Incorporating Teacher Effectiveness Into Teacher Preparation Program Evaluation

Journal of Teacher Education
63(5) 335–355

DOI: 10.1177/0022487112454437
http://jte.sagepub.com

Gary T. Henry[1], David C. Kershaw[2], Rebecca A. Zulli[3], and Adrienne A. Smith[4]

Abstract

New federal and state policies require that teacher preparation programs (TPP) be held accountable for the effectiveness of their graduates as measured by test score gains of the students they teach. In this article, the authors review the approaches taken in several states that have already estimated TPP effects and analyze the proposals for incorporating students' test score gains into the evaluations of TPP by states that have received federal Race to the Top funds. The authors organize their review to focus on three types of decisions that are required to implement these new accountability requirements: (a) selection of teachers, students, subjects, and years of data; (b) methods for estimating teachers' effects on student test score gains; and (c) reporting and interpretation of effects. The purpose of the review is to inform the teacher preparation community on the state of current and near term practice for adding measures of teacher effectiveness to TPP accountability practices.

Keywords

teacher preparation, value-added models, program evaluation, quantitative methods

Introduction

A new era of accountability for teacher preparation programs (TPP) is being ushered in across the country by recent federal and state policies, such as the Race to the Top (RttT) fund, which require that these programs be held accountable for producing effective teachers. New accountability reforms define effective teachers, at least in part, as those who produce higher student test score gains. In response, states are scrambling to create sophisticated databases that are able to link practicing teachers to their preparation programs as well as to the achievement data of the students they teach. Making these linkages is a necessary step for any TPP evaluation seeking to include estimates of the effects of TPP graduates on student test scores. However, the paucity of specific details in the RttT applications of successful states suggests that states are just beginning to grapple with the process of designing and implementing methods to estimate the effects of graduates of TPPs on their of students' test score gains. Although pioneering work exists in which the effects of TPPs on student test scores gains have been estimated, thereby proving that incorporating these measures into TPP evaluation can be done (see Boyd et al., 2009; Gansle, Noell, Knox, & Schafer, 2010; Henry, Thompson, Fortner, Zulli, & Kershaw, 2010; Henry et al., 2011; Noell, 2006; Noell & Burns, 2006, 2007; Noell, Porter, & Patt, 2007; Tennessee State Board of Education [TSBOE], 2009, 2010), the existing literature does not concisely identify the major decisions that states will face (nor the major options available) when incorporating student test score gains into TPP evaluations.

To begin to overcome this omission, we reviewed the scholarly literature, policy literature, and RttT documentation and identified many of the major decisions involved in producing test score gain estimates of TPP effectiveness. We developed a framework that separates the major decisions into three distinct categories: selection, estimation, and reporting. Selection decisions involve choosing which students, teachers, academic subjects, and academic years will be used to form the basis of TPP evaluations. Estimation decisions involve choosing and implementing methods for calculating TPP effects using student achievement data. Finally, reporting and interpretation decisions involve deciding how to present and aid the interpretation of TPP effectiveness estimates.

This article focuses on the ongoing work to create empirically grounded estimates of TPP effectiveness using student test scores that has been mandated by federal and state

[1]University of North Carolina at Chapel Hill, USA
[2]Slippery Rock University, PA, USA
[3]North Carolina State University, Raleigh, USA
[4]Horizon Research, Chapel Hill, NC, USA

Corresponding Author:
Gary T. Henry, Carolina Institute for Public Policy and Department of Public Policy, University of North Carolina at Chapel Hill, 122 Abernethy Hall, CB# 3435, Chapel Hill, NC 27599, USA
Email: gthenry@unc.edu

policies. However, future efforts also need to be directed toward creating additional measures of student learning and developing more comprehensive TPP evaluation frameworks. To be clear, this manuscript deals exclusively with the production of teacher effectiveness estimates as measured by student test score gains and does not attempt to present the options for combining these measures with other types of measures of TPP quality to form a comprehensive accountability system. We acknowledge the importance of developing additional measures, but it is beyond the scope of this article. Before describing important decisions in the process of using student test scores for TPP evaluation, we briefly discuss the recent policy initiatives for adding such evidence into TPP evaluation.

Emergence of Student Outcome–Based TPP Accountability

Historically, TPPs have been evaluated based primarily on the components of the preparation program itself, including required coursework, faculty who teach the courses, and the nature and types of experiences that preservice teachers receive. Traditionally, the primary vehicle for evaluation of preparation programs' capacity for these components has been through applying the standards and review processes of national accrediting organizations, such as the National Council for Accreditation of Teacher Education (NCATE, 2010). The NCATE, like other accrediting organizations, sets minimum quality standards in a number of areas. These areas include candidate knowledge, skills, and professional dispositions; assessment system and unit evaluation; field experiences and clinical practice; diversity; faculty qualifications; and unit governance and resources (NCATE, 2010). TPPs are only accredited by NCATE if they can demonstrate—through documentation and interviews—that they have developed their standards, policies, curricula, instruction, and assessment tools based on a research-grounded conceptual framework (NCATE, 2010). Notably, the accreditation process does not explicitly link teacher preparation to actual student achievement (Crowe, 2010). Accreditation focuses on evaluating the process of preparing teachers, not directly evaluating TPP graduates' instructional skills in the classroom or their ability to help their students learn.

In contrast, RttT and other calls for reform (Crowe, 2010) shifted the current policy environment by pushing states to make the achievement test score gains of the students of TPP graduates a central component of TPP accountability. For example, RttT called for

> (D)(4) Improving the effectiveness of teacher and principal preparation programs.
>
> (i) Link student achievement and student growth . . . data to the students' teachers and principals, to link this information to the in-State programs where those teachers and principals were prepared for credentialing, and to publicly report the data for each credentialing program in the State; and
>
> (ii) Expand preparation and credentialing options and programs that are successful at producing effective teachers and principals. (U.S. Department of Education, 2012, p. 19504-5)

The RttT student achievement–focused model of evaluation and accountability starkly contrasts with the traditional process-oriented TPP model of evaluation. Under RttT, states must develop and implement the capacity to reliably link student test scores to teachers to determine teacher effectiveness, and then to link these measures of teacher effectiveness back to the programs that prepared them to teach.

This recent emergence of the student learning evaluation approach was not necessarily due to a historical lack of interest in linking teacher preparation to student outcomes (Wilson, Floden, & Ferrini-Mundy, 2001). Rather, the statewide longitudinal data systems and analytical tools that link teachers to student outcomes on a large scale, which are needed to exercise the student learning evaluation approach, simply did not exist (National Research Council [NRC], 2010; Voorhees, Barnes, & Rothman, 2003).

Indeed, the recent policy initiatives pushing states to evaluate TPPs using student test scores coincide with advancements in state longitudinal data systems. Over the last two decades, states and school districts increased capacity to evaluate the effectiveness of TPP graduates in the classroom through the investment in longitudinal data systems that contain administrative data on students, teachers, and schools (Crowe, 2010; NRC, 2010). Without these necessary investments, states would not have the data necessary to assess how well the students of graduates from different TPPs and other credentialing pathways are performing in terms of increasing test scores.

Equally important for the emergence of incorporating student test score growth into TPP evaluation has been the development of quantitative methods that can be used to estimate the effect of individual TPPs on student achievement. Advances in two methodological areas are of particular value to states preparing to address the increasing accountability demands: the progress in calculating the effectiveness of individual teachers (see Aaronson, Barrow, & Sander, 2003; Kane, Rockoff, & Staiger, 2006; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004) and the recent evaluations of TPPs conducted in a handful of school districts and states (Boyd et al., 2009; Gansle et al., 2010; Henry et al., 2011; Henry, Thompson, Fortner, et al., 2010; Noell, 2006; Noell & Burns, 2006, 2007; Noell et al., 2007; TSBOE, 2009, 2010).

These developments are directly linked to prior studies in which researchers assessed the effectiveness of teachers through the application of the education production function (Coleman, 1966; Greenwald, Hedges, & Laine, 1996) that

focused on calculating the effects of different factors (home, school, and community) affecting student achievement including teachers with different characteristics. Education production function studies provide an estimate of the value of specific teacher characteristics (such as years of experience or certification status) in terms of student test score gains, controlling for other factors that also influence student achievement (Darling-Hammond, Berry, & Thoreson, 2001; Goldhaber & Brewer, 2000). Over the last few decades though, researchers have moved from describing the characteristics of effective teachers toward the identification of individual teachers' effects on student test score gains (Nye, Konstantopoulos, & Hedges, 2004). In the process, a number of researchers have developed and refined methodologies designed to isolate the effect of individual teachers on student achievement (Aaronson et al., 2003; Kane et al., 2006; Rivkin et al., 2005; Rockoff, 2004; Sanders & Rivers, 1996). Specifically, these researchers have developed various statistical models that attempt to estimate the unique contribution of a teacher to student learning above and beyond any learning that would be expected given a student's prior achievement and other individual student characteristics (e.g., socioeconomic status, mobility, and disability status), classroom context, and school context. Overall, these studies provided the "proof of concept" that researchers could link student achievement to teachers and generate estimates of teacher effectiveness based on student test score gains.

Greater sophistication in data systems and ability to estimate teacher effects has enabled researchers to begin to leverage these methodological advancements and data system improvements to evaluate teachers' performance and TPP (NRC, 2010). Although a number of studies attempt to tackle teacher preparation broadly through the analysis of certification or initial preparation of teachers (Boyd et al., 2006; Clotfelter et al., 2007, 2010; Darling-Hammond et al., 2001; Goldhaber & Brewer, 2000; Henry, Thompson, Bastian, et al., 2010; Kane, Rockoff, & Staiger, 2006), few researchers have attempted to generate numerical estimates of individual TPP effectiveness. To date, researchers have generated analyses that address TPP effectiveness for North Carolina (Henry, Bastian, & Smith, 2012; Henry, Thompson, Fortner, et al., 2010), New York City (Boyd et al., 2009), Louisiana (Gansle et al., 2010; Noell, 2006; Noell & Burns, 2006, 2007; Noell et al., 2007; Noell, Porter, Patt, & Dahir, 2008), Tennessee (TSBOE, 2009, 2010), and Florida (Florida Department of Education [FDOE], 2009a, 2009b, 2009c). These researchers used a diverse array of methods, models, students, subjects, teachers, and test scores to estimate the TPP effectiveness for their teachers. Table 1 provides an overview of how these researchers have chosen to estimate the effectiveness of TPPs, given the data limitations faced in each state.

A close look at the RttT applications of the RttT recipients reveals that few states provide concrete details about how the states will ultimately judge TPP effectiveness using student test score gains. Most states simply refer to using "student growth" or "student learning gains" to generate TPP effectiveness as part of their application. This reflects the fact that many of these states are at the beginning of the TPP evaluation process. Georgia was very forthright in this regard when they noted that they would "use a portion of RT3 resources, if awarded, to contract with a value-added model (VAM) provider to develop a statewide VAM" (The State of Georgia, 2010, p. 104). VAMs seek to isolate the effect of individual teachers on their students test scores by adjusting out the influence of other factors, always including the students' prior test scores and in most cases, also including other characteristics of the students, their families, their classroom context, and school context.

A few states, such as Massachusetts and Ohio, signal that they are further along in their plans to incorporate student test scores into the evaluation of TPPs. These states either mention specific methods for evaluating teacher effectiveness (i.e., student growth percentiles [SGP], The State of Massachusetts, 2010) or mention groups performing teacher effectiveness analyses in the state with known methodologies (i.e., Battelle for Kids who use a SAS Education Value-Added Assessment System [EVAAS]; The State of Ohio, 2010). Still, Ohio and Massachusetts provide few concrete details about how they will implement the methods as part of their TPP effectiveness evaluation and, in fact, fall short of committing to use these approaches as part of their TPP evaluations.

Finally, The State of North Carolina (2010, Section D, Appendix 26, p. 143) and The State of Tennessee (2010, p. D96) provided greater details about teacher or TPP effectiveness studies that have been done in their states. At the same time, the details provided by North Carolina and Tennessee in their RttT applications still do not sufficiently convey the detailed decisions that researchers must make when analyzing TPP effectiveness.

Table 2 lays out some of the language in the RttT proposals that the successful states used to describe their plans to evaluate and hold their TPPs accountable. To be sure, this table reflects the state's intention at the RttT proposal stage and only scratches the surface in terms of the plans to incorporate measures of TPP effectiveness and using these measures to improve teacher preparation in the RttT states. It is, however, instructive to review some of the language in their successful proposals. These proposals collectively indicate that the framework we present could inform the policy conversations as states begin to implement their plans.

Because the RttT applications are short on details—and because collecting, coding, and analyzing teacher preparation content data generates a substantial number of key decisions—the remainder of this article turns to existing policy and scholarly research to identify and understand the key decisions that must be faced when estimating TPP effectiveness.

Table 1. Select Examples of TPP Effectiveness Reports and Studies

State (study year)	Florida (2009)	The State of North Carolina (2010)	NYC (2009)	Louisiana (2007)	Louisiana (2010)	The State of Tennessee (2010)
Selection decisions						
TPP estimates for academic years	2007-2008	2005-2006 to 2007-2008	2000-2001 to 2005-2006 (varied)	2004-2005, 2005-2006	2005-2006 to 2008-2009	2009-2010
Grades/levels	3-8	3-12	3-8	4-9	4-9	4-12
Separate grade level analyses	No	Elementary, middle, and high school	No	No	No	EOG, EOC/gateway
Test scores assessed	Mathematics and reading	Mathematics, reading, science (HS), social studies (HS)	Mathematics and ELA	Math, science, and social studies	Math, science, English-language arts, reading, and social studies	EOG: Math, reading/language arts, science, and social studies; EOC: Algebra I, Biology I, English I, and English II
Student selection decisions	Unknown	No limiting criteria	Excludes retained students and grade skipping students	Excludes within-year movers, retained students	Excludes within-year movers, retained students	Students with at least 3 test scores
Teachers used in analyses	1st-year teachers	Teachers with less than 10 years of experience	1st- and 2nd-year teachers	All teachers	All teachers	All teachers
Minimum TPP teacher limits	10 teachers	10 teachers	40 teachers	10 teachers each year, 25 teachers all years	10 teachers each year, 25 teachers all years	5 teachers
Multiple teachers	Unknown	assigned equal weight	unknown	ELA-analysis not performed Math, science, social studies—unknown	Unknown	Weighted
Programs analyzed	10 Institutions	15 UNC traditional undergraduate prepared	Program analysis: 26 traditional programs at 18 state institutions; 4 NYC teaching fellows programs; Teach for America	3 alternative programs; 13 (now defunct) traditional programs	8 alternative programs and 8 traditional undergraduate program effects reported, 59 program effects reportedly estimated	41 institutions, Teach for America
Estimation decisions						
Methodological approach	Value table	Value-added model–direct estimation	Value-added model–direct estimation	Value-added model–direct estimation	Value-added model–direct estimation	Value-added model–aggregation of teacher quality estimates
Value-added model type	NA	Year-to-year with controls	Year-to-year with controls (school fixed effects), year-to-year with controls, and OLS	Year-to-year with controls	Year-to-year with controls	TVAAS/EVAAS
Teacher level controls	No	Yes (in some models)	Yes (in some models)	Yes	Yes	No

(continued)

Table 1. (continued)

State (study year)	Florida (2009)	The State of North Carolina (2010)	NYC (2009)	Louisiana (2007)	Louisiana (2010)	The State of Tennessee (2010)
Tests used to predict current achievement	NA	Reading and mathematics	Math and ELA	Math, science, ELA, and social studies	Math, science, ELA, and social studies	All available (for 5 years)
Teacher level controls	NA	Yes (in some models)	Yes (in some models)	Absences only	Absences only	No
Reference group	NA	All non-UNC undergraduate prepared teachers	Unclear	Experienced teachers, statistical comparisons with new teachers made	Experienced teachers, statistical comparisons with new teachers made	NA
Presentation of results	Percentage of TPP beginning teachers' students making learning gain bench marks	Table highlighting significant TPP effects; effectiveness estimates, days equivalent effects (where possible)	Effectiveness estimates	5 Performance bands group TPP by statistical relationship with veteran teachers and each other; effectiveness estimates	5 Performance bands group TPP by statistical relationship with veteran teachers and each other; effectiveness estimates	Table highlighting significant TPP effects; mean *t*-values; % of teachers in upper and lower quintiles

Note: NYC = New York City; TPP = teacher preparation programs; ELA = English-language arts; UNC = University of North Carolina; TVAAS = Tennessee Value-Added Assessment System; EVAAS = Education Value-Added Assessment System. HS = High school only; EOG = End-Of-Grade; EOC = End-Of-Course; OLS = ordinary least squares.

Table 2. Race to the Top Proposal Information

States	Snapshot of state's plan to assess teacher preparation programs
The State of Delaware, 2010	"Delaware's rigorous statewide educator evaluation system is based on the most respected standards for teaching and leading (Danielson's A Framework for Teaching and the Interstate School Leaders Licensure Consortium's standards for leaders). The system provides a multi-measure assessment of performance that incorporates student growth as one of five components. Rather than set a specific percentage that student growth must be weighted in the evaluation, these regulations go much further. They say that an educator can only be rated effective if they demonstrate satisfactory levels of student growth." (p. A-4)
The State of Tennessee, 2010	"Unlike most states, which likely are setting up their data systems in order to know which teacher preparation programs prepare the highest-achieving graduates, Tennessee can—and does already—perform this analysis considering teacher effect data, placement and retention, and Praxis scores. Our LEAs can, and do, optimize our new teacher supply by using these data to increase recruitment, selection and hiring from preparation programs whose teachers consistently achieve better outcomes. Tennessee already publicly reports this data for each credentialing program in the state." (p. 110)
The District of Columbia, 2010	"DC will use Race to the Top to deliver on the next phase of bold reforms. Specifically, the District will: 1. Identify teacher preparation programs that are not providing effective teachers and hold them accountable for their quality, providing them with specific feedback on the performance of their graduates to support targeted improvements, and revoking program approval after continued ineffectiveness, as necessary. Percentage of teacher preparation programs in the State for which the public can access data on the achievement and growth (as defined in this notice) of the graduates' students. 100%." (p. 92-93)
The State of Florida, 2010	"Supported by substantial data and statewide assessment systems, Florida measures growth and proficiency annually for each student (learning gains) in reading and mathematics in Grades 4 through 10 and reports these data at the school level as part of the state's accountability system. The state's longitudinal database links students with their teachers and courses, and teachers to teacher preparation programs and professional development. This linkage is currently used statewide to report individual teacher performance in the aggregate by school type, subgroup, and preparation program, but only cautiously, based on the knowledge that a more sophisticated measure of student growth is needed to further examine individual teacher performance." (p. 138)
The State of Georgia, 2010	"ACTIVITY (1): Create a Teacher Effectiveness Measure (TEM) for each teacher in the state and a Leader Effectiveness Measure (LEM) for each principal in the State. The TEM and LEM require linking student achievement and student growth data to the students' teachers and principals. This is the first necessary step in lining the information back to in-State teacher and principal preparation programs. ACTIVITY (2): Develop a Teacher Preparation Program Effectiveness Measure (TPPEM) and Leader Preparation Program Effectiveness Measure (LPPEM). The TPPEM and LPPEM include multiple components, including TEM and LEM of graduates aggregated by cohort, which provides the linkage between student growth data to in-State teacher and principal preparation programs. ACTIVITY (3): Calculate TPPEM and LPPEM and publish preparation program "report cards" (both traditional and alternative routes). Student growth data will be tracked as early as 2010-2011through value-added models, but the first full year of TEM/LEM implementation will not occur until SY2011-2012 (since the qualitative evaluation tool will be validated in 2010-2011 and launched in participating LEAs in 2011-2012). First TEM/LEM scores will be available in the fall of 2012; the earliest the State would have data to calculate TPPEM and LPPEM would be late 2012." (p. 143)
The State of Hawaii, 2010	"The State, HIDOE, and the Hawaii Teacher Standards Board (HTSB) share a vested interest in obtaining data about the effectiveness of teacher preparation programs and then acting on those data to ensure that the State Approved Teacher Education Programs (SATEP) and Administrator Certification for Excellence (ACE) programs are doing the best job possible at preparing Hawaii's teacher and principal corps. In addition, in response to requirements of the Federal Higher Education Opportunity Act (2008), local teacher preparation programs already have been demanding data linking student achievement to students' teachers and their respective preparation programs. Additionally, HTSB's Unit Performance Standards for State Approved Teacher Education Programs requires preparation programs to "collect and analyze data about program completer performance to evaluate and improve" the program. Hawaii leaders are working together to ensure its new data system collects and analyzes more relevant information about how well each program is preparing teachers and principals to be effective; they also are working to make sure this information is more widely distributed and easily understood, so it can be better used by policymakers in reviewing programs, by schools in making hiring decisions and by teacher and principal candidates in deciding on which preparation path will give them the best support." (p. 141)

(continued)

Table 2. (continued)

States	Snapshot of state's plan to assess teacher preparation programs
The State of Maryland, 2010	"All teacher preparation programs are evaluated on common performance criteria aligned with State and national outcomes; Maryland has closed one program and placed three others on probation for subpar performance. The State Board of Education adopted professional development standards to ensure quality across all professional development experiences, including induction. LEAs provide a teacher induction plan that follows beginning teachers through the tenure period. Additionally, Maryland will establish partnerships with the University System of Maryland to design a STEM teacher preparation program based on a proven national model, such as the National Math and Science Initiative's UTeach program. Partner institutions will commit to recruiting college students in their junior years for a specially designed model of instruction co-planned, implemented, and evaluated by the collaborative efforts of both the College of Arts and Sciences and the College of Education." (p. 51-52)
The State of Massachusetts, 2011	"Great Teachers and Leaders: The depth and breadth of the initiatives and strategies described in section D necessitate continued and consistent collaboration among ESE, EOE, DHE, UMASS, and other stakeholders, and ESE will coordinate these partnerships both during and beyond the four-year RTTT grant. For example, ESE will continue its partnership with DHE, institutions of higher education, and other partners to develop and embed measures of educator effectiveness into every component of the system; improve the content, quality, and structure of teacher preparation programs; and increase the diversity of the educator workforce. RTTT funding also will be allocated to the Readiness Centers to supplement the capacity of ESE to provide instructional and professional development services and to convene stakeholders to address cross-sector priorities." (p. 196)
The State of New York, 2010	"NYSED will partner with higher education institutions as they redesign their teacher preparation programs to align with the Department's new standards and performance-based assessments for teacher certification … NYSED will also publish transparent data profiles for all institutions that prepare teachers and principals that focus on the performance of students their graduates have taught. Leaders in educational policy—including superintendents, school board members, members of Congress and the State legislature, the Governor's office, and the Board of Regents—will have access to customized reports that provide information regarding K–12 program effectiveness, higher education program effectiveness, and the adequacy of teacher preparation programs. Information will be available to help inform discussions regarding teacher and administrator evaluation, as well as policy decisions regarding student performance and the achievement gap." (p. 18)
The State of North Carolina, 2010	"Ground-Breaking Study of UNC Teacher Preparation Programs. NC links student achievement and growth data to teacher preparation programs. The UNC General Administration (UNC-GA), in close partnership with constituent UNC institutions that prepare teachers and principals, has completed the first phase of a new value added accountability study of educator preparation programs (called *NC Teacher Quality Research*). Results from this first phase are outlined in *The Impact of Teacher Preparation on Student Learning in North Carolina Public Schools* (Henry, Thompson, Fortner, et al., 2010). A primary component of the study is a quantitative evaluation of the impact of teacher preparation program graduates on student learning at the elementary, middle, and secondary levels. This initiative—one of the first of its kind in the country—has begun the process of examining program impact across grade levels, content-area subjects, and subpopulations of students, as well as across nearly a dozen different "portals" of entry into the profession (e.g., alternative and out-of-State programs, in addition to traditional in-State routes). Future evaluations also will discern the impact of principals and other school-based professionals on student achievement and provide evaluations of their preparation programs." (p. 176)
The State of Ohio, 2010	"As part of ODE's longitudinal data system, teacher and principal effectiveness data will be computed annually and linked to teacher and principal preparation programs. Data will also shed light on achievement gaps and how such gaps or the absence of gaps connect to educator preparation programs. This information will be publicly reported through an annual Ohio Teacher Education Report Card and shared on the OBR website showing aggregate effectiveness ratings of graduates from Ohio programs and institutions. The reporting system will permit the public to view the aggregate rating distribution for all graduates by program and licensure area, as well as for specific years. The OBR report will highlight successful programs while also calling attention to programs that may consistently produce graduates who are unsuccessful in their positions or who fail to obtain their professional licensure." (p. D4-5)
The State of Rhode Island, 2010	"The state will be equally aggressive in holding teacher preparation programs accountable for the effectiveness of their graduates. Rhode Island will publicly report on the effectiveness of each educator preparation program's graduates. RIDE will use Race to the Top funds to create new educator preparation program report cards that include information on:

(continued)

Table 2. (continued)

States	Snapshot of state's plan to assess teacher preparation programs
	The impact of the program's graduates on student growth and academic achievement, as compared with all other teacher or principal (as appropriate) preparation programs in the state; The rate at which each program's graduates earn full Professional Certification, which under the new certification system (described in D (2) (iv)) will require evidence of effectiveness, by the end of their first three years of teaching; and The number of preparation programs' graduates working in Rhode Island schools, disaggregated by LEA and high/low-poverty and high/low-minority schools. These report cards will use a consumer-friendly format and will be available on the RIDE website to provide preparation programs, prospective teachers and employers, and the public a comprehensive, objective picture of the effectiveness of each preparation program's graduates. RIDE will also publish an annual statewide educator preparation report card that aggregates information on the performance of all preparation programs in the state." (D49-D50)

Identifying Major Decisions in TPP Evaluation

Presently, many RttT-funded states and other states hoping to incorporate measures of effectiveness into the evaluation of TPPs are struggling with undertaking the first steps of estimating TPP effectiveness. Fortunately, the empirical research cited above provides evidence that estimating a teacher's effectiveness is possible and can help to inform future efforts to incorporate student test score gains into the evaluation of TPPs. However, this literature also highlights limitations and challenges that will be faced when incorporating student test scores into TPP evaluations. The shortcomings identified from past and current efforts to link student test scores back to TPP can shed light on the major decisions, both methodological and policy-oriented, that will influence the ultimate nature and robustness of any such evaluation effort. This article will provide a synthesis of the relevant methodological, research, and RttT-related literature to carefully identify (as well as explain) the importance of many of the key decisions states and researchers have faced when attempting to generate a quantitative estimate of TPP effectiveness. We classify these decisions as falling into one of three domains: (a) selection, (b) estimation, and (c) reporting and interpretation.

Selection decisions refer to the choices that states that wish to implement an assessment of TPP effects on student test scores will need to make about the students, teachers, and subjects they will include in the evaluation. Estimation decisions involve the choices associated with selecting an analysis method that will be used to quantify TPP effectiveness. Each commonly used approach has a specific set of processes that generate a set of decisions. Reporting decisions refer to the choices that states who have undertaken TPP will need to make about what specific information is released and the manner in which the results of the evaluation (effect estimates) are presented. The following sections of this article will separately address each of these three types of decisions describing the choices researchers have made when faced with these decisions and identifying some of the consequences of the different choices.

Selection Decisions

Selection decisions involve deciding which students and teachers will be used to analyze TPP effectiveness as well as which content areas will be used to form the evaluation of the TPPs. Specifically, states must select (a) which subjects, grades, and academic years will be used to evaluate TPP effectiveness; (b) which students will be used to evaluate TPP effectiveness; and (c) which teachers will be used to evaluate TPP effectiveness. The choices made at the selection decision points ultimately define the comprehensiveness of the assessment of the TPP effects on student test scores. With the exception of including teachers too far removed from their initial training, researchers gain a more complete (unbiased) picture of TPP effectiveness by including more students, teachers, and subjects in their analyses.

Subjects, Grades, and Academic Years

Although RttT directly calls for connecting TPPs to student performance, the ability to generate teacher and TPP estimates are limited by the nature of the existing student assessment programs operating in the state. Obviously, if a state does not test students in a certain academic year, grades, or subjects, the state cannot use those years, grades, or subjects to estimate TPP effectiveness. For many states, requirements of No Child Left Behind (NCLB) established the assessment system parameters. NCLB required each state test students in mathematics and reading in Grades 3 through 8, and at least once in either Grade 10, 11, or 12 (PL 107-110). States were also required (by academic year 2007-2008) to test students in science one time in Grades 3 through 5, 6 through 9, and 10 through 12 (PL 107-110). Although all states should meet these basic requirements, states vary in the assessments they administer.

For example, Florida's Comprehensive Assessment Tests (FCAT) annually tests students learning in mathematics and reading in Grades 3 through 10, science in Grades 5, 8, and 11, and writing in Grades 4 and 8 (FDOE, 2005). Assessments for specific high school courses are not included. Consequently, analyses of TPPs that focus on training high school teachers

would largely be limited to teachers of students in Grades 9 through 11 and would not be tied to learning standards and objectives for specific high school courses (FDOE, 2005). In contrast, the California Standards Tests (CST) includes an English-language arts test for Grades 2 through 11, but the general mathematics test is only given to students between Grades 2 and 7 (California Department of Education [CDE], 2011). Beginning in Grade 7, students may end up taking additional mathematics CSTs, including Geometry, Algebra I, and Algebra II. High school geometry teachers and algebra teachers in California can be evaluated based on the specific course objectives, whereas all math teachers would be lumped together in any Florida teacher analysis. Note that Florida brought end-of-course (EOC) tests for high school students online in May 2011 (FDOE, n.d.).

The greater number of tested grades and subject areas tested, the greater the flexibility a state will have in creating detailed TPP analyses. For example, researchers in North Carolina took advantage of a diverse battery of tests designed to evaluate students' knowledge of the core curriculum at the elementary, middle, and high school levels. Henry, Thompson, Fortner, et al. (2010) used North Carolina end-of-grade tests to evaluate TPP effectiveness in preparing teachers of elementary and middle grades students in reading and mathematics classes, and EOC tests to estimate TPP effectiveness at promoting overall, mathematics, science, and English achievement for high school students.[1] Henry, Thompson, Fortner, et al. found that programs that were better than all other sources of teachers in one subject at one level of schooling often fell back in the pack—or even fell behind—at other levels and in other subjects. It is worth noting that other researchers have also found that TPPs housed in a single institution can have variable impacts on student achievement across different grade levels and subjects (TSBOE, 2009, 2010). Table 3 presents a list of the student assessments administered during academic year 2009-2010 in RttT recipient states.

Similarly, greater content in testing programs across grades and subject matter will yield a richer data set, which allows evaluators to obtain effectiveness estimates for a larger number of teachers in multiple areas of teacher preparation. The size of the pool of educators whose effectiveness can be assessed is directly related to the decisions concerning grade and content selection. Increasing the pool of educators (by maximizing grades and subjects tested) also means that the evaluation of TPP effectiveness would be based on a larger proportion of graduates. States with minimal assessment programs will end up estimating the program's effects on fewer program graduates. Reducing the number of teachers used in an analysis can reduce reliability (year-to-year stability in the numerical assessment of a TPP's effectiveness), can bias TPP effectiveness estimates (produce systematic differences between the estimated and actual effect of a TPP), and result in the omission of TPP that prepare teachers for untested grades and subjects such as early childhood education (Pre-K-3) or social studies.

In addition, although the availability of state testing in a particular academic year is a prerequisite for the academic year to be included in the analysis, researchers may not want to evaluate TPPs using all available data. After all, TPPs undergo leadership, faculty, and structural changes over time. For these reasons, researchers have either limited or varied the academic years used to evaluate TPP effectiveness. For example, Boyd and colleagues (2009) separately analyzed TPP effectiveness using different combinations of academic years to test the stability of the estimates. Gansle and colleagues (2010) refused to generate reports for many teacher-preparing institutions in Louisiana due to complete overhauls of TPPs at particular institutions in the state. Their concern was that many of Louisiana's current teachers who received training at these institutions in the past will not adequately represent the current state of teacher preparation at these institutions.

Students

Naturally, the students used to capture TPP effectiveness for RttT will exclude those students who did not take the state assessments selected for use in the TPP analyses. Similarly, the student pool will also exclude students who are unable to be matched to their teacher or (for most analyses) prior test score(s) (e.g., students in the first tested grade within the state or new public school students). The exclusion of these students is due to the need for prior test scores from which gains (or losses) can be calculated. However, some states require that students be matched to their test scores from more than 1 year. For example, Tennessee requires that all students included in the analysis have at least three prior test scores, which eliminates fourth-grade students (and their teachers) because these students only took two tests (third-grade reading and mathematics assessments) prior to the fourth grade and therefore lack a third test score.

In addition, some researchers have deliberately chosen to further limit the pool of students used to evaluate TPPs. Boyd and colleagues (2009) excluded students who skipped a grade as well as retained students. Similarly, Noell and colleagues (2007; Gansle et al., 2010) excluded retained students and students who moved within the school year. Researchers support their decisions to exclude students who skip a grade or are retained because they believe the difference between the prior year's test score and the current year's test score should be interpreted as qualitatively different from test scores of students on a normal grade progression. Other researchers opt to retain as many students in the analysis as possible. For example, Henry, Thompson, Fortner, et al. (2010) included grade skipping, retained, and within-year moving students in their analyses and adjusted for these students by including control variables for students who are

Table 3. Statewide Assessments of the General Student Population During Academic Year 2009-2010

	Delaware	District of Columbia	Florida	Georgia	Hawaii	Maryland	Massachusetts	New York	North Carolina	Ohio	Rhode Island	Tennessee[a]
End-of-grade exams (cell values are grades the test is offered)												
Writing[b]		4, 7, 10	4, 8, 10	3, 5, 8, 11			4, 7, 10				5, 8, 11	5, 8, 11
Reading/ELA	3-10	3-8, 10	3-10	1-8	3-8, 10	3-8	3-8, 10	3-8	3-8	3-8	3-8, 11	3-8
Mathematics	3-10	3-8, 10	3-10	1-8	3-8, 10	3-8	3-8, 10	3-8, 9-12	3-8	3-8	3-8, 11	3-8
Science–technology–engineering	4, 6, 8, and 11	5, 8	5, 8, 11	3-8	4, 6, 10	5, 8	5,8, 9/10 (multiple HS science/ technology subject tests)	4, 8	5, 8	5, 8	4, 8, 11	3-8
Social studies	4, 6, 8, and 11			3-8								3-8
End-of-course exams (generally offered Grades 9-12, but sometimes earlier)												
Math								X				
Science								X				
Geography								X				
U.S. history				X				X	X			X
Global history								X				
English				Ninth-grade literature and composition/ American literature and composition		X		X	English I			English I, II, III
Economics-civics				X					Civics and economics			
Geometry				Mathematics I and II					X			
Algebra I				Mathematics I and II	X	X			X			X
Algebra II				Mathematics I and II	X				X			X
Chemistry												X
Physics												X
Physical science				X					X			
Biology		X		X		X			X			X
Government						X		X				

Note: ELA = English-language arts.

[a]Tennessee measures mathematics, reading, science, and social studies with a single assessment in Grades 3 to 8.

[b]Writing/composition may be subsumed in an ELA test in states without an explicit writing assessment.

overage, underage, those who changed schools within year, and those who changed schools between years.

Overall, the decision to limit the students included in a TPP student outcome evaluation may have important consequences, if the excluded students are disproportionately more or less likely to be taught by graduates of some TPPs or if the graduates of some TPP are either more or less effective with the excluded students. Unfortunately, existing research does not specify what the impact, if any, these different decisions have had on TPP estimates in the studies reported above.

Teachers

Once a state decides which grades, subject areas, and students will be used to evaluate TPPs, their policy makers will need to decide which teachers will be included in the analyses. Researchers face several additional decision points that affect which teachers are included in TPP analyses: These decisions involve (a) choosing whether teachers of all experience levels are included in the analysis, (b) deciding how to handle students with multiple teachers, and (c) deciding which TPPs will be used in analyzing TPP effectiveness.

To address the notion that the influence of teachers' university preparation diminishes as teachers gain experience, researchers have often limited their analyses to teachers with a limited number of years of experience. For instance, North Carolina initially estimated effectiveness of teachers with less than 10 years of experience (Henry, Thompson, Fortner, et al., 2010) and subsequently limited the analysis to teachers with less than 5 years of experience (Henry et al., 2011) as the number of years of data in the longitudinal database increased. Using an even more restrictive approach, Boyd and colleagues (2009) based their estimates of TPP effectiveness on New York City teachers with 1 or 2 years of experience. Florida provided estimates of student learning for 1st-year teachers from virtually all TPPs within its borders (FDOE, 2005, 2009c). Restricting TPP analyses to teachers with limited experience increases the confidence that any significant program effect reflects the current TPP process and practices. However, this decision must be balanced by the requirement to include a sufficient number of teachers who graduated from the states' TPPs and teach tested grades and subjects in the state's public schools.

Note that not all TPP researchers restrict the pool of teachers used in the analysis. Researchers in Louisiana (Noell & Burns, 2006; Noell et al., 2007; Noell et al., 2008) and Tennessee (TSBOE, 2010) use teachers at all experience levels in the process of estimating TPP effectiveness for recent TPP graduates. In Louisiana, experienced teachers are used as the reference group and TPP effects are directly estimated for teachers in their first 2 years. The statistical model used in Tennessee allows the state to calculate teacher effects for teachers with all levels of experience (TSBOE, 2010). The teacher effects for Tennessee's beginning teachers from TPPs are compared with the effectiveness of more experienced teachers. More information on these approaches is provided in the section on estimation decisions.

Although focusing TPP estimates on a subset of novice teachers has advantages, this approach can substantially reduce the number of teachers used in the analysis of TPP effectiveness. In general, evaluators should be wary of interpreting estimates of programs when the number of teachers included in the analyses is small. Researchers have used different teacher count criteria for reporting purposes. Tennessee required a minimum of 5 teachers (TSBOE, 2010), North Carolina set the minimum teacher counts for each TPP to 10 (Henry, Thompson, Fortner, et al., 2010), Louisiana set the minimum at 25 (Noell & Burns, 2006; Noell et al., 2007; Noell et al., 2008), and Boyd and colleagues (2009) set 40 as the limit in their analysis of New York City TPPs.

In addition, the TPP evaluators must decide how to handle students with multiple teachers. For a variety of reasons, a student may experience multiple teachers within the same subject during an academic year. For example, some students have "pullout" teachers who provide additional help in certain subjects, whereas other students experience coteaching (Fuchs, Compton, Fuchs, Bryant, & Davis, 2008; Murawski & Lochner, 2011). In addition, some students take multiple classes assessed by the same test (e.g., English and Literature). All of a student's teachers in a particular subject area who serve as instructors during a particular year may impact a student's test performance to some degree. The key decision facing TPP evaluators is, "Should a teacher, linked to a TPP, be equally responsible for student achievement on a test when she is responsible for only a part of a student's instruction in a particular subject?"

Researchers have taken different approaches to dealing with students with multiple teachers that influence a single student test score. Henry, Thompson, Bastian, et al. (2010) weighted student test scores across all teachers who share influence of a particular student test score. For example, if a student had two reading teachers, each was credited with half of the student's reading achievement gains. The statistical models used by Tennessee (TSBOE, 2010)—and scheduled to be used by Ohio as part of RttT (State of Ohio, 2010)—similarly assign each teacher a weight equal to the proportion of a student's instructional time claimed by the teacher (Wright, White, Sanders, & Rivers, 2010).

In contrast, other researchers have gotten around this problem by avoiding estimating TPP effectiveness in subjects where a large number of students have multiple teachers. Noell and colleagues (2007) did not perform an English/language arts analyses in Louisiana in the 2004-2005 and 2005-2006 TPP evaluation due to the high prevalence of multiple teachers in English/language arts. The Louisiana report suggests the multiple-teacher problem was small enough as to not warrant special attention in mathematics, science, and social studies analyses (Noell et al., 2007). Gansle and colleagues (2010) subsequently circumvented this problem by

generating separate analyses for English-language arts and reading. Under an approach such as the Florida (FDOE, 2009a, 2009c) value table approach, each teacher would be fully accountable for a student's growth even if they share teaching responsibilities for a student with other teachers. It appears that this approach would effectively double-count students with multiple teachers in state estimates of TPP effectiveness.

Finally, researchers may also be faced with deciding which TPP programs will be used to estimate TPP effectiveness. RttT structures this decision for many states, as RttT requires states to evaluate traditional and alternative preparation programs at both public and private institutions within their borders (see section D(4) of RttT application). However, what is less clear is whether RttT requires states to evaluate TPP effectiveness using student achievement from students taught by teachers trained by *all* teacher preparation sources. In theory, a state might decide to estimate effectiveness for in-state TPPs using only students taught by in-state TPP graduates. Again, this is an important decision because the estimates of TPP effectiveness are relative rather than absolute and therefore will vary if teachers that are prepared in-state are more or less effective than those from out-of-state.

Although in-state TPPs often supply the majority of teachers to the labor pool, out-of-state sources often contribute substantial numbers. Teachers prepared in out-of-state TPPs make up nearly 30% of teachers in North Carolina (Henry, Thompson, Bastian, et al., 2010). Similarly, the Florida Committee on Pre-K-12 Education reported that out-of-state teachers made up approximately 46% of new teachers in Florida (The Florida Senate, Committee on Education Pre-K-12, 2009). Importantly, when out-of-state prepared teachers were included and identified in a NC teacher preparation entry portal (pathway) analysis, they performed worse than traditional, public state institution–trained teachers in several grades and subjects (Henry, Thompson, Bastian, et al., 2010). Furthermore, Henry, Thompson, Bastian, et al. (2010) found out-of-state teachers were particular less effective in elementary school reading and mathematics (Grades 3-5), where they are the largest source of teachers with less than 5 years of experience in elementary school models. Consequently, the omission of the out-of-state prepared teachers who were, on average, poorer performing teachers from the calculation of TPP effectiveness in North Carolina would, at minimum, make relative effectiveness of in-state TPPs appear worse. Omitting the out-of-state TPPs, which might be the source of additional teachers if in-state TPPs produced fewer teachers, could be an unanticipated negative side effect if the estimates of TPP effectiveness are used to make programmatic decisions.

Similarly, decisions must be made about including out-of-state alternative preparation programs, such as Teach for America, which supplies teachers who, according to some recent research, provide greater gains in student achievement than other sources of teachers (Boyd et al., 2009; Glazerman, Mayer, & Decker, 2006; Henry et al., 2012; Xu, Hannaway, & Taylor, 2011). Teach for America teachers only make up a tiny fraction of all teachers in any given state. For instance, Teach for America teachers only accounted for 0.3% of North Carolina teachers in academic year 2007-2008 (Henry, Thompson, Bastian, et al., 2010). Consequently, their omission would not likely greatly alter TPP estimates of effectiveness. However, for completeness and a sense of fairness, it may be considered advisable to include all other TPP, large and small, to provide a comprehensive view of traditional and alternative preparation programs from in-state and out-of-state rather than just comparing the in-state TPP with each other.

Selection decisions form the foundation of incorporating student test scores into TPP evaluations. There are many small decisions that can have significant impacts of the comprehensiveness and fairness of the TPP evaluation. The decisions concerning how much and which of the available data to be used will affect the choices for estimation, discussed next.

Estimation Decisions

With the selection decisions made, those interested in incorporating student outcome measures into the evaluation of TPPs will need to decide what type of analytical models will be used to produce quantitative estimates of TPP effectiveness. The goal for the estimation process is to isolate the effects of individual teachers on the test scores of their students. A central guiding criterion should be to choose an analytic approach that neither benefits nor adversely affects the TPP due to forces beyond the control of the programs, such as the choice of the type of schools in which their graduates choose to teach or the students assigned to the classes that they teach. However, to hold the programs accountable for the effectiveness of their graduates, it is important that the TPP be evaluated on both of the processes that they control that can affect teacher effectiveness: (a) the selection of candidates into the TPP and (b) the preparation provided to the teacher candidates. For this reason and consistent with the existing literature evaluating teacher effectiveness, we will not propose parsing the effects of selection into the TPP and preparation by the TPP, but, rather, focus this section on the decisions necessary to estimate the combined effects of selection and preparation.

The two estimation decisions require (a) choosing the analytic approach that will be used to calculate the estimates of teacher effectiveness and (b) selecting the specific statistical model that will be used to produce the estimates of TPP effectiveness. All the analytical models that are currently in use or under discussion use the differences between students' current test scores and prior test scores rather than the levels of their test scores to measure effectiveness. Focusing on student test score gains rather than the students' current scores concentrates attention on the influence of a particular teacher during the year that she or he teaches the students and avoids

penalizing teachers who teach students who have lower test scores when they enter her or his classroom. Although far from inclusive of all possible approaches, researchers have typically taken one of three broad analytic approaches to estimating TPP effectiveness (explanations to follow): (a) student growth model, (b) Value-Added Models (VAMs) that estimate the effectiveness of individual teachers and then averages individual teacher estimates for each TPP, or (c) VAMs that estimate each TPP's effectiveness directly. In the remainder of this section, we will present an overview of the three types of models and the associated estimation decisions. In the appendix in the online supplement located on the website referenced in the abstract, you will find a more detailed explanation of some of the approaches presented.

Student Growth Models

Student growth models generally refer to fairly straightforward models that attribute the total difference between a student's current test score and a prior test score to the teacher who taught the student the particular subject being tested during that academic year (Auty et al., 2008). In practice, student growth models typically use student test scores to classify students into performance categories using predetermined proficiency standards (Goldschmidt et al., 2005). Generally speaking, in these models, a numerical student growth estimate based on the change in each student's performance category is generated for each student, and these estimates are then aggregated by the teacher's TPP to form a comparison of TPPs. To date, two types of student growth models have been or are planned to be used to evaluate TPPs: Value Table/Transition growth models and SGP Models.[2] Florida is the only state known to have used a student growth model to estimate TPP effectiveness. Florida used a value table model (see the appendix in the online supplement located on the website referenced in the abstract for more detail on Florida's use of non-value-added growth models to estimate TPP effectiveness).

The attraction of using the value table approach for evaluating teachers or TPPs is its simplicity and transparency. Because it does not involve highly technical or sophisticated statistical modeling, virtually anyone can calculate and track student growth (Buzick & Laitusis, 2010). Still, this approach has some potentially serious limitations. In particular, this growth model does not account for student, classroom, or school characteristics that influence students' test score gains. If these characteristics are not balanced across teachers from different TPPs, there is a real danger that TPP effectiveness estimates will, for instance, incorrectly attribute differences in the types of students and schools in which graduates of a certain TPP teach to the TPP. If a particular TPP's graduates serve students less likely to make gains whereas another TPP's graduates serve students more likely to make higher gains, the differences in students taught by the graduates of these TPPs can be attributed to the TPP incorrectly (Horner, 2009; see for additional information on the limitation of value tables for estimating student growth as well as for estimating teacher and school effectiveness).

To date, the Florida results are the only widely known use of a non-value-added student growth model to estimate TPP effectiveness. However, Massachusetts' RttT Phase 2 application (section (D)(2)) explicitly proposed using another non-value-added growth model, a SGP model, to evaluate teacher effectiveness as part of the state's evaluation of teacher effectiveness using student growth (The State of Massachusetts, 2010). Importantly, the language Massachusetts uses to describe their plan to link "student growth" to TPPs (see section (D)(4)) suggests the state will use the results from the SGP as part of the state's TPP evaluation (see the appendix in the online supplement located on the website referenced in the abstract for more details about SGP).[3]

The SGP approach starts to address some of the limitations of the value table approach. By estimating student growth relative to their "academic peers," this approach may control for some of the student background and other contextual conditions that may influence the likelihood that a student makes learning gains. However, the research on the SGP approach does not explicitly test the degree to which the student characteristics or context influence TPP effectiveness estimates. If TPPs will ultimately be held accountable for the quality of their teachers based on these estimates, these issues must be addressed, as otherwise, a TPP may be held accountable for factors beyond their control (Tekwe et al., 2004). It is also important to point out that the main proponent of the SGP method recommends that states use this approach descriptively, not for causal attribution or punitive actions (Betebenner, 2009a, 2009b).

VAMs

A VAM is a growth model that uses a student's prior test performance and, in most cases, other student, classroom, and school variables to estimate student learning gains. The difference between non-VAM and VAMs is in how they attempt to divvy up responsibility for gains and losses in student test scores. Student growth models attribute all of the change in each student test scores or classification categories (proficient or below expectations, for example) to the student's teacher. VAMs attempt to apportion the total growth between the different factors that have been shown to influence student test scores, including the students and their families, the make up of their classes, their schools, and their teachers. These models aim to directly estimate the unique contribution of teachers, schools, and TPP to student achievement (Braun, 2005). These models often include additional factors that influence student achievement in an attempt to disentangle the unique effect attributable to teachers or their TPPs. As noted earlier, there are two types of value-added modeling approaches common in practice: (a) VAMs that directly estimate a TPP's effectiveness and

(b) VAMs that first estimate individual teacher effectiveness, then aggregate these effects to the TPP level.

Next, we lay out the aggregation and direct estimation of TPPs approaches. Before those explanations, it is important to mention the second layer of estimation decision making: the structure of the statistical model. Numerous variations in the structure of the VAM have been developed in recent years. Researchers have frequently implemented three distinct types of value-added statistical models to assess the effects of teachers on student achievement: (a) year-to-year VAMs with controls for student, classroom, and school characteristics, (b) year-to-year value-added fixed effects models, and (c) multiple-year VAMs (Boyd et al., 2009; Sanders & Horn, 1994, 1997). In theory, each of these models can be used with either the aggregation approach or the direct estimation approach. The year-to-year VAMs with controls for student, classroom, and school characteristics are often implemented using multilevel models to apportion the variability in test scores to three levels: students, classrooms, and schools. The year-to-year fixed effects models can be implemented with fixed effects at the student, teacher, or school levels. For example, school fixed effects models allow the analyst to control for all non-time-varying characteristics of the school because the teachers are compared only with other teachers within the school. However, what this means substantively is that school fixed effects models set teaching within the school as the standard for judging a teacher's effectiveness, rather than teaching within the state as the standard. This may distort the effectiveness estimates between teachers at higher performing schools and those at lower performing schools, deflate the effectiveness of teachers at the former and inflate the effectiveness at the later. To the best of our knowledge, the year-to-year VAM with student fixed effects has not been used to evaluate TPP effectiveness, but the year-to-year VAM with school fixed effects has been used for TPP evaluation.

VAMs: Aggregation Approach

In the aggregation approach to value-added modeling, the VAM is used to generate a teacher effectiveness estimate for each teacher. Then, these teacher estimates are averaged for all the teachers from each TPP to estimate the effectiveness of that TPP.

The process of estimating TPP effectiveness using the aggregation approach begins with selection of one of the three specific statistical model types listed above. Although a variety of VAMs could be used with the aggregation approach, in practice, states have only used Sanders and Horn's (1994) multiple-year value-added (mixed) model with the aggregation approach for estimating TPP effectiveness, which is commonly referred to EVAAS and formerly as TVAAS (Tennessee Value-Added Assessment System). The EVAAS value-added statistical model is a repeated measures, mixed model that uses all available tests scores from the past 5 years to estimate each teacher's contribution to growth in a student's test scores (Ballou, Sanders, & Wright, 2004; TSBOE, 2010; Wright et al., 2010). EVAAS estimates of teachers' effectiveness are based on the extent to which their students consistently exceed or fall below the district average gains for their grade and subject (Ballou et al., 2004). Once calculated, the teacher effects are averaged for all the teachers from a TPP and the averages are used to compare the performance of TPPs. Similar to some of the non-VAMs discussed previously, a major criticism of the EVAAS model is that it does not include other variables such as student, classroom, or school characteristics that may also affect student test scores. The extent to which these variables are adequately controlled by the EVAAS model which requires at least three prior test scores for every student in the database is only beginning to be empirically investigated. The complexity of the model also has led some to raise concerns about transparency (see the appendix in the online supplement located on the website referenced in the abstract for more information about the EVAAS model).

VAMs: Direct Estimation Approach

The final approach used by researchers to estimate TPP effectiveness is the VAM which directly estimates TPP effectiveness. This direct estimation approach departs from previously described value-added approaches by adding a TPP indicator variable for each TPP in the analysis to the model. The TPP indicator variable designates whether a teacher was or was not prepared by a particular TPP.[4] In this way, the effect of each TPP is estimated just like any other dichotomous independent variable in the model. That is, the coefficients on each TPP indicator variables provide estimates of the magnitude of the average gain (or loss) that students received on their test scores when taught by teachers from a particular TPP, adjusted for all other variables in the model. The magnitude of gain (or loss) is compared with students of teachers who were not prepared by any of the TPPs for which there is an indicator variable. For example, the models can be estimated with only one TPP indicator variable included in the analysis, which would compare teachers from the included TPP with all other teachers of the students with test scores included in the analysis. Researchers in North Carolina (Henry, Thompson, Fortner, et al., 2010), New York City (Boyd et al., 2009), and Louisiana (Noell, 2006) have used this approach using year-to-year VAMs with controls (and in some cases with fixed effects).

The choice of the omitted or reference group for the analysis poses some challenges. To calculate TPP effectiveness for each program in a state, the TPPs must each be compared with each other, teachers from other preparation routes, or some other group such as novice teachers from the TPPs to more experienced teachers. Regardless, the choice of reference group has implications for interpretation of the quantitative estimates of TPP effectiveness. Consequently,

researchers who use a VAM approach with direct estimation must thoughtfully choose a reference group against which the other TPPs are compared and think carefully about how to interpret the results—See the appendix in the online supplement located on the website referenced in the abstract for a detailed analysis of how Noell and colleagues (2008; see also Gansle et al., 2010; Noell & Burns, 2006; Noell et al., 2007; and Henry, Thompson, Fortner, et al., 2010) handled these issues.

Unlike the EVAAS model described previously, the work evaluating TPPs in North Carolina and Louisiana, and New York City which used VAM with direct estimation incorporated a wide array of controls at the student, classroom, and school levels (Boyd et al., 2009; Henry et al., 2011; Henry, Thompson, Fortner, et al., 2010; Noell, 2006).[5] These researchers use rich sets of covariates to isolate the unique contribution of each TPP on student achievement. Most of the variables used in the research on Louisiana, North Carolina, and New York City reflect a common set of student demographics and background characteristics (e.g., free lunch, disability, and limited English proficiency status), classroom peer characteristics, and school characteristics derived from data frequently found in state data systems (see the appendix in the online supplement located on the website referenced in the abstract for more details). As Boyd and colleagues (2009) noted and as we pointed out above, programs "supply high quality teachers by a combination of recruitment and selection of potentially excellent teaching candidates and by adding value to teaching ability" (p. 429). The analytical models attempt to hold TPPs accountable for both selection into the program and preparation by the program and therefore, the models omit most teacher characteristics other than the teachers' TPP to avoid bias in the estimates of effectiveness. In other words, the models do not include variables that have been used in other studies to examine teacher quality such as a teacher's SAT score or holding a graduate degree because all factors that may influence or be influenced by a teacher's selection into, and graduation from, a particular TPP should be captured in the TPP indicator variable. Noell and colleagues (2007) included a control for teacher absences, whereas Gansle and colleagues (2010) and Henry, Thompson, Fortner, et al. (2010) controlled for teacher experience and in the latter case, for teaching out-of-field. These variables, if uncontrolled, could bias the estimates of TPP effectiveness due to factors beyond the control of the programs, such as teachers being assigned to teach out-of-field.[6]

Hopefully, the description of estimation decisions that are required has illuminated the fact that these technical decisions have important policy implications that bear directly on what evaluators intend to hold TPPs accountable for. The choice of non-VAMs including the Value Table/Transition growth model or the SGP model implies that teachers and TPP are accountable for all the changes in student test scores from one year to the next. The VAMs using aggregation and VAMs using direct estimation are efforts to isolate the effects of teachers from other factors affecting student test score gains, including measurement error. In the most common VAM using aggregation, EVAAS, the full complement of available test scores for each student are used to isolate teachers' effects. In the VAMs using direct estimation, additional student, classroom, school, and some teacher characteristics, specifically, those deemed to influence student test scores and lie beyond the control of a TPP, are included. In truth, either the aggregation or the direct estimation of TPP effects can incorporate covariates or fixed effects, but the choices involve assumptions about the influences on student test scores, what TPPs are to be held accountable for, and to what the effects of TPPs are to be compared. Does a state want all TPPs held to a common, statewide standard or should the teachers be compared only with other teachers within their same schools (school fixed effects) to control for unmeasured school influences on student test scores and the nonrandom pattern of graduates of TPP programs taking positions in different types of schools? Does a state want to compare novice graduates of each TPP with more experienced teachers or with other novice teachers? If the latter is preferred, should all in-state TPP programs be compared with the same group (e.g., all novice teachers not prepared by in-state TPP programs), to each other, or to all teachers not prepared by the TPP? Ultimately, decisions made in answering these questions will define how the quantitative estimates of the TPP effects on student test scores should be interpreted. Reporting decisions, discussed next, should be sensitive to the implications of the estimation decisions to accurately convey the meaning of estimates of TPP effects on student test scores.

Reporting Decisions

As part of their RttT accountability responsibilities, RttT recipient states plan to publicly release the results of their TPP evaluations in the form of report cards. This implies that the reports releasing the information should be easily digestible by the public at large yet also provide sufficient detail to stakeholders about the magnitude of differences in TPP effectiveness. Prior reports vary in the way they present the findings, with some researchers reporting the actual numerical estimates of TPP effectiveness, others grouping programs into performance bands, and still others focused on whether each TPP meets some performance threshold (e.g., Are the effects of a TPP significantly different from the "comparison group" using traditional tests of statistical significance?).

TPP assessments in Louisiana, North Carolina, and New York City directly report the TPP effectiveness estimates. The Louisiana research typically presents the effect estimates along with a 68% confidence interval (Gansle et al., 2010; Noell & Burns, 2006; Noell et al., 2007; Noell et al., 2008). Boyd and colleagues (2009), who have examined

New York City TPPs, simply present a graph that plots effect sizes along with confidence intervals; however, they do not identify individual TPPs.

Henry and colleagues (2011) also reported quantitative estimates for North Carolina TPPs. However, because the tests used in these analyses were standardized to have a mean of zero and standard deviation of one, the TPP estimates are in standard deviation units and not easily interpretable. For this reason, they use information about test score standard deviations and average annual test score gains to convert some of these effects into "additional days of instruction" that assume learning occurs at the same daily rate throughout the school year. For example, an elementary mathematics student taught by a recent graduate of University of North Carolina-Asheville's TPP was estimated to have the equivalent of approximately 7.5 days of additional schooling than she or he would have had if taught by another novice teacher prepared outside the University of North Carolina system (Henry et al., 2011).

Because of the difficultly in interpreting the magnitude of an effect when presented as just a statistical estimate, TPP evaluations often report effect sizes or other metrics to aid in interpretation. Louisiana provides effect sizes that indicate how different new teachers from a specific TPP are from more experienced teachers. As stakeholders may also want to know how new teachers from one TPP compare with new teachers from other TPPs, Louisiana also classifies each TPP into one of five performance band levels based on the results of the analysis. These levels classify the programs based on their relative (to new and experienced teachers) outcomes. For example, a program "for which there is evidence that new teachers are more effective than experienced teachers" is coded to Level 1, whereas Level 5 is restricted to TPPs "whose effect estimate is statistically significantly below the mean for new teachers" (Gansle et al., 2010). This approach provides lots of ease to interpret information. However, important information can be lost using this approach, namely, "Are some of the programs whose effect estimate is below the mean for new teachers better or worse than others?"[7]

Reporting the effects of TPPs involves many decisions. What is the best metric for reporting the magnitude of the estimated effects? Are "equivalent days of schooling," standard deviation units, or the original scale scores most informative? Should programs be grouped into performance categories, tested for differences with one another, or tested for differences with a particular reference group, such as novice teachers who did not come from traditional in-state preparation programs at public institutions of higher education? As better methods of reporting information are developed, the formats and amount of information may be targeted to different audiences. Detailed graphical displays, perhaps like the "thermometer graphs," used in North Carolina may be informative for the faculty and staff of individual TPP (Henry et al., 2011), but state policy makers may value summary lists of TPPs by performance category more than detailed displays.

Conclusion

The complexity behind estimating and reporting TPP effects on student test scores goes beyond setting up the sophisticated data systems that link program graduates to practicing teachers to student test scores. Although the RttT Fund and other federal initiatives urge greater accountability for publicly funded TPPs, the decisions required to tie TPP graduates to the achievement of their students must be made with an understanding of the consequences of each of the myriad decisions involved. A series of selection, estimation, and reporting decisions are required, including determining what tests teachers can reasonably be held accountable for, which teachers should be included and which are excluded, whether highly technical VAMs will produce the most accurate estimates of the effects of TPPs on student performance or whether more transparent and simpler methods should be preferred, and how much information to report. Understanding the options as well as their strengths and weaknesses along with open dialogue on the insights and consequences of certain decisions will lead to better estimates of TPP effectiveness.

From the analysis of the descriptions of the various state systems in place and under development, we can discern a few important conclusions about the current efforts to include measures of TPP effects on student test scores. First, the overall goal of most of these efforts to date has been to obtain unbiased estimates of the overall effects of TPP on student achievement. The goal is causal inference, but without random assignment of teachers and students to schools and classes, numerous assumptions must be made (Reardon & Raudenbush, 2009) for us to believe that the causal effects of TPP have been isolated. However, as the purpose of the exercise is not to deliberatively manipulate (or assign) the "treatment," in this case the TPP, to be able to estimate the effect on a specified study population, but rather to study the effects in situ better and worse choices can be made. In place of the strict standard of random assignment, we can establish criteria that guide the selection of better choices. We believe that several criteria should be considered as states move forward with TPP effect estimates: (a) accuracy, (b) fairness, (c) transparency, and (d) inclusiveness.

Accuracy. First, we must accept that we do not have any means to measure or estimate a TPP program's *true effect* on student test scores. Therefore, we must rely on traditional criteria for our efforts such as reliability and validity. Effect estimates that are unreliable can produce year-to-year swings in program effect estimates that would provide little guidance about where performance problems exist. Although the true effect is illusive, it will be important to validate the effects on student test scores estimated for TPP evaluations

with other, independent measures of high quality instruction by TPP graduates, such as direct observation of teachers using observation instruments with desirable psychometric properties and student surveys that have been shown to measure high quality instruction or other positive attributes of effective teachers such as building and maintaining warm relationship with students.

Fairness. One way of viewing fairness is that the TPP are neither advantaged nor penalized by decisions of their graduates who are beyond their control. Preparing teachers who choose to teach in challenging schools or challenging students should not be a factor that affects the TPP effect estimates. All VAMs that are being used for the evaluation of TPPs attempt to isolate the effects of teachers and remove the influence on student test scores that are beyond the control of TPP. Some VAMs attempt to do this directly with student, classroom, and school covariates, and others do it with extensive student test scores, but it does not seem reasonable to expect that differences in students or schools have no bearing on teacher or TPP effectiveness. It is more difficult to discern whether the non-VAMs deal with these choices in a way that is fair to TPPs whose graduates choose more challenging students or schools and should along with the VAMs be investigated further in this regard.

Transparency. Transparency is an obvious, but difficult, criterion to satisfy because the most transparent approaches may fail the tests of accuracy and fairness, whereas approaches that seem accurate and fair require more complex modeling procedures that are opaque to many teachers and policy makers. Clear and understandable explanations will need to be offered when complex and sophisticated estimation approaches are used. Perhaps this suggests that the reporting burdens in terms of clarity and ease of interpretation will be greater when using VAMs than the non-VAMs. Ultimately, there may be a trade-off between accuracy and fairness on one hand and transparency on the other.

Inclusiveness. Finally, the inclusiveness criterion suggests that the current testing regime that has been put into place largely to meet NCLB requirements may fall short in the longer term and omit significant subprograms of larger traditional TPP, such as early education programs or special education. In addition, incorporating student test scores into the evaluation of TPP is only one of many measures that could be used to measure the effects of the graduates of these programs on student outcomes. Student engagement, graduation rates, and direct measures of high quality instruction are but a few of the potential measures that could be added to TPP evaluation for a more inclusive and comprehensive perspective on the program's effects on students. Although we believe it is reasonable to first attempt to evaluate TPPs utilizing the tests that are currently available, attention should be given to incorporating a broader set of teacher performance measures.

It is also important to note that the current efforts to incorporate student test scores into TPP evaluation partially fulfill one of the four purposes for evaluation, accountability/oversight, and do not directly or necessarily address the other three purposes for which evaluations are conducted: program improvement, assessment of merit and worth, or knowledge development (Mark, Henry, & Julnes, 2000). A TPP's overall effect on student test scores does not directly provide information concerning program improvement. To begin to perform this function, the effects of selection and retention processes will need to be distinguished from effects of the actual preparation processes to determine which of these processes are responsible for TPPs achieving their effects. Is the selection of high quality candidates or the preparation program itself that has the greatest effect on teachers' effectiveness or a combination of both? In addition, the extent to which variation in the programmatic components of TPP systematically relate to variations in teacher effectiveness will need to be explored. This has begun with research by Boyd et al. (2009; see also Boyd et al., 2006), who found that student test score gains associated with specific TPP components, such as exposure to specific content or the extent of clinical experiences. This type of research is in its nascent stage and, unfortunately, the data that will be needed to expand this effort is not currently available in many longitudinal data systems.

The overall assessment of merit and worth of TPPs requires much broader information than the effects of public in-state TPPs on student test scores. Assessing the true merit of TPP and their full worth to society requires understanding the counterfactual: What would student outcomes be in the absence of the TPPs being evaluated? How would teachers be prepared in the absence of these programs? How much would students learn and be able to do if the TPPs being evaluated did not exist? At a minimum, this suggests that equal attention be given to the effectiveness of all alternative TPP such as Teach For America, Visiting International Faculty, each of the various alternative or lateral entry programs supplying teachers in the state (whether they are located in the state), and state-to-state reciprocal licensing arrangements. However, this by itself will not be entirely adequate for assessing merit and worth of the traditional TPP at public and private institutions because we do not know how the teacher labor market would respond in the absence of these programs or about other outcomes of public schools such as graduation or participation as a citizen that are not attended to in the test scores.

Finally, we have much ground to cover for knowledge development, some of which has already been alluded to above. The empirical work to date has established "proof of concept" for incorporating student test scores into the evaluation of TPP. It shows that a sufficient signal can be found in

student test scores to reliably attribute effects to individual TPP even with the "noise" of confounding variables and nonrandom sorting of teachers and students into schools and classrooms. However, we have just begun to understand some of the differences in the approaches and models. Much technical work remains for assessing the accuracy and fairness of current methods and improving their transparency as well as developing measures that will increase the inclusiveness of TPP effects on valued student outcomes that go beyond test scores.

Acknowledgments

The authors gratefully acknowledge helpful comments and support from Alisa Chapman, Alan Mabe, Erskine Bowles, the Council of Education Deans of the University of North Carolina, Stephanie L. Knight, and two anonymous reviewers.

Authors' Note

The authors take full responsibility for the research, interpretations, and conclusions included in the manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. North Carolina offers end-of-course tests in Algebra I, Algebra II, biology, chemistry, English I, geometry, physical science, physics, U.S. history, and civics and economics courses. Student performance in English I, Algebra I and II, geometry, biology, chemistry, physics, and physical science were used in these analyses.
2. In response to No Child Left Behind and other initiatives, states have developed several types of growth models, including growth to proficiency models, value table/transition models, projection models, and student growth percentile (SGP) models. For more information on these approaches, see O'Malley et al., 2009.
3. Whether, or to what extent, Massachusetts will use the SGP scores to estimate teacher preparation program effectiveness is unknown by the authors at this time.
4. An indicator variable is a variable that takes on a value of 1 when, for instance, a person has a characteristic of interest and 0 when a person does not have a characteristic. For example, men would be coded 1 on a male indicator variable and women 0.
5. Proponents of the Education Value-Added Assessment System/Tennessee Value-Added Assessment System (EVAAS/TVAAS) argue and demonstrate (Ballou, Sanders, & Wright, 2004) that the use of multiple student test scores in the EVAAS/TVAAS models effectively subsumes many of these unobserved factors.
6. In an additional analysis, Henry, Thompson, Fortner, Zulli, and Kersham (2010) also incorporated teacher level covariates into some of their analyses, including SAT scores, high school rank, and high school grade point average, but these are initial attempts to parse the effects of preparation from selection and are not considered estimates of the overall effects of TPPs.
7. Note that in the Louisiana approach, placement in these levels is not always tied to statistical significance. In other words, TPP can be labeled as comparatively over or under achieving although they are not reliably different from the average TPP.

References

Aaronson, D., Barrow, L., & Sander, W. (2003). *Teachers and student achievement in the Chicago Public High Schools*. Washington, DC: Federal Reserve Bank of Chicago. (Working Paper Series: WP-02–28)

Auty, W., Bielawski, P., Deeter, T., Hirata, G., Hovanetz-Lassila, C., Rheim, J., . . . Williams, A. (2008). *Implementer's guide to growth models*. Retrieved from The Council of Chief of State School Officers website: http://www.ccsso.org/Documents/2008/Implementers_Guide_to_Growth_2008.pdf

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment for teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37-65.

Betebenner, D. (2009a). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 24*, 42-51.

Betebenner, D. (2009b). *A primer on student growth percentiles*. Retrieved from http://www.cde.state.co.us/cdedocs/Research/PDF/Aprimeronstudentgrowthpercentiles.pdf

Boyd, D. J., Grossman, P., Lankford, H., Loeb, S., Michelli, N. M., & Wyckoff, J. (2006). Complex by design: Investigating pathways into teaching in New York City schools. *Journal of Teacher Education, 57*, 155-166.

Boyd, D. J., Grossman, P., Lankford, H., Loeb, S., Michelli, N. M., & Wyckoff, J. (2009). Teacher preparation and student achievement. *Education Evaluation and Policy Analysis, 31*, 416-440.

Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models* (Tech. Rep.). Princeton, NJ: Educational Testing Service.

Buzick, H. M., & Laitusis, C. C. (2010). Using growth for accountability: Measurement challenges for students with disabilities and recommendations for research. *Educational Researcher, 39*, 537-544.

California Department of Education. (2011). *2010 STAR Test Results: About 2010 STAR*. Retrieved from http://star.cde.ca.gov/star2010/aboutSTAR.asp

Clotfelter., Ladd, H., & Vigdor, J. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review, 26*(6), 673-682.

Clotfelter., Ladd, H., & Vigdor, J. (2010). Teacher credentials and student achievement in high school: A cross-subject analysis with student fixed effects. *The Journal of Human Resources, 45*(3), 655-681.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity.* Washington, DC: USGPO.

Crowe, E. (2010). *Measuring what matters: A stronger accountability model for teacher education*. Washington, DC: Center for American Progress. Retrieved from http://www.americanprogress.org/issues/2010/07/pdf/teacher_accountability.pdf

Darling-Hammond, L., Berry, B., & Thoreson, A. (2001). Does teacher certification matter? Evaluating the evidence. *Educational Evaluation and Policy Analysis, 23*(1), 55-77.

The District of Columbia. (2010). *Race to the Top Proposal, Phase 2*. Retrieved from the U.S. Department of Education website: http://www2.ed.gov/programs/racetothetop/phase2-applications/index.html

Florida Department of Education. (2005). *FCAT Handbook: A resource for educators*. Tallahassee, FL: Florida Department of Education.

Florida Department of Education. (2009a). *20072008 FCAT learning gains results*. Retrieved from http://www.tampabay.com/blogs/gradebook/sites/tampabay.com.blogs.gradebook/files/images/typepad-legacy-files/54957.june-2009-teacher-quality-data.pdf

Florida Department of Education. (2009b). *Memorandum: Data on program completers' impact on K-12 student learning to meet continued approval requirements*. Retrieved from http://www.tampabay.com/blogs/gradebook/sites/tampabay.com.blogs.gradebook/files/images/typepad-legacy-files/54991.11-20-09teacher-preparation.pdf

Florida Department of Education. (2009c). *Overall performance of 2007-08 teacher preparation program completers teaching reading and mathematics Grade 4-10 during 2008-09*. Retrieved from http://www.tampabay.com/blogs/gradebook/sites/tampabay.com.blogs.gradebook/files/images/typepad-legacy-files/54991.tq_deans_list_0809-kh.pdf

Florida Department of Education. (n.d.). Florida End-of-Course (EOC) Assessments. Retrieved from http://fcat.fldoe.org/eoc/

The Florida Senate, Committee on Education Pre-K-12. (2009). *Teacher quality: Issue Brief 2010-313*. Retrieved from http://archive.flsenate.gov/data/Publications/2010/Senate/reports/interim_reports/pdf/2010-313ed.pdf

Fuchs, D., Compton, D. L., Fuchs, L. S., Bryant, J., & Davis, G. N. (2008). Making "secondary intervention" work in a three-tiered responsiveness-to-intervention model: Findings from the first-grade longitudinal reading study of the national research center on learning disabilities. *Reading and Writing, 21*(4), 413-436.

Gansle, K. A., Noell, G. H., Knox, R. M., & Schafer, M. J. (2010). *Value added assessment of teacher preparation in Louisiana: 2005-2006 to 2008-2009*. Retrieved from http://regents.louisiana.gov/assets/docs/TeacherPreparation/2010VATechnical082610.pdf

Glazerman, S., Mayer, D., & Decker, P. (2006). Alternative routes to teaching: The impacts of Teach for America on student achievement and other outcomes. *Journal of Policy Analysis and Management, 25*, 75-96.

Goldhaber, D. D., & Brewer, D. J. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis, 22*(2), 129-145.

Goldschmidt, P., Roschewski, P., Choi, K., Auty, W., Hebbler, S., Blank, R., & Williams, A. (2005). *Policymakers' guide to growth models for school accountability: How do accountability models differ?* The Council of Chief of State School Officers. Retrieved from http://www.ccsso.org/Documents/2009/Guide_to_United_States_2009.pdf

Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). The effect of school resources on student achievement. *Review of Educational Research, 66*, 361-396.

Henry, G. T., Bastian, K. C., & Smith, A. A. (2012). Scholarships to recruit the "best and brightest" into teaching: Who is recruited, where do they teach, how effective are they, and how long do they stay? *Educational Research, 41*(3), 83-92.

Henry, G. T., Thompson, C. L., Bastian, K. C., Fortner, C. K., Kershaw, D. C., Marcus, J. V., & Zulli, R. A. (2011). *UNC Teacher Preparation Program Effectiveness Report*. Chapel Hill: Carolina Institute for Public Policy, University of North Carolina at Chapel Hill.

Henry, G. T., Thompson, C. L., Bastian, K. C., Fortner, C. K., Kershaw, D. C., Purtell, K. M., & Zulli, R. A. (2010). *Portal report: Teacher preparation and student test scores in North Carolina*. Chapel Hill: Carolina Institute for Public Policy, University of North Carolina at Chapel Hill.

Henry, G. T., Thompson, C. L., Fortner, C. K., Zulli, R. A., & Kershaw, D. C. (2010). *The impact of teacher preparation on student learning in North Carolina public schools*. Chapel Hill: Carolina Institute for Public Policy, University of North Carolina at Chapel Hill.

Horner, M. (2009). *Quantifying student growth: Analysis of the validity of applying growth models to the California Standards Test*. Retrieved from http://digitallibrary.usc.edu/assetserver/controller/item/etd-Horner-2920.pdf;jsessionid=81F14E4D9DD8AB29986304A535E41942

Kane, T., Rockoff, J., & Staiger, D. (2006). *What does certification tell us about teacher effectiveness? Evidence from New York City*. NBER Working Paper 12155. Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w12155

Mark, M. M., Henry, G.T., & Julnes, G. (2000). *Evaluation: an integrated framework for understanding, guiding, and improving public and nonprofit policies and programs.* San Francisco, CA: Jossey-Bass.

Murawski, W. W., & Lochner, W. W. (2011).Observing co-teaching: What to ask for, look for, and listen for. *Intervention in School and Clinic, 46*, 174-183.

National Council for Accreditation of Teacher Education. (2010). *Professional Standards for the Accreditation of Teacher Preparation Institutions*. Washington, DC: National Council for the Accreditation of Teachers.

National Research Council. (2010). *Preparing teachers: Building evidence for sound policy*. Washington, DC: National Academies Press.

Noell, G. H. (2006). *Technical report of: Assessing teacher preparation program effectiveness: A pilot examination of value added approaches (2006)*. Retrieved from http://www.laregentsarchive.com/Academic/TE/technical_report.pdf

Noell, G. H., & Burns, J. L. (2006). Value-added assessment of teacher preparation: An illustration of emerging technology. *Journal of Teacher Education, 57*, 37-50.

Noell, G. H., & Burns, J. M. (2007). *Value added teacher preparation assessment overview of 2006-07 study*. Retrieved from http://regents.louisiana.gov/assets/docs/TeacherPreparation/NarrativeDescriptionof2006-07ValueAddedStudy10.24.07.pdf

Noell, G. H., Porter, B. A., & Patt, R. M. (2007). *Value added assessment of teacher preparation in Louisiana: 2004–2006*. Retrieved from http://regents.louisiana.gov/assets/docs/TeacherPreparation/VAATPPTechnicalReport10-24-2007.pdf

Noell, G. H., Porter, B. A., Patt, R. M., & Dahir, A. (2008). *Value added assessment of teacher preparation in Louisiana: 2004-2005 to 2006-2007*. Retrieved from http://www.laregentsarchive.com/Academic/TE/2008/Final%20Value-Added%20Report%20(12.02.08).pdf

Nye, B., Konstantopoulos, S., & Hedges, L. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*, 237-257.

O'Malley, K., Auty, W., Bielawski, P., Bernstein, R., Boatman, T., Deeter, T., . . . Blank, R. (2009). *Guide to United States Department of Education Growth Model Pilot Program 2005-2008*. The Council of Chief of State School Officers. Retrieved from http://www.ccsso.org/Documents/2009/Guide_to_United_States_2009.pdf

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy, 4*(4), 492-519.

Rivkin, S., Hanushek, E., & Kain, J. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417-458.

Rockoff, J. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review, 94*(2), 247-252.

Sanders, W. L., & Horn, S. (1994). The Tennessee value-added assessment system (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*, 299-311.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement.* Retrieved from the University of Tennessee Value-Added Research and Assessment Center website: http://heartland.org/sites/all/modules/custom/heartland_migration/files/pdfs/3048.pdf

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee value-added assessment system: A quantitative, outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools* (pp. 137-162). Thousand Oaks, CA: Corwin Press.

The State of Delaware. (2010). *Race to the Top Proposal, Phase 1*. Retrieved from the U.S. Department of Education website: http://www2.ed.gov/programs/racetothetop/phase1-applications/index.html

The State of Florida. (2010). *Race to the Top Proposal, Phase 2*. Retrieved from the U.S. Department of Education website: http://www2.ed.gov/programs/racetothetop/phase2-applications/index.html

The State of Georgia. (2010). *Race to the Top Proposal, Phase 2*. Retrieved from the U.S. Department of Education website: http://www2.ed.gov/programs/racetothetop/phase2-applications/index.html

The State of Hawaii. (2010). *Race to the Top Proposal, Phase 2*. Retrieved from the U.S. Department of Education website: http://www2.ed.gov/programs/racetothetop/phase2-applications/index.html

The State of Maryland. (2010). *Race to the Top Proposal, Phase 2*. Retrieved from the U.S. Department of Education website: http://www2.ed.gov/programs/racetothetop/phase2-applications/index.html

The State of Massachusetts. (2010). *Race to the Top Proposal, Phase 2*. Retrieved from the U.S. Department of Education website: http://www2.ed.gov/programs/racetothetop/phase2-applications/index.html

The State of Massachusetts. (2011). *MCAS student growth percentiles: Interpretive guide*. Retrieved from the U.S. Department of Education website: http://www.doe.mass.edu/mcas/growth/InterpretiveGuide.pdf

The State of New York. (2010). *Race to the Top Proposal, Phase 2*. Retrieved from the U.S. Department of Education website: http://www2.ed.gov/programs/racetothetop/phase2-applications/index.html

The State of North Carolina. (2010). *Race to the Top Proposal, Phase 2*. Retrieved from the U.S. Department of Education website: http://www2.ed.gov/programs/racetothetop/phase2-applications/index.html

The State of Ohio. (2010). *Race to the Top Proposal, Phase 2*. Retrieved from the U.S. Department of Education website: http://www2.ed.gov/programs/racetothetop/phase2-applications/index.html

The State of Rhode Island. (2010). *Race to the Top Proposal, Phase 2*. Retrieved from the U.S. Department of Education website: http://www2.ed.gov/programs/racetothetop/phase2-applications/index.html

The State of Tennessee. (2010). *Race to the Top Proposal, Phase 1*. Retrieved from the U.S. Department of Education website: http://www2.ed.gov/programs/racetothetop/phase2-applications/index.html

Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M. E., Roth, J., . . . Resnick, M. B. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics, 29*, 11-36.

Tennessee State Board of Education. (2009). *Report card on the effectiveness of teacher training programs*. November Report. Retrieved from http://www.tn.gov/sbe/TeacherReportCard/2009/2009%20Report%20Card%20on%20Teacher%20Effectiveness.pdf

Tennessee State Board of Education. (2010). *Report card on the effectiveness of teacher training programs*. December Report. Retrieved from http://www.tn.gov/sbe/Teacher%20Report%20Card%202010/2010%20Report%20Card%20on%20the%20Effectiveness%20of%20Teacher%20Training%20Programs.pdf

Voorhees, R. A., Barnes, G., & Rothman, R. (2003). *Data systems to enhance teacher quality*. Denver, CO: State Higher Education Executive Officers.

U.S. Department of Education. (2010). Overview information; Race to the Top Fund. *Federal Register, 75*(71), 19496-19531.

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.

Wright, S. P., White, J. T., Sanders, W. L., & Rivers, J. C. (2010). *SAS® EVAAS® Statistical Models (white paper)*. Cary, NC: SAS Institute.

Xu, Z., Hannaway, J., & Taylor, C. (2011). Making a difference? The effects of Teach For America in high school. *Journal of Policy Analysis and Management, 30*(3): 447-469.

About the Authors

Gary T. Henry holds the Duncan MacRae '09 and Rebecca Kyle MacRae Professorship of Public Policy in the Department of Public Policy and serves as a fellow with the Carolina Institute for Public Policy and the Frank Porter Graham Institute for Child Development at the University of North Carolina at Chapel Hill. He specializes in education policy, educational evaluation, teacher quality research, and quantitative research methods.

David C. Kershaw is an assistant professor in the Department of Political Science at Slippery Rock University of Pennsylvania. Dr. Kershaw's research areas include education policy, political behavior, and research methods.

Rebecca A. Zulli, PhD, is project manager for the ATOMS Project in the Department of Elementary Education at NC State University. Her research focuses on the evaluation of educational innovation and intervention in the areas of teacher preparation, professional development, and student achievement.

Adrienne A. Smith, PhD, is a research associate at Horizon Research, Inc. Her research focuses on educational policy, program evaluation, and research methods.

Errata

Journal of Teacher Education
64(1) 107

DOI: 10.1177/0022487112470806
http://jte.sagepub.com

Knight, S. L., Edmondson, J., Lloyd, G. M., Arbaugh, F., Nolan Jr., J., Whitney, A. E., & McDonald, S. P. (2012). Examining the complexities of assessment and accountability in teacher education. *Journal of Teacher Education, 63*, 301-303. (Original doi: 10.1177/0022487112460200)

The above-referenced editorial explains that technical information for several articles would be placed in online-only appendices, and that links to the appendices would be provided in abstracts. However, links to the appendices were omitted from the abstracts, and for one article (by Gansle, Noell, and Burns) the appendix was included within the article as a normal appendix, rather than separated into an online-only appendix. The online appendices for these articles can be found at http://JTE.sagepub.com/supplemental; by viewing the November/December 2012 issue online at http://JTE.sagepub.com/content/63/5.toc; or by visiting the respective articles at http://JTE.sagepub.com/content/63/5/318 and http://JTE.sagepub.com/content/63/5/335.

Gansle, K. A., Noell, G. H., & Burns, J. M. (2012). Do student achievement outcomes differ across teacher preparation programs? An analysis of teacher education in Louisiana. *Journal of Teacher Education, 63*, 304-317. (Original doi: 10.1177/0022487112439894)

The appendix for the above-referenced article should have been provided as a separate online-only appendix, without inclusion in the print issue. A link to the online-only appendix should have been provided in the abstract.

Plecki, M. L., Elfers, A.M., & and Yugo, N. (2012). Using evidence for teacher education program improvement and accountability: An illustrative case of the role of value-added measures. *Journal of Teacher Education, 63*, 318-334. (Original doi: 10.1177/0022487112447110)

Henry, G. T., Kershaw, D. C., Zulli, R. A., & Smith, A. A. (2012). Incorporating teacher effectiveness into teacher preparation program evaluation. *Journal of Teacher Education, 63*, 335-355. (Original doi: 10.1177/0022487112454437)

In each of the above-referenced articles, the abstract should have included a link to the article's online appendix, which is available at http://JTE.sagepub.com/supplemental. The online appendix can also be found by viewing the November/December 2012 issue online at http://JTE.sagepub.com/content/63/5.toc or by viewing the respective articles online at http://JTE.sagepub.com/content/63/5/318 and http://JTE.sagepub.com/content/63/5/335.