NATIONAL COMPREHENSIVE CENTER FOR TEACHER QUALITY

Teacher Preparation Program Evaluation for Accountability and Improvement

Webinar

Thanks for joining!

The webinar will begin shortly.

September 2012

LEARNING POINT Associates®
An Affiliate of American Institutes for Research®
ETS®
VANDERBILT®

During This Webinar, You Will:

- Gain knowledge of the most commonly used levers to hold teacher preparation programs accountable.
- Examine methods to evaluate programs using evidence of the quality of program processes (e.g., program selection, content, and structure) and evidence of impact.
- Learn what states at the forefront of change are doing to evaluate programs.
- Gain insight into the latest research examining the extent to which teachers who are prepared by different teacher-preparation programs differ in effectiveness.

Today's Presenters

- **Jane Coggshall, Ph.D.,** American Institutes for Research/TQ Center
- **Lauren Bivona,** American Institutes for Research/TQ Center
- **Dan Reschly, Ph.D.,** Vanderbilt University/TQ Center
- **George Noell, Ph.D.,** Professor, Louisiana State University
- **Cory Koedel, Ph.D.,** Assistant Professor, University of Missouri–Columbia

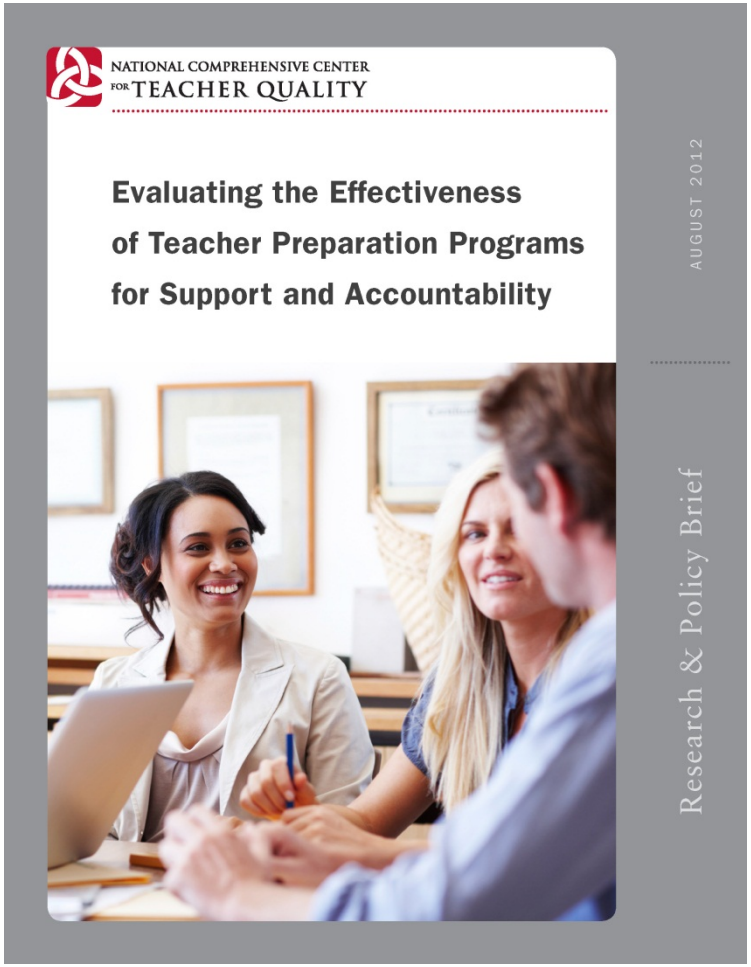
NATIONAL COMPREHENSIVE CENTER FOR TEACHER QUALITY

Evaluating the Effectiveness of Teacher Preparation Programs for Support and Accountability:

Webinar Presentation

Jane Coggshall, American Institutes for Research
Lauren Bivona, American Institutes for Research
Dan Reschly, Vanderbilt University

September 2012



http://www.tqsource.org/publications/TQ_RandP_BriefEvaluatingEffectiveness.pdf

Evaluation for What?

- Accountability
  - Program rankings/ratings (public transparency)
  - Shuttering failing programs
  - Rewarding/scaling up successful programs
- Program Improvement
  - Illuminating strengths and weaknesses
  - Learning from successful programs

Evaluation for Whom?

- Grades PK–12 students and parents
- Prospective teacher candidates
- Employers (school and district leaders)
- Program administrators and faculty
- State education leaders
- State policymakers (legislators, governors, state boards)
- Federal policymakers

Calls for Reform

- State accountability levers
  - Approval
  - Accreditation
  - Certification
- Federal accountability levers
  - Title II HEA reporting requirements
  - Race to the Top program requirements
- Accountability through transparency
  - National Council on Teacher Quality, *U.S. News & World Report*

Examples of Options for Evaluating Teacher Preparation Programs

- Input measures
  - Faculty qualifications
  - Enrollment data
- Outcome measures
  - Teacher effectiveness
  - Teacher practice
  - Survey results
- Process measures
  - Selectivity of programs
  - Program content
  - Program structure

Outcome Measures

- Student achievement (e.g., VAMs)
- District teacher performance evaluation results
- Surveys of principals/employers
- Surveys of graduates
- Data on hiring, placement, and persistence of graduates
- Teacher candidate knowledge and skills

Opportunities for Using Outcome Measures

- Can provide authentic evidence of effectiveness/impact
- May incentivize programs to build stronger partnerships with school districts
- May provide districts with more specific information on potential effectiveness of new hires

Challenges of Using Outcome Measures

- Program size matters
- Insufficient existing state data system capacity
- Signals of effectiveness versus in-depth information
- Timeliness of data collection and analysis

Example—Principal Surveys

- Texas Education Agency's *First Year Principal Survey*
  - Piloted in 2010; revised and administered statewide in May 2011
- Opportunities
  - Particular strengths and weaknesses of graduates can be detailed; principal evaluations tend to align with VAM scores; can be used for teachers in nontested subjects
- Challenges
  - Little variability in responses; better training for principals needed; disaggregated reporting more useful for prep programs

Learning From States: What We've Learned

- States are in different stages of implementation.
- Use of multiple measures is widespread.
- States must wrestle with tradeoffs associated with measures.
- Stakeholder engagement and collaboration matter.

Learning From States: Ongoing Challenges

- Building capacity in order to use VAM
- Accumulating sufficient data to present to the public
- Providing actionable data
- Locating and leveraging expertise
- Organizing stakeholder meetings

Process Measures

- Reviews of candidate selection processes
- Syllabi reviews
- Document reviews
- Surveys about student teaching and clinical experiences

Strengths of Process Measures

- Provide more information to policymakers and faculty on the research-based content and structures of prep programs
- Can identify possible gaps in the content or program coursework and clinical work
- Can help faculty to understand and implement research-based program coursework and clinical work

Evaluating the Content of Teacher Preparation Programs Example

TQ Center Innovation Configurations:

- Content is crucial in teacher preparation.
- Identify evidence-based practices in key areas such as reading, mathematics, classroom management.
- Rating scale from none to full implementation (e.g., reading comprehension strategies)
- Used in OSEP 325T projects to update and strengthen teacher preparation

Challenges of Process Measures

- Research base is not robust enough to build assessment for accountability based on process measures.
- May discourage innovation.
- May require complex qualitative measures that are difficult to score reliably across programs.
- Do not always accurately capture what actually happens in preparation programs.

Questions?

Jane Coggshall, Ph.D.
jmcconnochie@air.org

1000 Thomas Jefferson Street NW
Washington, DC 20007-3835
Phone: 877-322-8700 or 202-223-6690
Website: www.tqsource.org

Evolution of Louisiana's Assessment Systems for Teacher Preparation

George H. Noell, Ph.D.

Redesign in Louisiana
The Blue Ribbon Commission

quality

Louisiana's Teacher Preparation Programs: Four Levels of Effectiveness

Level 4: Effectiveness of Growth in Student Learning (Value-Added Teacher Preparation Program Assessment)

Level 3: Effectiveness of Impact (Teacher Preparation Accountability System)

Level 2: Effectiveness of Implementation (NCATE—Comprehensive Assessment System)

Level 1: Effectiveness of Planning (Redesign of Teacher Preparation Programs)

How Redesign Led to Louisiana's Assessment

- Accepting the proposition that teachers matter
  - Corollary: Preparation programs have roles in recruitment, selection, and preparation.
- Recognizing
  - K–12 data suggested needs for change.
  - Formal teacher preparation data did ***not*** suggest problems.
- No data connecting preparation and K–12 learning outcomes

Challenges in Assessing Teacher Preparation

- The challenge of measures
  - Achievement versus opinions
- Geography
- Heterogeneous schools and classes
- Data management
- Technical issues
- ***The plausible counter factual*** (Rubin)

Evolution of the Work in Louisiana

- Stakeholder engagement around analysis and responses
  - Concerns about families: study
  - Concerns about class composition: propensity study
  - Concern about attendance: included variable
- Stakeholder engagement around process
  - Providing data ahead of public release
  - Engagement at the Chancellor/President level
  - How data are presented

Some Key Policy Issues

- Coping with repeated observations
- How a teacher preparation assessment interacts with K–12 assessments
- Consideration of key potential predictors
- Minimum *n* to report results
- Standard setting and corrective action

Moving From Data to Actions

- Challenges to effective action
  - Specificity of information: VAM is a global indicator.
    - Subgroup and descriptive data
  - Resources
  - Will and structural barriers
  - Timeline to effect
- Program changes and changed results
  - ***Reading example***

What Is Louisiana Working On Now?

- Integrating VAM into a more comprehensive Title II accountability system
- Integrating our legislatively mandated teacher level value added system with our preparation system
  - Revisiting all decisions in light of this change
- Building systems to support continuous improvement

Teacher Preparation Programs and Teacher Quality: Are There Real Differences Across Programs?

Cory Koedel, Eric Parsons, Michael Podgursky, and Mark Ehlert

Motivation

- Evaluations of teacher preparation programs (TPPs) using student achievement data are increasingly popular.
  - A major component of successful Race to the Top applications (Crowe, 2011)
  - Strongly advocated by the U.S. Department of Education
  - This paper began as a typical TPP evaluation.

Evaluation Models

- The models being used to evaluate TPPs are value-added models or variants of value-added models.
- Basic idea: Construct a forecasting model for students' year-t achievement that depends on prior achievement and other information. The predictor of interest is the TPP for the teacher assigned to the student in year t.
  - Key question: Do students taught by teachers from some TPPs systematically exceed or fall short of their predicted performance?

Example

$$Y_{ijst} = Y_{ijs(t-1)}\delta_1 + X_{ijst}\delta_2 + S_{ijst}\delta_3 + TPP_{ijst}\theta + \gamma_s + \varepsilon_{ijst}$$

Y_{ijst} = test score for student *i* with teacher *j* at school *s* in year *t*.

X_{ijst} = student characteristics (race, gender, free/reduced-price lunch status, etc.).

S_{ijst} = school characteristics (analogous to student characteristics, but aggregated).

TPP_{ijst} = vector of variables that indicate TPP by which teacher *j* was certified.

Missouri Data

- Teachers certified between 2004 and 2008 and observed teaching in 2008–09 in a Missouri public elementary school, teachers' classrooms observed for up to three years (through 2010–11)
- Main Analysis
  - "All" programs: 24 TPPs with more than 15 teachers observed in the analytic data, 1,309 teachers in 656 different elementary schools (60,000+ student-year records attached)
  - Large programs: 12 TPPs with 50+ teachers (produced three fourths of the teachers in our sample)
    - On average, there were more than 80 teachers per program in our data for the large programs, which is more than in the reports in Louisiana and Tennessee, and in notable research studies (e.g., Boyd et al., 2009). We do not have a small sample.

Results From Missouri

TPP Effects. Programs with at least 50 New Teachers. Model B. Math.

Program	Effect Estimate
Program 1	0.030
	(0.028)
Program 2	-0.002
	(0.033)
Program 3	0.029
	(0.032)
Program 4	0.069
	(0.038)
Program 5	-0.011
	(0.027)
Program 6	0.000
	(0.029)
Program 7	0.005
	(0.033)
Program 8	0.034
	(0.042)
Program 9	-0.013
	(0.025)
Program 10	0.022
	(0.032)
Program 11	0.002
	(0.029)
Program 12	0.000
	(0.025)

Unadjusted St Dev	0.024
Unadjusted Range	0.083
Adjusted St Dev (Adj 1)	0.020
Adjusted Range (Adj 1)	0.069

Results From Missouri

TPP Effects. Programs with at least 50 New Teachers. Model B. Math.

Program	Effect Estimate
Program 1	0.030
	(0.028)
Program 2	-0.002
	(0.033)
Program 3	0.029
	(0.032)
Program 4	0.069
	(0.038)
Program 5	-0.011
	(0.027)
Program 6	0.000
	(0.029)
Program 7	0.005
	(0.033)
Program 8	0.034
	(0.042)
Program 9	-0.013
	(0.025)
Program 10	0.022
	(0.032)
Program 11	0.002
	(0.029)
Program 12	0.000
	(0.025)

Unadjusted St Dev	0.024
Unadjusted Range	0.083
Adjusted St Dev (Adj 1)	0.020
Adjusted Range (Adj 1)	0.069

The Problem
(Not Very Exciting But Important)

- There is an important clustering structure to the data for TPP evaluations.
  - Key issue: Each teacher is observed teaching multiple students and/or classrooms, but these students and classrooms ***are not*** independent observations with regard to the effectiveness of that teacher from the TPP.
    - This clustering structure is well supported by a large body of research on the importance of differences in teacher quality across individual teachers.
    - The most important level of clustering in TPP analyses is at the ***teacher level***.

REAL Results From Missouri

TPP Effects. Programs with at least 50 New Teachers. Model B. Math.

Program	Effect Estimate
Program 1	0.047
	(0.028)
Program 2	-0.010
	(0.030)
Program 3	-0.006
	(0.028)
Program 4	-0.067
	(0.032)
Program 5	0.022
	(0.026)
Program 6	0.004
	(0.028)
Program 7	0.023
	(0.038)
Program 8	-0.016
	(0.043)
Program 9	0.005
	(0.027)
Program 10	0.011
	(0.033)
Program 11	0.049
	(0.033)
Program 12	0.000
	(0.026)

Unadjusted St Dev	0.031
Unadjusted Range	0.116
Adjusted St Dev (Adj 1)	**0**
Adjusted Range (Adj 1)	**0**

Why Isn't This Working?

There is too much variation in teacher performance within programs, too little between.

- Large-program math models:
  - Change in R^2 for student-achievement model when we add TPP indicators (preferred specification): 0.001
  - Change in R^2 for student-achievement model when we add individual-teacher indicators (preferred specification): 0.047
  - Ratio: 0.019
  - That is, variation in teacher quality across TPPs explains just 1.9 percent of the total variation in test scores attributable to differences across individual teachers.

Why Isn't This Working?

- Because there is so much variation in quality within programs and so little between, the data environment is challenging. The data requirements to perform this type of evaluation—speaking in the abstract and purely from a statistical perspective—are substantial. Even in states where the data are of high quality, we will generally not have enough data to do a good job of this given the clustering structure.
- The most interesting and surprising finding for us is that the differences between teachers from different TPPs are so small. This is at the root of the statistical issue—the models are being asked to statistically distinguish very small program differences with lots of noise floating around.

A Brief Note on Selection

Table 4. Average ACT Scores by University for 11 Public Universities Included in Our Evaluation.

	Average ACT Scores		
	All Graduates	Graduates with Education Major	Observed Elem Teachers
Univ of Missouri-Columbia*	26.3	25.7	24.2
Univ of Missouri-Kansas City	25.3	23.8	22.8
Missouri State Univ*	24.7	24.1	21.7
Missouri Southern State Univ*	23.9	24.0	21.1
Univ of Missouri-St. Louis*	23.7	23.0	22.0
Southeast Missouri State Univ*	22.9	23.2	21.9
Northwest Missouri State Univ*	22.6	22.7	22.8
Univ of Central Missouri*	22.6	22.5	20.8
Missouri Western State Univ*	22.5	23.3	21.0
Lincoln University	21.4	22.0	21.0
Harris-Stowe State University	19.3	19.8	18.8
Variance of ACT Scores Across Universities	3.68	2.17	1.93
Range of ACT Scores Across Universities	7.0	5.9	5.4

Important Caveat

Our analysis looked only at “traditional TPPs.” Evaluations that consider Alt-Cert programs, or more generally, that compare more heterogeneous groups of programs, may find differences that were not present in our study.

Should We Give Up?

- My opinion: No
- Why?
  - The current statistical problem is driven in large part by the fact that the real differences between TPPs are small.
  - Given that there has never been an evaluation system that monitored output, this is perhaps unsurprising.
  - The mere presence of an evaluation system moving forward may prompt positive changes in TPPs as programs attempt to move up in the rankings. This could improve students' short-term and long-term outcomes in meaningful ways.
  - Even if these types of evaluations are not particularly informative now, they do no harm (if people understand the magnitudes of differences between programs). Keeping the systems in place leaves open the option that they may grow into a more valuable recruitment and monitoring tool in the future.

Key Takeaways

- The differences in teacher quality between teachers from different TPPs are small. In Missouri, we cannot reject the null hypothesis that the differences we see in our models are driven purely by noise in the data—that is, there are no real differences.
- This does not mean that these types of evaluations cannot become more useful over time, so all hope is not lost.
- However, it is important that educational administrators not put undue weight on the rankings produced by these models. The rankings are noisy, and the implied magnitudes of the differences between programs are typically very small.
  - The data indicate that if you rank the TPPs from best to worst based on estimates from these types of models, there are still many teachers from the lowest ranking program who will perform better than many teachers from the highest ranking program.

TQ Center Resources

- TQ Connection: Making Links between Teacher Preparation and Educator Effectiveness. Innovation Configurations (http://www.tqsource.org/connection/)
- *Teaching as a Clinical Practice Profession: Implications for Teacher Preparation and State Policy* (http://www.tqsource.org/publications/clinicalPractice.php)
- Systems That Last: Great Teachers and Leaders for America's Schools (http://www.tqsource.org/whatworks/wwc12systemsthatlast/)
  - **Sessions:**
    - **Enhancing Teacher Preparation, Development, and Support: Where Have We Been, and Where Are We Going?**
    - **Concurrent Session I.C: Preparing Tomorrow's Teachers and Leaders**