February 2013

Education Innovation Institute
POLICY BRIEF

Bringing education to life.
UNIVERSITY of NORTHERN COLORADO

TEACHER PREP EVALUATION: FINDING THE RIGHT MEASURES, PART I

Highlights

The Ultimate Goal
To provide effective teachers for all students.

The Problems
Critics charge that too many teacher preparation programs (TPPs) lack academic rigor and fail to prepare teachers for a rapidly changing profession. But establishing the causal effect of a specific TPP on teacher quality is hard because the data aren't clean and teachers are not randomly assigned to TPPs or the schools where they teach. Moreover, effective teaching takes many forms and experts do not agree on how best to measure effectiveness.

Symptoms of Problems
Pressure is growing for greater accountability and higher academic standards in TPPs, including in admissions. Many accountability systems being proposed do not deal adequately with the diversity and complexity of the prep universe and are not firmly rooted in research.

Popular Responses
There has been a growing push for data collection – by federal and state agencies, accreditors, nonprofits, and prep institutions themselves – though much of it measures inputs instead of outcomes such as classroom performance.

Possible Unintended Consequences
Reliance on a narrow range of accountability measures, such as standardized tests, could pressure TPPs to narrow their curriculum and deemphasize innovation. It also may discourage research into the influence of harder-to-capture measures on student learning, such as ability to transmit a love of learning. Concerns about the validity of test score growth as a primary measure of effective teaching could undercut faith in TPP evaluation processes and results.

Keep Thinking
The information gathered through accountability systems needs to serve two purposes: to ensure that programs operate at a high level and to provide data to programs that will help them improve. Because the ultimate goals of TPP evaluation are very similar to those for teacher licensure and evaluation, policy makers should align the standards and measures of these three systems. Because evaluations must measure a diverse array of factors affecting teacher preparation and data must come from many sources, experts with a range of specializations should contribute to the design.

Key Issues and Findings from the Research

A growing number of studies are finding that good teachers are key to improved student learning, generally measured through test scores. This conclusion has led to increased scrutiny of teachers along with the programs that train them and widespread calls for more rigorous evaluation of both. In Colorado, two recent laws – SB 10-036 and SB 11-245 –put teacher preparation programs (TPPs) under the microscope, with a focus on assessing the performance of program graduates. The laws further instruct the Colorado Commission on Higher Education to recommend a new TPP evaluation system by December 2013.[1, 2] Carrying out the bill's charge will not be a cut and dried undertaking. Identifying data to accurately measure the effectiveness of the widely varying TPPs in Colorado while also accounting for the many factors that affect instructional quality in K-12 classrooms is enormously complicated. As Wilson, Floden and Ferrini-Mundy[3] note in their extensive review of research on teacher prep, "While no one argues for teachers who are less qualified, there are serious disagreements about what it means to be well qualified and what it takes to prepare teachers well" (p. 1).

Even if policy makers, practitioners, and researchers agree that K-12 student performance is the correct outcome measure for evaluating TPPs, mapping the range of differences in teacher effectiveness to systematic differences in TPPs must consider a large array of intervening factors. And even if evaluations reveal trends and correlations between student learning and TPPs, causality is extremely hard to prove because teachers are not randomly assigned to TPPs or the schools where they teach. Moreover, demonstrating causality requires a chain of evidence linking TPP training to graduates' teaching practices to student learning, which can be very hard to establish.[4,5,6] These obstacles mean anyone using the evaluation findings must assess how well the analytical methods control for variables that can distort the results. Only when such precautions are sufficient can analysts hope to untangle contextual questions such as how and why certain courses or field experiences produce superior K-12 teaching and whether results vary by grade level and student demographics. To be most useful, evaluations must assess the success of a TPP's graduates while also providing enough detailed feedback to help the program build on its strengths and improve weaknesses.[7]

Colorado's SB10-036 directs the state to report on the effectiveness of TPPs for teachers in their first three years including, at a minimum, correlations between TPPs and student academic growth and teacher placement, mobility and retention. The emphasis on student academic growth echoes another recent law, SB10-191, which mandates creation of a new teacher evaluation system that bases half the rating on student growth, using test scores and other measures, and the rest on classroom observations. Student growth is measured through the Colorado Growth Model, which compares each student to peers on rate of growth over time and progress toward meeting state standards. Several other states calculate growth using statistical models called Value-Added Measurement (VAM), which use a student's year-to-year growth to assess how much value a teacher has added in one year. A few states, including Louisiana and Tennessee, already use versions of VAM to assess TPPs and several others have or are building the data capacity to do so.[8] Research and reports from those states reveal how hard it is to design models that are complete, valid, fair, and workable for the educators using them.

*An **effective teacher** under Colorado's new evaluation system is one who meets state standards for student growth and professional practice, demonstrating "the knowledge, skills, and commitments that ensure equitable learning opportunities and growth for all students."*

The public demand for high-quality teachers makes it crucial to find valid and reliable TPP evaluation measures. This two-part set of briefs does not provide a roadmap for evaluating TPPs. Rather, they explore the challenges in evaluating an enterprise as diffuse, multi-faceted, and heterogeneous as teacher preparation. Part I describes recent legislation in Colorado, discusses why the large number of TPPs and differ-

ences among programs make it hard to compare their effectiveness, explains why establishing causality is hard, and considers the challenges in data capacity. Part II discusses the limitations and growing body of findings of the best recent research, examines nascent evaluation efforts in other parts of the country, and makes recommendations for policy at all levels. Both parts seek to stimulate critical thinking about the core mission of TPP evaluation and the possible unintended consequences of placing too much emphasis on program components that are easy to measure and too little on elements that are hard to measure but important for producing successful teachers. The ultimate goal is to help policy makers make informed decisions about the likelihood a given measure will lead to improved student performance while also giving TPPs information and tools to improve. Whenever possible these briefs base assumptions, theories, and recommendations on research findings, and will be clear about where rigorous relevant research does not exist.

Both parts of this brief devote considerable space to discussions of TPP research because 1) TPP evaluations are a form of research and 2) research findings should inform the selection of evaluation measures in order to reduce the influence of politics, business, and popular trends as much as possible. Both parts also explain the importance of establishing causality and the difficulty of doing so. A key challenge in any TPP research is that isolating the elements of prep programs that lead to improved student learning requires dealing with layers of outcomes. As Cochran-Smith & Zeichner[5] wrote:

> *"This kind of research depends on a chain of causal evidence with several critical links: empirical evidence demonstrating the link between teacher preparation programs or structures and teacher candidates' learning (i.e., candidates' knowledge growth, skills and dispositions); empirical evidence demonstrating the link between teacher candidates' learning and their practices in actual classrooms; and empirical evidence demonstrating the link between the practices of graduates of teacher preparation programs and what their pupils learn. Individually each one of these links is complex and challenging to estimate. When they are combined, the challenges are multiplied" (p. 3).*

Scholars have further warned that no unified theory exists to suggest how these individual elements and layers interact to explain how and why certain experiences in a teacher's training lead to certain outcomes by her students several years later.[4]

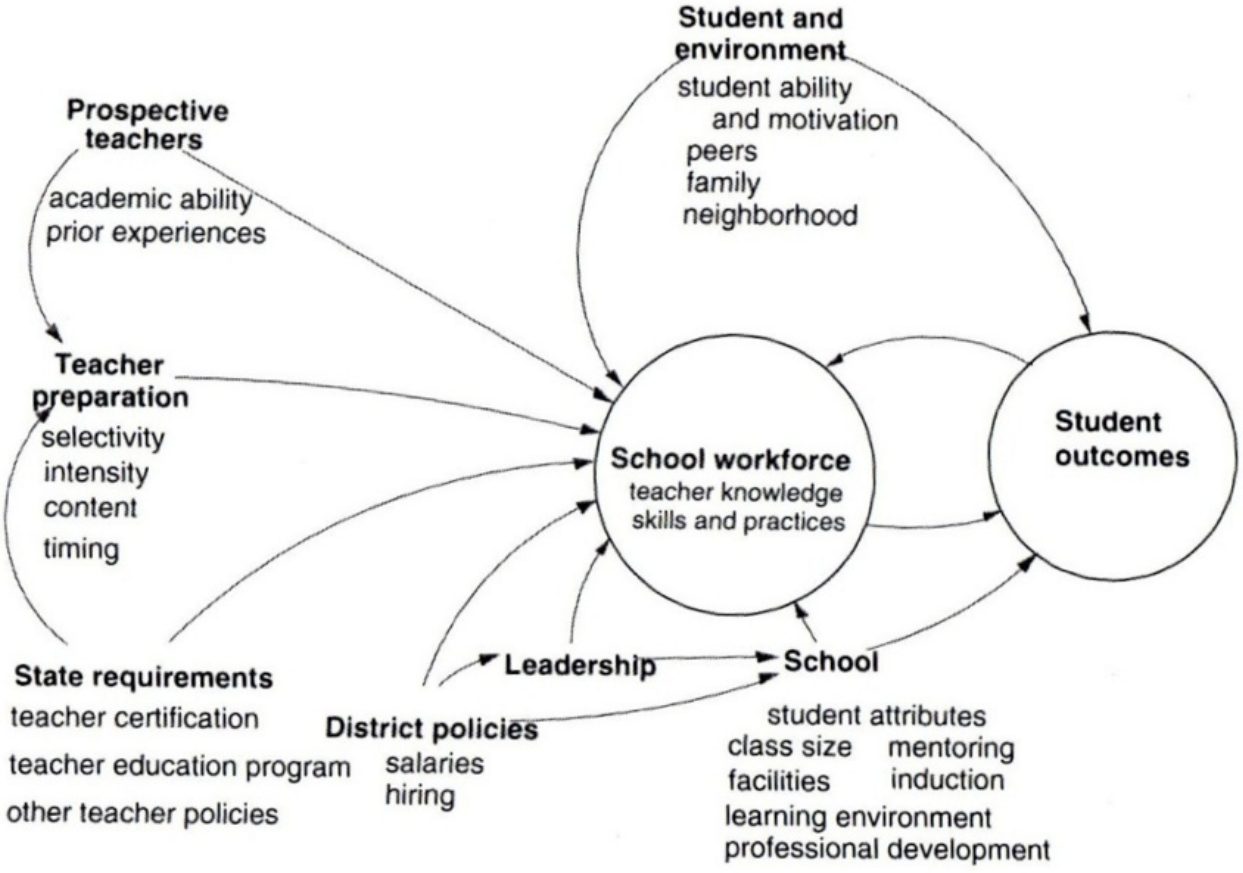
Selection bias is a major impediment to establishing causality. Teacher candidates choose TPPs for specific reasons. They later select – or are selected by – K-12 schools where they teach for a different set of reasons.[4] Further, many TPPs send a majority of new graduates to only a few districts that are either nearby or have characteristics emphasized by the TPPs, such as rural or inner-city locations.[9] In addition, teacher prep candidates bring attitudes, expectations, and dispositions about education that may persist even if the TPP stresses different priorities and values.[5] These layers of choices and influences – as opposed to random assignment – make it hard to isolate the effects of the TPP.

Only a replicable experimental design using random assignment can establish causality beyond dispute. That's because random assignment of adequate numbers of teacher candidates helps ensure that the differences among candidates in various TPPs occur by chance and do not drive their enrollment in chosen programs. Obviously, random assignment is impractical in most settings. While statistical techniques can control for many confounding variables, they cannot control for characteristics that are hard to observe and quantify, such as degree of motivation and personal attitudes, which may influence a teacher candidate's success. A 2010 National Research Council book[4] on the relationship between teacher prep research and policy notes the difficulty in establishing a chain of causality: "Disentangling the role that teachers play in influencing student outcomes is difficult, and understanding the ways in which teacher education influences student outcomes is much more difficult" (p 22). The book also notes the practical and logistical barriers to using random assignment in education.

Factors like selection bias and influence of unobserved variables make it hard to prove a TPP's impact on K-12 student outcomes.

Identifying the causal impact of TPPs is made even more difficult by the great amount of variation among them. There are more than 2,000 TPPs in the U.S., preparing about 235,000 completers each year. Seventy-one percent are traditional programs based at institutions of higher education (IHEs) and another 21 percent are alternative programs based at IHEs. Only 8 percent are alternative programs not based at a college or university.[10] Within those three broad program structures there is enormous diversity. Programs vary in size from a few students to thousands. Their admissions standards, average age of students, curricular offerings, graduation requirements, pedagogical frameworks, guiding philosophies (i.e. social justice, emphasis on rural schools, etc.), instructional methodologies, field experiences, types of districts where graduates are placed, and expectations for faculty research all vary as well.[9, 10] In addition, TPPs allow for specialization in a large array of content areas, at multiple grade levels, and for different types of students (i.e. special education, gifted, Limited English Proficient, etc.). Field experiences and student teaching are similarly diverse in areas ranging from their duration to processes for selecting supervising teachers to terms with participating districts to the degree to which field work is linked to content and methods courses. While this diversity of programs is important to a robust system of teacher preparation, it also makes comparisons among them challenging.

Isolating how much teacher's effectiveness arises from training or specific school features is also complicated. Features like induction programs, professional development, leadership quality, adequacy of funding, curricular and instructional requirements, and characteristics and needs of students all influence existing measures of teacher effectiveness.[4, 5, 6] (See Figure 1 from the 2010 National Research Council Book.)



Source: NRC, 2010, p. 23.[4] Adapted from Boyd, D., Grossman, P., Lankford, H., Loeb, S., Michelli, N., and Wyckoff, J. (2006). Complex by design: Investigating "Pathways into Teaching in New York City Schools," Journal of Teacher Education, 57, 155-166.

The Challenge of Comparing Across Diverse Programs

Consider the challenge of finding evaluation measures to isolate the prep program effects for two hypothetical teachers. One is the first in his family to attend college and chose a TPP at an urban university that draws applicants mostly from local populations. The majority of teachers this institution produces work at local inner-city schools where most students have educational disadvantages. The second is from a middle-income family, attended a TPP at a flagship state university, and now teaches at a suburban school serving students from backgrounds similar to hers. These examples illustrate the pitfalls of basing program evaluations heavily on limited outcomes like student test scores or designing them primarily for a certain type of TPP at, say, elite universities. Easy-to-quantify measures like student test scores don't begin to capture the degree to which teachers were prepared to succeed in very different school environments. Evaluation models with inadequate flexibility and controls could end up penalizing TPPs that send graduates to teach in challenging schools.

Some advocates of greater accountability for TPPs focus on two key gatekeepers into the teacher pipeline: entrance standards and licensure requirements. Selectivity is a sticky issue when designing evaluation measures because it can distort outcome measures. Without adequate controls it is difficult to tell whether the high performance of a highly selective program's graduates is due to the training they received or to their academic ability. One way to deal with this is to compare programs only within bands of selectivity criteria. One problem with increasing selectivity sharply is that admissions criteria based on measures like SAT scores and GPA are not strong predictors of teaching effectiveness. Disadvantaged students are also much less likely to excel on these measures, and excluding such students from teaching programs becomes a policy issue if states are also trying to increase the diversity of their teaching forces. Similar dynamics apply to licensure requirements that rely heavily on tests of content knowledge. Poor performance can disqualify people who have skills not measured by the tests but that are more strongly associated with effective teaching. Some researchers recommend allowing principals to consider the test score as an important signal to be weighed against other observed strengths.[11]

> *"…in the study of teacher effects, it is what teachers do in the classroom and the knowledge they bring to bear that produces student learning gains, not simply what and how much they know, the attitudes they hold, or what they learned in a methods class (although those things also are important). It is how they use that knowledge and how deftly they enact the practices they were taught that affects how high their students achieve" (Goe & Coggshall, 2007, p. 4).*[6]

Data and Data Capacity

One complexity in creating a TPP evaluation system that uses student test score growth is the need for a well-designed system for collecting and maintaining longitudinal data that links 1) K-12 teachers to their students, schools, and TPPs and 2) links data from K-12 to higher education and the workforce. Colorado's experience illustrates the importance of coordinating the timing between building this data capacity and requiring sophisticated analyses. The Legislature required the creation of individual identification numbers for teachers just a year before passing SB 10-036 requiring TPP evaluations and SB 10-191 establishing a new teacher evaluation system. Both laws demand an accurate link between students and the teachers accountable for their achievement growth. While a teacher identifier is necessary for establishing this link, the additional challenge of establishing consistent statewide practices around student-teacher links should not be underestimated.

Building the longitudinal data infrastructure is a long, time-consuming process. Dozens of difficult decisions must be made, such as how to establish a teacher of record for a student who moves frequently or when a team of teachers has joint responsibility for student learning.[8] When evaluating student growth, how many years of data are needed to accurately and fairly rate a teacher's performance? What measures will ensure

a student growth model that doesn't penalize teachers whose students struggle academically for reasons outside their control? A recent project by researchers in Colorado illustrates the difficulty of even the early task of assigning each teacher to the TPP that recommended him for licensure. The researchers had to create a new "completer" variable by matching human resources data from the Colorado Department of Education (CDE) with TPP data reported to the Department of Higher Education, and then asking each TPP to verify the matches.[12] Complications included incomplete data collection and/or retention by universities, teacher candidates who attended multiple TPPs, and changes in universities' data management systems that made data for some years inaccessible. The labor required to carry out these basic steps is just one example the challenges states face in putting together full TPP evaluation systems with multiple measures.

About the Education Innovation Institute

The Education Innovation Institute, created in 2009 by the Colorado General Assembly, identifies and interprets the nation's best research on current education issues to help shape policy and reform. It is housed at the University of Northern Colorado, a leader in teacher education since 1889. For more information about EII and its work, visit www.unco.edu/eii.

Questions for Thought

- *Can a single formula or set of measures accurately and reliably assess the quality of all teacher preparation programs (TPPs) given their complex mission to educate teachers to succeed with students from diverse backgrounds in schools that vary greatly?*
- *Should TPPs generate their own evaluation measures, subject to approval by the state?*
- *Should measures for each program fit within a set of baseline fundamental standards required of all TPPs to provide consistency across the country? Who should develop and oversee implementation of these standards? Federal or state governments? Accreditors? Some other independent organization?*
- *Should TPP evaluations include outcome measures, such as performance in student teaching, low turnover among graduates, and satisfaction of labor-market goals, including placement in hard-to-staff schools and fields?*
- *If evaluations grade TPPs on selectivity, should they consider only academic measures, such as GPA, or should they also include harder-to-measure attributes such as communication skills, a desire to work with underprivileged students, and commitment to teaching?*
- *How accurately can evaluations measure the full range of imparted skills if they focus on the first few years, when teachers are least confident and capable but the direct effects of training are easiest to isolate?*
- *Should TPPs be evaluated only on training for initial licensure or should other training missions, such as professional development and specialized endorsements, be included?*
- *In what level of detail should TPP evaluation results be made public?*

References

NOTE: The citations below are hyperlinked to their source. Simply click on the blue citation title.

1. Colorado Senate Bill 10-036 (2010) Concerning requiring the Department of Education to prepare an annual report analyzing educator preparation program effectiveness using data obtained through the educator identifier system.

2. Colorado Senate Bill 11-245 (2011) Concerning educator preparation programs at institutions of higher education.

3. Wilson, S., Floden, R., & Ferrini-Mundy, J. (2001). Teacher Preparation Research: Current Knowledge, Gaps, and Recommendations. Washington, DC: Center for the Study of Teaching and Policy.

4. National Research Council (2010) Preparing Teachers: Building Evidence for Sound Policy. Committee on the Study of Teacher Preparation Programs in the United States, Center for Education. Washington, DC: The National Academies Press.

5. Cochran-Smith, M. and Zeichner, K.M. (2005) Studying Teacher Education: The Report of the AERA Panel on Research and Teacher Education. Mahwah, N.J.: Lawrence Erlbaum Associates.

6. Goe, L. and Coggshall, J. (2007) Teacher Preparation > Teacher Practice > Student Outcomes Relationship in Special Education: Missing Links and New Connections. National Comprehensive Center for Teacher Quality.

7. Plecki, M.L., Elfers, A.M., and Nakamura, Y. (2012) Using Evidence for Teacher Education Program Improvement and Accountability: An Illustrative Case of the Role of Value-Added Measures. Journal of Teacher Education, XX(X) 1–17.

8. Data Quality Campaign (2011) Hot Topic: Improving Teacher Preparation.

9. University of Colorado-Boulder, University of Colorado-Denver, & University of Northern Colorado (2012a) An Initial Exploration of Colorado-Trained Teachers: Providing Context for Outcomes-Based Teacher Preparation Program Evaluation.

10. U.S. Department of Education (2011a) Preparing and Credentialing the Nation's Teachers: The Secretary's Eighth Report on Teacher Quality Based on Data Provided for 2008, 2009 and 2010.

11. Goldhaber (2006) Everyone's Doing it, but What Does Teacher Testing Tell Us about Teacher Effectiveness? National Center for the Analysis of Longitudinal Data in Education Research.

12. University of Colorado-Boulder, University of Colorado-Denver, & University of Northern Colorado (2012b) Enhancing Colorado Data Systems: Linking Teachers to Preparation Programs.