February 2013

Education Innovation Institute
POLICY BRIEF

Bringing education to life.
UNIVERSITY of NORTHERN COLORADO

TEACHER PREP EVALUATION: FINDING THE RIGHT MEASURES, PART I 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.

Issues with Prep Program Research

Colorado is in the midst of figuring out implementation of two recent laws – SB 10-036 and SB 11-245 –that require increased evaluation and reporting on teacher preparation programs (TPPs), with a focus on assessing the performance of program graduates. The laws charge the Colorado Commission on Higher Education with recommending a new TPP evaluation system by December 2013.[1, 2] Choosing the right measures of effectiveness is crucial for accurate and fair evaluations, but it is a complicated undertaking. For starters, standards and measures must recognize the staggering array of TPP program structures and curricular requirements and account for the many factors inside and outside the K-12 classroom that affect a teacher's performance and students' achievement.

Experts advocate basing standards and measures on research findings. But rigorous empirical research measuring the effect of TPPs on a teacher's performance and student achievement is in its infancy and thus provides only thin – and often inconclusive – findings on some questions related to performance and little to none on others.[3, 4] Research clearly linking preparation to classroom practices to student learning is rare.[5] These limitations present obstacles for creating evidence-based evaluation measures. In noting the dearth of rigorous research, Wilson, Floden and Ferrini-Mundi[3] made recommendations for improving future investigations, including ensuring that research reports include clear links between programmatic elements and student achievement.

> *"Research on teacher preparation, like other education research, should contribute to our understanding of how to improve student achievement. . . . To help practitioners and policymakers see the contributions of the research, reports should make the connections to student achievement explicit, using measures of teacher knowledge, skill, and practice that are thought important for effective teaching" (p. 33).*

Until recently, research on TPPs focused mostly on inputs, such as course content, requirements for student teaching and graduation, and whether programs are alternative or traditional.).[3, 4, 6] Much of the extant research is not widely applicable because it is based in a single school, district, or state.[3] Some scholars further argue that the advancement of TPP research depends on linking it much more tightly to research on both teaching and higher education to fully incorporate the complexities of teaching and the political, institutional, and labor market contexts that affect program operations.[7]

"Rigorous research" *is a somewhat subjective term. In a 2002 book the National Research Council[8] lists principles to guide education research design:*

- *Pose significant questions that can be investigated empirically*
- *Link research to relevant theory*
- *•Use methods that permit direct investigation of the question*
- *Provide a coherent and explicit chain of reasoning*
- *Replicate and generalize across studies*
- *Disclose research to encourage professional scrutiny and critique*

Among quantitative methods, randomized controlled trials (RCTs) are preferred. When RCTs are not possible, as is often the case in education, quasi-experimental methods are acceptable if they adequately control for selection bias.

One major obstacle to large-scale quantitative research on TPPs has been that states are only now building data systems capable of tracking student and teacher performance over multiple years while also linking K-12 students to teachers and teachers to a range of TPPs.[9] Similarly, education researchers have only recently embraced statistical models capable of dealing with selection issues, and even now their use is limited.

Though still small, the body of rigorous research linking TPPs to student outcomes has grown in recent years with the emergence of statewide longitudinal databases, statistical methods to measure student growth and control for selection issues and student, teacher, and school characteristics, and teacher evaluation tools that use multiple methods and seek to address reliability issues. The most common method is to aggregate student academic

growth scores, such as value-added measures (VAM) and Colorado's growth model, for groups of students taught by each teacher to calculate that teacher's effect on their academic progress. Those calculations of teacher effects are then aggregated for all recent graduates to assess a TPP's effectiveness. The use of such calculations to evaluate teachers and TPPs has stirred intense debate among scholars. Many charge that the knowledge and skills measured by achievement tests don't come close to the full range of teaching and learning outcomes expected in a high-performing classroom, and that such results reveal nothing about why a certain outcome occurred or how programs might be improved. Another charge is that growth models shouldn't drive high-stakes evaluations until questions about their accuracy and validity are satisfied. Despite these limitations, many experts believe that student growth can capture a teacher's effectiveness more accurately than can static measures like a single year of student test scores or inputs like licensure exam scores, prep courses taken, or the length of time spent student teaching.

A challenge in evaluating the effectiveness of TPPs is that variables from multiple sources affect teaching and learning in K-12 classrooms.

Existing rigorous research that is relevant to the requirements of SB 10-036[1] illustrates well several challenges discussed in this brief, including measuring the complexities of teaching and learning. It also provides useful guidance for future research on the strengths and pitfalls of different analytical methods. For example, weak effect sizes in several studies raise red flags about basing judgments of program effectiveness on test score growth. However, while some studies found only small effects for TPPs on student test score growth, those effects were larger than for some other factors, such as certain teacher characteristics, demonstrating the complex interplay of multiple variables on student performance. Our review of the literature underscores the need for continued support of high-quality research and nuanced interpretation (and application) of the results in the interest of eventually accumulating enough evidence for a deeper understanding. It also makes clear the need to adopt provisions for ongoing review and change in evaluation systems as the research base grows. See the example from Missouri below as an example.

This brief describes only selected research from a narrow range, focusing on papers that use rigorous methods and datasets with student-teacher links to measure the effects of TPPs on student outcomes or instructional practices considered particularly conducive to learning. It excludes entire areas of inquiry, including comparisons of traditional and alternative TPPs, and the effects of inputs such as licensure, course-taking, academic performance, and program accreditation on various definitions of teaching quality.

Researchers using VAM models on data from Louisiana and Washington state found differences in the effectiveness of teachers from various TPPs based on student test score growth. The effect sizes were small in both studies, and the Louisiana project found that the magnitude varied across subjects tested. The authors acknowledge that their database is inadequate to analyze features of different TPPs that contribute to the classroom success of graduates, the most valuable information from a program improvement perspective.[9] In the Washington study, the program indicators accounted for less than 1 percent of total variation in student achievement but the authors noted that even that level was greater than the effect for teacher characteristics such as gender and race, and in mathematics it was greater than teacher credentials. The authors also noted that the results did reveal interesting trends, including that while some selective institutions did not produce particularly effective teachers other less-selective schools produced highly effective teachers.[10]

However, a third study using VAM on Missouri data found greater differences in the effects of graduates on K-12 student growth within university-based TPPs than between them.[11] The authors attributed their contradictory findings to the failure of other researchers to make a technical adjustment (to the clustering of standard errors). The authors also note that their findings can have important policy implications. "We conclude, therefore, that TPP rankings based on commonly-used value-added models are, at present, of little value to state departments of education, TPP accreditation agencies, and K-12 school administrators. If it is not made clear to K-12 administrators that the substantive differences that separate TPPs in the rankings are small, the rankings could lead to suboptimal hiring decisions because almost all of the variation in teaching effectiveness occurs within training programs" (p. 7).

A fourth study that looked at VAMs of recent graduates

of all 159 TPPs in Texas noted multiple problems in trying to sort TPPs into reliable categories of effectiveness.[12] For example, in addition to finding that TPPs generally accounted for only a small part of differences in K-12 student growth, the researchers also found that effectiveness varied across circumstances. A TPP with highly effective graduates at one grade level or in one subject could have ineffective graduates in different grades and/or subjects. The authors mused that "value-added evaluation can offer only limited and uncertain feedback to [TPPs], and that feedback to [TPPs] may be a limited and indirect way to improve teacher quality" (abstract).

A mixed-methods study of New York City teachers that linked features of TPPs with teachers' effectiveness illustrates the utility of augmenting administrative data with other information sources.[13] The study linked the teachers to their students' value-added scores on achievement tests and to data about their TPPs, including program documents describing structure and content, interviews with program personnel and field experience supervisors, surveys of faculty teaching methods courses, and surveys of current students and graduates. The findings suggested that a focus on certain program features – greater oversight of student teaching and required capstone project – was identified with greater student achievement for first-year teachers. The analysis finds variation in average value-added scores for participating institutions and finds that, on average, programs linked with higher score gains in math also were associated with higher gains in English Language Arts.

A large Florida study illustrates the difficulty of establishing causality in the absence of random assignment. The authors included a raft of variables and statistical approaches to reduce selection bias to measure the effects of teachers' experience and pre-service and in-service training on value-added scores of their students. Pre-service prep was measured by the number of credits in education and subject content courses, with a variable for college entrance exam scores included to account for pre-existing ability. The analysis of TPP elements yielded few statistically significant results. Only courses in classroom management registered a strong effect on outcomes of one group of teacher candidates – those who later taught high school reading – while statistics credits had a much smaller positive effect on the scores of students taught by middle school mathematics teachers. Similarly, entrance exam scores showed no effect on teachers' later success.[14] A second Florida study illustrates the difficulty of comparing the effectiveness of TPPs using student growth scores because teacher placement in schools is not random and TPPs often send many graduates to the same K-12 schools, as is the case in Colorado. The success of using school fixed effects to control for differences in schools depends on whether teachers from multiple TPPs teach in a given school and the amount of overlap that exists among teachers from different programs across schools.[15]

New research is assessing whether different statistical models produce different results for schools, teachers, and students of various characteristics.

Finally, a growing body of research is digging deeper into how the different statistical details of various student academic growth models can affect a teacher's effectiveness rating and the stability of that rating from year to year. The sometimes-nuanced findings of these studies give policy makers a deeper understanding of the challenges, strengths, and trade-offs of aggregating teacher ratings to evaluate TPPs. Some researchers are comparing different models to see whether they produce similar ratings for the same teachers or whether teachers with certain characteristics are more likely to be rated incorrectly.[16, 17] Others examine the relationship between a teacher's effectiveness rating and factors such as school, student, and teacher characteristics and how many years of score data are used.

One recent study that applied different growth models to the same teachers found that most would receive similar ratings using any of the models – except teachers with high percentages of low-income students with poor test scores, a vulnerable group in need of good teachers.[16] Two 2012 studies examined differences in the rate at which teachers improve their skills over five to 10 years, with one paper finding faster gains among teachers in K-12 schools with a strong professional environment.[18, 19] When some part of a teacher's effectiveness rating is attributable to student characteristic and the school work climate, it's hard to know how much to then attribute to the TPP. As with all research cited in this brief, findings of these studies should be duplicated by additional, preferably causal, studies. But they do indicate the need for policy makers to follow research developments closely and adjust policy when findings indicate the need.

Current Practices and Policies

Current forms of accountability – federal reporting requirements, state approval processes, voluntary accreditation, and teacher licensure – are widely considered inadequate. Even when persistent and serious weaknesses are identified, few programs are forced to close.[4, 6, 20, 21, 22] A plan issued by the U.S. Department of Education[22] calls for overhauling federal reporting requirements to focus on three broad categories of data: achievement growth of K-12 students taught by TPP graduates; job placement and retention rates; surveys of graduates and their principals. In addition, the Race to the Top grant competition rewards states that tie teacher evaluations to student achievement growth, tie that information to the in-state TPPs, and expand TPPs and licensure options that produce effective teachers. Some advocates are further calling for the federal government to step in and create a plan for evaluating and improving teacher prep.

When revising educator licensure, retooling TPP accountability or creating longitudinal databases, states should include information to help TPPs improve.

Many states, including Colorado, are also stepping up state oversight and accountability of TPPs as they build the data capacity to track teachers' performance over time. In 2011, the Data Quality Campaign[23] reported that a majority of states have the data capacity to link teachers to student achievement but many do not have policies and practices in place to ensure that links and data are high quality and reliable. DCQ further reported that only five states automatically shared teachers' performance with their TPPs each year.

Licensure has typically served as a minimal gatekeeper, with states setting criteria that often include little more than TPP completion information, criminal background checks, and tests of basic skills and content knowledge in subjects that candidates seek to teach. A few organizations are working on alternate, performance-based models, however. For example, the American Association of Colleges of Teacher Education (AACTE) and Stanford University formed a partnership to create a new assessment of teacher candidates' readiness for teaching. Working with 25 states, including Colorado, they are creating "a common framework for defining and measuring a set of core teaching skills and performances that form a valid and robust vision of teacher competence."[24] The assessments include two parts: formative assignments embedded in program courses, which can include student teaching; and portfolios of teaching artifacts and commentaries that are subject-specific and can include a range of elements from lesson plans to video clips of instruction to analyses of student learning. Field testing was initiated in spring 2012 with plans for national distribution in the 2012-13 school year. In Colorado, a report from the New Teacher Project[25] recommends basing both initial licensure and renewal on demonstrated performance instead of completion of a TPP or credit hours.

What's Next?

Clearly, designing evaluation systems for a complex enterprise like teacher preparation should be undertaken cautiously. It requires careful analysis and broad-based input from practitioners and researchers who understand the ways children learn, the evolving demands of classroom teaching, and ways to mitigate the effects of noisy data, like test scores. With a goal of creating a data system that links specific prep courses to documented teaching practices to student performance outcomes, designers will need to identify multiple measures for each step.[5] The measures chosen will inevitably influence the future shape of TPPs so should provide information programs can use to improve.

States that are in the early phases of creating longitudinal databases and creating links between students and teachers and K-12 and higher education may want to consider phasing in their TPP evaluation systems as their data capacity matures. Steps could be to establish which TPPs are responsible for each credential a teacher holds; require TPPs to submit detailed information about every candidate's academic and field experiences; establish teachers of record for each K-12 student each year; link new teachers to first jobs, induction programs, and subsequent jobs. Concurrent with that multi-year build-out, states can continue cre-

ating and fine-tuning educator licensure and evaluation systems that are aligned with each other and the prep evaluations and designed to measure both instructional competence and student performance.

Evaluation models must be able to separate the sources of information and practices exhibited by a teacher, from prep coursework to student teaching to norms of the school that hires her. And it must account for differences in family backgrounds and communities that affect K-12 student performance to avoid penalizing teachers for working with disadvantaged and special needs children.

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.

Recommendations for Policy Makers at All Levels

- *Provide funding and incentives to encourage more high-quality research to determine which of the many elements involved in teacher preparation have the greatest positive effects on teaching practices and K-12 student learning. Even studies with findings limited to a single location or measure of student performance can contribute to a growing understanding of how all the pieces and layers fit together to explain links between TPPs, teaching practices, and student learning.*
- *Given the limitations of extant research, take time to assess the strength of findings and rigor of study designs before using them to justify evaluation measures.*
- *Involve teacher prep faculty and K-12 teachers in designing the measures. Seek out experts who know the research on student learning, know what good teaching looks like, and fully understand the current demands of teaching.*
- *Align standards for evaluating TPPs with standards for educator licensure and evaluation, and ensure that instruments used for those processes measure critical outcomes related to student achievement.*
- *As states build longitudinal databases they will want to include variables and links to allow for tracking the performance of teacher candidates through TPPs and certification and into their first and subsequent jobs.*
- *Measures of successful teaching should reach much deeper than test-score growth of program graduates' K-12 students and should be informed by the highest-quality quantitative and qualitative research available.*
- *Give careful consideration to how best to deal with the wide variation among TPPs, which range from traditional university-based programs to online alternative certification. We recommend the following:*
  - *Hold all types of TPPs to the same accountability and licensure standards. Teacher candidates and hiring districts should be able to compare and predict performance across programs.*
  - *Consider the potential impact on diversity in the teaching profession when accounting for different admissions standards and the varying levels of academic preparation that candidates bring to TPPs.*
- *Consider the possible unintended consequences of shifting the focus from inputs to outcome measures. Accepting a full shift implies that any means justify acceptable ends.*

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. 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.

5. Goe, L. (2006) Teacher Preparation > Teacher Practice > Student Outcomes Relationship in Special Education: Missing Links and New Connections. National Comprehensive Center for Teacher Quality.

6. Levine, A. (2006) Educating School Teachers. The Education Schools Project.

7. Pam Grossman, P. and McDonald, M. (2008) Back to the Future: Directions for Research in Teaching and Teacher Education. American Educational Research Journal, 45(1), 184 –205.

8. National Research Council (2002). Scientific Research in Education. Committee on Scientific Principles for Education Research. R.J. Shavelson and L. Towne, Eds. Washington DC: The National Academies Press.

9. Gansle, K.A., Noell, G.H. and Burns, J.M. (2012) Do Student Achievement Outcomes Differ Across Teacher Preparation Programs? An Analysis of Teacher Education in Louisiana. Journal of Teacher Education.

10. Goldhaber, D., and Liddle, S. (2011) The Gateway to the Profession: Assessing Teacher Preparation Programs Based on Student Achievement. Center for Education Data & Research.

11. Koedel, C., Parsons, E., Podgursky, M., and Ehlert, M. (2012) Teacher Preparation Programs and Teacher Quality: Are There Real Differences Across Programs? CALDER Working Paper No.79.

12. Osborne, C., von Hippel, P.T., Lincove, J., and Mills, N. (2012, October). The Size and Reliability of Teacher Training Effects in Texas. Paper presented at the Association for Public Policy Analysis and Management 2012 Fall Research Conference, Baltimore, MD.

13. Boyd, Donald J., Grossman, Pamela L., Lankford, Hamilton, Loeb, Susanna, Wyckoff, James (2009) Teacher Preparation and Student Achievement. Educational Evaluation and Policy Analysis, 31(4), 416–440.

14. Harris, D.N., & Sass, T. (2011). Teacher training, teacher quality and student achievement. Journal of Public Economics, 95 798–812.

15. Mihaly, K., McCaffrey, D., Sass, T., and Lockwood, J.R. (2012) Where You Come From or Where You Go: Distinguishing Between School Quality and the Effectiveness of Teacher Preparation Program Graduates. National Center for the Analysis of Longitudinal Data in Education Research.

16. Goldhaber, D., Walch, J., and Gabele, B. (2012, October). Does the Model Matter? Exploring the Relationship Between Different Student Achievement-Based Teacher Assessments. Paper presented at the Association for Public Policy Analysis and Management 2012 Fall Research Conference, Baltimore, MD.

17. Guarino, C.M., Reckase, M.D., and Wooldridge, J.M. (2012) Can Value-Added Measures of Teacher Performance Be Trusted? Institute for the Study of Labor (Germany).

18. Kraft, M., and Papay, J. (2012, October). Estimating and Explaining Differences in Returns to Experience across Teachers. Paper presented at the Association for Public Policy Analysis and Management 2012 Fall Research Conference, Baltimore, MD.

19. Atteberry, A., Loeb, S. Wyckoff, J. (2012, October). Returns to Teacher Experience in Early Career Years. Paper presented at the Association for Public Policy Analysis and Management 2012 Fall Research Conference, Baltimore, MD.

20. Aldeman, C., Carey, K., Dillon, E., Miller, B., and Silva, E. (2011) A Measured Approach to Improving Teacher Preparation. Education Sector.

21. Crowe, E. (2010). Measuring What Matters: A stronger Accountability Model for Teacher Education. Washington, DC: Center for American Progress.

22. USDOE (2011b) Our Future, Our Teachers: The Obama Administration's Plan for Teacher Education Reform and Improvement.

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

24. edTPA (n.d.) American Association of Colleges of Teacher Education website.

25. The New Teacher Project (2012) Making Licensure Matter. Colorado Department of Education.