Journal of Teacher Education
63(5) 318–334

DOI: 10.1177/0022487112447110
http://jte.sagepub.com
SAGE

Using Evidence for Teacher Education Program Improvement and Accountability: An Illustrative Case of the Role of Value-Added Measures

Margaret L. Plecki[1], Ana M. Elfers[1], and Yugo Nakamura[1]

Abstract

In this article, the authors consider what can be learned from limited forms of evidence, for purposes of accountability and improvement of teacher education programs. They begin with a review of recent research on how evidence has been used to examine the effectiveness of teacher preparation and development. Using empirical evidence from a state with limited data capacity, they illustrate what can be learned from value-added measures as one form of evidence. As a case in point, the value-added scores for fifth-grade teachers are used to answer the question: To what extent are teachers' years of experience and the institutions from which they obtained their teacher training related to student achievement? The authors conclude with a discussion of the use of evidence by shifting the focus of accountability from simply responding to external requirements to developing internal practices that generate knowledge for improvement, and argue for collective responsibility among multiple stakeholders.

Keywords

school/teacher effectiveness, teacher research, educational policy, quantitative research

Introduction and Overview of the Study

Having a well-qualified teacher in every classroom is a centerpiece of education reform. Although this goal can be simply stated, crafting and implementing strategies for its accomplishment is vastly more complex. Clearly, teacher preparation has a prominent role to play in addressing the challenge of improving the quality of teaching and learning, especially for students who have been historically underserved. In recent years, increasing attention has been paid to the quality of educator preparation programs, with a growing numbers of stakeholders participating in the debate over how to hold schools of education accountable to prepare high-quality educators. Efforts to measure the impact of preparation programs on K-12 learning have become a routine part of the conversation within an evidence-based approach. However, because student success is affected by factors beyond the educator's or preparation program's control, finding meaningful measures has proven difficult. Given these complexities, the use of multiple forms of evidence within a broader framework has the potential to respond to issues of accountability and continuous program improvement in teacher education.

Schools of education operate in an increasingly politicized landscape. Researchers, policy makers, and practitioners have been calling for the redesign of teacher education over the past several decades (Cochran-Smith, 2001; Darling-Hammond, 1997, 2006b; Duncan, 2010; Goodlad, 1991; Labaree, 1995, 2004; National Council for Accreditation of Teacher Education [NCATE], 2010; Soltis, 1987). A blue ribbon panel assembled by the NCATE called for "turning the education of teachers upside-down" to build "an entire system of excellent programs" throughout the nation (NCATE, 2010). Improvement strategies have included strengthening partnerships between teacher preparation institutions and local school districts and offering alternative routes to certification. Recent efforts to reshape teacher education have placed a central emphasis on the systematic collection of evidence to inform decisions about how teacher preparation can best be improved, including data about teaching practice, student learning, and persistence in the teaching profession (Crowe, 2010; Ludlow et al., 2011; Wilson, Floden, & Ferrini-Mundy, 2001; Wineburg, 2006).

Federal and state policies focused on teaching effectiveness have intensified the need for robust and reliable outcome

[1]University of Washington, Seattle, USA

Corresponding Author:
Margaret L. Plecki, University of Washington, Box 353600, Seattle, WA 98195, USA
Email: mplecki@uw.edu

measures (National Research Council, 2010). States have begun to respond by building state data systems assisted in part by the federal Statewide Longitudinal Data Systems Grant Program (SLDS). Since 2005, 41 states and the District of Columbia have received at least one SLDS grant. This effort has enhanced the ability of state agencies to integrate and manage educational data from a variety of sources. A recent proposal by the U.S. Department of Education to change the accountability requirements for teacher education programs under Title II of the Higher Education Act (HEA) reflects this move toward evidence-based accountability (Sawchuck, 2011). If approved, this policy would require schools of education participating in Title II funding to report some measure of the achievement of students taught by program graduates, the placement of program graduates in high needs schools and subjects, and employer satisfaction with the quality of graduates. Given the sweeping changes being considered for improving teacher education, the need for comprehensive measures of effectiveness becomes even more pronounced (Darling-Hammond, 2006a; Goe & Holdheide, 2011).

Although a necessary initial step, simply acquiring new and improved data sets and systems will not be sufficient. Many teacher preparation institutions routinely gather information on aspects of their program; however, this information often is not used for continuous improvement and accountability, and measures typically are not consistent across institutions (Cochran-Smith & The Boston College Evidence Team, 2009; Wineburg, 2006). For data to productively inform program improvement, it must be integrated in a systematic way, with the capacity to add and amend data elements over time. Due to the expansion of approaches to teacher training, information collected should reflect different candidate experiences and the varying programs, purposes, and missions of institutions. In addition, attention should be paid to the validity and reliability of various measures, and the development of new forms of evidence where robust measures may be lacking (Cochran-Smith & The Boston College Evidence Team, 2009; Goe, Bell, & Little, 2008; Wineburg, 2006).

Multiple stakeholders, including state agencies and local partners, also share responsibility for preparing educators and must grapple to solve problems that are complex and not always within their realm of control. As such, the improvement of teacher training demands a systematic examination of factors along the continuum of an educator's career. Crowe (2010) writes, "No single measure—no matter how powerful the findings—is enough to gauge all the relevant components of teaching quality or program effectiveness" (p. 14).

The challenge for schools of education is to rigorously respond to the call for accountability and collect and analyze data for purposes of program improvement, even given limited capacity and less than ideal conditions. In this article, we consider what can be learned from limited forms of evidence, for purposes of accountability and program improvement, and explore what it might look like given the full breadth of the call. To situate our work, we begin with a literature-based discussion and a review of recent research on how evidence has been used to examine the effectiveness of teacher preparation and development. We seek to shift the focus of accountability from simply responding to external state or federal reporting requirements to internal practice that can generate knowledge for use in local programs and more broadly.

Second, using empirical evidence from a state with limited data capacity, we illustrate what can be learned from value-added measures as one form of evidence. As a case in point, the value-added scores for fifth-grade teachers in a single state context are used to answer the question, "To what extent are teachers' years of experience and the institutions from which they obtained their teacher training related to student achievement?" Due to data limitations, this approach does not answer specific program improvement questions; however, it does allow us to explore the territory and provides insight for future directions as data quality and capacity improve.

Finally, the concluding section discusses the implications of our inquiry and suggests ways to use evidence for improving teacher education and development. We also argue for collective responsibility among teacher education institutions, professional organizations, and state and local agencies as they respond to the demand for increased accountability.

Forms of Evidence Needed for Systematic Improvement and Accountability

Teacher education institutions face at least four kinds of challenges when making decisions about program improvement: (a) recruiting the types of candidates that meet state and regional labor market needs; (b) designing program content and related field experiences that produce candidates who are well prepared to begin their teaching career; (c) helping to place graduates in subject areas, grade levels, and schools where they are most needed, and helping to provide induction support; and (d) engaging in partnerships with districts and schools in the design and delivery of high-quality professional learning in the initial years, as well as ongoing professional learning throughout a teacher's career. In each of these areas, characteristics and teaching practices of candidates and their preparation programs can potentially be identified as forms of evidence for program improvement and accountability. Unfortunately, much of the research to date has been conducted using a narrow range of available data elements—often collected for other purposes. As such, they provide only a tentative understanding of program and graduate outcomes. Responding to the call for internal and external accountability means identifying and collecting evidence that appropriately addresses actual problems of

practice or poses specific questions. In contrast to mere compliance with external standards, proactive engagement with data can lead to finding "workable solutions to wider programmatic challenges" (Peck, Gallucci, & Sloan, 2010, p. 460).

Evidence about applicants. The recruitment of potential teacher candidates provides a starting place to consider sources of data for program improvement. Even before embarking on recruitment, it is critically important for teacher preparation institutions to understand the kinds of teachers needed in the schools and districts that hire the majority of their graduates. Selective recruitment involves considering ways in which the applicant pool can be diversified and efforts targeted to applicants willing to work in challenging school contexts and in potential shortage areas (e.g., mathematics, science, special education, English as second language [ESL], or bilingual education). Research on teacher qualifications (data often considered when making recruitment decisions) offer mixed findings with regard to teacher effectiveness. Such measures typically include undergraduate major, grade point average (GPA), graduate education, selectivity of the undergraduate institution, and college entrance examinations. Some studies have found that teachers with stronger academic backgrounds produce larger gains for their students (e.g., Boyd, Lankford, Loeb, Rockoff, & Wyckoff, 2008; Clotfelter, Ladd, & Vigdor, 2007; Rockoff, Jacob, Kane, & Staiger, 2011), whereas others have not (e.g., Constantine et al., 2009; Harris & Sass, 2009a). The relationship between content knowledge and teacher effectiveness, particularly in the area of mathematics, has been documented in several studies (Clotfelter et al., 2007; Goldhaber & Brewer, 1997), though differences in the proxy used for content knowledge may be a determinant (Hill, Rowan, & Ball, 2005). In summarizing models of teacher effectiveness, Harris and Rutledge (2010) identified cognitive ability as a predictor. Some (e.g., Boyd et al., 2008; Rockoff et al., 2011) argued for the use of a broader set of variables when considering teaching qualifications.

Evidence about program features and candidates' preparation. An important source of evidence can be found in data collected about candidates while they are immersed in their preparation experiences. As professional preparation activities are increasingly delivered at school sites and initial pathways into teaching expand, systematic feedback from school partners can also be used to inform program development. Information about candidates' coursework, performance assessments, field placements, and evaluations they receive from those who work in participating schools and districts could all be incorporated into an evidence-based framework. Again, research about various programmatic elements and teaching effectiveness has been mixed, in part due to the paucity of available measures.

Pecheone and Chung (2006) found that performance assessments provide a measure of individual teacher competence for the purposes of licensure and a tool for teacher learning and program improvement. Peck et al. (2010) documented substantive changes in the way one program operated using performance assessments, including increased alignment of concepts and practices across program experiences. However, Harris and Sass (2009a) found no evidence that teachers' preservice training influenced student performance when linked to teachers' inservice training, their college coursework and majors, and precollege entrance exam scores.

Boyd, Grossman, Lankford, Loeb, and Wyckoff (2005) used teacher preparation as a measure of teacher quality by examining different pathways into teaching in New York City. Their findings suggest differences in teacher quality among teachers prepared in different ways, though it was unclear whether the differences were due to the preparation and support teachers received or the characteristics of individuals who entered the profession through different pathways. In a subsequent study, these researchers combined administrative data sets with detailed descriptors of the structure and content of each program (Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2009). The study found variation in the average effectiveness of teachers from different programs and identified features of the program that can make a difference in the outcomes for students. The researchers note that study data and methods are imperfect and results are suggestive rather than clearly establishing cause. In particular, the study found that more effective 1st-year teachers were produced when teacher preparation was directly linked to the classroom work of 1st-year teachers. Another study that examined alternative routes to teacher certification found that variation in student achievement was not strongly linked to preparation route or other measured teacher characteristics (Constantine et al., 2009).

An early study that used value-added measures to examine the efficacy of teacher preparation programs was conducted by Noell and his colleagues using student and teacher data from Louisiana (Noell, Porter, Patt, & Dahir, 2008). These researchers assessed the efficacy of teacher preparation programs and examined possible explanatory factors using pooled data across 3 academic years. They found differences across performance bands developed to describe teacher preparation programs and found that teachers not certificated in the content in which they were teaching were less effective than those who were content certified. More recently, researchers have used statewide data from other states to conduct comparisons of the graduates from various teacher preparation programs (e.g., North Carolina, Henry et al., 2011; Tennessee, Tennessee Higher Education Commission and State Board of Education, 2011; Washington state, Goldhaber & Liddle, 2011). The primary focus of many of these studies, however, is on individual candidate characteristics with much less attention paid to examining the training that goes on within programs. Few of these studies engaged in original data collection from sources that were not routinely available and this dearth of data severely limits what can be learned for purposes of program improvement.

Evidence about novice teachers. As teacher preparation institutions are under increasing pressure to ensure that they adequately prepare their graduates to face challenging classroom assignments, they need information about the nature of these initial placements. State administrative data sets and surveys increasingly are used to track where teacher education graduates find employment, the characteristics and learning needs of the students in those schools, and the extent to which teachers move from one classroom setting to another (e.g., Donaldson & Johnson, 2010; Hanushek, Kain, & Rivkin, 2004; Luekens, Lyter, & Fox, 2004; National Center for Education Statistics, 2005). As alternative route programs offer other avenues for teachers to enter the profession, it is important to consider the placement, retention, and mobility of teachers from these programs compared with traditionally prepared graduates (Donaldson & Johnson, 2010). These types of analyses enable us to understand the characteristics of a state's beginning teacher workforce and suggest ways in which future teachers can be better prepared for the school contexts in which they are likely to begin their careers.

For years, researchers have noted the importance of providing high-quality learning opportunities focused on the unique needs of novice teachers (Johnson et al., 2001). For example, Smith and Ingersoll (2004) found that beginning teachers who had mentors from the same subject field were less likely to change schools or leave teaching after the 1st year. In a review of 15 studies on the effects of induction on various outcomes for beginning teachers, Ingersoll and Strong (2011) found empirical support for the claim that teacher mentoring programs, in particular, have a positive impact. The exception to this trend was a large randomized controlled trial that investigated the impact of comprehensive induction (Glazerman et al., 2010). This study found some positive effects on student achievement but no effects on either teacher retention or teachers' classroom practices. However, methodological concerns have been raised regarding differences between the treatment and control groups, as well as the outcome measure used for teachers' classroom practices.

Teaching experience is perhaps the only factor that is consistently related to student outcomes. The relationship between teacher experience and student outcomes is most pronounced in the first several years of experience (Boyd et al., 2008; Clotfelter et al., 2007; Hanushek, Kain, O'Brien, & Rivkin, 2005). As we have seen, many studies consider the relationship between teachers' characteristics, such as experience and measures of student achievement, but relatively few have explored the potential relationship between program features, instructional practices, and student outcomes.

Evidence about the teacher workforce. Although teacher preparation programs are most closely linked to graduates and novice teachers, it is also important to understand the characteristics and needs of the overall teacher workforce in a state or region. Issues of teacher supply and demand have a direct bearing on decisions regarding the number of teachers and type of preparation needed in a given time period or region (Elfers, Plecki, & Knapp, 2006; Guin, 2004).

The observational component of teacher evaluations has received renewed focus as states consider multiple ways to consider teacher effectiveness. Using a rigorous system of teacher evaluation based on a structured observational protocol, several studies have illustrated how teacher evaluations, such as Danielson's (1996) *Framework for Teaching*, are related to student achievement, and in some cases, correlated with value-added measures (Kane, Taylor, Tyler, & Wooten, 2010; Kimball, White, Milanowski, & Borman, 2004; Milanowski, 2004). In a pilot study, Grossman and her colleagues (2010) used CLASS, a structured observational protocol (Pianta, Hamre, Haynes, Mintz, & La Paro, 2006) in addition to teacher logs and student work in secondary English/language arts instruction, and found consistent evidence that high value-added teachers have a different profile of instructional practices than low value-added teachers. Jacob and Lefgren (2008) found that principals can effectively identify those teachers who produce the largest and smallest standardized achievement gains in their schools but have far less ability to distinguish between teachers in the middle of the distribution, whereas Harris and Sass (2009b) found principals' subjective evaluations and teacher valued-added measures positively correlated. They suggest that, in some cases, the addition of principals' ratings can more accurately gauge teacher performance than either measure alone.

Teachers often return for additional training at colleges and universities to pursue leadership opportunities and further develop their skills in K-12 settings. The ability to assess formal professional learning activities provides another opportunity to understand and plan for teachers' career development. In a review of nine rigorous studies examining how professional development affects student achievement, researchers found that teachers who receive substantial professional development can boost student achievement (Yoon, Duncan, Lee, Scarloss, & Shapley, 2007). This is an area deserving further attention, particularly given variation in district efforts to support teachers through curriculum, professional development, and teacher leadership activities.

By looking across the continuum of teacher development, we have identified ways in which the complexities of teaching and learning might be more fully captured. These studies are instructive in identifying elements important to address in the improvement of professional preparation. However, improvement occurs in a context that is increasingly focused on accountability pressures that emanate from state and federal policies.

A few states have begun to proactively examine available sources of evidence from teacher education programs and their graduates, and others have committed to similar efforts

through the federal Race to the Top program. One might argue that given imperfect data systems and measures of teacher and program effectiveness, in what ways should this be taken into account when moving forward on an accountability and program improvement agenda? There is an inherent tension between those whose primary role is regulation (e.g., state agencies, national accreditation bodies) and an institution's professional responsibilities toward internal accountability for program improvement. Ideally, internal and external accountability efforts would interact and be mutually reinforcing. Although the regulator's interests and purposes overlap in some areas with the goals of internal accountability, by design they are insufficient to address many aspects of program improvement. Yet, regulatory bodies share responsibility for the evidence base, and teacher preparation programs and external regulators need a richer set of measures that can inform their mutual and independent purposes.

In the second half of this article, we provide an empirical example of the use of available data to explore how new forms of evidence can be used for program improvement and accountability. Given that value-added measures are part of many accountability models, we examine what can be learned from value-added analyses within the constraints of limited state capacity, and to advance efforts toward internal and external accountability.

Value-Added Measures as a Form of Evidence

Value-added models are increasingly being used by researchers interested in developing more accurate and reliable measures of teachers' and schools' contribution toward student academic achievement. Specifically, the models can answer questions regarding the contribution of a particular teacher or school to student performance compared with that of the average teacher or school, how much of the change in student performance can be attributed to students taught by one teacher or school rather than another, and the effectiveness of an individual teacher or school at producing growth in student achievement. Thus, there are several ways in which value-added models can help inform accountability and improvement goals.

As previously discussed, a number of studies have used value-added models in examining issues related to teacher preparation (Boyd et al., 2005; Boyd et al., 2009; Clotfelter et al., 2007; Harris & Sass, 2009a; Noell et al., 2008). The studies to date have used a variety of models to derive value-added measures, and researchers have noted that significant challenges exist with respect to data availability, quality, and comprehensiveness. In the next section of this article, we conduct a value-added analysis of fifth-grade teachers in Washington state to illustrate the challenges and opportunities associated with using value-added measures for purposes of accountability and program improvement in teacher preparation.

Research Questions and Purpose

Using available data in Washington state and accounting for variation in student background factors, we conducted analyses that investigated two empirical questions.

Research Question 1: What differences in value-added scores are found by years of teaching experience?

Research Question 2: What differences are found in value-added scores for teacher education graduates by teacher preparation institution attended?

We focus on examining whether differences in teacher value-added scores exist by type of teacher preparation institution attended and years of teacher experience because both of these factors have been part of the debate about policies aimed at improving teaching and learning. For example, calls to improve the design and delivery of teacher preparation programs beg questions as to whether some types of preparation are associated with improved student learning. Similarly, questions about the role of experience are also of interest, specifically with respect to whether the positive effects associated with gaining additional years of experience plateau after a certain point in a teacher's career. To date, the research has not yet provided conclusive evidence regarding either of these two issues.

We address these research questions to present an empirical case that illustrates the following issues: (a) how value-added measures can extend the base of evidence for improving teacher education programs, (b) the measurement gaps that currently exist when attempting to articulate differences among teacher education program features and practices, and (c) the concerns that need to be addressed when examining both the results and the policy implications of value-added analyses and their specific application to accountability issues.

Data Sources

The data used in this analysis are derived from administrative records maintained by the Washington State Office of the Superintendent of Public Instruction (OSPI). For this study, we linked several sets of data to form a database of student, teacher, and school records. Student information included gender, race/ethnicity, grade level, participation in the Free or Reduced-Price Lunch program (FRPL), primary language spoken, participation in learning assistance programs, disability status, school and district location, and test scores derived from student assessment records. The teacher data came from administrative data sets of personnel and certification records. For each teacher, the available data included gender, race/ethnicity, age, years of teaching experience, highest degree, recommending agency of the initial teaching certificate, and school and district location. Unfortunately, data about teachers' achievement (e.g., SAT

scores) or academic major prior to enrolling in a teacher preparation institution were not available in state data sets. During the period under investigation (2006-2008), the vast majority of candidates prepared in Washington's 22 institutions completed their training as part of traditional undergraduate or graduate programs. Statewide data about the specific characteristics of the various teacher education programs (e.g., nature of coursework and fieldwork) were not available.[1]

Based on our prior working knowledge of many of these data sets, and a review of newly collected data elements, it became evident that the quality and availability of the specific data elements would substantially affect decisions about the types of analyses that were feasible. At present, state tests are the only uniform, empirical indicators available to show what students are learning from their teachers statewide. In addition, similar to that which is found in most states, student assessment data in Washington is available only in certain grades (Grades 3 through 8, and 10) and for specific subjects in those grades (reading, math, writing, and science). This excludes from analysis teachers who do not in teach in these grades or subjects. Some fields provided generally reliable information, whereas others had considerable missing or incomplete data. For example, most crucial to these analyses was a link between the students and the teachers who taught them. At the time the analyses were conducted, the only link between students and teachers was the teacher name associated with the student's assessment data. In some cases, the teacher name was missing or incomplete in the student records or the students did not have a unique code which enabled linking their assessment records over multiple years. This was particularly problematic for student and teacher records beyond the sixth grade.

To create a link between teacher and student records, we matched teachers by code to their school and by teacher name to their students' test scores.[2] By mapping backward, we were able to determine other student characteristics. These procedures allowed us to identify math and reading teachers for at least 71% of all students beginning Grade 4 across a 2-year period from 2006-2007 to 2007-2008. This level of matched data is the highest we were able to obtain and is similar to that which has been encountered in similar studies (Clotfelter et al., 2007; Goldhaber & Liddle, 2011). Therefore, we selected fourth- and fifth-grade data for math and reading to use the most complete data available.[3]

The combined data set used in this analysis consisted of a cohort of students over 2 years (fourth and fifth graders) with a total sample size of 51,161 students for mathematics and 50,988 students for reading. The fifth-grade teacher sample size included 2,874 teachers for reading and 2,864 teachers for math in the 2007-2008 academic year.[4]

The dependent variable is a standardized test score on the Washington Assessment of Student Learning (WASL) in reading or math for each student. The raw scores on each student's scores were normalized with a mean observation of 0 and a standard deviation of 1. This standardization made possible test comparisons across grades and years.

The student and teacher level covariates used in the models are presented below. We arrived at these covariates through an iterative process by constructing models that examined all student and teacher variables that were uniformly collected and have been shown in the literature to affect measures of student learning. For students, the covariates were measures of gender, ethnicity, and poverty.[5] For teachers, the covariates were years of experience, gender, ethnicity, and possession of an advanced degree. The descriptive statistics for the variables used in the analysis are provided in Table 1. Table 2 provides the variable description and categorization and the summary statistics of the dependent variable and the student and teacher level covariates used in the analysis.[6]

We use this data to estimate value-added measures in reading and mathematics for fifth-grade teachers. We then examine whether differences in value-added measures exist by teacher experience and by the institutions where these teachers obtained their initial teacher training. In doing so, we conduct separate analyses that focus specifically on novice teachers.

Model Selection and Specification

Value-added models range from single level linear regression models based on least squares estimation to more complex linear mixed models that correct for the various limitations of the single level least squares. These models often use either maximum likelihood or Bayesian methods to help correct for biases in the estimates and provide for more reliable hypotheses testing. The types of models that are used depend in part on the number of years of available data. There is also a debate about whether to use fixed or random effects. Fixed effects assign a dummy variable for each student, whereas random effects define a distribution. We selected a random effects model because we wanted to use more flexible estimation methods such as maximum likelihood and Bayesian estimates, making the model more parsimonious and efficient.[7]

In this study, we adopt a univariate (single outcome) nonclassified two-level hierarchical linear model. This model reduces the multivariate data to a univariate outcome and also reduces the potentially cross-classified membership structure to a full nested model by linking the student outcome with only one teacher or school (Clotfelter et al., 2007; Lockwood, Doran, & McCaffrey, 2003). Such strategies are most commonly used for their practicality and intuitive interpretation. We follow the hierarchical modeling methods of Raudenbush and Bryk (2002).

Separate univariate models for reading and mathematics achievement were analyzed. We begin the analysis with a null unconditional model to assess the amount of variation

Table 1. Summary Statistics

	M	*SD*	Mode	Minimum	Maximum
Dependent variables					
Fifth-grade reading scores	414.6	26.01	—	283.0	475.0
Fifth-grade math scores	408.5	37.70	—	219.0	550.0
Student level variables					
Gender (reading)	0.49	0.50	0 (51%)	0	1
Gender (math)	0.49	0.50	0 (51%)	0	1
Ethnicity (reading)	4.25	1.32	5 (66%)	0	5
Ethnicity (math)	4.25	1.32	5 (66%)	0	5
Free or reduced-price lunch eligibility (reading)	0.41	0.49	0 (59%)	0	1
Free or reduced-price lunch eligibility (math)	0.17	0.37	0 (83%)	0	1
Fourth-grade reading score	410.8	20.56	—	302.0	475.0
Fourth-grade reading score (grand mean centered)	0.00[a]	20.56	—	−108.8	64.2
Fourth-grade math score	406.1	41.45	—	213.0	550.0
Fourth-grade math score (grand mean centered)	0.00[a]	41.45	—	−193.1	143.9
Teacher level variables					
Years of experience (reading)	13.79	9.83	—	0	45
Years of experience (reading, grand mean centered)	0.00[a]	9.83	—	−13.8	31.2
Years of experience (math)	13.70	9.85	—	0	45
Years of experience (math, grand mean centered)	0.00[a]	9.85	—	−13.7	31.3
Novice status (less than 2 years of experience) (reading)	0.10	0.29	0 (90%)	0	1
Novice Status (less than 3 years of experience) (reading)	0.14	0.34	0 (86%)	0	1
Novice status (less than 2 years of experience) (math)	0.10	0.30	0 (90%)	0	1
Novice status (less than 3 years of experience) (math)	0.14	0.35	0 (86%)	0	1
Highest education degree (reading)	0.67	0.47	1 (67%)	0	1
Highest education degree (math)	0.67	0.47	1 (67%)	0	1
Gender (reading)	0.75	0.43	1 (75%)	0	1
Gender (math)	0.74	0.44	1 (74%)	0	1
Ethnicity (reading)	0.93	0.25	1 (93%)	0	1
Ethnicity (math)	0.94	0.24	1 (94%)	0	1
Teacher preparation institutions	*na*				

[a]Number is rounded to 0 as decimal point exceeds negative 10th power.

Table 2. Variable Description and Categorization

	Coding and categories
Dependent variables	
Fifth-grade reading scores	Continuous variable
Fifth-grade math scores	Continuous variable
Student level variables	
Gender	1 = *student is female*, 0 = *male*
Ethnicity	1 = *student is Native American*, 2 = *Asian*, 3 = *Black*, 4 = *Hispanic*, 5 = *White*
Free or reduced-price lunch eligibility	1 = *student is on FRPL*, 0 = *not on FRPL*
Fourth-grade reading score	Continuous variable
Fourth-grade math score	Continuous variable
Teacher level variables	
Teacher experience	Continuous variable
Teacher novice status	Teacher has less than 2, 3 or 5 years of experience = 1, otherwise = 0
Highest education degree	1 = *teacher has a masters or doctorate*, 0 = *bachelors and others*
Gender	1 = *teacher is female*, 0 = *male*
Ethnicity	1 = *teacher is White*, 0 = *otherwise*
Teacher preparation institutions	Coded 0 to 13, with some institutions/campuses combined to generate sufficient sample size

Note: FRPL = Free or Reduced-Price Lunch program.

that exists in value-added scores among teachers in the data set (Model 1). The presence of significant differences in between-teacher performance in Model 1 allows us to continue with our multilevel analysis. Next, we use a value-added model that builds on Model 1 by taking student background factors into account (Model 2). Finally, exploratory models of the value-added measures are analyzed to examine teacher level characteristics that are associated with their value-added scores (Model 3). The mathematical specifications of these three models are provided in Appendix A (available online at http://JTE.sagepub.com/supplemental).

We do not aim to make causal statements about the teacher value-added measures or the effects of teacher experience or teacher preparation institutions on student achievement. Rather, the intention is to advance beyond the conventional descriptive measures by observing various forms of association (or confounding effects) between these variables and student learning by means of sequential regression described above.

Analysis and Results

The results from the analysis of the null unconditional model (Model 1) found that, when partitioning the total variation in students' performance in our cohort of fifth-grade teachers, there was evidence of significant random teacher effects and between teacher variation. The range of differences between means for individual teachers compared with the state grand mean was found to be from a minimum of −0.56 to a maximum of 37.36 points for reading and from a minimum of −73.45 to a maximum of 70.28 points for math. The results from Model 1 provided us with the initial justification to further explore the underlying determinants of teachers' value-added scores with more complex statistical analyses.[8] Thus, we proceeded to a value-added analysis that takes into account students' prior achievement and background characteristics (Model 2). In Model 2, we control for the following student level factors: gender, race/ethnicity, poverty status, and prior fourth-grade achievement scores. After controlling for these student factors, we find that there are fifth-grade teachers who have statistically significant value-added scores.[9]

Next, we sought to explore why some teachers have higher value-added scores than others. Model 3 asks whether differences in value-added scores are associated with differences by years of teaching experience or by type of teacher preparation institution attended. First, we explored teacher experience by examining the full range of years of experience for all teachers in our sample. We then conducted additional analyses that were limited to those teachers with less than 3 years of experience and those teachers with less than 2 years of experience.[10]

Teacher experience and value-added scores. We first examine the relationship between teacher experience and value-added scores with our entire sample of teachers. Results of this analysis are provided in Table 3.

In Table 3, columns 1 and 3 describe results obtained using only years of experience as the single teacher characteristic of interest. The results depicted in columns 2 and 4 include other teacher characteristics in addition to years of experience. Our analysis of this data supports the claim that there is a significant relationship between teacher experience and teacher value-added scores. As seen in columns 2 and 4, the positive relationship between teacher experience and student outcome measures does not change in magnitude and significance even after taking teacher gender, education level, and race/ethnicity into account. For reading, teacher experience is found to be the only significant factor in predicting teacher value-added scores. For math, in addition to teacher experience, female teachers have higher value-added scores than males (0.73 points) and White teachers have higher value-added scores than teachers with other ethnic backgrounds (1.83 points).

Looking at the variance components in columns 2 and 4, the residuals between teacher variance (σ_a^2) are smaller than the estimates of 22.78 for reading and 50.17 for math observed in Model 2 (see Appendix B, available online at http://JTE.sagepub.com/supplemental.). This means that the parameter variation in teacher value-added scores (the random intercept) has been explained by the teacher variables. Comparing the between teacher variance estimates across models, the proportion of variance between teachers' value-added scores explained by Model 3 are 0.8% for reading and 0.5% for math.

We also conducted an analysis that recodes the continuous teacher experience variable into a binary variable where teachers with 0 to 2.9 years of experience are given a value of 1 and teachers with 3 or more years of experience with value of 0. Results are displayed in Table 4.

Our analyses indicate that novice teachers with less than 3 years of experience have a lower value-added measure by 1.11 points for reading and 1.75 points for math. These estimates are robust even after controlling for other teacher characteristics. The proportions of variance between teachers are 0.7% for reading and 1.1% for math. Recoding the teacher novice variable to teachers with less than 2 years of experience, novice teachers have even lower value-added measures by 1.37 points for reading and 1.78 points for math. This finding is generally consistent with the positive effect of teacher experience noted in Table 3, but also hints at the possibility of a nonlinear effect of teacher experience.

To test whether teacher experience has an increasing or diminishing effect on teacher value-added measures, the squared term of teacher experience was modeled (see Table 5).

As shown in columns 1 and 3, the coefficient of the squared experience term has a significant negative value of −0.0026 for reading and −0.0036 for math. This negative squared term implies a diminishing effect of teacher experience over teacher's value-added scores. That is, as teachers

Table 3. Teacher Experience and Value-Added Measures

	Reading						Math					
	Column 1			Column 2			Column 3			Column 4		
	Coefficient	*SE*	*t*	Coefficient	*SE*	*t*	Coefficient	*SE*	*t*	Coefficient	*SE*	*t*
Intercept	414.82	0.39	1,064.2	413.86	0.65	633.4	405.93	0.46	874.9	403.52	0.86	468.8
Student characteristics												
Student is female	0.28	0.15	1.8	0.28	0.15	1.8	2.44	0.18	13.4	2.44	0.18	13.4
Student is Native American (base)	—	—	—	—	—	—	—	—	—	—	—	—
Student is Asian	1.99	0.46	4.4	2.00	0.46	4.4	4.48	0.55	8.2	4.49	0.55	8.2
Student is Black	−0.16	0.49	−0.3	−0.15	0.49	−0.3	−1.16	0.59	−2.0	−1.14	0.59	−1.9
Student is Hispanic	−1.24	0.42	−3.0	−1.23	0.42	−2.9	0.17	0.51	0.3	0.20	0.51	0.4
Student is White	1.77	0.38	4.7	1.76	0.38	4.7	1.83	0.45	4.1	1.82	0.45	4.1
Student on free or reduced-price lunch	−4.00	0.18	−21.8	−3.99	0.18	−21.8	−1.79	0.27	−6.7	−1.79	0.27	−6.7
Student's fourth-grade grand mean centered score	0.85	0.00	205.8	0.85	0.00	205.8	0.72	0.00	290.2	0.72	0.00	290.2
Teacher characteristics												
Teacher's experience (grand mean centered)	0.04	0.01	3.6	0.04	0.01	3.5	0.03	0.02	2.0	0.03	0.02	1.9
Teacher has a masters or doctorate degree				0.20	0.26	0.8				0.23	0.35	0.6
Teacher is female				0.21	0.28	0.8				0.73	0.38	1.9
Teacher is White				0.73	0.49	1.5				1.83	0.68	2.7
Between- and within-teacher variance												
Between-teacher variance (σ_a^2)	22.64	4.76		22.60	4.75		50.17	7.08		49.92	7.07	
Within-teacher variance (σ_y^2)	292.30	17.10		292.30	17.10		414.90	20.37		414.87	20.37	
Intraclass correlation coefficient	0.07			0.07			0.11			0.11		
BIC value[a]	436,698			436727			456,846			456,867		
Number of student observations	50,988			50,988			51,161			51,161		
Number of teacher observations	2,874			2,874			2,864			2,864		

[a]BIC refers to the Bayesian Information Criterion used to assess model fit.

gain more experience, their value-added scores still increase, but at a decreasing rate.[11] We find that teacher experience peaks at approximately 24.91 years for reading and 20.96 years for math.[12] This diminishing and plateau effect of teacher experience remains significant even after taking other teacher characteristics into account (see columns 2 and 4).

Teacher preparation institutions and value-added scores. Our second research question explored whether differences in teacher value-added scores existed by the preparation institutions from which teachers earned their initial certification.[13] As described earlier, we compare results from institutions in Washington state with results obtained for teachers who did not receive their teaching credential from an institution located in Washington.[14]

In our examination of differences by teacher preparation institution attended, we first included all teachers in our sample (n = 2,864), followed by a separate analysis that examined only teachers with less than 5 years of experience (n = 778). In the analysis of all teachers in our sample, we found significant differences in value-added scores across Washington state institutions.[15] We conduct the second analysis in an attempt to restrict our analysis to the more recent programs, as many teacher preparation programs have altered their content and practices in recent years.

Table 4. Teacher Novice Status (Less Than 3 Years of Experience) and Value-Added Measures

	Reading						Math					
	Column 1			Column 2			Column 3			Column 4		
	Coefficient	*SE*	*t*	Coefficient	*SE*	*t*	Coefficient	*SE*	*t*	Coefficient	*SE*	*t*
Intercept	414.98	0.39	1,056.7	414.04	0.66	627.2	406.18	0.47	866.9	403.94	0.87	464.4
Student characteristics												
Student is female	0.28	0.15	1.8	0.28	0.15	1.8	2.44	0.18	13.4	2.44	0.18	13.4
Student is Native American (base)	—	—	—	—	—	—	—	—	—	—	—	—
Student is Asian	1.99	0.46	4.4	2.00	0.46	4.4	4.48	0.55	8.2	4.49	0.55	8.2
Student is Black	−0.17	0.49	−0.3	−0.16	0.49	−0.3	−1.15	0.59	−1.9	−1.13	0.59	−1.9
Student is Hispanic	−1.24	0.42	−2.9	−1.23	0.42	−2.9	0.20	0.51	0.4	0.22	0.51	0.4
Student is White	1.77	0.38	4.7	1.76	0.38	4.7	1.83	0.45	4.1	1.82	0.45	4.1
Student on free or reduced-price lunch	−4.00	0.18	−21.8	−3.99	0.18	−21.8	−1.79	0.27	−6.7	−1.79	0.27	−6.7
Student's fourth-grade grand mean centered score	0.85	0.00	205.8	0.85	0.00	205.8	0.72	0.00	290.3	0.72	0.00	290.3
Teacher characteristics												
Teacher has less than 3 years of experience	−1.11	0.35	−3.2	−1.06	0.36	−3.0	−1.75	0.47	−3.7	−1.71	0.48	−3.5
Teacher has a masters or doctorate degree				0.13	0.26	0.5				0.02	0.36	0.0
Teacher is Female				0.20	0.28	0.7				0.73	0.38	1.9
Teacher is White				0.74	0.49	1.5				1.79	0.68	2.6
Between- and within-teacher variance												
Between-teacher variance (σ_a^2)	22.66	4.76		22.62	4.76		49.87	7.06		49.62	7.04	
Within-teacher variance (σ_y^2)	292.31	17.10		292.31	17.10		414.91	20.37		414.89	20.37	
Intraclass correlation coefficient	0.07			0.07			0.11			0.11		
BIC value[a]	436,701			436,730			456,836			456,858		
Number of student observations	50,988			50,988			51,161			51,161		
Number of teacher observations	2,874			2,874			2,864			2,864		

[a]BIC refers to the Bayesian Information Criterion used to assess model fit.

To examine whether significant differences exist for teacher education graduates with less than 5 years of teaching experience, we reconstructed the data set to include only teachers with 0 to 4.9 years of experience. Results of this analysis are displayed in Table 6.[16]

As can be seen in column 1 of Table 6, teacher preparation programs at Institutions 8 and 9 have significant positive associations relative to non-Washington institutions on the value-added reading scores for teachers with less than 5 years of experience. Compared with teachers trained outside of Washington state, Institutions 8 and 9 add 2.31 and 2.7 higher points on their teachers' value-added scores in reading.[17] As shown in column 2, this finding is robust even after taking other teacher characteristics into account. However, none of the institutions have significant associations with recent graduates' value-added scores in math. Furthermore, compared with the case of reading, although not significant, more institutions have a negative effect and underperform compared with the teachers trained outside of Washington state.[18] As depicted in column 4, Institutions 4, 8, 9, and 12 were the only institutions with a positive coefficient, but these results were not statistically significant.

These findings suggest that variation exists across teacher education programs when measuring the effectiveness of their graduates using value-added scores. This is consistent with the outcomes of similar studies conducted in recent years (Boyd et al., 2005; Goldhaber & Liddle, 2011; Henry et al., 2011; Noell et al., 2008). Our findings about the positive relationship between teacher experience and student achievement, especially in the early years of teaching, is also consistent with multiple prior studies, though the plateau effect found in our study occurs later than that found in similar research studies (Aaronson, Barrow, & Sander, 2007; Clotfelter et al., 2007; Croninger, Rice, Rathbun, & Nishio, 2007).

Table 5. Teacher Experience Squared and Value-Added Measures

	Reading						Math					
	Column 1			Column 2			Column 3			Column 4		
	Coefficient	*SE*	*t*	Coefficient	*SE*	*t*	Coefficient	*SE*	*t*	Coefficient	*SE*	*t*
Intercept	413.80	0.47	881.0	412.95	0.69	600.9	404.89	0.59	690.5	402.68	0.91	444.2
Student characteristics												
Student is female	0.28	0.15	1.8	0.28	0.15	1.8	2.44	0.18	13.4	2.44	0.18	13.4
Student is Native American (base)	—	—	—	—	—	—	—	—	—	—	—	—
Student is Asian	2.00	0.46	4.4	2.01	0.46	4.4	4.49	0.55	8.2	4.49	0.55	8.2
Student is Black	−0.15	0.49	−0.3	−0.15	0.49	−0.3	−1.15	0.59	−1.9	−1.13	0.59	−1.9
Student is Hispanic	−1.23	0.42	−2.9	−1.22	0.42	−2.9	0.19	0.51	0.4	0.21	0.51	0.4
Student is White	1.77	0.38	4.7	1.76	0.38	4.7	1.83	0.45	4.1	1.82	0.45	4.1
Student on free or reduced-price lunch	−4.00	0.18	−21.8	−3.99	0.18	−21.8	−1.79	0.27	−6.7	−1.79	0.27	−6.7
Student's fourth-grade grand mean centered score	0.85	0.00	205.8	0.85	0.00	205.7	0.72	0.00	290.2	0.72	0.00	290.2
Teacher characteristics												
Teacher year(s) of experience	0.13	0.04	3.0	0.12	0.04	2.8	0.15	0.06	2.6	0.14	0.06	2.3
Teacher year(s) of experience squared	−0.0026	0.00	−2.1	−0.0024	0.00	−1.9	−0.0036	0.00	−2.1	−0.0032	0.00	−1.9
Teacher has a masters or doctorate degree				0.13	0.26	0.5				0.12	0.36	0.3
Teacher is female				0.19	0.28	0.7				0.70	0.38	1.8
Teacher is White				0.71	0.49	1.5				1.80	0.68	2.6
Between- and within-teacher variance												
Between-teacher variance (σ_a^2)	22.56	4.75		22.52	4.75		50.05	7.07		49.81	7.06	
Within-teacher variance (σ_y^2)	292.31	17.10		292.31	17.10		414.89	20.37		414.88	20.37	
Intraclass correlation coefficient	0.07			0.07			0.11			0.11		
BIC value[a]	436,704			436,734			456852			456,874		
Number of student observations	50,988			50,988			51,161			51,161		
Number of teacher observations	2,874			2,874			2,864			2,864		

[a]BIC refers to the Bayesian Information Criterion used to assess model fit.

However, there are many important variables that we were unable to include in this exploratory study due to a lack of uniform data. Examples include differences across individual programs within the same institution, variation in the individual program features across institutions, participation in alternative routes to certification, and issues of selectivity in the recruitment of teacher candidates. Thus, these results are illustrative, and given significant limitations in state data capacity and the need for further development of value-added methods, use of these measures have limited application at this time. Consequently, it is premature to use results from this analysis to draw conclusions about program quality or provide individual ranking and evaluation of institutions solely based on value-added measures.

Discussion of value-added measures as a source of evidence. This study illustrates that there is potential in using value-added models as an additional form of evidence that can inform our understanding of the effectiveness of teacher preparation programs in producing teachers who can positively affect student learning. Value-added measures can assist in the identification of teachers who potentially may be more effective, as well as those who may be in need of extra assistance and support. A shown in Figure 1, our analysis can discriminate the potentially more effective and the potentially least effective teachers in terms of value-added scores. However, it is important to note that for the vast majority of teachers, our value-added analyses do not provide any statistically significant evidence of effectiveness, as the results for the majority of teachers are not statistically significant from the state grand mean.

The use of value-added models represent a relatively new phenomenon in education, and initial efforts have been accompanied with a number of theoretical, statistical, and practical issues (Braun, 2005; Harris, 2009, 2011). In addition, there is a need across most state data systems to improve the accuracy of data. Analyses involving teacher preparation institutions should include a variety of program variables, including the nature and length of field placements, program characteristics and types (e.g., traditional or alternative), and methods to address issues of selectivity. Perhaps the most perplexing matter concerns the need to improve the comprehensiveness of the measures of student growth, as well as other important factors that affect student learning, such

Table 6. Teacher Institutions and Value-Added Measures for Novice Teachers With Less Than 5 Years of Experience

	Reading						Math					
	Column 1			Column 2			Column 3			Column 4		
	Coefficient	*SE*	*t*	Coefficient	*SE*	*t*	Coefficient	*SE*	*t*	Coefficient	*SE*	*t*
Intercept	412.76	0.86	479.0	411.26	1.29	317.9	403.67	1.06	381.7	402.80	1.68	239.1
Student characteristics												
Student is female	0.11	0.30	0.4	0.11	0.30	0.4	2.28	0.35	6.5	2.28	0.35	6.5
Student is Native American (base)	—	—	—	—	—	—	—	—	—	—	—	—
Student is Asian	2.39	0.89	2.7	2.38	0.89	2.7	3.92	1.07	3.7	3.92	1.07	3.7
Student is Black	0.13	0.95	0.1	0.10	0.95	0.1	−0.87	1.14	−0.8	−0.87	1.14	−0.8
Student is Hispanic	−1.07	0.81	−1.3	−1.01	0.82	−1.2	−0.81	0.98	−0.8	−0.77	0.98	−0.8
Student is White	2.13	0.75	2.8	2.09	0.75	2.8	1.68	0.90	1.9	1.65	0.90	1.8
Student on free or reduced-price lunch	−3.88	0.36	−10.8	−3.85	0.36	−10.7	−0.96	0.49	−1.9	−0.94	0.49	−1.9
Student's fourth-grade grand mean centered score	0.87	0.01	109.4	0.87	0.01	109.4	0.72	0.00	150.6	0.72	0.00	150.5
Teacher characteristics												
Teacher's grand mean centered yrs of experience				0.23	0.14	1.6				0.48	0.19	2.5
Teacher has a masters or doctorate degree				0.43	0.53	0.8				−0.40	0.72	−0.6
Teacher is female				−0.36	0.55	−0.7				0.34	0.74	0.5
Teacher is White				1.48	0.88	1.7				0.67	1.20	0.6
Teacher is trained outside of Washington State (base)	—	—	—	—	—	—	—	—	—	—	—	—
Teacher is trained at Institution 1	−2.05	2.14	−1.0	−1.77	2.14	−0.8	−3.55	2.84	−1.3	−2.99	2.84	−1.1
Teacher is trained at Institution 2	0.21	2.17	0.1	0.26	2.18	0.1	−4.27	2.96	−1.4	−3.79	2.97	−1.3
Teacher is trained at Institution 3	−2.25	1.81	−1.2	−1.91	1.82	−1.1	−2.73	2.59	−1.1	−2.67	2.59	−1.0
Teacher is trained at Institution 4	−0.14	0.83	−0.2	0.17	0.85	0.2	0.55	1.14	0.5	0.65	1.17	0.6
Teacher is trained at Institution 5	−0.86	1.41	−0.6	−0.71	1.41	−0.5	−0.77	1.94	−0.4	−0.63	1.93	−0.3
Teacher is trained at Institution 6	−1.18	0.87	−1.4	−0.91	0.89	−1.0	−1.60	1.21	−1.3	−1.53	1.24	−1.2
Teacher is trained at Institution 7	0.07	1.07	0.1	0.10	1.07	0.1	−0.09	1.43	−0.1	−0.06	1.42	0.0
Teacher is trained at Institution 8	2.31	1.01	2.3	2.72	1.02	2.7	−0.10	1.40	−0.1	0.35	1.42	0.3
Teacher is trained at Institution 9	2.70	1.47	1.8	2.61	1.46	1.8	1.15	1.91	0.6	1.00	1.91	0.5
Teacher is trained at Institution 10	1.81	1.85	1.0	1.85	1.85	1.0	−1.72	2.55	−0.7	−1.38	2.56	−0.5
Teacher is trained at Institution 11	0.56	0.87	0.6	0.85	0.88	1.0	−1.26	1.19	−1.1	−0.92	1.20	−0.8
Teacher is trained at Institution 12	0.17	0.85	0.2	0.28	0.88	0.3	−0.36	1.17	−0.3	0.33	1.21	0.3
Teacher is trained at other Washington Institutions 13	−0.82	0.87	−0.9	−0.24	0.89	−0.3	−1.77	1.20	−1.5	−1.24	1.23	−1.0
Between- and within-teacher variance												
Between-teacher variance (σ^2_a)	21.68	4.66		21.43	4.63		48.77	6.98		48.09	6.93	
Within-teacher variance (σ^2_y)	288.23	16.98		288.16	16.98		406.00	20.15		406.01	20.15	
Intraclass correlation coefficient	0.07			0.07			0.11			0.11		
BIC value[a]	113,640			113,671			119,815			119,846		
Number of student observations	13,269			13,269			13,429			13,429		
Number of teacher observations	778			778			778			778		

[a]BIC refers to the Bayesian Information Criterion used to assess model fit.

student attendance, class size, professional development, resource levels, leadership, and school working conditions.

As these models continue to be examined, additional variables need to be considered in model refinement so that a more robust depiction of the contexts in which teachers work can be included. These models can also be enhanced with assessments of teaching performance such as those being developed through the efforts of the Teacher Performance Assessment Consortium. It will be important for value-added models to consider the variance in expectations for teacher performance, as teachers gain experience and further develop their professional abilities.

Conclusions

The review of the empirical research identified common elements that are important to consider in examining measures of effectiveness in teacher training and development. As discussed earlier, some of these include teaching experience, candidate selectivity, subject matter knowledge, and specific program features of professional preparation. From the exercise of using a state's existing data capacity to explore some of these elements, we learned that differences exist among teacher preparation institutions on some student learning outcomes, though the extent to which program features or other mitigating

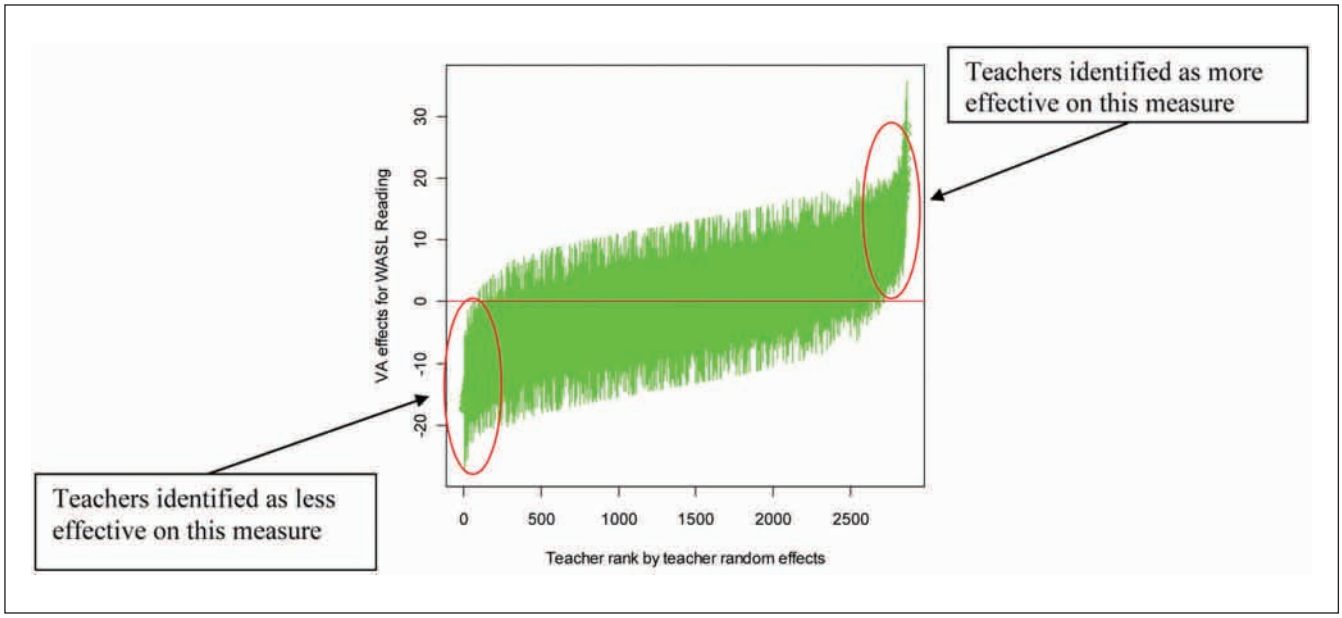

Figure 1. Washington: Ordered teacher value-added effects
Note: WASL = Washington Assessment of Student Learning.

factors may contribute to differences is often unknown. Consistent with other studies, we found compelling evidence to suggest that teaching experience matters to student achievement, particularly in the earlier years of a teacher's career.

Although these findings provide some potentially useful information related to the impact of teacher education programs on aspects of teaching quality, few definitive statements can be made. To date, most researchers have been able unable to disentangle the effects of programs from the characteristics and abilities of teacher education graduates and the settings in which they teach (Zeichner & Conklin, 2005). This suggests the need for better existing data and the development of new measures that accurately reflect substantive differences among programs. A richer data set can help address possible biases with existing measures. One application of value-added measures for institutions of higher education may be found in analyzing specific programs regarding malleable factors and education outcomes that could contribute to the identification and development of potentially beneficial interventions for teacher training. Early efforts to provide value-added information to institutions show some promise in prompting program changes (Sawchuck, 2012). It can also shed light on the relative contribution of these factors in raising student performance and the possibility of significant interactions between measures. In addition, longitudinal statewide analyses provide an opportunity for more robust and inclusive investigations.

Legitimate concerns have been raised as to the fairness and appropriate use of particular metrics for purposes of accountability. For teacher preparation programs, there remains the question of the extent to which an institution can influence a particular outcome, and ultimately the fairness in being held accountable for things over which one may have little control (Harris, 2011). Accountability and improvement efforts are not well served by simple measures used out of convenience. If these efforts are to provide direct guidance about the quality of teacher preparation, then additional data will be required—data that are useful to institutions and more accurately reflect their work. This requires a deeper examination of the variety of elements that need to be taken into account, and an acknowledgment that teacher preparation programs vary in their individual features and purposes. Finally, consideration must be given to the appropriate use of data for particular purposes, whether accreditation, licensure, or improvement. The tension between knowing that measures are imperfect, and the call for their use in particular forms of public accountability remains divisive.

> Simplistic rankings of programs and schools have appeal to policymakers and the public. However, the profession has not articulated clearly what an appropriate system of accountability for schools and preparation programs should look like and how that system can best lead to improvement in outcomes for students. (Cibulka, 2011, p. 2)

The singular focus on the unique contribution of individual teachers is insufficient to fully inform the improvement of teacher preparation and the development of high-quality teaching.

Not unlike what has happened with accountability in the K-12 sector, responses by institutions of higher education have been varied. Some have seized the opportunity to clarify program goals, focus on candidate practices and assessments, and develop better data systems, whereas others have engaged in symbolic compliance without significant change (Bell &

Youngs, 2011). Much of this is mediated by institutional capacity and available resources. In this regard, collective efforts to articulate how progress toward excellent teaching can be identified and measured is worth consideration.

The challenge for preparation programs is to do the hard work of collectively identifying and developing measures that better reflect unique program features and seek to build capacity for reliable data collection. This will require the cooperation and agreement among programs and institutions about what elements matter, and which can and should be consistently and reliably obtained across settings. State agencies and accreditation bodies also share in the responsibility to improve the measures often used for student and teacher learning and find more robust and comprehensive ways of assessing the effectiveness of teaching. This suggests the need for collective responsibility on the part of teacher preparation institutions, professional organizations, state agencies, and K-12 districts and schools for the quality, availability, and appropriate uses of data for purposes of internal and external accountability.

The development and implementation of a comprehensive and coherent approach to the collection and use of evidence in the improvement of teacher preparation programs provides two additional benefits: increased transparency and joint accountability. First, the sharing and use of evidence from multiple sources can help increase public understanding of the complexities of recruiting, preparing and supporting the next generation of teachers, and can help policy makers engage in debates that are informed by evidence. Second, the development of a consistent base of evidence can help specify the ways in which the variety of state, regional, and local institutions involved in teacher preparation and development share accountability for ensuring that all students have access to high-quality teaching. The use of an evidence-based approach for the ongoing improvement of teacher education programs can and should have a significant impact on the quality of teaching and learning.

Author's Note

Appendices A and B are available online at http://JTE.sagepub.com/supplemental.

Declaration of Conflicting Interests

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

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was partially supported through a research effort under the Teachers for a New Era (TNE) project at the University of Washington, funded by the Carnegie Corporation of New York, the Annenberg Foundation and the Ford Foundation.

Notes

1. Although a few institutions prepared candidates in programs that were recognized by the State of Washington as alternative routes to teacher certification, these candidates comprised approximately 3% of all candidates statewide.
2. Although we cross-checked the data to verify that each individual associated with a student's test scores was actually a classroom teacher at the school, we cannot say with absolute certainty that the identified teacher was the student's actual teacher or another teacher in the same school who served as the test proctor.
3. This decision is consistent with prior similar studies.
4. A careful investigation was conducted to identify sources of missing data. Hand matching was attempted wherever possible. Most of the missing data was due to a lack of a linking variable for either teachers or students and not due to missing outcome or explanatory variables. This limited the possibility of conducting imputation methods. Misspelled teacher names were tracked and corrected.
5. For students, we also examined the possibility of including primary language spoken at home, disability status, and participation in a supplemental education program but these covariates were not included as they did not change our estimates.
6. The student ethnicity variable consists of five categories as described in the table. The category of "Native American" was used as the baseline group for interpreting the coefficient estimates for other categories. That is, the estimated coefficients are interpreted relative to the baseline group. In addition, the continuous variables such as student prior score and teacher experience were grand mean centered to provide easier and more realistic interpretation of the parameter estimates.
7. We recognize that there is an ongoing debate among researchers about the selection of value-added model features and approaches.
8. Additional details and displays of the results obtained from Model 1 are provided in Appendix B.
9. Appendix B contains additional details of the results from Model 2.
10. See Appendix B for additional details of the Model 3 analyses.
11. This is represented by the concave shape of value-added scores over years of teaching experience.
12. To obtain this finding, we took the first derivative of the regression equation with respect to experience and solved for the maximum point (by setting the first derivative to zero and solving for experience).
13. Due to small sample sizes, data from some of the smaller institutions and branch campuses were combined in this analysis. Thus, the number of institutions displayed in our analysis is less than the total number of institutions in the state that issue teaching credentials.
14. The decision to compare each institution's value-added results to results obtained from those who were not trained in Washington state is consistent with analyses used in several similar

studies (Goldhaber & Liddle, 2011; Henry et al., 2011). This decision does not impact the ability to compare results between Washington state institutions.

15. Details of this analysis are provided in Appendix B.
16. In this analysis, teacher preparation institutions are not identified by name.
17. Institutions 8 and 9 include a state university with corresponding branch campuses and a private university. One offers exclusively graduate training for teacher preparation, while the other offers both undergraduate and graduate training. Both are considered either selective or more selective under Carnegie classifications for undergraduate education.
18. The institutions that underperform on value-added measures include private and state universities with undergraduate and graduate education programs. Selection bias is expected to be a factor. This suggests the need to identify and develop measures that better reflect program features.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics, 25*(1), 95-135.

Bell, C. A., & Youngs, P. (2011). Substance and show: Understanding responses to teacher education programme accreditation processes. *Teaching and Teacher Education, 27*(2), 298-307.

Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J. (2005). *How changes in entry requirements alter the teacher workforce and effect student achievement*. Albany, NY: Teacher Policy Research.

Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis, 31*(4), 416-440.

Boyd, D. J., Lankford, H., Loeb, S., Rockoff, J. E., & Wyckoff, J. (2008). The narrowing gap in New York City teacher qualifications and its implications for student achievement in high-poverty schools. *Journal of Policy Analysis and Management, 27*(4), 793-818.

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

Cibulka, J. G. (2011). The new normal: Challenges and opportunities for the transformation of teacher preparation. *Quality Teaching: A Publication of the National Council for the Accreditation of Teacher Education, 21*(1), 1-5.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). *How and why do teacher credentials matter for student achievement?* (National Center for Analysis of Longitudinal Data in Education Research, Working Paper No. 2). Washington, DC: The Urban Institute.

Cochran-Smith, M. (2001). Learning to teach against the (new) grain. *Journal of Teacher Education, 52*(1), 3-4.

Cochran-Smith, M., & The Boston College Evidence Team. (2009). Reculturing teacher education: Inquiry, evidence, and action. *Journal of Teacher Education, 60*(5), 458-468.

Constantine, J., Player, D., Silva, T., Hallgren, K., Grider, M., & Deke, J. (2009). *An evaluation of teachers trained through different routes to certification: Final report* (NCEE 2009-4043). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Croninger, R., Rice, J., Rathbun, A., & Nishio, M. (2007). Teacher qualifications and early learning: Effects of certification, degree, and experience on first-grade student achievement. *Economics of Education Review, 26*(3), 312-324.

Crowe, E. (2010). *Measuring what matters: A stronger accountability model for teacher education*. Washington, DC: Center for American Progress.

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.

Darling-Hammond, L. (1997). *Doing what matters most: Investing in quality teaching*. New York, NY: The National Commission on Teaching and America's Future.

Darling-Hammond, L. (2006a). Assessing teacher education: The usefulness of multiple measures for assessing program outcomes. *Journal of Teacher Education, 57*(2), 120-138.

Darling-Hammond, L. (2006b). Constructing 21st century teacher education. *Journal of Teacher Education, 61*(5), 300-314.

Donaldson, M. L., & Johnson, S. M. (2010). The price of misassignment: The role of teaching assignments in Teach for America teachers' exit from low-income schools and the teaching profession. *Educational Evaluation and Policy Analysis, 32*(2), 299-323.

Duncan, A. (2010). Teacher preparation: Reforming the uncertain profession. *Education Digest, 75*(5), 13-22.

Elfers, A., Plecki, M., & Knapp, M. (2006). Teacher mobility: Looking more closely at "the movers" within a state system. *Peabody Journal of Education, 81*(3), 94-127.

Glazerman, S., Isenberg, E., Dolfin, S., Bleeker, M., Johnson, A., Grider, M., & Jacobus, M. (2010). *Impacts of comprehensive teacher induction: Final results from a randomized controlled study* (NCEE 2010-4027). Washington, DC: U.S. Department of Education.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved from http://www.tqsource.org/publications/EvaluatingTeachEffectiveness.pdf

Goe, L., & Holdheide, L. (2011). *Measuring teachers' contributions to student learning growth for nontested grades and subjects*. Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved from http://www.tqsource.org/publications/MeasuringTeachersContributions.pdf

Goldhaber, D. D., & Brewer, D. J. (1997). Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *Journal of Human Resources, 32*(3), 505-523.

Goldhaber, D. D., & Liddle, S. (2011). *The gateway to the profession: Assessing teacher preparation programs based on student achievement*. Bothell, WA: Center for Education Data and Research.

Goodlad, J. (1991). Why we need a complete redesign of teacher education. *Educational Leadership, 49*(3), 4-10.

Gottardo, R., & Raftery, A. E. (2009). Bayesian robust variable and transformation selection: A unified approach. *Canadian Journal of Statistics, 37*(3), 361-380.

Grossman, P., Loeb, S., Cohen, J., Hammerness, K., Wyckoff, J., Boyd, D., & Lankford, H. (2010). *Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores* (Working Paper No. 16015). Cambridge, MA: National Bureau of Economic Research.

Guin, K. (2004). Chronic teacher turnover in urban elementary schools. *Education Policy Analysis Archives, 12*(24), 42. Retrieved from http://epaa.asu.edu/epaa/v12n42/

Hanushek, E. A., Kain, J. F., O'Brien, D. M., & Rivkin, S. G. (2005). *The market for teacher quality* (Working Paper No. 11154). Cambridge, MA: National Bureau of Economic Research.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2004). Why public schools lose teachers. *Journal of Human Resources, 39*(2), 326-354.

Harris, D. H., & Sass, T. (2009a). *Teacher training, teacher quality and student achievement* (CALDER Working Paper No. 3). Washington, DC: The Urban Institute.

Harris, D. H., & Sass, T. (2009b). *What makes for a good teacher and who can tell?* (CALDER Working Paper No. 30). Washington, DC: The Urban Institute.

Harris, D. N. (2009). Would accountability based on teacher value added be smart policy? An examination of the statistical properties and policy alternatives. *Education Finance and Policy, 4*(4), 319-350.

Harris, D. N. (2011). *Value-added measures in education: What every educator needs to know*. Cambridge, MA: Harvard Education Press.

Harris, D. N., & Rutledge, S. A. (2010). Models and predictors of teacher effectiveness: A comparison of research about teaching and other occupations. *Teachers College Record, 112*(3), 914-960.

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.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*(2), 371-406.

Ingersoll, R. M., & Strong, M. (2011). The impact of induction and mentoring programs for beginning teachers: A critical review of research. *Review of Educational Research, 81*(2), 201-233.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers: Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26*(1), 101-136.

Johnson, S. M., Birkeland, S., Kardos, S., Kauffman, D., Liu, E., & Peske, H. (2001). *Retaining the next generation of teachers: The importance of school-based support* (Harvard Education Letter Online). Retrieved from http://www.edletter.org/current/support/shtml

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2010). *Identifying effective classroom practices using student achievement data* (Working Paper No. 15803). Cambridge, MA: National Bureau of Economic Research.

Kimball, S. M., White, B., Milanowski, A. T., & Borman, G. (2004). Examining the relationship between teacher evaluation and student assessment results in Washoe County. *Peabody Journal of Education, 79*(4), 54-78.

Labaree, D. (1995). *How to succeed in school without really learning*. New Haven, CT: Yale University Press.

Labaree, D. (2004). *The trouble with ed schools*. New Haven, CT: Yale University Press.

Lockwood, J. R., Doran, H., & McCaffrey, D. F. (2003). Using R for estimating longitudinal student achievement models. *R News, 3*(3), 17-23.

Ludlow, L. H., Pedulla, J., Reagan, E., Enterline, S., Cannady, M., & Chappe, S. (2011). Design and implementation issues in longitudinal research. *Educational Policy Analysis Archives,* 19(11). Retrieved from http://epaa.asu.edu/ojs/article/view/802

Luekens, M. T., Lyter, D. M., & Fox, E. E. (2004). *Teacher attrition and mobility: Results from the Teacher Follow-up Survey, 2000–01* (NCES Rep. No. 2004-301). Washington, DC: U.S. Department of Education, National Center for Educational Statistics.

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education, 79*(4), 33-53.

National Center for Education Statistics. (2005). *The condition of education 2005* (NCES Rep. No. 2005-094). Washington, DC: U.S. Government Printing Office.

National Council for Accreditation of Teacher Education. (2010). *Transforming teacher education through clinical practice: A national strategy to prepare effective teachers*. Washington, DC: Author.

National Research Council. (2010). *Preparing teachers: Building evidence for sound policy* (Committee on the Study of Teaching Preparation Programs in the United States). Washington, DC: The National Academies Press.

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

Pecheone, R. L., & Chung, R. R. (2006). Evidence in teacher education: The performance assessment for California teachers (PACT). *Journal of Teacher Education, 57*(1), 22-36.

Peck, C. A., Gallucci, C., & Sloan, T. (2010). Negotiating implementation of high-stakes performance assessment policies in teacher education: From compliance to inquiry. *Journal of Teacher Education, 61*(5), 451-463.

Pianta, R. C., Hamre, B. K., Haynes, N., Mintz, S. J., & La Paro, K. M. (2006). *Classroom assessment scoring system (CLASS) manual:*

Middle/secondary version pilot. Charlottesville: University of Virginia.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: SAGE.

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one? *Education Finance and Policy, 6*(1), 43-74.

Sawchuck, S. (2011, September 30). Momentum builds for teacher education overhaul. *Education Week*. Available from http://edweek.org

Sawchuck, S. (2012, February 7). "Value added" concept proves beneficial to teacher colleges. *Education Week*. Available from http://edweek.org

Smith, T. M., & Ingersoll, R. M. (2004). What are the effects of induction and mentoring on beginning teacher turnover? *American Educational Research Journal, 41*(3), 681-714.

Soltis, J. F. (1987). *Reforming teacher education: Impact of the Holmes Group report*. New York, NY: Teachers College Press.

Tennessee Higher Education Commission and the State Board of Education. (2011). *2011 Report card on the effectiveness of teacher training programs*. Nashville, TN: Author.

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.

Wineburg, M. S. (2006). Evidence in teacher preparation: Establishing a framework for accountability. *Journal of Teacher Education, 57*(1), 51-64.

Yoon, K. S., Duncan, T., Lee, S. W.-Y., Scarloss, B., & Shapley, K. (2007). *Reviewing the evidence on how teacher professional development affects student achievement* (Issues & Answers Report No. 033). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest.

Zeichner, K. M., & Conklin, H. G. (2005). Teacher education programs. In M. Cochran-Smith & K. Zeichner (Eds.), *Studying teacher education: The report of the AERA panel on research and teacher education* (pp. 645-735). Mahwah, NJ: Lawrence Erlbaum.

About the Author(s)

Margaret L. Plecki is associate professor in educational leadership and policy studies at the University of Washington, Seattle. Her teaching and research interests include teaching quality, leadership, resource allocation, and education policy.

Ana M. Elfers is research assistant professor at the College of Education, University of Washington, Seattle. Her research and teaching focus on education policy and issues of teaching quality and the teacher workforce.

Yugo Nakamura is a graduate research assistant in educational leadership and policy studies at the University of Washington, Seattle. His research focuses on the use of value-added models for the improvement of education policy.