State Teacher Preparation Program Accountability Systems: A Review of the Literature

By

Edward Crow
Co-founding Partner
Teacher Preparation Analytics, LLC

October 1, 2016

Collecting, organizing, and reporting the information needed to understand the quality of teacher preparation programs are not easy tasks. Despite some progress over the last five years in improved state data systems, data collection (and use of data) about teacher preparation programs and graduates is quite fragmented and incomplete. One consequence is the absence of systematic and reliable information about the knowledge, skills, and effectiveness of program graduates outside of a few states (Louisiana, Texas, Florida, Tennessee), a small number of universities that invested their own resources in this work (e.g., New York University, Virginia), and research projects making effective use of access to state datasets (CALDER and Pathways are the best known of these)[1].

Where they exist at all most indicators other than student achievement are proxies for the concepts, knowledge, and behaviors they claim to measure. And measures of teacher effectiveness are available today for only about one-third of all teachers. Challenges related to the availability and quality of data encompass almost everything about teacher preparation: the characteristics of entering students, their experiences and performance in preparation programs, outcomes such as teaching performance, pupil learning, persistence in teaching, and how teaching context may or may not affect these and other outcomes.

This state of affairs is in some ways a reflection of the field itself, where there is still too little agreement on the knowledge and skills that graduates should have and be able to demonstrate in the classroom. Where agreement can be found—on "standards", "competencies", and "dispositions"—it exists mostly at a level of abstraction where the concept is so general as to be often non-observable and not measurable in reliable and valid ways. These problems have consequences for accreditation policies and practices, for research about teacher education, for state oversight of preparation programs, for systematic and consistent reporting about preparation programs, and for the efforts of programs to assess their own effectiveness. These issues are compounded by the lack of strong data systems able to collect and share results *within* states, much less across state lines.

What's Needed Now: Program Performance Indicators and Good Data Systems

[1] In 2013 the Data Quality Campaign reports that seventeen states now link student performance data to preparation programs. Using the data for reporting still appears to occur in just a handful of states.

Selection**:** Academic Strength of Students Admitted to Preparation Programs

There is good evidence to support a positive relationship between measures of teacher academic ability and teacher effectiveness. For traditional undergraduate programs, available measures of academic ability include high school and college grade point averages, high school rank in class, and standardized test scores on the ACT and SAT (and the GRE for graduate programs).

Studies about the academic ability of teachers find that that measures of teacher verbal ability are associated with student achievement and there is some evidence for the same connection between teacher mathematical ability and K-12 math achievement. Studies also find links between program or institutional selectivity and K-12 student achievement. A key finding from several pieces of research is that teacher academic ability is especially important to student learning outcomes for at risk students.[2]

It is also relevant to note that studies in multiple states find that the academic ability of those seeking to enter teaching has improved somewhat since the mid-1990s—particularly as measured by SAT scores, average SAT percentile distribution, undergraduate GPA, and—to a lesser extent—selectivity of undergraduate institution. There is no evidence that these improvements result from changes to teacher preparation programs; they appear to be influenced by state and federal policy changes in some states. These improved academic profiles of teacher candidates and new teachers suggest that program selection criteria emphasizing academic ability measured through standardized tests and prior performance will affect fewer potential applicants than might have been the case 20 years ago when the overall academic profile of teacher candidates and new teachers was weaker than it now appears to be. Other professions also employ standardized tests as a component of selection into professional education (e.g., the LSAT, GMAT, GRE, and MCAT tests). Studies consistently find that these tests have predictive validity for subsequent academic success in the professional education programs for which they are used. Perhaps because these professions do not have

[2] For more on the research about academic ability and teacher preparation, see *CAEP Standard 3.2: Research, Study and Analysis,* a report to the CAEP by Teacher Preparation Analytics LLC, which can be retrieved from http://caepnet.org/about/news-room/caep-board-clarifies-refines-caep-standa.

internal debates about whether academic ability is relevant to success in professional education or in professional practice—and because selection into professional education programs in these fields is highly competitive—accreditation standards do address selection procedures but do not specify performance levels on the standardized tests universally employed to determine who should be admitted.

In the international context, a McKinsey consulting report found that the highest performing national school systems in the world recruited teachers from the top-third of the college graduating class. This study further claimed that 23% of new teachers in the United States come from the top third.[3]

While there continues to be debate among teacher educators about the importance of selection standards for admission into teacher education—and recognizing that the research literature has mixed findings on the relevance of teacher academic ability to key program outcomes , measures of these traits are built into state program oversight policies, national accrediting standards, program admission requirements, and—sometimes—preparation program progression standards.[4] It now appears that a declining number of educators and policy makers advocate lowering academic performance standards as entry requirements into teacher education programs.

Selection: Potential for Teaching

Preparation programs, school districts, and national organizations like Teach for America (TFA) all seek to measure individual attitudes and values that may predict suitability for and success in teaching. These attributes are also known as "dispositions" in the teacher education world and there is recent research linking beliefs or values to measures of teaching quality or teacher effectiveness.[5] Where solid evidence does exist, the findings hold some promise for pre-

[3] August, Kihn, and Miller, 2010. They define the top-third of US college students in terms of ACT, SAT, and GPA. (Ibid., p. 10)

[4] Here for instance, students already enrolled in a teacher preparation program may have to earn a certain GPA to be admitted to student teaching and may need to earn a minimum grade to be recommended for state certification by the program.

[5] For example, see *Screen Twice, Cut Once: Assessing the Predictive Validity of Teacher Selection Tools,* by Dan Goldhaber, Cyrus Grout, and Nick Huntington-Klein, retrieved from http://cedr.us/papers/working/CEDR%20WP%202014-9.1.pdf. Another interesting approach is being taken by TeacherMatch (https://www.teachermatch.org/). TeacherMatch uses a 75-item assessment of

screening applicants to preparation programs as is done routinely in other professional fields and employment recruitment. For teacher education programs, there does not appear to be independent evidence for the reliability or validity of instruments such Gallup Insight or the "STAR teacher pre-screener" sold through the Haberman Foundation.[6]

Despite the largely missing research base for the value of instruments like Gallup Insight and the Haberman pre-screener, there is reason to believe that programs could make effective use of protocols that seek to determine "goodness of fit" between an applicant seeking admission and the career that she or he hopes to join. In recent years, Angela Duckworth and her colleagues at the University of Pennsylvania have developed, tested, and made wide use of the "Grit Scale" to gauge potential for success in teaching and other fields. Duckworth describes grit[7] as "the tendency to sustain interest in and effort toward very long-term goals." Research studies have found a link between individual-level measures of "grit" and outcomes for teachers who are prepared through Teach for America[8].

Teach for America (TFA) screens applicants for a number of traits and attitudinal characteristics that it argues are associated with effective teaching.[9] During its highly competitive and multi-stage recruitment process, TFA screens applicants for their: demonstrated previous achievement; perseverance through challenges; critical thinking skills; capacity to motivate and influence other people; the applicant's organizational ability; how well applicants understand and strongly support the TFA vision for high quality teaching; and evidence for respect of students and families in low-income communities. The organization argues that its measures of these characteristics predict successful teaching and persistence in very challenging school settings. Unfortunately for the development of national indicators of these traits, TFA's measures and research about their usefulness are not in the public domain.

The American Psychological Association (APA) task force of education and

---

attitudes, cognitive ability, and teaching skills as a tool for recruiting, screening, and developing teacher talent in collaboration with schools and districts.

[6] For more information, see http://www.habermanfoundation.org/starteacherprescreener.aspx. Apart from studies and papers published by the test developer, however, there is no independent evidence for the reliability or predictive validity of this instrument.

[7] More information at https://sites.sas.upenn.edu/duckworth/pages/research-statement.

[8] Duckworth, A. L., Quinn, P. D., & Seligman, M.E.P. (2009). Positive predictors of teacher effectiveness. *Journal of Positive Psychology, 19,* 540-547.

[9] A general description of the TFA selection process and what it claims to screen for can be found at http://www.teachforamerica.org/why-teach-for-america/who-we-look-for.

measurement experts examined issues associated with indicators and measures of teacher education program quality,[10] Including research literature on "early-stage selection or screening tools in development that have shown preliminary evidence of validity for predicting candidates' competence in classroom interactions". While APA found " few if any systematic uses of such instruments in teacher preparation, however, and very little, if any, validity data that predict competence in the classroom or are useful for making selection decisions".[11]

Diversity of Admitted Candidates and Program Completers

Policy leaders and teacher educators support the idea that the teaching force should be diverse, not only to provide opportunities for talented individuals but also because of the increasing diversity of the K-12 student population in the United States. Currently, about 84% of US K-12 teachers are white, 7% are African-American, and 6% are Hispanic. Men comprise 16% of the K-12 teaching population.[12] The demographic composition of the K-12 student population is far more diverse than that of the teacher workforce.

Most preparation programs collect information about the demographic composition of applicants, admitted students, and program graduates. This data—particularly any comparisons between demographics of admitted students and completers—is not widely shared outside the program. Through annual reporting to the U.S. Department of Education, states provide information that facilitates construction of comparison tables like this one:

California Racial/Ethnic Distribution				
Ethnic Group	**Enrollees** (TTPs)	**Students** (host IHEs)	**K-12 Students** (state)	**K-12 students** (national)
Amer Indian/Alaska Native	0.7%	0.7%	0.7%	1.3%
Asian or Pacific Islander	10.0%	24.5%	11.6%	5.0%

[10] *Assessing and Evaluating Teacher Preparation Programs*. Washington, DC: American Psychological Association, 2013. Discussion of issues related to selection measures and their use is on pages 8-9 of the report's working draft.

[11] Ibid, p.8. The APA group cited the work of Jamil et al. as having potential promise. See Jamil, F., Hamre, B., Pianta, R., & Sabol, T. (2012). *Assessing teachers' skills in detecting and identifying effective interactions in classrooms.* Manuscript submitted for publication.

[12] See C. Emily Feistritzer, 2011. *Profile of Teachers in the U.S., 2011*. Washington, DC: National Center for Education Information. Despite the name of the organization, this is a private entity that collates information from various governmental and professional organizations.

.

Black or African American	6.0%	6.1%	6.9%	16.6%
White	57.1%	44.8%	27.0%	53.4%
Hispanic/Latino, any race	21.9%	23.8%	50.4%	23.0%
Two or more races	4.2%	0.1%	3.4%	0.7%

This information is the most recent summary available through the US Department of Education, and it covers the 2009-10 academic year. States could also report information about the demographic composition of newly licensed teachers; across the country, however, large percentages of newly licensed teachers in some states were prepared through teacher education programs in other states.

There is little evidence from research showing empirical relationships between teacher demographics and K-12 student outcomes, although the "race and ethnicity of teachers is related to race and ethnicity of students.[13] One study found a positive relationship between teacher ethnicity and pupil outcomes in mathematics but other analyses reported no or negative linkages.[14] Nonetheless, charting and reporting on the demographic composition of entering and exiting preparation program candidates is a policy concern in every state. Current data and reporting resources are not adequate to support universal and reliable indicators on this subject, but given the diverse composition of US school enrollment and of the adult population, we think it is reasonable to include demographic measures of those admitted to and graduating from every preparation program.

Quality of Preparation: Content and Pedagogical Content Knowledge

Preparation program accreditation and accountability are taking steps to improve the quality of information about the content knowledge and professional knowledge of teacher candidates and program graduates. The problem is finding measures in both areas that are strong, credible, and useful indicators. Praxis and similar tests have been used by the states for many years, but few outside the profession see these tests—in their current incarnations (paper-and-pencil, non-performance based)—as credible indicators of candidate or new teacher knowledge. Many inside the profession share these doubts.

[13] Karen Zumwalt and Elizabeth Craig, 2005. "Teachers' Characteristics: Research on the Demographic Profile." In Cochran-Smith and Zeichner, Studying Teacher Education. PUBLISHER, p. 125.
[14] Aaronson et al. 2007. Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago Public High Schools. *Journal of Labor Economics, 25*(1), 95-135.

Indicators of content and pedagogical content (content knowledge for teaching) knowledge can be built up to program-level measures by aggregating or averaging the scores earned by individuals. Since 1998's Title II of the Higher Education Amendments (HEA), program-passing rates on various tests provide another window into the quality of preparation. The current problem with these indicators is three-fold:

- Test content: tests now in use do not measure knowledge and skills that can be tied to important teaching outcomes.
- Passing scores: the test score threshold for success is established by states and is generally set for political (or supply-demand) reasons unrelated to teaching ability or effectiveness.
- Interstate comparisons: unlike other professions, virtually every state uses its own set of teacher knowledge and skills tests; where two states use the same test(s), passing scores are often set at different points.

A recent report noted that more than 1100 teacher tests are in use across the fifty states, with over 800 content knowledge tests alone. Even when two or more states employ the same test of content or professional knowledge, the states set different passing scores.[15] According to the US Department of Education, 97% of all test-takers in the United States get passing scores on the current panoply of teacher tests.[16]

These are serious problems for quality control and consistent reporting by accreditors, states, and others trying to understand the quality of preparation programs and their graduates. The tests themselves have little demonstrable relationship to the knowledge, skills, and teaching performance required in today's schools. For example, the expert panel convened by the National Research Council reported in *Testing Teaching Candidates: The Role of Licensure Tests in Improving Teacher Quality* (Washington, DC: National Academy of Sciences, 2001):

> "Teacher licensure tests focus on the knowledge and skills identified by panels of educators as critical for entry into the profession. They cover the material considered to be *minimally* necessary for beginning teaching. Teacher licensure tests are not designed to distinguish moderately qualified teachers from highly

[15] Crowe, "Measuring What Matters: A Stronger Accountability Model for Teacher Education," p. 8.
[16] "Preparing and Credentialing the Nation's Teachers: The Secretary's Eighth Report on Teacher Quality, Based on Data Provided for 2008, 2009, and 2010." (Washington: U.S. Department of Education), p. 54.

qualified teachers. They are not constructed to predict the degree of teaching success a beginning teacher will demonstrate" (NRC, p. 47).

Other researchers report "little evidence that…a relationship exists between teachers' scores on such tests and their teaching success."[17] Candidates who cannot pass these tests probably should not have been admitted to a program in the first place, and programs with low pass rates should be closed. But other than using teacher test data to set a much higher quality floor than is currently the case in any state, the licensing tests now in use do not measure outcomes relevant to the academic success of K-12 students or their schools.

Making headway on this challenge would be a significant contribution to teaching quality in the United States *and* would help to enhance the professional status of teachers and the programs that produce them. A recent report from the Council of Chief State School Officers (CCSSO) shows that states may be ready for real reform in this regard. The Chiefs are calling for a multi-state effort to develop "innovative licensure assessments" that include evidence about teacher impact on student achievement. Their report also argues for state program approval standards that address a program's ability to produce graduates who positively affect student learning.[18]

CCSSO argues for a range of indicators to measure teaching and program quality—including observation data, pupil achievement measures, surveys of graduates and school leaders, program retention rates, and placement into hard to staff teaching positions. Perhaps of equal importance, the Chiefs call for state data on preparation programs, disaggregated in various ways, to be provided to accreditors.

Relevant to this discussion is the research conducted by Deborah Ball and her colleagues at the University of Michigan indicate at least two empirically discernable subdomains within "pedagogical content knowledge" (PCK) - *knowledge of content and students* and *knowledge of content and teaching* - and an important subdomain of "pure" content knowledge unique to the work of teaching - *specialized content knowledge* - which is distinct from the *common content*

[17] Suzanne Wilson and Peter Youngs, "*Research on Accountability Processes in Teacher Education.*" In Marilyn Cochran-Smith and Kenneth Zeichner, eds., Studying Teacher Education. Washington, DC: AERA, 2005, p. 592.

[18] This discussion draws on "Our Responsibility, Our Promise: Chief State School Officers' Task Force Report on Transforming Educator Preparation and Entry into the Profession." (Washington: Council of Chief State School Officers, 2012).

knowledge needed by teachers and nonteachers alike.[19] The assessment of PCK – also called "content knowledge for teaching" (CKT) - is needed to provide some assurance that the candidate pedagogical mastery of content that a beginning teacher might well be called upon to teach in his or her very first semester on the job. A TPA 50-State analysis for CCSSO found that no state employs such a comprehensive assessment of PCK, but a new test being currently piloted for elementary grades by ETS – part of its NOTE series – does assess a much broader range of content-based teaching knowledge.[20]

To be useful as program indicators of candidate knowledge and skills—and of graduate knowledge and skills—uniform reporting means that the same tests should be used for every program, no matter what state it is located in, accompanied by uniformly high passing cut scores applied nationally, no matter what an individual state might establish as its own passing score.[21] Building a better system of content and pedagogical content knowledge measures can draw from the experience of other professions when it comes to tests of content knowledge and professional knowledge.[22]

Demonstrated Teaching Skills for Candidates and Graduates

The classroom teaching performance of candidates and program graduates is a key outcome to use as a quality measure.[23] For graduates, it also may be useful (and cost effective) to explore

[19] Deborah Loewenberg Ball, Mark Hoover Thames, and Geoffrey Phelphs. Content Knowledge for Teaching: What Makes it Special? Journal of Teacher Education, November/December 2008 vol. 59 no. 5 389-407

[20] Pearson. Teacher Licensure Testing and Performance Assessment. Website. Accessed 9-20-16 at: http://www.pearsonassessments.com/teacherlicensure.html

[21] A recent AFT report, "Raising the Bar: Aligning and Elevating Teacher Preparation and the Teaching Profession" (Washington: American Federation of Teachers, 2012) calls for a "universal assessment process for entry" but says nothing about passing standards or establishing a common passing score across all states.

[22] Engineering, accountancy, nursing, and medicine operate with uniform state accountability standards and requirements. In nursing, for instance, the NCLEX-RN is accepted by every state as the single licensure test that determines whether or not a program graduate is granted a license to practice nursing. Every state uses the same passing standard, and pass rates are tied to program accountability for more than 1200 professional nursing programs in the United States (https://www.ncsbn.org/nclex.htm). There is a similar story in engineering. All states employ the same battery of tests for would-be engineers, and every state employs the same passing score (see http://www.ncees.org/Exams.php). The profession of accountancy follows a similar pattern, with all states using the same four-part Uniform CPA Examination and passing scores (see http://www.bls.gov/oco/ocos001.htm#training).

ways to collect multiple measures of teacher performance from employers and mentor teachers on the performance of beginning teachers. That is to say, as more districts do better evaluation of their teachers, these school- or district-based data may be good sources of information for programs about graduate's performance as teachers, if the district will share with the program their findings about graduates of the program who teach in the district. While the value of this information would be affected by the quality of district-based evaluation mechanisms, it might be worth looking into as a source of data. Similarly, as states implement statewide annual teacher performance evaluation systems that include measures of teaching ability and student learning, these data could serve as indicators of teaching skills.

There are several challenges associated with use of statewide teacher evaluation measures for a national reporting system: the first, of course, is that each state will have its own system which limits cross-state comparability for graduates of different programs. More recently, states have outsourced their teacher evaluation practices to individual districts—allowing each district to develop its own mix of evaluation components and its own metrics. This development comes on top of the fact that state-level teacher evaluations in each state give different weights to measures within the overall evaluation score. This is particularly the case with the weight assigned to student learning outcomes. Although student achievement is the most important indicator of teaching and school quality, there appears to be a "race to the bottom" among states in assigning a weight to achievement results.

Classroom observation and assessment of on-the-job teaching should be regarded as a key measure of quality because no single measure tell us all we need to know about a program or its graduates. Some *programs* now employ classroom observation to gauge development of requisite knowledge and teaching skills by their teacher candidates, suggesting there might really be two performance-related measures here for outcomes-focused teacher education programs: performance of candidates *during* the program and their performance as teachers of record upon completion of the program. The Key Quality Indicators framework developed by Teacher Preparation Analytics includes both uses of this measure.

---

[23] Currently measures of teaching skills can be more easily collected from a larger number of candidates and teachers (or from representative samples of candidates and program graduates across the grades and subject areas). Data on student performance are less widely available because most teachers are assigned to untested subjects and grades.

Such data would help the program faculty and administrators identify knowledge and skill sets that make a difference in the professional practice of their candidates and graduates. Classroom assessment results can highlight areas for individual candidate improvement, and for preparation programs that provide induction support to new teachers, teaching assessment findings can flag areas where continued development of teaching skills would improve a graduate's overall effectiveness in the classroom and persistence in teaching.[24] Widespread implementation of a classroom teaching performance outcome measure would be a major step in providing robust and relevant evidence about the connection between teacher preparation and student achievement.

It is important to bear in mind, however, that a system of quality classroom observation must support fair judgments based on reliable and valid findings for individual teachers and for groups of teachers.[25] Not all classroom teaching observation protocols are the same. It appears as though few of those now used by teacher education programs (including most of those mandated by state regulations) meet standards of rigor. Candidates, graduates, programs, and the public deserve "validated, standardized observational assessments of teachers' classroom instruction and interactions."[26]

The edTPA instrument being placed into use by a number of states may be one way to measure the teaching skills of candidates. Within a state that uses edTPA for all candidates, results are comparable for completers for all programs in that state. Across states, however, edTPA may have less value because states are adopting different passing scores. If actual scores from test-takers were accessible, it would be possible to construct a fully comparable cross-state measure.

Although there have been studies that have corroborated the predictive validity of the edTPA for teaching effectiveness, the studies have not been unequivocally positive. A recent study by Daniel Goldhaber and colleagues, for example, notes that edTPA scores for teachers in grades 3-8 can (depending upon other variables) be correlated with teaching success in mathematics, but

---

[24] Gary T. Henry, Fortner, C. Kevin, and Bastian, Kevin C. (2012) "The Effects of Experience and Attrition for Novice High School Science and Mathematics Teachers" Science 335, 1118-1121.
[25] Laura Goe, Courtney Bell, and Olivia Little. "Approaches to Evaluating Teacher Effectiveness: A Research Synthesis" (Washington: National Comprehensive Center for Teacher Quality, 2008), p. 9.
[26] Robert Pianta and Bridget Hamre, "Conceptualization, Measurement, and Improvement of Classroom Processes: Standardized Observation Can Leverage Capacity" (Educational Researcher 38 (2), 2009), p.109.

this is not the case for reading. Simply passing the edTPA, however, is significantly predictive of teaching success in reading but not in mathematics. The study suggests that this problem is likely a function of the fact that simply passing or failing the assessment may not be as predictive as actual scores—further evidence that pass rates are not effective measures.[27]

Other measures of classroom teaching performance have been implemented and tested far more extensively than the edTPA. These include the CLASS instrument developed at the University of Virginia and the Danielson Framework for Teaching (FFT). Both instruments[28] were core components of the large scale Measures of Effective Teaching (MET) project that collected multiple sources of information about thousands of teachers across the United States.[29] From the perspective of using the quality of classroom teaching as a program performance measure, our growing knowledge base creates opportunities for innovative approaches to linking teacher performance to the programs that prepared them. Fortunately, there is a growing number of quality classroom observation instruments available.[30] National studies and pilot projects are building a foundation of knowledge for using classroom observation as a program outcome. Two large studies have produced relevant findings by examining links between observation instruments and pupil learning gains through videotaped observations of many teachers.[31] Similarly, the edTPA initiative conducted pilots in 21 states with 7,000 teacher candidates from cooperating university preparation programs, with the focus on teaching skills while candidates are still in their programs.[32] Advocates of edTPA hope it will be a reliable and valid source of performance information, but edTPA results need to be reported publicly by program and performance measures for program graduates are still needed.

---

[27] Dan Goldhaber, James Cowan, and Roddy Theobald, "Evaluating Prospective Teachers: Testing the Predictive Validity of the edTPA" National Center for Analysis of Longitudinal Data in Educational Research, (CALDER), Working Paper 157, May 2016

[28] Other instruments also were employed in the study: PLATO and MQI.

[29] For more on the MET project, see http://www.metproject.org/reports.php.

[30] See the discussion of these issues can be found in Pianta and Hamre, "Conceptualization, Measurement, and Improvement of Classroom Processes: Standardized Observation Can Leverage Capacity," p. 111; as well as Goe, Bell, and Little, "Approaches to Evaluating Teacher Effectiveness: A Research Synthesis," p. 22.

[31] The Understanding Teacher Quality initiative examines six instruments through videotaped observations of 450 teachers, while the Measures of Effective Teaching (MET) project has videotapes for about 3700 teachers. For more information about the Understanding Teacher Quality project see http://www.utqstudy.org/index.html; more about the MET effort can be found at http://www.metproject.org/index.php.

[32] See http://edtpa.aacte.org/about-edtpa.

To be useful as program quality indicators, observational findings for individual program graduates have to be aggregated and summarized for all the graduates of a specific program in order to constitute a program outcome indicator. An alternative strategy would require large enough samples of graduates to produce reliable findings. Some programs do this on their own, using the evidence to guide candidate development and for program improvement. MET and other efforts can provide useful lessons, especially as states and districts implement higher quality teacher evaluation systems. Tapping these state and district datasets for program purposes (not to evaluate individual graduates) can be a productive focus supporting development of strong program quality measures.

Employer and graduate satisfaction with preparation programs

Employer and graduate satisfaction with teacher preparation programs offer two outcome measures that are already being used by a growing number of programs. By themselves, these measures would clearly not be enough to capture the performance or impact of a program. Combined with indicators of student achievement, classroom teaching, and persistence in the profession, however, the feedback of graduates and those who hire them offers a comprehensive picture. APA's 2013 task force on teacher preparation program improvement and accountability also cited the potential utility of surveys: "Given their established utility with in-service teachers, surveys can be very useful as a program evaluation tool with former teacher candidates within a year of graduation and several years after graduation".[33]

Where these surveys are used, graduates are contacted to find out how well their program prepared them to teach, and some programs solicit similar feedback from principals or other district-based employers of their graduates. Many who talk with schools or school district about teacher hiring hear anecdotes about the graduates of various programs. Some report that a particular provider's graduates are so good in the classroom that they would hire every one of them. Other HR offices or principals are less positive, saying they would never hire someone from such-and-such a program. Districts and schools act on these feelings, but they do not constitute systematic feedback about program or teacher quality.

[33] APA task force, p. 29.

Surveys and their response rates must meet standards of quality to yield reliable results. In addition to survey quality and adequate response rates, few programs have the ability (or the will) to track their graduates into employment. This is another area where better state data systems—and cross-state collaboration—would be beneficial. Besides the efforts of individual programs to survey graduates and their employers, there are multi-program or statewide feedback surveys that can be tapped as models. Since 1998, the North Carolina State Board of Education has produced an annual IHE Performance Report on program graduates and employer assessment of all state and private teacher preparation programs, with results made availably to the public on-line at: http://www.ncpublicschools.org/ihe/reports/. In New York, the Pathways Project implemented follow-up surveys of preparation program graduates and of first- and second-year teachers who had completed programs in the Pathways research initiative (see http://tinyurl.com/ybgufex). The Pathways survey findings has contributed rich contextual information about program features, the organization of clinical practice in a variety of preparation programs, and the extent to which preparation of teachers was "coherent" in ways that strengthened the capacity of program graduates to be successful teachers.[34] In Chicago the Consortium on Chicago School Research conducted surveys of Chicago Public School (CPS) teachers prepared by multiple programs in the area (http://tinyurl.com/yeabgel. These surveys were not envisioned as ends in themselves, but as useful sources of information to support research and program improvement.

A reliable set of outcomes measures that include survey findings requires data systems that allow all programs to locate their graduates in the districts and schools where they are employed as teachers. It is certainly more feasible for states to collect and disseminate than for about 2000 individual programs to develop their own surveys and go off in search of employment data. Moreover, survey quality and response rates must be high enough to allow programs, states, and accreditors to be confident about inferences made from the responses. For feedback measures to be useful to programs, employers, and others the surveys ought to be conducted annually or no less frequently than every other year. Longer intervals between surveys mean that findings will be "stale" as an indicator of program performance and as a program improvement tool.

K-12 student perceptions of their teachers

[34] Research reports and published studies using this information can be accessed at http://tinyurl.com/5w9ak7.

Student surveys as an indicator of teaching quality provide another way to measure program performance, through its impact on K-12 schools. The Measures of Effective Teaching (MET) project reported in 2010 that student perceptions about instruction were related to teaching effectiveness.[35] For example, MET reported that "student perceptions of a given teacher's strengths and weaknesses are consistent across the groups of students they teach. Moreover, students seem to know effective teaching when they experience it: student perceptions in one class are related to the achievement gains in other classes taught by the same teacher."[36] MET reports that the strongest student perceptions as explanations for learning outcomes are a "teacher's ability to control a classroom and to challenge students with rigorous work." School administrators concerned about the classroom management skills of new teachers as well as parents worried that too many teachers have low expectations for their children would understand the meaning of these findings.

MET argues that student perceptions are an "inexpensive way" to construct a teaching quality indicator that can supplement other measures. Of course, the quality of this indicator depends on the instrument used to capture student attitudes. MET employed a survey developed by Ronald Ferguson and his colleagues at the Tripod Project for School Improvement. There are seven dimensions to this instrument: Care, Control, Clarify (teacher explanations, student understanding), Challenge, Captivate (student interest), Confer (teacher questioning), and Consolidate (teacher feedback). A sample item shows the flavor of the survey: "In this class, the teacher expects nothing less than our full effort." This MET report found statistically significant relationships between some Tripod student responses and teacher value added scores in ELA and mathematics.[37]

Here again, the 2013 APA task force had insights into the quality and use of student surveys in connection with preparation program improvement and accountability:

> Student surveys of teacher effectiveness have considerable support in the empirical literature. Scores of constructs based on observable behaviors are internally consistent and stable, are related to achievement outcomes in both college and K-12 students, are more highly correlated with student achievement than are teacher self-ratings and ratings by principals, and distinguish between more and less effective teachers identified using other metrics. Moreover, student surveys can be particularly

[35] "Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project." (Seattle: Bill and Melinda Gates Foundation, 2010), p.7.
[36] Ibid., p. 9.
[37] Ibid., p. 25-27.

useful in formative evaluation contexts because the scores can isolate areas in which teachers need to improve[38].

Data quality and access issues

Implementing a student perceptions survey as an indicator of program quality will require an instrument that meets standards of rigor. Programs may use locally developed instruments for internal purposes, but an approved tool with known properties will be required for national reporting. Obtaining survey results will require the cooperation of schools and districts, and there are precedents for this. New York City and the Chicago Public Schools are among the districts that already conduct student surveys on a regular basis. The MET project had the cooperation of six school districts: Charlotte-Mecklenburg, Dallas ISD, Denver, Hillsborough County (Tampa), Memphis, and New York City.

Distributing, collecting, and analyzing student surveys would be a large logistical task. State data systems could be used to aggregate the data from different schools and link findings to the graduates of specific preparation programs, just as they will have to do for other outcomes measures. The state systems or consortia like the Texas-based CREATE could perform these tasks as well as managing a reporting platform for public dissemination of findings[39].

Program Completers: Impact on K-12 Students

To many, the most important preparation program outcome is teacher effectiveness—defined as the extent to which program graduates help their K-12 students to learn. Since high quality instruction is the main in-school driver for student achievement, it makes sense that teacher effectiveness measures ought to be a central outcome. Today, however, only a few states have elevated teacher effectiveness as a core expectation or outcome for preparation programs. Louisiana uses value-added analyses of student academic performance to make decisions about the quality of every public or private "traditional" or other pathway into teaching.[40] A few years ago, Florida began measuring and ranking its teacher education programs according to the

[38] APA task force, p. 27.
[39] CREATE is a unique Texas-based organization that works with university-based teacher preparation programs across the state. See http://www.createtx.org/content.php?p=6.
[40] For more information on Louisiana's system as well as the policies and research behind its development, see State of Louisiana, Board of Regents, "Teacher Education Initiatives," available at http://tinyurl.com/27y5fzg.

learning gains demonstrated by K-12 students taught by program graduates.[41] And Texas has announced a program accountability policy that, like Florida and Louisiana, includes program graduate impact on K-12 learning as a core indicator.[42] Tennessee and North Carolina have published studies linking prep programs to student achievement results but neither state uses the information for accountability or program improvement.[43]

Louisiana has had the longest track record as a state in using teaching effectiveness as a required preparation program outcome. It is still unclear how Florida and Texas will implement their policy focus on this outcome, and the work of the Race to the Top states (including Florida) with student achievement as a program outcome has not yet produced any publicly accessible reports of program performance.[44] Aside from preparation program outcomes, however, many more states are building or implementing teacher evaluation systems in which student achievement has a central role. These evaluation policies and practices require sophisticated district-level data systems but they also can tap state-level data systems that are fed from the districts. Indiana, Michigan, Missouri, Ohio, and Washington State are among the states with some experience at the state level linking teachers with their pupils to calculate "value-added" results. To date, none of these states has published any results from this work but researchers have been able to use the data for studies of preparation program or program graduate effectiveness[45].

States have relatively little experience with implementation of teacher effectiveness as a preparation program outcome, but at least 20 states have taken steps in this direction (19 Race to the Top states, including Louisiana in Round 3, plus Texas). Whether or not program faculty and administrators share this state goal, analyses and judgments will be made about programs in these states based on their performance on this indicator. This poses opportunities as well as

[41] More on Florida's efforts can be found at http://tinyurl.com/yjwd8md.
[42] Details on the Texas approach come from Senate Bill 174 and Chapter 229 of the Texas Administrative Code, both adopted in 2009. See http://tinyurl.com/yz9jmfg.
[43] By this we mean that neither state's education agency has yet found a way to incorporate the results of these analyses into their program approval processes or into decisions about which programs should be authorized to remain open.
[44] Tennessee began producing its annual performance reports and posting them on a website before receiving its Race to the Top grant. It's not clear whether the state will actually do anything with this information.
[45] For examples of this work, see the state partners and research papers available through the Center for Analysis of Longitudinal Data in Education Research (CALDER). State partners are at http://www.caldercenter.org/, and a sample of papers about teacher effectiveness can be accessed at http://www.caldercenter.org/publications/publications-teachers-and-principals.cfm.

challenges: improved state data systems are needed to link teacher and student data; effective confidentiality and privacy policies are crucial; and analysis of K-12 testing data must be careful to use appropriate statistical models.[46]

There is also a robust literature on the use of value-added measures. Studies address methodological challenges associated with estimating teacher effectiveness, appropriate use of findings for accountability and program improvement, limitations of this approach to measuring teaching quality, and strategies for improving VAM research and reports.[47]

Many preparation program graduates in these states and across the country teach grades and subject areas that are not tested by the states; one estimate is that about two-thirds of teachers fall into this category. A major challenge, therefore, is to develop learning outcomes for students of teachers in these untested subjects and grades. CAEP and others interested in this problem can tap work underway by Race to the Top states that face the same problem and are trying to address it.

With respect to data systems needed to collect and analyze teacher effectiveness information, most states can link student and teacher data in their K-12 system, but they are not able to tie employed classroom teachers back to their in-state preparation programs. This will need to be worked out for accreditation and accountability, and it's also needed for programs themselves to acquire, use, and report information on the teacher effectiveness of their graduates.

Despite the challenges, value-added analyses and growth model calculations of student learning are becoming more common as states and districts work out ways of measuring student outcomes in order to improve them. Expanded use of these analytical strategies has stimulated efforts to improve the student tests that function as dependent variables, and it seems safe to say that the nation will see further work to refine the analytical methods used to determine the impact of teachers on the academic achievement of their pupils.

Meeting state and district needs for teachers

Production of new teachers in high demand fields is a program outcome also highly relevant to the needs and interests of schools and their students. Florida and New York include production

[46] Goe et al., "Approaches to Evaluating Teacher Effectiveness: A Research Synthesis."
[47] Good sources that summarize the issues and challenges include: Goldhaber's recent paper for the Carnegie Knowledge Network (http://tinyurl.com/lj3qt2g), and numerous studies or reports from the National Center for Longitudinal Data in Education Research (http://www.caldercenter.org/publications.cfm).

of teachers in high-need fields as an explicit focus of Race to the Top. Employment as an outcome measure is part of the Race to the Top strategy for Florida, Massachusetts, New York, Ohio, Rhode Island, and Tennessee.[48] It's important to note here that Massachusetts, New York, and Rhode Island will use these production and employment numbers as part of beefed-up accountability systems. The other states simply report on them.

As measurable program outcomes, production and employment outcomes require comprehensive state-level data about program graduates. The state data systems needed for measuring teacher effectiveness as a program outcome—linking K-12 students, their teachers and schools to the programs producing these teachers—would also be necessary to capture information on the production of new teachers in demand fields such as STEM subjects, special education, and ESL.[49]

Program completion, teacher retention and employment

Two outcomes related to the impact of preparation programs on K-12 schools are: how long graduates persist in teaching and where they are employed as teachers.[50] Similarly it is reasonable to track program completion rates to gauge the proportion of entering teacher candidates who complete their course of study and obtain certification to be a classroom teacher. It also makes sense to disaggregate these program completer statistics by gender, ethnicity, and subject area. Obtaining accurate reports of program graduates who enter teaching is very difficult. Few programs follow their graduates once a degree has been awarded or certification is recommended to the state. In some states a significant proportion of preparation program graduates seek and find employment in other states.

[48] For Florida, this means the production of new teachers in science, mathematics and other STEM subject employed in difficult to staff subjects and schools; New York targets—but doesn't define—"shortage" subject areas. And all the states with program graduate employment indicators focus their attention on high need schools.

[49] As an example, to address teacher needs in Georgia, the University System of Georgia (USG) created a structure for identifying historical and anticipated teacher needs, by licensure area, in all Georgia districts. This was data that USG institutions were encouraged to reference in considering campus teacher education productivity goals.

[50] Through their Race to the Top work, some states have added an indicator for the subject areas taught by program graduates, hoping to create incentives and pressure on programs to concentrate output in fields like special education, ESL, and STEM, while reducing chronic overproduction in a field like elementary education.

Studies and reports over the last decade have documented the impact of teacher turnover on schools and students.[51] As the Consortium on Chicago School Research noted in 2009, "High turnover rates produce a range of organizational problems for schools…thwart efforts to develop a professional learning community among teachers and make it difficult to develop sustained partnerships with the local community."[52]

It has been widely reported that teacher turnover is a serious problem in low-achieving schools that have high proportions of poor and minority students. Teacher effectiveness studies show, however, that positive teacher impact on student achievement grows as teachers gain experience (up to a point), which mean that teacher turnover thwarts student academic performance. Research also indicates that preparation matters when it comes to teacher effectiveness.[53] It is particularly important where candidates obtain their clinical experience during preparation, and it matters how a program's clinical component is organized and supported by faculty so that graduates become effective teachers.[54]

And yet high rates of teacher turnover persist despite the claims of many teacher preparation programs that their graduates are specifically prepared for challenging schools.[55] K-12 schools are already held accountable for the consequences of teacher turnover: high rates of turnover lead to weaker student academic gains than would otherwise occur. Preparation programs are not solely responsible for turnover or for its solution, but given the causes and consequences of

---

[51] Studies and reports on teacher turnover include work by NCTAF in *No Dream Denied* and their 2007 study of teacher turnover in five school districts (see http://tinyurl.com/22wasx), work by Smith and Ingersoll (2004), and the study of turnover in Illinois by White et al. (2008). More recently, the Consortium on Chicago School Research provided a very detailed analysis of teacher turnover and its impact of particular schools and students. See Elaine Allensworth, Stephen Ponisciak, and Christopher Mazzeo, "The Schools Teachers Leave: Teacher Mobility in Chicago Public Schools" (Chicago: Consortium on Chicago School Research, University of Chicago, 2009).

[52] Allensworth et al., "The Schools Teachers Leave."

[53] Boyd et al., "Teacher Preparation and Student Achievement"; Harris and Sass, "Teacher Training, Teacher Quality and Student Achievement;" the essays in Dan Goldhaber and Jane Hannaway, eds., "Creating a New Teaching Profession". (Washington, DC: The Urban Institute Press, 2010); and

[54] Pam Grossman et al., "Constructing Coherence: Structural Predictors of Perceptions of Coherence in NYC Teacher Education Programs." Journal of Teacher Education (2008) 59**;** Donald Boyd, Pamela Grossman, Hamilton Lankford, Susanna Loeb, and James Wyckoff*,* "Teacher Preparation and Student Achievement." Educational Evaluation and Policy Analysis, (2009) 31 (4); and Matthew Ronfeldt, "Where Should Student Teachers Learn to Teach? Effects of Field Placement School Characteristics on Teacher Retention and Effectiveness." *Education Evaluation and Policy Analysis*, March 2012, 34 (1).

[55] Many programs don't know very much about whether their graduates become teachers or how long they stay in the profession. And few know whether their graduates teach in the kinds of schools the program believes it has trained them for.

teacher turnover, persistence in teaching is a program outcome that can help to align the interests of producers and employers.[56]

Why should persistence rates matter as a program outcome? How can preparation programs address teacher persistence rates? At least five states are working through Race to the Top on teacher persistence as a preparation program indicator. CAEP has argued strongly for the "clinical residency" model of teacher preparation, for programs "that are fully grounded in clinical practice and interwoven with academic content and professional courses."[57] Programs that take (or have taken) significant steps to implement a well-designed clinical residency model are likely to produce graduates whose experiences in a really rigorous clinical approach to preparation will provide them with the knowledge, skills, and teaching experience to survive school environments that are less than ideal. Better teacher preparation along these lines, plus improved school working conditions are probably the keys to teacher retention.

Some programs do track the persistence rates of their own graduates. But a reliable strategy to acquire data on persistence as a program outcome requires data systems that enable all programs to locate their graduates in the schools and districts where they teach. Thanks to the federally funded State Longitudinal Data System (SLDS) initiative, such systems are becoming more common in the states.[58] Data system availability and functionality, however, doesn't mean that states or programs actually track their graduates and analyze persistence rates. Making persistence rates a strong operational outcomes indicator will require programs and states to work together to gather and share the data.

It is worth saying again here that the use of persistence rates as a program outcome does not mean that preparation programs are solely responsible for teacher turnover. But turnover rates will not improve until producers and employers have incentives to focus on the problem. It seems likely that public confidence in teacher education will be improved when programs take public ownership of this issue.

DATA SYSTEM LANGUAGE

[56] See the discussion in Gary T. Henry, Fortner, C. Kevin, and Bastian, Kevin C. (2012) "The Effects of Experience and Attrition for Novice High School Science and Mathematics Teachers" Science. 335, 1118-21.

[57] "Transforming Teacher Education Through Clinical Practice: A National Strategy to Prepare Effective Teachers." (Washington: National Council for Accreditation of Teacher Education, 2010), p. ii.

[58] These systems are becoming increasingly common in the states, as discussed below and in Appendix A.

Data collection systems useful for capturing information about outcomes and available for sophisticated analyses can be developed and tapped for program assessment, policy analysis, and continuous improvement. This kind of system can also help to build an evidence base for what works in teacher preparation. For all this to occur, however, a robust data collection system must be in place (such as those that Race to the Top states are building or adapting) to generate mainly aggregate measures of preparation program outcomes from individual-level data, or from datasets with links between files containing information about students, teachers, schools, and preparation programs. Data elements, data collection protocols, and management of the system(s) by multiple parties[59] have to be configured to produce accurate data.

To understand measures of program quality as well as outcomes-focused teacher preparation, these are the data system linkages that matter most:

- School link to teachers
- School link to pupils
- Classroom-level data: classes, teachers, and pupils
- Pupil individual identifier
- Pupil demographics
- Pupil test data
- Pupil link to teachers
- Pupil link to classes
- State certification data for teachers
- Teacher employment records
- Teacher individual identifiers linked to schools, pupils, and EPP candidate identifiers (such as university IDs or SSNs)

As just one example, information on individual schools and employed teachers is necessary to calculate persistence rates in teaching for program graduates. The National Commission on Teaching and America's Future (NCTAF) described three types of teacher turnover.[60] It is not easy to determine whether a specific individual teacher has left teaching entirely, but data about

[59] Such as universities and university systems, state agencies, schools and school districts, and federal government (IPEDS, Core of Common Data, other NCES resources).

[60] NCTAF defined teacher turnover in terms of: ***Leavers,*** or teachers employed in a classroom-teaching role in a school in Year 1 and not employed as classroom teachers *in any district* in Year 2; ***Within-District Movers,*** teachers employed in a classroom teaching role in a school in Year 1 who are employed as classroom teachers at a different school *in the same district* in Year 2, or "cross-school, within-district movers;" and ***Cross-District Movers,*** who are teachers employed in a classroom teaching role in a school in Year 1 who are employed as classroom teachers at a different school *and in a different district* in Year 2.

teacher employment at school and district levels are needed to calculate and report the most widely used measures of persistence and turnover.

Given the wide range of information needed on teachers, students, and schools, a system that meets these conditions will probably be a compatible set of independent databases maintained by different parties and linked through common identifiers. Examples already exist, such as the one developed through North Carolina's Education Research Data Center (http://www.childandfamilypolicy.duke.edu/project_detail.php?id=74). The Data Quality Campaign's "Essential Elements" and "10 state actions to ensure effective data use" provide an overview of how comprehensive data systems need to work if they are to be useful (see http://www.dataqualitycampaign.org/build/elements and http://www.dataqualitycampaign.org/build/actions).

Given how teacher preparation programs actually work in practice, the best system configuration for teacher education would use interstate data system linkages to cope with mobility of teacher candidates and program graduates across state lines. As Secretary Duncan has reported to Congress, 20 percent of initial teaching licenses in 32 states were granted to new teachers prepared for the classroom in a different state from the one granting the license to teach. In another 12 states and the District of Columbia, programs in other states prepared 40% of initially certified teachers.[61]

The optimal system—a comprehensive data system at the national level—is highly unlikely ever to be available. Efforts to construct such a system in the last decade were blocked in Congress but there is renewed support for a unit-record data system that would have comparable data from all states. Within specific states, universities, or school systems, missing pieces include large chunks of relevant data, ability to link datasets with common identifiers, barriers constructed at every level in the name of "privacy", and technical problems with hardware, software, or staffing capacity. Even so, individual states can develop and implement high quality longitudinal data systems, and the states can work together to find ways of sharing data in compatible formats so that graduates from programs in one state can be located in data systems of others states where they are employed as teachers[62].

---

[61] See U.S. Department of Education, "The Secretary's Sixth Annual Report on Teacher Quality," page 23. Retrieved from https://title2.ed.gov/secReport11.asp.

[62] Here again, the 2013 Data Quality Campaign report suggests that progress is being made within states. See http://www.dataqualitycampaign.org/.

Gaps in data system components and dataset linkages are gradually being bridged. But they still exist. In spite of these challenges, some information needed for solid answers about preparation program outcomes already exists. Examples include the Pathways Project, AIR's Center for the Analysis of Longitudinal Data for Education Research (CALDER), the Texas CREATE initiative, the California State University system[63], and the data systems behind publications on preparation program effectiveness from states like Louisiana, Florida, North Carolina, Tennessee, and Texas.

[63] While the CSU system has developed and published numerous studies and reports over the past decade or so, there is no evidence that this information has been used to improve programs or close weak programs.