s model for preparing the data that are attributed to the EPP’s preparation program.** EPPs describe how the data are produced, and their interpretation of the data.

Responsible state data systems make information transparent to describe:

- **The state teacher evaluations that are sent to EPPs, including:**
 - a) The psychometric soundness of the assessments taken by students and the alignment of those assessments with the State’s curriculum, and
 - b) Other sources of information in the teacher evaluation that complement that on P-12 student learning, such as employer satisfaction, teacher classroom observations, candidate satisfaction with preparation, and other relevant measures.
- **The P-12 students from whom the data come:**
 - c) The proportion of the EPP’s completers for whom P-12 student growth measures are available and the extent to which the reported completers are representative of all completers from the EPP programs,
 - d) The degree of attrition from prior to current performance measures of P-12 students that would influence interpretations of the data, and
 - e) The manner by which pupil data are linked with teachers to judge the accuracy of the associated teacher data (scores should only be used for P-12 students who are actually taught by the EPP’s completers).

⁵³ Ewell, P. (2013). *Report of the data task force to the CAEP Commission on Standards and Performance Reporting* (p. 2). Washington, DC: CAEP.

- **The state's practices in reporting the data:**
 - f) The level of state disaggregation of data so that relevant information is available for specific preparation fields,
 - g) The state criteria used to establish the minimum number of completers for whom data are provided to the EPP,
 - h) The state's decisions as to the number of years after preparation program completion that a completer's performance is associated with their preparation,
 - i) The state's practice in flagging possible biases or misrepresentation in the results,
 - j) The disaggregations provided by the state that permit comparisons for prior P-12 student performance, and
 - k) The disaggregations provided by the state that permit comparisons for completers teaching in similar situations, such as special education, disability, English Language Learners, attendance, and giftedness.
- **Document the EPP's analysis and evaluation of information provided on P-12 student learning, addressing such factors as the following:**
 - **Characteristics and patterns in the data, such as:**
 - a) The stability of the data over time,
 - b) Identification of trends or associations with program or policy features that are observed,
 - c) Separating, to the extent possible, recruitment efforts from program actions, and
 - d) Adjusting, to the extent possible, for the years of experience of teachers for whom data are reported.
 - **Interpretations of the data, such as:**
 - e) Comparisons of P-12 student learning results for the EPP with other EPPs in the state, or with the range in performance across all providers in the state;
 - f) EPP explanation of why P-12 learning results may be high or low based on EPP placements and other factors related to their mission, noting relevant factors such as the location of typical employment sites; and
 - g) Explanation of the relationships that confirm or question P-12 student learning results, based on other evidence (especially other evidence on program impact such as employer surveys; completer retention and career trajectory; structured teacher observations; and P-12 student data).
- **Judge the implications of the data and analyses for the preparation program, consider appropriate modifications, and describe EPP actions to revise the curriculum or experiences in preparation.**

iii. EPPs that do not have access to state P-12 student learning data and EPPs that are supplementing state or district data with data on subjects or grades not covered:
--

- The EPP creates data similar to those described in point ii, above, in conjunction with student assessment and teacher evaluations conducted *in school districts* where some portion of its completers are employed.
 - This type of EPP study could be phased in. For example, initially the EPP would create an appropriate design, then conduct a pilot data collection and analysis, then make refinements and further data collection.

- The EPP could maintain a continuing cycle of such studies, examining complete performance in different grades and/or subjects over time.
- In two years, by 2016, all EPPs should at least have a design in place and pilot data collection under way.
- The case study guide in section 6 on p. 18 of this document provides additional information relevant to constructing P-12 learning information from district sources.

Appendix I

Applying Principles of “good evidence” To typical accreditation measures

The seventy-nine measures included in the appendix to the CAEP Standards⁵⁴ adopted by the CAEP Board of Directors on August 29, 2013 are of different kinds, as described below:

- **Examinations.** Prospective teachers take examinations in the course of their training and in order to be licensed to practice. Dimensions of interest for examinations include their content coverage, the level at which this content is tested, the depth at which various kinds of knowledge is probed (which affects the duration of the examination), how items are structured (e.g. constructed response or multiple choice), and whether or not responses can be compared across test-taking populations (degree of standardization). The results of examinations can be reported on an absolute basis or in the form of Value-Added Measures (VAM).
- **Surveys.** Students in teacher preparation programs are frequently surveyed as they progress and after they are employed. Dimensions of interest for surveys strongly resemble those for examinations except that items are self-reported. Another important coverage dimension involves the extent to which survey items are directed at actions or behaviors, or are self-reports on knowledge or skill outcomes. This is important because students are generally more accurate commentators on the former than the latter. Surveys are also administered to the employers of teachers and to their students to help provide evidence of the effectiveness of teacher training.
- **Observations.** Observations of teacher candidates in field placements and of newly employed program graduates are also used as quality measures. Dimensions of interest parallel those of surveys but employ an element of peer judgment embodied in a trained observer using a well-developed observational protocol.
- **Statistics.** Various behavioral statistics are used as outcome measures for teacher training programs. The most prominent examples are completion rates and job placement rates. Dimensions of interest include the outcome of interest (the numerator), the population to which the rate applies (its denominator), and the time period over which the calculation is run (e.g. “one-and-a-half times catalog length of program” or “within one year”).
- **Curricular Features.** The CAEP standards address various aspects of teacher training curricula, so some of the proposed measures address descriptive aspects of the programs themselves such as the extent to which students are taught assessment methods or reflect upon professional expectations. Dimensions of interest here are the aspect of the program in question and the extent or duration of coverage.
- **Case Studies.** Where quantitative measures are unavailable, the CAEP Standards call for qualitative investigations termed “case studies” (for example, “case studies of districts in which a large number of program graduates are employed”). Dimensions of interest include the question to be investigated through the case study, baseline conditions, the intervention or phenomenon of interest, the goal of the intervention, observed results, and implications for action.

⁵⁴ Council for the Accreditation of Educator Preparation (CAEP). (2013). *CAEP Accreditation Standards*. Washington, DC: CAEP. Retrieved from http://caepnet.files.wordpress.com/2013/09/final_board_approved1.pdf

To illustrate how the principles can be used, this Appendix applies each of the principles to three measures: licensure passage rates, employment rates, and case studies of districts where a large number of program graduates are employed.

Validity and Reliability

- Examinations example: Licensure Passage Rates. Although state licensure examinations differ by provider, the two main test vendors have established rigorous standards of test development and regularly report statistics on validity and reliability. As a result, this measure fully meets the principle.
- Statistics example: Employment Rates. These are also well-defined measures that are valid and reliable so long as they are properly calculated. Because of the latter, they should be examined carefully for threats to validity and reliability such as exclusions from the denominator or changed conditions over repeated annual measures.
- Case Studies example. Case studies are a bit more problematic with respect to this principle because their validity depends on the type of information collected and the extent to which the same procedures are used across cases and over time. This will naturally be a peer judgment.

Relevance

- Examinations example: Licensure Passage Rates. Insofar as licensure tests faithfully reflect content knowledge and knowledge of pedagogical practice, program completers' performance on them constitutes relevant information about the quality of the program.
- Statistics example: Employment Rates. Many things can affect employment rates that are not directly related to program quality including local job market conditions or the general state of the economy. As a result, this measure does not fully meet the principle.
- Case Studies example. So long as the topics addressed in the case study are selected to reflect important dimensions of program performance—for example, the ability of program graduates to effect learning growth in their pupils or their demonstration of ethical and professional practices—the principle of relevance is met.

Verifiability

- Examinations example: Licensure Passage Rates. Licensure examinations are designed and administered by testing organizations that can be queried about the way these examinations operate. Similarly, the state authorities responsible for administering them can be more or less transparent in how they calculate passage rates. Both of these will affect verifiability.
- Statistics example: Employment Rates. Under ideal conditions, these are collected by state agencies on behalf of EPPs using surveys or unit record wage databases using known calculation rules. If these rules are documented, states' application of them can be independently audited to determine compliance, thus rendering them verifiable. Where surveys are conducted by EPPs themselves, such documentation may or may not be present.
- Case Studies example. The contents of case studies can also, in principle, be audited if their construction is fully documented. But this is a good deal more complicated than for program completion rates. As a result, case studies will probably not fully meet the principle.

Representativeness

- Examinations example: Licensure Passage Rates. The examined population is the population of interest with the denominator representing the parent population. So long as these populations are correctly constituted, the measure is representative.
- Statistics example: Employment Rates. These will typically be representative but when measured over time, they may not be if labor market conditions change.
- Case Studies example. These must be carefully examined to determine if the districts chosen are typical of many districts to which the program sends graduates as employees. This could be done if the requisite documentation was supplied.

Cumulativeness

- Examinations example: Licensure Passage Rates. Although these are precise, they are sole source measures so there is not much ability to triangulate results. One indication of “weight of evidence” might be recording performance over time or correlating test score performance with other evidence of academic achievement on the part of program graduates such as grades in course examinations or portfolios.
- Statistics example: Employment Rates. The same situation largely applies here as to licensure passage rates. A possible exception is if this information is collected both by surveys and by tapping wage record databases.
- Case Studies example. Case study results will be bolstered by the presence of other measures such as teacher observations or student surveys that show similar conclusions. They also can be examined for consistency over time. Finally, the construction of case studies themselves is important: if they involve mutually reinforcing lines of inquiry or investigation, their credibility is strengthened.

Fairness

- Examinations example: Licensure Passage Rates. These are supplied by third parties, so there should normally be little opportunity for bias. Where this could occur is when EPPs themselves report these rates.
- Statistics example: Employment Rates. If states collect these data on behalf of EPPs using surveys or by tapping wage record databases, the measure should be unbiased. Again, if EPPs themselves conduct the surveys, bias could enter.
- Case Studies example. Because they are conducted by EPPs entirely and are used to advance a particular quality claim, case studies are unlikely to be entirely unbiased.

Stakeholder Interest

- Examinations example: Licensure Passage Rates. These can be looked at from a number of stakeholder perspectives including the prospective candidate who wants to know her or his chances of success, the program as a measure of the quality of preparation in the areas tested, and the employer who wants to know how well potential employees are prepared.
- Statistics example: Employment Rates. The case is similar to licensure passage rates.

Appendix II

The eight annual reporting measures: Developing more useful measures

When fully developed, how will the eight annual reporting measures be defined? It may be premature to pose that question since the state of measurement in preparation is so frequently EPP specific, so it may be some time before an “ideal” system can emerge. Still, in the interest of transparency, this Appendix describes CAEP’s preferences as to characteristics of data that would make them more direct and powerful for all users. These preferences will shape the annual CAEP reporting measures and other CAEP actions.

The view of CAEP, as the educator preparation accreditor, is that any measures will be better and stronger if they are more user friendly. The CAEP proposition is that measures will be more user friendly if similar topics or practices or results of EPPs can be compared and contrasted. That requires that the measures, themselves, have rigorous definitions, that statistical descriptors be the same, that protocols for gathering data (e.g., time periods covered, reporting dates) be established, that some standardized assessments be used commonly so that performance benchmarks can anchor the preparation data. With that perspective, CAEP’s descriptions of ideal directions for measures are set out in the paragraphs, below.

THE FOUR MEASURES OF PROGRAM IMPACT

1. Impact on P-12 learning and development

IDEAL—Pre-service candidate impact on P-12 student learning is evaluated through recurring formative assessments and in some standardized culminating assessment that includes explicit demonstration of P-12 student learning. EPP practices that integrate pre- and post-instruction P-12 student learning into edTPA or the ETS pre-service portfolio are among examples.

In-service performance is assessed through state and/or local district teacher evaluations:

- that include information on P-12 student impact appropriately attributed to each teacher;
- the evaluation models are generally understood and accepted by technical, administrative and policy representatives;
- there are appropriate adjustments for prior P-12 student learning;
- there are appropriate adjustments for characteristics of the schools and/or students in which the teacher is employed; and
- there are additional measures such as classroom observation, teacher surveys and student surveys.

EPPs routinely make use of these data when they are provided by the state, or seek them out when they are available from school districts and can be reported and analyzed in similar ways. EPPs routinely supplement these data with special studies of teachers in grades or subjects not covered by accessible state or district teacher evaluations.

2. Indicators of teaching effectiveness

IDEAL—One or two classroom observation measures are commonly used by most EPPs. These are validated for teacher evaluation purposes (e.g., through the MET study). Reviewers are trained and external to the program.

CAEP state partnership protocol arrangements examine the potential for such measures across the state.

3. Results of employer surveys, including retention and employment milestones

IDEAL—CAEP collaborates with states with the objective of creating common employer satisfaction surveys that explicitly link preparation satisfaction with various elements of preparation that are important in accreditation. As a result:

- One or two surveys are commonly administered for employers of new teachers in their first, second and third year of teaching and results are returned to the EPP;
- Questions address employer satisfaction with completers preparation along particular dimensions of preparation (similar, perhaps, to those recently created for Ohio and Missouri);
- State Education Agencies (SEAs), State Higher Education Executive Officers (SHEEOs) and state employment agencies enter into agreements to provide employment and retention data on all first, second, or third year teachers—alternatively, employers provide these data to EPPs; and
- EPPs make comparisons with self-selected or state-selected peers and CAEP develops benchmark performances for national reporting.

4. Results of completer surveys

IDEAL—One or two surveys are commonly administered for new teachers in their first, second, and third year of teaching and results are returned to the EPP. Questions address satisfaction of completers with particular aspects of preparation (similar, perhaps, to those recently created for Ohio or Missouri). These data are tracked over time to indicate trends. EPPs make comparisons with self-selected or state-selected peers and CAEP develops benchmark performances for national reporting.

THE FOUR MEASURES OF PROGRAM OUTCOME AND CONSUMER INFORMATION

5. Graduation rates

IDEAL—EPP statistical records have capacity to follow individual students longitudinally from admission to completion and at least three years thereafter. For each entering cohort, statistics are derived on those who dropped out or were counseled out, those who completed the full program and certification, and those employed.

From these data the completion rate is calculated as the number of completers divided by the number of admitted candidates in a cohort. Dropouts and counseled out candidate rates will be calculated similarly.

EPPs make comparisons with self-selected or state-selected peers and CAEP develops benchmark performances for national reporting.

6. Ability of completers to meet licensing (certification) and any additional state requirements

IDEAL—State licensure tests are closely aligned with InTASC, Common Core, college and career ready, and CAEP standards. They have many common features that make them at least partially aligned, and they are scored so that comparisons can be made. CAEP would require an 80% pass rate on either the first or second administration for completing candidates. The statistic is defined as number of licenses earned by completers in a cohort divided by number of admitted candidates in the cohort. Trends are reported for three to five years.

EPP statistical records have capacity to follow individual candidates longitudinally from admission to completion and at least three years thereafter. These records include data on licensure test taking and results.

EPPs compare their results with self-selected or state-selected peers and CAEP publishes national data with benchmark performances for groups of EPPs.

7. Ability of completers to be hired in education positions for which they were prepared

IDEAL—EPPs report the completer employment status as of September 1 after preparation program termination, disaggregated by:

- a. employed in position for which trained/ admitted cohort;
- b. employed in any other education position/ admitted cohort;
- c. enrolled in continuing education/ admitted cohort;
- d. other employment/ admitted cohort; and
- e. other or not employed/ admitted cohort.

The statistic would be defined as the number in each of the a-through-e categories divided by the number of completers used in item 5, graduation rates. CAEP would use these data to develop national benchmarks on groups of similar EPPs.

8. Student loan default rates and other consumer information

IDEAL—Student loan default rates are calculated from U.S. Department of Education data by extracting the EPP information from the institution-level report. This is one of several indicators of “consumer information” that include additional measures created by EPPs such as:

- Cost of attendance using some common methodology,
- Beginning salary of completers based on official employer records and trends over time, and
- Placement location patterns for completer cohorts, with trends over time.

Appendix III

Strengths and Limitations of Commonly Used Measures of Teacher Preparation Programs, National Academy of Educationⁱ

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

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

TABLE 4-1 Continued

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

TABLE 4-1 Continued

Measure	Strengths	Limitations
<i>Textbooks</i>	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
<i>Course offerings and required hours</i>	Easy to collect Easy to make comparisons across programs	Does not indicate actual quality of instruction in TPP courses
<i>Number of required content courses</i>	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		
<i>Fieldwork policies including required hours</i>	Easy to collect Easy to make comparisons across programs	Does not indicate actual quality of fieldwork experience
<i>Qualifications of fieldwork mentors</i>	One aspect of quality of fieldwork experience	Little empirical evidence links characteristics of mentors to their success in teacher preparation
<i>Surveys of candidates</i>	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
<i>Records from observations of student teaching</i>	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		
<i>Percentage of faculty with advanced degrees, part-time, adjunct, etc.</i>	Easy to collect	Many instructors of teacher candidates are in departments other than education and tend not to be included in the evaluation
	East to make comparisons across programs	
	Face validity—TPP faculty should have appropriate expertise and credentials	Little empirical evidence to support connection to effective teacher preparation
Effectiveness in Preparing Candidates Who Are Employable and Stay in the Field		
<i>Pass rates and/or average scores on licensure tests</i>	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)
<i>Hiring and retention data</i>	Important to potential candidates; face validity	Influenced by numerous geographic and non-TPP factors
		May be inaccurate and/or difficult to collect; have to track graduates post-TPP

TABLE 4-1 Continued

Measure	Strengths	Limitations
Success in Preparing High-Quality Teachers		
<i>Teacher performance or portfolio assessments administered near end of program</i>	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
<i>Ratings of graduates by principals/employers</i>	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
<i>Value-added models</i>	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

ⁱ Feuer, M. J., et al. (2013). *Evaluation of teacher preparation programs: Purposes, methods, and policy options* (pp. 84-88). Washington, DC: National Academy of Education. Retrieved from http://www.naeducation.org/xpedio/groups/naedsite/documents/webpage/naed_085581.pdf