Article

The Politics and Statistics of Value-Added Modeling for Accountability of Teacher Preparation Programs

Journal of Teacher Education
2014, Vol 65(1) 24–38

DOI: 10.1177/0022487113504108
jte.sagepub.com
SAGE

Jane Arnold Lincove[1], Cynthia Osborne[1], Amanda Dillon[1], and Nicholas Mills[1]

Abstract

Despite questions about validity and reliability, the use of value-added estimation methods has moved beyond academic research into state accountability systems for teachers, schools, and teacher preparation programs (TPPs). Prior studies of value-added measurement for TPPs test the validity of researcher-designed models and find that measuring differences across programs is difficult. This study is the first to examine the reliability and usefulness of a value-added model for TPPs developed through a collaborative stakeholder process and mandated by state law for use in accountability. Based on the experience of developing a test-based metric for Texas TPPs, our results suggest that although value-added results are statistically robust, accountability status for individual programs is very sensitive to decisions about accountability criteria, the selection of teachers, and the selection of control variables.

Keywords

educational policy, education reform, school/teacher effectiveness, elementary teacher education

Introduction

Despite lingering questions about reliability and validity, value-added estimation methods are quickly moving from academic research settings to practical education settings, where they are being used as accountability measures for schools and teachers (Harris, 2011). New policies promote the use of value-added models (VAMs) to measure the effectiveness of teacher preparation programs (TPPs) through their effects on the performance of students of program graduates. The U.S. Department of Education has proposed that states use value-added measures in external assessments of training programs, and Race to the Top (RttT) requires that TPP quality be measured with student outcomes (Crowe, 2010; U.S. Department of Education, 2011a). The Council for Accreditation of Educator Preparation (CAEP) has proposed new requirements for accredited programs to use value-added measures in internal assessments (CAEP Standards for Accreditation of Educator Preparation, 2011). Louisiana, Florida, Tennessee, and North Carolina have already published the results of state-mandated assessments of individual TPPs based on performance of students of graduates (Gansle, Noell, Knox, & Schafer, 2010; Henry, Kershaw, Zulli, & Smith, 2012; Henry, Thompson, Fortner, Zulli, & Kershaw, 2010; Noell & Burns, 2006, 2007; Noell, Burns, & Gansle, 2011; Noell, Porter, & Patt, 2007; Tennessee State Board of Education, 2009, 2010).

Value-added measures of TPP effectiveness are the product of an extensive statistical modeling process that requires many analytic choices (Henry et al., 2012). For obvious reasons, the published results of statewide analyses cited above include only one set of estimates with each state's model based on a unique set of choices made by researchers or policy makers. Academic research on value-added estimation suggests that decisions regarding selection, estimation, and interpretation can influence results (Armour-Garb, 2009; Ballou, Sanders, & Wright, 2004; Guarino, Reckase, & Wooldridge, 2012; Harris, 2011; Kane, McCaffrey, Miller, & Staiger, 2013; Koedel, Parsons, Podgursky, & Ehlert, 2012; Mihaly, McCaffrey, Sass, & Lockwood, in press; Rothstein, 2009; Schochet & Chiang, 2010). When VAMs are a component of accountability, it is important if choices made in the research or policy process influence outcomes for individual TPPs in terms of public perception of program quality or consequences for state accreditation.

A second issue with the transition of VAMs from research to accountability is the importance of stakeholder input in the

[1]University of Texas, Austin, USA

Corresponding Author:
Jane Arnold Lincove, University of Texas, P.O. Box Y, Austin, TX 78713, USA.
Email: janelincove@austin.utexas.edu

political process to develop an accountability system. Despite enthusiasm among policy makers, it is unclear whether a VAM designed in a laboratory or academic setting will be acceptable in a policy setting.[1] Because VAMs are highly technical, stakeholders without a background in statistics may find them difficult to comprehend. However, a credible and legitimate accountability system must include stakeholder input and provide results that are useful to consumers and programs (Plecki, Elfers, & Nakamura, 2012). Advocates of test-based accountability argue that results should be clearly communicated to facilitate program improvements (Crowe, 2010). Although researchers are concerned with issues such as clean data and theoretically sound estimation, stakeholders are concerned about the practical consequences of publishing results.

This article seeks to clarify the implications of using VAMs in TPP accountability by comparing results based on different estimation choices. Our data and research questions are derived from the process of developing a pilot accountability measure of TPP effectiveness in Texas. Importantly, the choices we test were identified through a stakeholder participation process where representatives voiced concerns about how programs in Texas would be held accountable for the performance of students of graduates. Comparing VAM results across different modeling choices provides an assessment of whether stakeholder preferences can influence accountability classification for individual programs, effectively stacking the deck for or against certain types of programs. We address two research questions:

> **Research Question 1:** Are estimates of TPP value-added statistically reliable across data selection and estimation choices that matter to stakeholders? Specifically, do results change for individual programs if we include different samples of teachers or different control variables?
> **Research Question 2:** Are interpretations of value-added results for accountability sensitive to the choices of stakeholders? That is, are the same programs identified as having positive and negative effects on student performance if we modify the strategy for estimation and classification?

We find that VAM scores are highly statistically reliable across several choices of data selection and estimations. These results would reassure researchers that estimates are robust to the tested choices. However, when we apply different strategies to translate VAMs into accountability classifications, results for individual TPPs are highly sensitive to the selection of criteria for assigning accountability status and choices concerning sample selection and estimation.

This study presents important new evidence on the use of value-added measures in a state accountability system. Specifically, this study is the first to present central issues of concern for TPP stakeholders. We find that stakeholders are concerned with fairness in how VAMs are calculated and how results are interpreted and shared with the public. Second, this study is the first to illustrate how decisions in value-added modeling can influence quality determinations in an accountability system.

Review of Research

In theory, linking TPP accountability to student performance will increase the quality of teacher training by holding programs responsible for the performance of their graduates in the classroom (Crowe, 2010). Advocates argue for new state accountability measures based on student test scores and public disclosure of results to spur improvements in program quality (Crowe, 2010; Greenberg, McKee, & Walsh, 2013). New state laws and federal recommendations calling for test-based measures of TPPs have proceeded despite a lack of consensus among experts regarding the validity of value-added measures for teachers (Armour-Garb, 2009; Baker et al., 2010; Guarino et al., 2012; Harris, 2011; Rothstein, 2009; Schochet & Chiang, 2010).

There are two types of research specific to measurement of TPP effects: a growing economic literature that tests the capacity of VAMs to measure differences in TPP effects (Goldhaber & Liddle, 2012; Koedel et al., 2012; Mihaly et al., in press) and a policy-oriented teacher education literature that documents and reports state programs that measure differences in TPP effects (Gansle et al., 2010; Henry et al., 2012; Henry et al., 2010; Noell & Burns, 2006, 2007; Noell et al., 2007; Tennessee State Board of Education, 2009, 2010). The former focuses on developing and testing the validity of different theoretical models in a laboratory setting using convenient data, whereas the latter focuses on providing accurate public information on TPP quality and consequences for TPP accreditation. Economic studies in a research setting often generate multiple results based on different modeling choices, and typically find that results are sensitive to choices regarding complex empirical issues such as how researchers address the clustering of TPP effects within teachers (Koedel et al., 2012) and how teachers are assigned to schools based on where they received their training (Mihaly et al., in press). Importantly, many academic studies of TPP effects question whether measurable differences in TPP effects exist at all (Goldhaber & Liddle, 2012; Koedel et al., 2012; Mihaly et al., in press; Osborne, Von Hippel, Lincove, & Mills, 2013). Studies in a policy setting begin with the assumption that TPP effects exist and can be measured and focus on describing and defending a single set of choices believed to be the best fit to a state's data and policy objectives (Henry et al., 2012; Noell & Burns, 2006).

Policy makers often overlook the challenges of estimating differences in TPP effects on student achievement. Henry et al. (2012) present three categories of challenges, each requiring numerous decisions by researchers. The first is selection, which involves decisions regarding which students and teachers are included in the estimation. The second is

estimation, which involves decisions about the empirical model for the VAM. The third challenge, interpretation of results, is where political decisions can have the greatest influence. VAM models produce a continuous distribution of scores around a reference value that must be selected by researchers (Noell & Burns, 2006). Advocates of state accountability promote systems that provide clear identification of TPP quality based on VAMs. Crowe (2010) calls for state accountability based on, "a set of clear signals about program quality that policymakers can understand." As an example, the National Center for Teacher Quality (NCTQ) rates TPPs based on meeting specific standards by assigning each program zero to four stars (Greenberg et al., 2013). To apply this type of tiered rating system to value-added results would require subjective decisions about the definitions and cutoffs for each tier. Inevitably, individual TPPs will fall near these cutoffs. We found no prior research that tests the implications of setting different criteria for the interpretation of VAM results for accountability.

As a group, the economic studies bring into question the capacity of VAMs to accurately measure the effects of TPPs on student performance and the sensitivity of VAMs to researcher choices. It is our objective to fill a gap in the literature on TPP effectiveness by applying the empirical methods of academic studies to a context of state accountability. This study contributes to our understanding of the practical use of VAMs for TPPs by estimating the statistical and practical implications of an important set of choices generated by stakeholders.

Texas Policy Context

Texas has one of the largest and most diverse markets for teacher preparation. In all, 152 Texas TPPs are accredited by the State Board for Educator Certification (SBEC). Texas's decentralized TPP market allows for many types of programs including traditional 4-year undergraduate programs and alternative programs offered by universities, local education agencies, nonprofits, and for-profit firms. In response to growing concerns about quality, the 2009 Texas Legislature established new standards of TPP accountability, one of which was the influence of a program's graduates on student performance on state standardized tests during the first 3 years following certification. Combined with three other standards (pass rates on state certification exams, feedback from school administrators, and the quality of field supervision), the test-based measure of TPP performance would, by law, be publicly available for consumers. The legislation did not specify how the measure would be estimated or used for accreditation. In 2010, the Texas Education Agency (TEA) contracted with researchers at the LBJ School of Public Affairs at the University of Texas at Austin to develop a value-added measure and help determine how the results would be used.[2]

To promote validity and transparency, TEA and the researchers agreed that the process should include ongoing collaboration with stakeholders. Two characteristics of the Texas process are particularly important. First, prior to contracting for the development of a value-added measure, TEA had convened a group of TPP stakeholders to write a principal survey designed to solicit feedback on new teachers. This group consisted of representatives of statewide organizations, as well as some TPPs. Second, many Texas stakeholders believed that TEA's effort to produce a value-added measure for TPPs would be quickly followed by an effort to produce a value-added measure for individual teachers. This raised awareness and concern among teacher groups despite the focus of the current legislation on teacher training only.

When the research project began, membership in the existing stakeholder group was extended to additional organizations through TEA and the research team. Independent of TEA, the Texas Association of Colleges of Teacher Education (TACTE) encouraged its members to participate through a direct appeal from the association president. Any group that requested an invitation from the researchers was welcome to participate. The group quickly expanded to include 50 individual TPPs and four TPP associations, as well as associations of teachers and school administrators (Appendix A provides a full list of participants).

The stakeholder process was funded through TEA's contract with the researchers and managed independently by the research team. To maintain independence, researchers scheduled, hosted, and led meetings and controlled meeting agenda and all communication between the research project and stakeholders. TEA representatives attended meetings to present information and answer questions but did not actively participate in debate or decision making. The researchers hosted six meetings of the stakeholder group during the 18-month process, with each meeting lasting approximately 4 hr. These meetings were well-attended (60-100 attendees), and attendance grew throughout the process.

At each meeting, researchers presented the stakeholders with specific questions that needed to be addressed based on academic research of VAMs and practical issues with Texas's data. For example, stakeholders were asked to discuss which teachers should be included in estimation, how to resolve ambiguity in student-teacher assignments, what control variables to include in the estimation, and how VAM results should be shared with the public. When issues were too complex for large-group discussions, stakeholders were divided into tables of 6 to 12 members to engage in small-group discussions with at least one research team member observing at each table. Researchers encouraged additional input outside of meetings through one-on-one meetings, phone conversations, and email. Whenever possible, researchers tested recommendations from stakeholders with Texas data and reported results at the next meeting. Stakeholders were never given access to individual program results, but were presented with the larger implications of modeling options. The objective was to obtain a consensus

regarding stakeholder views or, absent a clear consensus, evidence of the distribution of preferences.

Some issues were too technical for the diverse members of the stakeholder group. The researchers also convened a statistical advisory subgroup that included technically trained members of the stakeholder group, additional statisticians and methodologists from the faculty of TPPs, and academic researchers who specialize in educator quality in Texas. The statistical advisory group participated in five all-day meetings where researchers presented findings and empirical issues, followed by discussions of the statistical and policy implications of different choices. This expanded the stakeholder process to include basic discussions of value-added modeling and its potential uses, and more technical discussions of complex issues.

Stakeholder Concerns and Choices

Texas stakeholders were generally supportive of efforts to measure the performance of students of TPP graduates. However, there were serious concerns that information shared with the public be accurate, fair, and useful. For this study, we selected a subsample of the major issues raised by the stakeholder group. All are issues that will be confronted by any state attempting to develop a similar metric. Under the category of selection, we test the effects of stakeholder requests to exclude groups of teachers. Under the category of estimation, we test the importance of choices related to the inclusion of student, classroom, and campus covariates requested by stakeholders. Finally, we test the implications of these choices for interpretation of results for accountability, applying different classification criteria suggested by stakeholders. We recognize the importance and relevance of many other choices (see Osborne et al., 2012, for a more inclusive list of stakeholder debates and discussions in Texas). The issues presented here are meant to illustrate a set of choices, and the implications of those choices, that will be increasingly relevant as more states implement test-based accountability for TPPs.

Under the category of selection, we test the implications of excluding two groups of new teachers: probationary teachers and teachers who are teaching outside the area that their TPP trained them to teach. Probationary teachers are in their first year of teaching, but still undergoing training. Some stakeholders argued that attributing student outcomes for these partially trained teachers to a TPP would be unfair. In Texas, teachers can receive a certification-by-exam in any area, regardless of the training program they attended, if they complete the required coursework. There were strong concerns that TPPs should not be accountable for teachers who chose the certificate-by-exam path to teach beyond the scope of their training.

Under the category of estimation, we test the implications of the selection of covariates in the estimation of TPP effects. This is perhaps the most controversial issue related to implementation of value-added measures in accountability systems (McCaffrey, Lockwood, Koretz, & Hamilton, 2003). The simplest VAM predicts TPP effects using only the student's prior test performance as a control for baseline performance. This estimation measures the average growth of students of TPP graduates compared with the average growth of all students. However, typical growth rates are not the same for all groups of students. VAMs that control for student demographics would estimate average TPP effects based on typical growth within a demographic subgroup. This could be sufficient if the distribution of student backgrounds was similar across TPPs. However, there is evidence from Texas and other states that student assignments are associated with a teacher's TPP (Boyd, Lankford, Loeb, Rockoff, & Wyckoff, 2008; Mihaly et al., in press; Osborne et al., 2012). This can occur if graduates of TPPs sort into campuses of different qualities (e.g., if graduates of a well-regarded TPP earn jobs in a wealthy district) and if principals sort new teachers into classrooms based on the TPP they attended (e.g., assigning a new teacher from a well-regarded TPP to a more challenging class). When this occurs, value-added estimates cannot separate TPP effects from correlated student, peer, campus, and neighborhood effects without control variables. Prior studies investigate the implications of adding control variables selected by researchers (Baker et al, 2010; Kane et al., 2013; McCaffrey et al., 2003) and find that campus and classroom covariates are particularly important for removing bias.

Louisiana's TPP VAM researchers selected control variables based on a statistical procedure to identify significant covariates (Noell & Burns, 2006). In Texas, relevant covariates were identified through the stakeholder process and review of available data. After reviewing academic evidence, the objective of the stakeholders was to identify variables that were beyond the control of TPPs and that might influence the estimation of individual TPP effects through the teaching assignments of graduates. It was considered unfair to hold TPPs accountable for issues such as where teachers obtain jobs, how principals assign teachers to classrooms, and how schools and districts support (or fail to support) new teachers through professional development.

Stakeholders identified five constructs that could theoretically bias estimation of differences in TPP effects on student performance. S*tudent demographics* (race, gender, and socioeconomic status) relate to how students from different groups typically perform. S*tudent experiences* relate to whether a student has special needs or circumstances (such as pull-out special education or low attendance) that limit the teacher's influence on performance. *Campus demographic and performance aggregates* relate to the socioeconomic status of the neighborhood and community and can indicate the level of social supports in the community. *Classroom aggregates* relate to peer effects on student performance. These four constructs are similar to theoretical influences discussed in prior literature (Baker et al., 2010; Kane et al., 2013;

McCaffrey et al., 2003). However, Texas stakeholders selected a more extensive set of variables to represent each construct in the value-added estimation. The fifth construct, *campus climate*, relates to the quality of school leadership and supports for new teachers. This construct is not tested in prior research on VAMs but reflects the strong preference of many stakeholders to avoid attributing to TPPs the effects of district policies, school leadership, or teacher supports. Detailed variable lists for each construct are provided in Appendix B.

The VAM

There are many strategies to estimate the effects of TPPs on student performance. To test the effects of specific stakeholder choices, it is necessary to hold other estimation choices constant with a single model. We test the effects of stakeholder decisions using a fixed-effects approach that resembles the models used in economic studies of TPP effects (e.g., Goldhaber & Liddle, 2012; Koedel et al., 2012; Mihaly et al., in press). The advantage of this model is that it provides a direct estimate of the TPP effect with a standard error to measure precision.[3] We model student performance on state standardized tests as a function of the student's prior test performance and the TPP of the student's teacher. Our base empirical model controls for prior test score, student characteristics, and teacher experience:

$$\text{current test} = \text{constant} + b_1 \times \text{prior test} + b_2 \times \text{student characteristics} + b_3 \times \text{teacher experience} + \text{TPP effect},$$

where each student is assigned to the TPP that trained his teacher. By controlling for prior test performance, the estimation of the TPP effects reflects the contribution (or value-added) common to teachers from a TPP that pushes a student above or below her expected score based on past performance. To measure the effects of choices regarding selection of teachers on estimation, we use the base model to estimate value-added scores for four teacher samples proposed by stakeholders: (a) all new teachers, (b) excluding out-of-area teachers, (c) excluding probationary teachers, and (d) excluding both out-of-area and probationary teachers.

To estimate the effects of decisions regarding covariates, we incrementally add covariate groups until we reach a full specification with covariates for all five constructs. Comparison of value-added scores across estimations also requires that effects are measured in relation to the same reference point (Henry et al., 2012; Noell & Burns, 2006). We measure TPP effects as the standardized distance from the grand-weighted mean of all TPP effects.[4] To avoid bias introduced when TPP effects are clustered within teachers, we estimate robust standard errors for clustering at the teacher level (Koedel et al., 2012).

Our dataset, provided by TEA, includes all students in Texas public schools during the 2010-2011 school year.[5] Using these data, we linked each student to a math and reading teacher. Data on teaching certificates and TPPs were merged through the SBEC database. Based on the Texas law, we identified new teachers as those in the first 3 years of teaching. The student dataset includes a comprehensive set of demographic variables, as well as the history of a student's enrollment, attendance, language and special education designations, and test performance in Texas public schools. Campus and classroom aggregates were constructed using data for all students in the group. Additional campus climate variables, such as geography and state accountability status, were merged from the state's public accountability system known as the Academic Excellence Indicators System (AEIS).

The nature of value-added estimation provides some limitations on how student performance can be used for accountability. It is not possible to include all grade levels in a VAM score. Each student must have at least one current test score and one prior observation of student performance. This eliminates all students in untested grades (pre-K to 2) and the first tested grade (Grade 3). Although stakeholders were not comfortable with the exclusion of these untested grade levels from any measure of program quality, there is no data in Texas to calculate value-added in early grades.

We hold constant a set of researcher-driven choices that facilitate identification of the effects of other choices. For example, it is problematic to link performance outcomes to the correct TPP if students have multiple teachers in a subject (Henry et al., 2012). To minimize potential data problems while maintaining a consistent student sample, estimates in this study are based on self-contained classrooms. This maximizes the likelihood that the teacher of record is the teacher who contributed to student performance in both tested areas (reading and math) and therefore those outcomes are attributed to the correct TPP. Grades 4 and 5 are the only tested elementary grades in Texas with prior year tests, so all other grades are excluded. We also exclude students who have no prior test score and students who took the Modified, Alternate, or Spanish TAKS (Texas Assessment of Knowledge and Skills) exams.[6] To create comparable scores across tests, TAKS scale scores are standardized within subject, grade, and year with a mean of zero and standard deviation of one.

Data were available for 507,070 fourth and fifth graders assigned to 22,503 teachers. Of these, 83,184 students were taught by new teachers. The number of fourth- and fifth-grade math teachers associated with a TPP ranges from 1 to more than 300. The number of students associated with a TPP ranges from 1 to more than 8,900. The dataset used to estimate TPP VAMs excludes students with missing data and those not in self-contained classrooms, for a total of 81,667 student observations. Reported results are limited to 141 TPPs that have at least 5 teachers in the dataset.

Table 1 displays summary statistics for Texas fourth and fifth graders by the type of teacher. Reflecting the diverse

Table 1. Mean Values (*SD*) of Student Characteristics by Teacher Certification for Texas Fourth and Fifth Graders.

	All teachers	New teachers	New teachers–permanent certificate	New teachers–probationary certificate	New teacher–originally certified in area	New teacher–teaching out-of-area
Math TAKS 2010	0.003 (0.998)	−0.119 (1.054)	−0.091 (1.038)	−0.256 (1.119)	−0.122 (1.054)	−0.099 (1.050)
Math TAKS 2011	0.004 (0.998)	−0.130 (1.061)	−0.101 (1.046)	−0.275 (1.123)	−0.133 (1.063)	−0.110 (1.054)
Reading TAKS 2010	0.002 (0.999)	−0.105 (1.053)	−0.082 (1.038)	−0.220 (1.116)	−0.107 (1.052)	−0.095 (1.053)
Reading TAKS 2011	0.003 (0.999)	−0.122 (1.056)	−0.098 (1.042)	−0.239 (1.115)	−0.125 (1.057)	−0.099 (1.050)
Female	0.498 (0.500)	0.499 (0.500)	0.498 (0.500)	0.503 (0.500)	0.499 (0.500)	0.498 (0.500)
Free/reduced lunch	0.565 (0.496)	0.629 (0.483)	0.614 (0.487)	0.703 (0.457)	0.629 (0.483)	0.626 (0.484)
Black	0.130 (0.337)	0.150 (0.357)	0.145 (0.352)	0.176 (0.380)	0.151 (0.358)	0.141 (0.348)
Hispanic	0.472 (0.499)	0.517 (0.500)	0.507 (0.500)	0.564 (0.496)	0.519 (0.500)	0.503 (0.500)
Asian	0.039 (0.193)	0.037 (0.189)	0.038 (0.192)	0.030 (0.171)	0.037 (0.188)	0.037 (0.190)
Pacific Islander	0.001 (0.034)	0.001 (0.036)	0.001 (0.038)	0.001 (0.025)	0.001 (0.037)	0.001 (0.029)
Native American	0.004 (0.066)	0.004 (0.062)	0.004 (0.062)	0.004 (0.063)	0.004 (0.062)	0.004 (0.060)
Multiple races	0.002 (0.044)	0.002 (0.043)	0.002 (0.044)	0.001 (0.037)	0.002 (0.043)	0.002 (0.041)
Teacher experience	2.72 (0.75)	1.06 (0.81)	1.21 (0.77)	0.33 (0.60)	1.05 (0.81)	1.12 (0.81)
No. of students	570,070	83,184	69,200	13,984	72,665	10,519
No. of teachers	22,503	3,526	2,930	596	3,094	432
No. of EPPs	141	141	130	78	138	94

Note. TAKS = Texas Assessment of Knowledge and Skills.

demographics of Texas, students are 47% Hispanic, 13% Black, and 57% on free/reduced lunch. Approximately 15% of students were taught by a new teacher, and average experience of new teachers was slightly more than 1 year. Students of new teachers performed below the state average in math and reading and were more likely to be minorities and on free/reduced lunch than the state average. Seventeen percent of new teachers were on a probationary certificate. Probationary teachers were assigned to students who were even farther below the state average in test performance. Thirteen percent of new teachers were teaching outside the area recommended by the TPP.[7] These teachers were assigned to students who were similar to the average for other new teachers. These differences in students taught by different types of teachers suggest that choices regarding teacher and covariate selection in value-added modeling can influence results, as teaching assignments do vary for teachers with different types of certificates.

Statistical Implications of Value-Added Modeling Choices

Research Question 1 examines the statistical implications of the stakeholder concerns described above by estimating a VAM for each choice and examining the correlation of results across estimations. We present results here based on the 2011 math TAKS.[8] To test the effect of sample selection, we estimated TPP effects for four different samples of fourth- and fifth-grade students. The core analytic sample includes all students of new teachers. Assessing the reliability of TPP effects across teacher selection choices requires that the same sample of TPPs be used for all estimations. Some TPPs have only probationary or out-of-area teachers; thus, tests of reliability are estimated using a stable subsample of 126 TPPs that have five or more teachers in each subsample.

We define statistical reliability as, "consistency with which results occur" (Triola, 1997). For this study, we examine consistency of a TPP's estimated value-added score across samples and covariates. We calculate the Pearson correlation coefficient of VAMs across different pairs of choices as a measure of this association. As a rule of thumb for interpretation, coefficients greater than .50 are moderately positive, coefficients greater than .70 are high positive, and coefficients greater than .90 are very high positive (Hinkle, Wiersma, & Jurs, 2002).

Pearson correlations (r statistics) of TPP value-added scores across the four samples for math are displayed in Table 2. VAM results excluding out-of-area teachers have a very high association with results including all teachers (r = .95). Results excluding probationary teachers have a weaker association, but correlations still meet the high positive standard (r = .86). The weakest association is between the sample that excludes probationary teachers and the sample that excludes out-of-area teachers, but there is a high positive association (r = .80). Overall, VAMs estimated with different samples are highly or very highly associated, but never perfectly associated (r = 1.0). A VAM estimated from one sample is never perfectly predictive of a VAM estimated with an alternate sample.

Correlations across results estimated with different covariates are displayed in Table 3. VAMs estimated with different sets of covariates all have very high positive

Table 2. Correlations of TPP Value-Added Scores Across Selection Choices.

Samples	All new teachers	Exclude out-of-area	Exclude probationary	Exclude out-of area and probationary
All new teachers	1.000			
Exclude out-of-area	.950	1.000		
Exclude probationary	.857	.798	1.000	
Exclude out-of area and probationary	.846	.852	0.966	1.000

Note. Estimates control for prior test score, student demographics, and teacher experience. Includes 126 TPPs with at least five teachers in each sample. All correlations are significant at $p < .001$. TPP = teacher preparation program.

Table 3. Correlations of TPP Value-Added Scores Across Estimation Choices.

	Base model	Add student experiences	Add campus aggregates	Add classroom aggregates	Add campus climate
Base model	1.000				
Add student experiences	.996	1.000			
Add campus aggregates	.977	.986	1.000		
Add classroom aggregates	.974	.984	.998	1.000	
Add campus climate	.965	.974	.985	.987	1.000

Note. Base model estimates control for prior test score, student demographics, and teacher experience. Analytic dataset includes 126 TPPs with at least five new teachers. All correlations are significant at $p < .001$. TPP = teacher preparation program.

Table 4. Correlations of TPP Value-Added Scores Across Selection and Estimation Choices.

	All new teachers	Exclude out-of-area	Exclude probationary	Exclude out-of area and probationary
Base model	1.000	.950	.857	.846
Add student experiences	.996	.949	.865	.855
Add campus aggregates	.977	.933	.848	.842
Add classroom aggregates	.974	.926	.842	.835
Add campus climate	.965	.922	.850	.845

Note. Base model estimates control for prior test score, student demographics, and teacher experience. Estimates down the vertical access include all new teachers and alter covariates. Estimates along the horizontal access alter the teacher sample. Analytic dataset includes 126 TPPs with at least five teachers that meet criteria for each sample. All correlations are significant at $p < .001$. TPP = teacher preparation program.

associations with each other (ranging from $r = .97$ to $r > .99$). Even with the addition of the stakeholders' highly inclusive full list of covariates, correlations with the base model still exceed the standard for a very high positive association. Table 4 illustrates the effect of changing selection and covariate choices. Again, all correlations achieve the high positive or very high positive standard even when selection and covariate choices are changed simultaneously (ranging from $r = .84$ to $r > .95$).

This evidence suggests a high to very high level of statistical reliability among value-added scores estimated with different teacher samples and covariates. Researchers would consider the TPP VAMs robust to the selection and estimation decisions tested here, and in a research setting, we might conclude that stakeholder preferences were largely irrelevant to the estimation of the VAMs. We next examine whether this conclusion is problematic in an accountability setting.

Accountability Implications of Value-Added Modeling Choices

Research Question 2 addresses how stakeholder choices influence accountability determinations. Value-added scores are not easy to interpret. This is illustrated in Figure 1, which displays the range of value-added scores for 141 TPPs estimated with the base model and the sample of all new teachers for fourth- and fifth-grade math. Each circle is a single TPP. Each TPP effect reflects how a program is predicted to influence student performance. The full range of TPP effects is about one standard deviation on the TAKS test, and most scores are close to the mean of zero. It is not immediately apparent from this distribution which TPPs—if any—are substantively better or worse.

There are numerous options for interpretation and no clear guidance from national policy or Texas statute. North

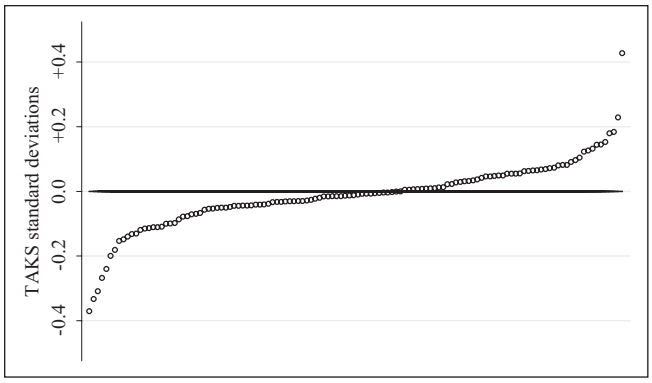

Figure 1. TPP value-added scores for math.
Note. Value-added scores for 141 TPPs using sample of all new teachers and base model. Value-added scores are centered at a mean of zero.
TPP = teacher preparation program.

Carolina's TPP effectiveness scores were published as continuous values (Henry et al., 2010). Tennessee looked at the distribution of individual teacher VAMs across quintiles to determine whether, for each TPP, statistically more teachers were in the highest than the lowest quintile (Tennessee State Board of Education, 2010). Quintiles and other statistically determined bins create cut-points for quality classifications. Teachers just above or below a cut-point (in this case the border of a quintile) are classified as different despite being mathematically similar. Other states used standard errors of TPP estimates to identify programs that were statistically high-performing or statistically low-performing (American Institutes for Research, 2011; Noell & Burns, 2006).

For illustration here, we select a simple and plausible accountability structure that sorts TPPs into three groups—low, average, and high. Texas stakeholders were wary of any system (such as ranking or listing continuous values) that did not acknowledge the presence of estimation error in the value-added scores. There were particular concerns that statistically insignificant differences in scores would affect public perception or accreditation status. We test the influence of three decision rules for sorting that were proposed by stakeholders. All three compare individual scores to average performance (which we set at a value-added score of zero). The first rule, most often applied in academic research, uses a 95% confidence interval. Low TPPs are significantly below zero, average TPPs are statistically equal to zero, and high TPPs are significantly above zero. Some stakeholders supported a more rigid standard, as the 95% confidence interval is expected to misidentify 5% of TPPs. The second rule applies a more conservative 99% confidence interval.

The third strategy, which had the most support among stakeholders, grew from concerns that the number of TPP graduates would be too influential in a system based on statistical significance. For example, a large TPP might have a statistically significant score, even if the size of the effect was very small, while a small TPP might have a statistically insignificant score even if the effect size was large. Stakeholders preferred comparison with an absolute effect size indicative of a meaningful or "educationally significant" contribution to student performance. The stakeholders selected 0.25 standard deviations from the grand mean as a measure of an educationally significant effect size. There was some precedent for this cut-point. TEA's Best Practices Clearinghouse set 0.25 as the minimize effect size associated with a "practice with strong statistical evidence" (TEA, 2011), and the national What Works Clearinghouse defines 0.25 as an effect that is "substantially important" (U.S. Department of Education, 2011b). Our third sorting rule identifies high performance as scores greater than 0.25 standard deviations, low performance as scores less than −0.25 standard deviations, and average effects as any values in between.

The three criteria are imposed on the distribution of TPP VAM scores in Figure 2. TPPs with high or low ratings by each criterion are highlighted with dark circles. It is apparent that statistical significance identifies some TPPs with small effect sizes as high- or low-performing, while TPPs with larger effects are classified as average. The educational significance criterion identifies some TPPs as low-performing with effects that are large but not significantly different than zero.

Table 5 displays the number of TPPs identified as low-performing, average, or high-performing across the three sorting rules. We take 95% confidence as the base criteria and measure changes in classification as we move from 95% confidence to an alternate rule. The vertical axis of each table displays the distribution of TPPs under the 95% confidence rule. The horizontal axis shows the distribution of the same TPPs under an alternative rule. The diagonal from top left to bottom right counts programs that have the same classification under both rules. Counts off the diagonal indicate programs that change classification.

The 95% confidence rule classifies 17 TPPs as low-performing (12%), 22 as high-performing (16%), and the remaining 102 as average (72%). With the transition from 95% confidence to 99% confidence, 7 programs (5%) improve classification, moving from negative to average, and 6 programs (4%) decline, moving from positive to average.

Comparing the 95% confidence rule to the stakeholders' measure of educational significance (an effect size of ±0.25), 37 programs (26%) change classification. Educational significance classifies only 4% of the programs as either low- or high-performing. Of the 22 programs that are classified as high-performing under the 95% confidence rule, all but 2 decline to average under the educational significance rule, and of the 17 programs that are low-performing under the 95% confidence rule, all but 2 improve to average under the educational significance rule. Two additional programs classified as average under 95% confidence decline to low-performing under educational significance, an indication of

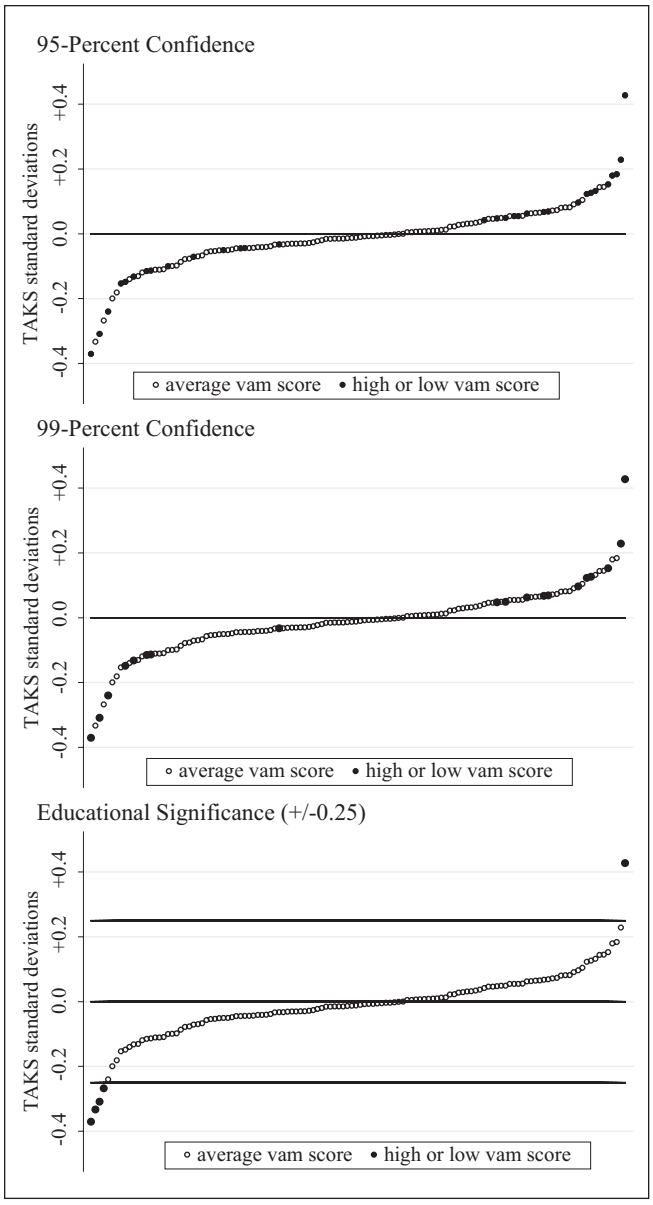

Figure 2. TPP value-added scores with different criteria for sorting.
Note. TPP = teacher preparation program.

large, negative value-added scores that were not statistically different from zero. While the educational significance rule protects against identifying small effects as important, it is so conservative that it provides little insight to differentiate TPPs.

We next examine the implications of sample and estimation choices holding the criterion for interpretation constant. Table 6 shows the distribution of TPP status for the sample of all new teachers compared with alternative samples. We apply the 95% confidence rule to classify program effects. In addition to programs that change status, changing the teacher sample also excludes some TPPs from analysis if they do not have five teachers left in the sample. Importantly, these

Table 5. Comparison of TPP Accountability Status Across Different Criteria.

95% Confidence	Negative	Average	Positive	Total
		99% confidence		
Negative	10	7	0	17
Average	0	102	0	102
Positive	0	6	16	22
Total	10	115	16	141
		Educational significance		
Negative	2	15	0	17
Average	2	100	0	102
Positive	0	20	2	22
Total	4	135	2	141

Note. 95% confidence criteria identifies programs as high-performing/low-performing if the estimated TPP effect is significantly greater/less than zero with 95% confidence ($p < .05$). 99% confidence criteria identifies programs as high-performing/low-performing if the estimated TPP effect is significantly greater/less than zero with 99% confidence ($p < .01$). Education significance criteria identifies programs as high-performing/low-performing if the estimated TPP effect is greater/less than 0.25 standard deviations from the mean. TPP = teacher preparation program.

Table 6. Comparison of TPP Accountability Status Across Teacher Selection.

All new teachers	Negative	Average	Positive	Excluded	Total
	Exclude out-of-area				
Negative	14	2	0	1	17
Average	3	96	3	0	102
Positive	0	1	19	2	22
Total	17	99	22	3	141
	Exclude probationary				
Negative	13	2	0	2	17
Average	5	88	3	6	102
Positive	0	3	16	3	22
Total	18	92	19	11	141
	Exclude out-of-area and probationary				
Negative	11	3	0	3	17
Average	6	84	5	7	102
Positive	0	3	14	5	22
Total	17	90	19	15	141

Note. Identifies programs as high-performing/low-performing if the estimated TPP effect is significantly greater/less than zero with 95% confidence ($p < .05$). TPP = teacher preparation program.

programs would have no accountability classification. Excluding out-of-area teachers results in three programs (2%) dropping from classification. Nine programs (6%) remain in the estimation but change classification. Contrary to the assumption that including out-of-area teachers would harm TPPs, five programs improve and four decline.

Excluding probationary teachers has a more profound effect with 11 programs (8%) excluded, and 13 programs (9%) changing classification. Interestingly, the exclusion of probationary teachers, who have received the least training,

Table 7. Comparison of TPP Accountability Classification Across Covariate Selection.

Base model	Negative	Average	Positive	Total
	Add student covariates			
Negative	15	2	0	17
Average	2	99	1	102
Positive	0	1	21	22
Total	17	102	22	141
	Add campus covariates			
Negative	15	2	0	17
Average l	1	98	3	102
Positive	0	2	20	22
Total	16	102	23	141
	Add classroom covariates			
Negative	15	2	0	17
Average	0	99	3	102
Positive	0	1	21	22
Total	15	102	24	141
	Add campus climate covariates			
Negative	13	4	0	17
Average	2	97	3	102
Positive	0	3	19	22
Total	15	104	22	141

Note. Identifies programs as high-performing/low-performing if the estimated TPP effect is significantly greater/less than zero with 95% confidence ($p < .05$). TPP = teacher preparation program.

is not always beneficial. Of the 13 programs that change classification, 5 improve and 8 decline. Excluding probationary and out-of-area teachers compounds this instability in interpretation of results. Fifteen programs (11%) are excluded, and 17 programs change classification (12%). Thus, despite high statistical correlations, the choice of teacher sample can influence up to 1 in 5 TPPs with either a reclassification or exclusion from classification.

Table 7 illustrates the instability of classifications across estimations with different covariates. We begin (on the vertical axis) with a base model that controls for prior test score, student demographics, and teacher experience. We then incrementally add covariate groups. All estimates in Table 7 include the same sample of all new teachers, so no TPPs are excluded. Each covariate addition results in at least six changes in classification. As predicted by prior research, campus-level variables induce the greatest number of changes. Eight programs (6%) change classification with the addition of *campus aggregates*, and 12 programs (9%) change status with the addition of *campus climate*. It is notable that classification is most sensitive to the group of covariates selected to reflect campus climate, a category of variables that has not been included in prior academic studies of TPP effects.

As a final illustration of the implications of estimation strategies, Figure 3 illustrates the relationship between VAM scores and classification. The scatterplots display VAM scores from two estimations for TPPs that are significantly low- or high-performing using the 95% confidence rule in at least one of the two estimations (those that are always average are excluded to avoid overcrowding the figures). Programs that change classification across two estimations are represented by open circles. Programs that have the same accountability status in both estimations are represented by plus signs. Figure 3 illustrates that changes in classification are often associated with very small changes in value-added scores, while changes in effect sizes for other TPPs may not result in any change in classification. This suggests that the classification changes tabulated in Tables 6 and 7 are due to small (or sometimes no) changes in estimated value-added scores. The decision rule of educational significance provides the most stable results, because it depends only on effect size and not statistical significance.

It is important to note that all the results here are based on some simplifications to facilitate comparisons. In a practical setting, we would not limit estimation to fourth- and fifth-grade self-contained classes, and VAMs for multiple subjects would be estimated. It is likely that the statistical and practical implications of stakeholder choices would be compounded in a comprehensive evaluation system that included all grade levels and subjects.

Discussion

In a pure research setting, issues of sample selection and covariates would be settled through theory development and testing. In an accountability setting, these issues must be addressed with consideration of stakeholder preferences and policy goals. While a national agenda moves toward requiring test-based accountability for TPPs, the process of developing a test-based metric in Texas illustrates that stakeholders have strong preferences about how TPP effects are calculated, and the policy implications of these choices are profound. Meanwhile, the Florida value-added accountability system, in which design elements were selected by legislative mandate instead of through stakeholder input, is facing stakeholder lawsuits questioning the measure's validity (O'Connor, 2013). This study illustrates that stakeholders have profound and reasonable concerns about value-added modeling in a research setting and its application to accountability. We also illustrate how accountability classifications for individual programs can depend on decisions made in the process. Although not tested here, it is likely that other issues such as data quality and test characteristics are also influential in value-added results. The growing emphasis on including value-added measures in educational accountability requires that statistical expertise and political processes be connected so that (a) stakeholders can understand value-added measures and participate in their development and (b) experts are sensitized to the policy implications of data limitations and analytic strategies.

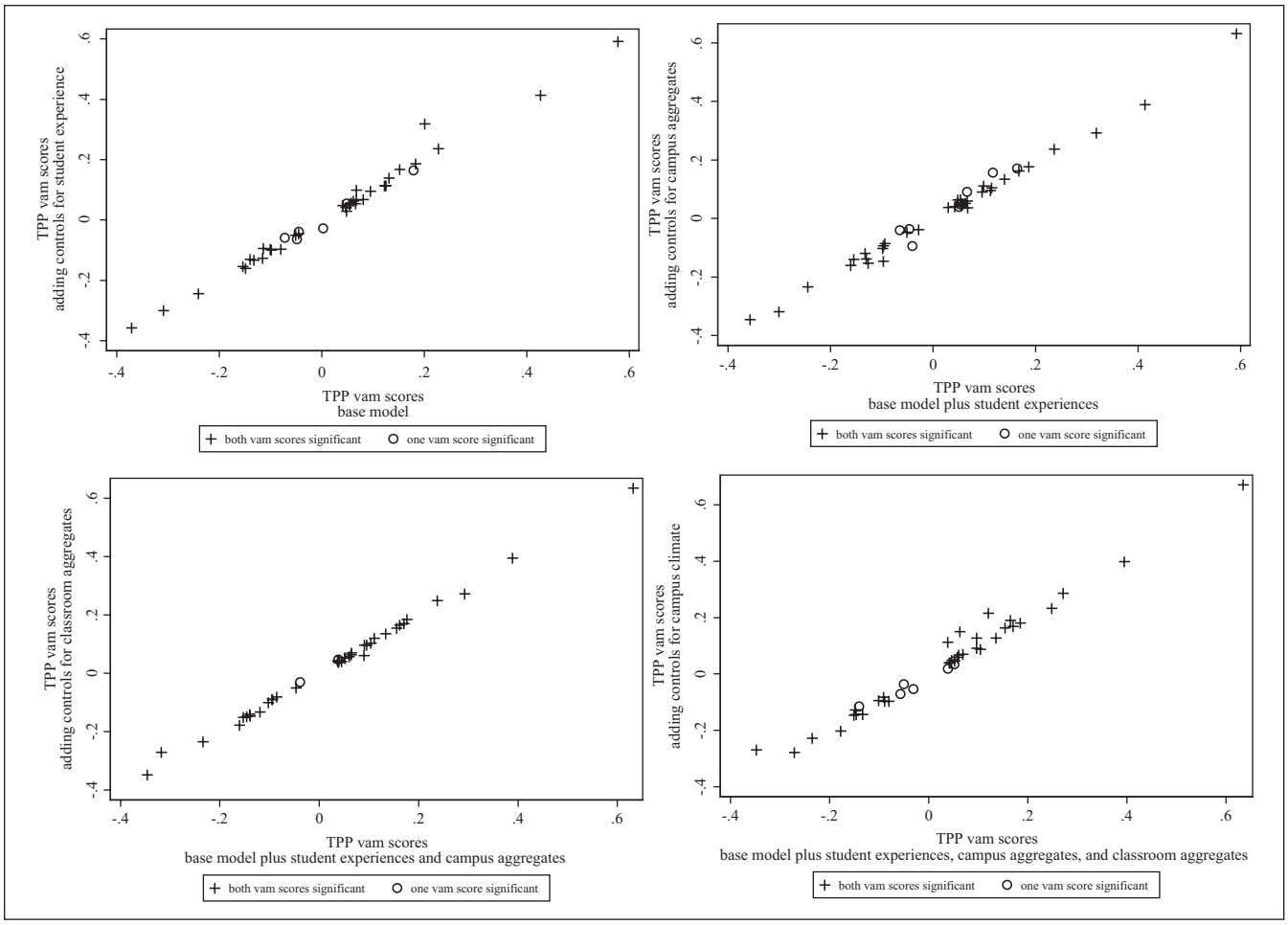

Figure 3. Scatterplots of TPP value-added scores for math, base model versus models with additional covariates.
Note. Base model includes controls for prior test score, student demographics, and teacher experience. Sample includes all new teachers. Results are displayed for TPPs that have at least one statistically significant VAM score at 95% confidence level. TPP = teacher preparation program.

From a research standpoint, the decisions discussed here may seem trivial. TPP effects are statistically reliable across samples and estimations. From a policy and accountability standpoint, however, each change to the base model and teacher sample resulted in reclassification of at least one TPP, with some decisions changing the status of up to 20% of programs. These results are supported by other academic studies that illustrate problems with the precision of VAMs (Koedel et al., 2012; Mihaly et al., in press). Although other states have not published results based on alternative empirical choices, it is likely that results, and accountability consequences, would vary there as well. By combining the empirical approach of academic studies and the context of accountability in Texas, this study provides evidence that problems in estimation of VAMs can lead to different practical interpretations with potentially high-stake consequences.

Our analysis also highlights the ambiguity of value-added metrics in accountability systems without clear criteria to identify success and failure. Policy makers often expect value-added measures to tell us which programs, schools, or teachers are good or bad. They also assume that value-added results provide useful feedback to programs. In practice, value-added scores create a highly aggregated measure that can be interpreted in different ways. The number of TPPs identified as low-performing in our estimations varies from a high of 17, based on a 95% confidence interval, to a low of 4, based on our stakeholder definition of educational significance. More importantly, different TPPs are identified based on different criteria. The sensitivity of accountability status to sampling and modeling also depends on the rigor of the criteria for identifying negative effects. States will need to grapple with the difficult trade-off between the risk of misidentifying a TPP (which will occur 5% of the time with a 95% confidence rule) and the lack of information provided by an accountability system that identifies almost all programs as average. Reliance on VAMs to determine program quality also limits the scope of evaluation. No information is provided for programs that specialize in untested grades (typically early elementary and upper high school grades) or untested subjects (such as fine arts or vocational courses),

and no information is provided about which specific training methods and strategies are the most effective.

States implementing value-added measures will also need to consider the stakes attached to accountability. Hill (2009) argues that high stakes for teachers are inappropriate when estimates lack statistical validity. For TPPs operating in a competitive market, even seemingly low-stakes accountability, such as posting value-added scores or rankings on a consumer website, could have significant consequences. A low rating could induce potential teacher candidates to enroll elsewhere. More importantly, in a competitive teacher job-market, rankings could influence school district hiring decisions and the job-market prospects of graduates. At the same time, it will be a slow and costly process for TPPs to change their programs based on student performance results. Adding higher stakes, such as threats to state accreditation status, only increases the importance of validity and reliability.

Although value-added measurement is intended to provide an objective measure of the quality of TPPs, the Texas experience highlights the ambiguity of use in accountability systems. Texas TPPs were eager to receive information on the performance of their graduates but also concerned about the fairness and transparency of results. This analysis suggests that modeling and sampling decisions have important policy implications for accountability and points to the lingering subjectivity of TPP assessment. Due to the high level of necessary statistical expertise, it is common for states to contract with private research firms to create TPP VAMs (American Institutes for Research in Florida and Texas; EVAAS in Tennessee). In this context, stakeholders and policy makers may be even more detached from decisions that influence accountability. Given the ambiguity regarding the relationship between research decisions and results, value-added measures for accountability may not be worth the price or controversy they create. In the future, we can determine whether early adopters of test-based accountability for TPPs see improvement in student performance through changes in teacher training that does not occur in other states. We predict that programs will have difficulty using state-produced VAMs to generate program improvements that influence student achievement, and a long-term investment in value-added measures for TPPs is unlikely to be an effective strategy to improve teacher training.

Appendix A

Members of the Texas Stakeholder Group

A+ Texas Teachers
Abilene Christine University
Alamo Community College
Angelo State University
Association of Texas Professional Educators
Austin College
Austin Community College
Consortium of State Organizations for Texas Teacher Education
Dallas Independent School District
Education Deans of Independent Colleges & Universities of Texas
Education Service Center—Region IV
Education Service Center—Region VI
Education Service Center—Region XII
Education Service Center—Region XIII
Education Service Center—Region XX
Hardin-Simmons University
Harris County Department of Education
Houston Federation of Teachers
Huston-Tillotson University
iTeachTexas/K&L Gates, L.L.P.
Lamar University
Pflugerville Independent School District
Prairie View A&M University
Round Rock ISD/Canyon Vista Middle School
Sam Houston State University
Southwest Conference on Language Teaching
Southwestern Adventist University
Southern Methodist University
St. Edwards University
St. Mary's University
Stephen F. Austin University
Tarleton State University
Texas A&M International University
Texas A&M University
Texas A&M University—Commerce
Texas A&M University—Corpus Christi
Texas A&M University—San Antonio
Texas Alternative Certification Association
Texas American Federation of Teachers
Texas Association of Colleges for Teacher Education
Texas Association of School Boards
Texas Association of School Personnel Administrators
Texas Association of Secondary School Principals
Texas Charter School Association
Texas Christian University
Texas Classroom Teachers Association
Texas Coordinators for Teacher Certification Testing
Texas Elementary Principals and Supervisors Association
Texas Lutheran University
Texas State Teachers Association
Texas State University—San Marcos
Texas Tech University
The New Teacher Project/Texas Teaching Fellows/Teach for America
Trinity University
University of Houston
University of Houston—Clear Lake
University of Houston—Victoria
University of the Incarnate Word
University of Mary Hardin-Baylor
University of North Texas

University of St. Thomas
University of Texas at Arlington
University of Texas at Austin
University of Texas at Permian Basin
University of Texas at Tyler
West Texas A&M University

Appendix B

Summary Statistics for Additional Covariate Groups.

Student covariates	*M*	*SD*
No. of years on free/reduced lunch	1.928	1.685
LEP	0.128	0.334
No. of years on LEP	0.597	1.252
Special education	0.045	0.208
High-inclusion special education	0.010	0.100
Low-inclusion special education	0.001	0.022
Days present	167.04	18.12
Days enrolled	172.21	17.50
Skipped a grade	0.019	0.135
Repeated a grade	0.084	0.278
TAKS tested with accommodations	0.018	0.132
Ever took Spanish TAKS	0.027	0.162
Campus aggregates	***M***	***SD***
% free/reduced lunch	0.643	0.273
% Black	0.135	0.167
% Hispanic	0.525	0.298
% LEP	0.265	0.209
% special education	0.809	0.029
Classroom aggregates	***M***	***SD***
% free/reduced lunch	0.640	0.288
% Black	0.154	0.201
% Hispanic	0.512	0.318
% LEP	0.147	0.232
% special education	0.084	0.084
Mean math TAKS 2011	−0.119	0.526
Class size	16.766	6.177
Campus climate variables	***M***	***SD***
Campus TAKS pass rate 2011	0.819	0.173
District TAKS pass rate 2011	0.834	0.064
% teacher turnover	0.147	0.115
No. of principals since 2007	1.496	0.744
% disciplinary referrals	0.388	0.778
% state funds	0.448	0.172
Campus failed state accountability	0.007	0.086
District failed state accountability	0.027	0.163
Urban	0.159	0.366
Rural	0.025	0.158
Charter schools	0.052	0.222

Note. LEP = Limited English proficiency; TAKS = Texas Assessment of Knowledge and Skills.

Acknowledgments

The authors are grateful to the Texas stakeholders and members of the Project for Educator Effectiveness and Quality (PEEQ) technical advisory committee; Jerel Booker, Priscilla uino, and other Texas Education Agency (TEA) staff; Paul von Hippel, Laura Bellows, and other PEEQ staff and graduate assistants. We thank Jesse Rothstein, Cory Koedel, and anonymous reviewers for helpful comments on earlier drafts. This research is independent of TEA and does not reflect the views of the the agency, its staff, or the State of Texas.

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) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by the Texas Education Agency through the Project for Educator Effectiveness and Quality (PEEQ) at the Center for Health and Social Policy at the University of Texas at Austin.

Notes

1. For example, Florida's teacher value-added results are being challenged in court by teacher groups (O'Connor, 2013).
2. See Osborne et al. (2012) for a full description of the teacher preparation programs (TPP) accountability system in Texas and detailed discussion of the development of Texas's pilot metric.
3. Alternative state accountability models use a hierarchical linear model (Louisiana), aggregation of teacher value-added scores (Tennessee), and student percentile growth models (Colorado). Of these three, the Louisiana model is the most similar in that it directly estimates a TPP effect with a standard error. The latter two options were rejected by Texas stakeholders (and are also inappropriate for the empirical work here) because they do not provide a standard error to construct a statistical confidence interval around the estimated effect size.
4. A common alternative in value-added research is to compare TPPs with a common single omitted TPP. This approach is less useful in accountability setting, because one TPP needs be selected as the omitted group and therefore would serve as the benchmark for TPP performance rather than receiving its own accountability measure.
5. Many studies of teacher and TPP value-added models recommend using multiple years of data to establish a stable estimate of a teacher or program effect. Texas only had 1 year of student-teacher linked data when the state mandated a TPP metric. Therefore, our results (and Texas's accountability system) are based on only a single year of data.
6. The Modified, Alternate, and Spanish TAKS (Texas Assessment of Knowledge and Skills) are scored on different scales that are not compatible with value-added estimation.
7. There are several ways that elementary schoolteachers could teach outside their recommended area. For example, a teacher certificated for Grades 6 to 8 can take online courses and test into a Grades 4 to 8 certificate. Texas also has several elementary certificates that do not include Grades 4 to 5. For example, a teacher originally certified for an early primary certificate (Grades K-2) could test for a K-6 certificate without additional coursework.

8. Results for reading are often different than math for individual TPPs, but implications for reliability of estimates and stability of accountability classifications are similar.

References

American Institutes for Research. (2011, August 1-2). *Florida's Value-Added Model: Technical assistance workshop*. Presentation to the Student Growth Implementation Committee, Orlando, FL. Retrieved from http://www.fldoe.org/committees/pdf/august12tammpres.pdf

Armour-Garb, A. (2009). Should "value added" models be used to evaluate teachers? *Journal of Policy Analysis and Management*, *28*, 692-712.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., . . .Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers* (Vol. 278). Washington, DC: Economic Policy Institute.

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

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.

CAEP Standards for Accreditation of Educator Preparation. (2011). In *Council for the Accreditation of Educator Preparation*. Retrieved from http://caepnet.org/accreditation/standards/

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

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*. Baton Rouge: Louisiana State University.

Goldhaber, D., & Liddle, S. (2012, January). *The gateway to the profession: Assessing teacher preparation programs based on student achievement* (CALDER Working Paper 65). Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

Greenberg, J., McKee, A., & Walsh, K. (2013). *Teacher prep review: A review of the Nation's Teacher Preparation Programs 2013*. Washington, DC: National Council of Teacher Quality. Retrieved from http://www.nctq.org/dmsView/Teacher_Prep_Review_2013_Report

Guarino, C., Reckase, M., & Wooldridge, J. (2012, June). *Can value-added measures of teacher education performance be trusted* (Working Paper., Vol. 18). East Lansing: Michigan Education Policy Center, Michigan State University.

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

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*(5), 335-355.

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: University of North Carolina at Chapel Hill, College of Arts and Sciences, Carolina Institute for Public Policy.

Hill, H. C. (2009). Evaluating value-added models: A validity argument approach. *Journal of Policy Analysis and Management*, *28*(4), 700-709.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2002). *Applied statistics for the behavioral sciences* (5th ed.). Stamford, CT: Cengage.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013, January). *Have we identified effective teachers? Validating measures of effective teaching using random assignment* (Research Paper, MET Project). Seattle, WA: Bill & Melinda Gates Foundation.

Koedel, C., Parsons, E., Podgursky, M., & Ehlert, M. (2012, July). *Teacher preparation programs and teacher quality: Are there real difference across programs?* (CALDER Working Paper No. 79). Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003).*Evaluating value-added models for teacher accountability* (Monograph). Santa Monica, CA: RAND Corporation.

Mihaly, K., McCaffrey, D., Sass, T., & Lockwood, J. R. (in press) Where you come from or where you go: Distinguishing between school quality and the effectiveness of teacher preparation program graduates. *Journal of Education Finance and Policy*.

Noell, G. H., Burns, J. M., & Gansle, K. A. (2011). *Value-added teacher preparation programs in Louisiana: 2007-08 to 2009-10*. Baton Rouge, LA: Louisiana Board of Regents.

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: 2004–2007*. Retrieved from http://www.regents.state.la.us/Academic/TE/Value%20Added.htm

O'Connor, J. (2013, March 15). What the Florida teacher evaluation lawsuit could mean for other states, *State Impact*. Retrieved from http://stateimpact.npr.org/florida/2013/05/15/what-the-florida-teacher-evaluation-lawsuit-could-mean-for-other-states/

Osborne, C., Lincove, J. A., Von Hippel, P., Mills, N., Dillon, A., & Bellows, B. (2012, June). *The Texas report: Education preparation programs' influence on student achievement* (Project on Education Effectiveness and Quality). University of Texas at Austin.

Osborne, C., Von Hippel, P., Lincove, J. A., & Mills, N. (2013, March). *The small and unreliable effects of teacher preparation programs on student test scores in Texas*. Presented at the Conference of the Association of Education Finance and Management. New Orleans, LA.

Plecki, M. L., Elfers, A. M., & Nakamura, Y. (2012). Using evidence for teacher education program improvement and accountability: An illustrative case of the role of value-added measures. *Journal of Teacher Education*, *63*(5), 318-334.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education*, *4*(4), 537-571.

Schochet, P. Z., & Chiang, H. S. (2010, July). *Error rates in measuring teacher and school performance based on student test score gains* (No. NCEE 2010-4004). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

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

Tennessee State Board of Education. (2010). *Report card of the effectiveness of teacher training programs*. Retrieved from http://www.tn.gov/thec/Divisions/fttt/account_report/2011reportcard/report_card.shtml

Texas Education Agency. (2011). *How to interpret effect sizes. Best practices clearing house*. Retrieved from http://www.tea.state.tx.us/best_practice_standards/how_to_interpret_effect_sizes.aspx

Triola, M. (1997). *Elementary statistics* (7th ed.). Reading, MA: Addison Wesley.

U.S. Department of Education. (2011a, September). *Our future, our teachers: The Obama administration's plan for teacher education reform and improvement*. Author. Retrieved from http://www.ed.gov/sites/default/files/our-future-our-teachers.pdf

U.S. Department of Education. (2011b, September). *What works clearinghouse: Procedures and Standards Handbook (Version 2.1)*. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v2_1_standards_handbook.pdf

Author Biographies

Jane Arnold Lincove is an assistant professor at the LBJ School of Public Affairs at the University of Texas at Austin and codirector of the Project for Educator Effectiveness and Quality.

Cynthia Osborne is an associate professor at the LBJ School of Public Affairs at the University of Texas at Austin and director of the Project of Educator Effectiveness and Quality and the Child and Family Research Partnership.

Amanda Dillon is a graduate of the Masters of Global Policy Studies program at LBJ School of Public Affairs at the University of Texas at Austin and research associate for the Project for Educator Effectiveness and Quality.

Nicholas Mills is a graduate of the Masters of Public Affairs program at the LBJ School of Public Affairs at the University of Texas at Austin. He worked as a research associate for the Project for Educator Effectiveness and Quality from 2010 to 2012.