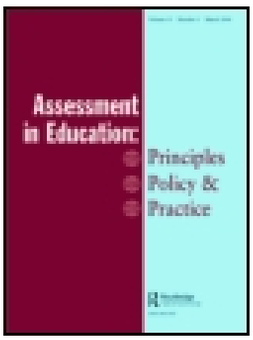

Routledge
Taylor & Francis Group

Assessment in Education: Principles, Policy & Practice

ISSN: 0969-594X (Print) 1465-329X (Online) Journal homepage: http://www.tandfonline.com/loi/caie20

Value-added assessment of teacher preparation programs in the United States: a critical evaluation

Carla M. Evans & Jade Caines Lee

To cite this article: Carla M. Evans & Jade Caines Lee (2016): Value-added assessment of teacher preparation programs in the United States: a critical evaluation, Assessment in Education: Principles, Policy & Practice, DOI: 10.1080/0969594X.2016.1255180

To link to this article: http://dx.doi.org/10.1080/0969594X.2016.1255180

Published online: 14 Nov 2016.

ASSESSMENT IN EDUCATION: PRINCIPLES, POLICY & PRACTICE, 2016
http://dx.doi.org/10.1080/0969594X.2016.1255180

Routledge
Taylor & Francis Group

Value-added assessment of teacher preparation programs in the United States: a critical evaluation

Carla M. Evans and Jade Caines Lee

Education Department, University of New Hampshire, Durham, NH, USA

ABSTRACT

The purpose of this study is to critically evaluate value-added accountability measures currently enacted in the United States at the federal and state levels to assess teacher preparation programme (TPP) effectiveness. We draw on Newton and Shaw's framework for the evaluation of testing policy to evaluate the technical quality and social acceptability of using K-12 student test scores to assess TPP effectiveness. Through six guiding questions, we examine the assumptions and arguments that support value-added assessment, culminating in overall judgments about the acceptability of implementing (or continuing to implement) the testing policy. Findings suggest policy-makers may have more pragmatic concerns about the efficiency of value-added assessment, while TPPs may have more theoretical concerns about the validity of value-added assessment. The relevance of this evaluation approach to improving policy-related decision-making will also be discussed.

ARTICLE HISTORY
Received 2 November 2015
Accepted 27 October 2016

KEYWORDS
Value-added assessment; value-added models; teacher preparation; programme evaluation; educational policy; education reform

Introduction

Accountability demands for teacher preparation programmes (TPPs) are rapidly expanding worldwide (Cochran-Smith, 2013; Conway, 2013; Ell & Grudnoff, 2013; Furlong, 2013). More specifically, teacher education in the United States is under increased scrutiny and must meet heightened accountability metrics more than ever before (Darling-Hammond, 2010; Ginsberg & Kingston, 2014; Wilson & Youngs, 2005). Sharp criticisms in the United States claim that TPPs[1] have failed to prepare teachers to improve student achievement outcomes (Crowe, 2010; U.S. Department of Education [USDOE], 2011). For example, it is currently common practice within many state educator evaluation systems to link K-12 students' standardised test scores to determinations regarding teacher effectiveness (Ehlert, Koedel, Parsons, & Podgursky, 2016). All but a few states use some form of a student growth measure in evaluating educator effectiveness (Collins & Amrein-Beardsley, 2014). One type of methodology, value-added modelling (or VAMs), attempts to measure teachers' impact on student achievement from one year to the next using large-scale standardised testing data (Koedel, Mihaly, & Rockoff, 2015). This represents a seismic shift from previous

CONTACT Carla M. Evans carla.m.evans@gmail.com

teacher evaluation paradigms that relied almost solely on principal observations of teacher performance (Weisberg, Sexton, Mulhern, & Keeling, 2009).

Yet, despite extensive critique of the use of VAMs in measuring teacher effectiveness (American Educational Research Association [AERA], 2015; American Statistical Association [ASA], 2014; Amrein-Beardsley, 2014; Baker et al., 2010; Berliner, 2014; Lavigne, 2014; Newton, Darling-Hammond, Haertel, & Thomas, 2010; Paufler & Amrein-Beardsley, 2013; Rothstein, 2009; Schochet & Chiang, 2012),[2] over a dozen states plan to use the technique to examine the efficacy of TPPs in raising student achievement (Sawchuk, 2012). This is highly problematic without a more thorough examination and evaluation of associated claims.

Therefore, the purpose of this paper is to critically evaluate the value-added accountability measures currently enacted at the U.S. federal and state levels to assess TPP effectiveness[3] (referenced hereafter simply as value-added assessment). The basic premise of all value-added assessment of TPPs is that variance in K-12 student gains on standardised achievement tests can be attributed to the quality of teacher training a teacher received when other variables are controlled or adjusted. Quantifying the long-term impacts of teacher training on K-12 student learning outcomes in this way, however, may risk over-simplifying complex interactions (National Research Council, 2010).

It is important to note that we do not attempt to use all of the evidence in support of value-added assessment of TPPs; rather, we investigate the problematic assumptions that undergird arguments for this testing policy[4] (what we are calling a critical evaluation), especially those that do not support the technical adequacy or social acceptability of value-added assessment. As such, this evaluation examines (1) evidence related to the validity of inferences from value-added assessment; (2) the relationship of standardised test scores to specific decisions about TPP effectiveness; and 3) the potential high-stakes consequences for TPPs (such as loss of accreditation and federal funding).

The policy context

Education in the U.S. is primarily funded by individual states. Only about 10% of all K-12 public school funding comes from the federal government (U.S. Department of Education [USDOE], 2012). Despite this small financial role, however, states often rely heavily on federal competitive grant contributions because of state education budget cuts and increased demands for improved student performance (Center on Education Policy, 2011). More recently, states have had significant opportunities to apply for additional federal monies and have done so, despite the accountability measures attached. For example, as of 2014, the U.S. Department of Education (USDOE) awarded nineteen out of fifty states more than $4.35 billion to spark education reforms, including TPP reform (U.S. Department of Education [USDOE], 2014). Known as *Race to the Top*, this federal competitive grant programme required states to commit to improving TPP effectiveness (U.S. Department of Education [USDOE], 2009). States were awarded points, amongst other incentives, based on the extent to which they linked K-12 student achievement data to teachers and the TPPs in which they attended. This is just one way in which the federal government incentivises the way states use student growth data to evaluate TPPs (Chiang et al., 2011).

In addition, the United States Congress recently amended federal regulatory policy for teacher preparation providers. States are now required to provide data to the federal

government on 'aggregate learning outcomes of PK-12 students taught by new teachers… using student growth or teacher evaluation measure or both' for each TPP in the state (20 U.S.C. 1022d, p. 7). Similarly, new national accreditation standards for TPPs now require programmes 'to demonstrate the impact of their graduates on student learning, classroom instruction, and employer satisfaction' (Council for the Accreditation of Educator Preparation [CAEP], 2013, p. 13).

Some states plan to use these data for TPP accountability and improvement purposes (Coggshall, Bivona, & Reschly, 2012; Crowe, 2010), while other states already publish results of value-added assessment for public accountability purposes (Gansle, Burns, & Noell, 2011; Patterson & Bastian, 2014). Given these implications (as well as many others), a more rigorous way in which to evaluate value-added assessment must be explored.

Research on value-added assessment of TPPs

Literature review

To date, research on value-added assessment of TPPs falls into three main topical areas: econometric, policy and critical. According to Lincove, Osborne, Dillon, and Mills (2014), most econometric-oriented studies on TPP effectiveness test the ability of VAMs to measure differences in TPP effects using different theoretical models. These studies highlight methodological problems and aim to find the best models to fit the data in order to refine methodological practice and limit bias; they also typically provide multiple results that are sensitive to different model specifications and choices (Boyd, 2006; Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2009; Goldhaber, Liddle, & Theobald, 2013; Koedel, Parson, Podgursky, & Ehlert, 2012; Mihaly, McCaffrey, Sass, & Lockwood, 2013; Sass, 2011).

Policy-oriented studies, on the other hand, document and report on state programmes that measure differences in TPP effects. These studies are based on the assumption that TPP effects exist and can be measured (Lincove et al., 2014). They attempt to provide accurate public information on TPP quality for accountability and programme improvement purposes (Center for Teacher Quality, 2007; Gansle, Noell, & Burns, 2012; Henry et al., 2011, 2014; Kukla-Acevedo, Streams, & Toma, 2009; Lincove et al., 2014; Mason, 2010; Noell, Gansle, Patt, & Schafer, 2009; Osborne, 2012; Patterson & Bastian, 2014; Plecki, Elfers, & Nakamura, 2012; Tennessee Higher Education Commission, 2014).

Finally, in critical-oriented studies, a position on the relationship between TPPs and value-added approaches is taken in order to (1) highlight problems conceptually and/or methodologically using value-added models to estimate TPP effects, (2) suggest or report on alternative TPP effectiveness measures, and (3) inform policy decisions (Amrein-Beardsley, Barnett, & Ganesh, 2013; Cochran-Smith, Piazza, & Power, 2013; Cochran-Smith et al., 2016; Floden, 2012; Ginsberg & Kingston, 2014; Knight et al., 2012; Tatto et al., 2016; Zeichner, 2011). In general, these studies aim to draw awareness to concerns about the unintended consequences that may result from implementing value-added assessment.

Although these three research areas related to value-added assessment have made significant contributions to the teacher education, education policy, and educational measurement fields, there is still a need to examine the theory of action and research evidence that supports value-added assessment. In addition, causal links between TPP quality and K-12 student standardised test scores are complex and have not yet been substantiated (National

Research Council, 2010). For example, large-scale standardised tests were designed to assess student achievement outcomes, but because test scores are now being used to estimate TPP effectiveness, the evidence and arguments used to support the interpretations of K-12 student test scores for this proposed use must be evaluated (Kane, 2006; Messick, 1989).

Our study adds to already existing research in a unique way by synthesising across the three strands of literature in order to evaluate the use of VAMs to estimate TPP effectiveness. While there is a wealth of literature evaluating the use of VAMs to assess teacher effectiveness, the arguments and assumptions undergirding these claims have yet to be thoroughly examined in the research on TPP effectiveness. Therefore, we ask two overarching research questions to guide this evaluation:

(1) Is it *technically possible* to make more accurate decisions about TPP effectiveness by using VAMs to incorporate K-12 student test scores into the decision-making process?
(2) Is it *socially acceptable* to incorporate K-12 student test scores into the decision-making process about TPP effectiveness? In other words, have all the intended and unintended consequences been considered?

In order to evaluate the technical possibilities and social acceptability of this testing policy, we use six guiding questions that lead to overall evaluative judgments about the use of value-added assessment as a teacher education reform and improvement policy. We conclude by describing the relevance of this evaluation approach to improving policy-related decision-making, and make recommendations for future areas of research.

Conceptual framework

In our work, we draw on Newton and Shaw's (2014) framework for the evaluation of testing policy. We chose this framework because it includes a range of varying concerns, including the technical and consequential evidence related to value-added assessment. Doing so provides a holistic evaluation of testing policy based in validity theory. We recognise there is an on-going debate in the field over the use of the term validity and what is included (or excluded) from a validity evaluation (Baker, 2013; Borsboom, 2015; Cizek, 2015; Kane, 2015; Markus, 2015; Moss, 2015; Newton & Shaw, 2015; Shepard, 1997; Sireci, 2015). To avoid quarrelling over terminology, we do not discuss validity explicitly, but rather, following Newton and Shaw (2014), we situate our work in a comprehensive evaluation framework. In constructing this framework, Newton and Shaw (2014) argue that an overarching theory of the evaluation of testing policy must consider several dimensions, including: the evaluation of measurement, the evaluation of decision-making, and the evaluation of secondary policy objectives and side effects (see Figure 1). We use six guiding questions as dimensions in order to explore the technical quality and social acceptability of value-added assessment. Each dimension is first considered individually, then relationally, in order to derive overall judgments.

Technically possible? Evaluating the technical quality of testing policy

The overarching question with regards to the evaluation of technical quality is whether or not it is *technically possible* to make more accurate decisions about TPP effectiveness by

	Evaluation of Measurement	***Evaluation of Decision-Making***	***Evaluation of Secondary Policy Objectives and Side Effects***
Technical Quality	**Dimension 1** *Is it possible to measure the desired attribute using the test?*	**Dimension 2** *Is it possible to make more accurate decisions using test scores?*	**Dimension 3** *Is it possible to achieve a range of secondary impacts by implementing the testing policy?*
Social Acceptability	**Dimension 4** *Is it feasible to measure the desired attribute using the test?*	**Dimension 5** *Is it harmful to make decisions using test scores?*	**Dimension 6** *Is it fair to achieve secondary impacts by implementing the testing policy?*
Overall Judgment *Is it acceptable to implement (or continuing implementing) the testing policy?*			

Figure 1. A framework for the evaluation of testing policy. Note. Adapted from Newton and Shaw (2014).

incorporating K-12 student test scores into the decision-making process using VAMs. This evaluation is based on the premise that the technical quality of value-added assessment can be theorised independently of its social acceptability, and that it is important to do so in an evaluation (Newton & Shaw, 2014). The purpose of the technical quality evaluation is to construct and appraise the arguments and assumptions underlying three claims (see Figure 1).

Dimension 1: is it possible to measure the desired attribute using the test?

Many policy reports argue that there should be a positive relationship between the quality of a TPP and positive impact on K-12 student learning outcomes (Brabeck et al., 2014; CAEP, 2013). It is presumed that higher quality TPPs would tend to produce teachers who have a positive impact on K-12 student test scores. While it is hard to argue with the claim that TPPs should have a positive impact on K-12 student learning, the flow of impact from the quality of teacher training received to K-12 student learning outcomes is much more complicated (Diez, 2010; National Research Council, 2010). As a recent statement from the AERA (2015) explains:

> At first blush, it might seem commonsensical that VAM scores of novice teachers or leaders aggregated back to their preparation programs could serve as a basis for comparison. However, such use presents further challenges since those teachers and leaders are working in a wide range of schools, grades, and districts. Important differences in those settings, including variations in student populations, curricula, class sizes, and resources, as well as in the quality of

> induction and mentoring, contribute to differences in educators' performances and, therefore, are confounded with differences in the efficacy of their training programs. (p. 2)

When appraising the technical quality of a testing policy one might raise questions about whether it is possible to measure TPP effectiveness using VAMs to incorporate K-12 student test scores. For example, in claiming that TPP effectiveness can be measured using K-12 student test scores, some may assume: (1) there is significant variation between TPPs in terms of the teacher effectiveness of programme graduates (i.e. TPP effects exist), and (2) if TPP effects exist, they can be isolated from other factors that may bias value-added estimates of TPP effectiveness. In order to evaluate the technical quality of this testing policy, we must examine these two assumptions in detail.

Do TPP effects exist?

Some studies find small to no differences in TPP effectiveness across VAM specifications (Boyd, 2006; Goldhaber et al., 2013; Koedel et al., 2012; Plecki et al., 2012). In these studies much of the variation in TPP effectiveness occurs within programmes, not between programmes (Koedel et al., 2012). This is significant because it implies that what varies is not the quality of teacher training received, but the quality of programme graduates who attend each programme. Additionally, some studies find small variation in TPP effects for only first-year teachers because TPP effects decay over time and may be conflated with in-service teacher training (Boyd, 2006; Goldhaber et al., 2013). Given these findings, which are a small subset of a much broader literature base, assumptions related to the existence of TPP effects can and should be challenged.

Can TPP effects be isolated?

The second major assumption is that TPP effects can be isolated from other factors that may bias value-added estimates of TPP effectiveness (Baker et al., 2010; Berliner, 2014; Rothstein, 2009). There are, however, numerous nonrandom factors that do not allow TPP effects to be isolated (Meyer et al., 2014). For example, the recruitment and selection of teachers into and out of TPPs is a complex endeavour. There are inherent differences in teacher applicants including varying personal characteristics, ability, and experience levels that may be natural and have nothing to do with the programme they choose or the programme quality. In other words, some TPPs may seem more effective because they are able to attract higher quality teacher candidates through more rigorous recruitment and selectivity procedures (Boyd et al., 2009). In addition, teacher applicants may apply to and attend a TPP for specific reasons, including: geographic convenience, price point, and reputation of the programme. Furthermore, an interaction might exist between an applicant's ability level and a TPP's selection criteria, which could bias TPP effects and threaten the validity of TPP effect estimates.

Not only is it nonrandom how teachers enter programmes, it is also nonrandom how teachers enter teaching positions and schools/districts (Mihaly et al., 2013). There are many factors that influence where a teacher ends up teaching that are nonrandom, including: type of TPP attended (Amrein-Beardsley et al., 2013), geography (Mihaly et al., 2013), and even labour market complications (Floden, 2012). Researchers in one state, for example, found value-added assessment not feasible in part because 'schools tend to hire disproportionately from a single TPP' (Kukla-Acevedo et al., 2009, p. 15).

One of the most common critiques of VAMs is the nonrandom sorting of students to classrooms (Berliner, 2014; Koedel & Betts, 2009; Paufler & Amrein-Beardsley, 2013; Rothstein, 2009). This is when students are placed into classrooms and with teachers for particular reasons that are not under the teacher's control. There is no evidence to date that these nonrandom student placement effects can be statistically mitigated – 'only imperfectly and with an unknown degree of success' (AERA, 2015, p. 2). Student placement may bias TPP effects and is a potential threat to the validity of TPP effect estimates. In sum of dimension 1, there is no clear evidence that it is technically possible to use value-added assessment to estimate TPP effectiveness.

Dimension 2: is it possible to make more accurate decisions using test scores?

Another question that appraises the technical quality of the testing policy relates to the primary ways in which the test scores will be used to make consequential decisions. For example, '*Is it possible to make more accurate decisions about TPP effectiveness for programme accountability and/or programme improvement purposes using VAMs to incorporate K-12 student test scores*?' Programme accountability focuses on the summative, high-stakes use of data to compare or rank TPPs in order to make decisions regarding federal funding, state approval and/or programme reaccreditation. Programme improvement, on the other hand, focuses on the formative, low-stakes use of data to inform TPP inquiry and curricular redesign. Decision-making related to programme accountability assumes that TPP effect estimates are accurate; whereas, decision-making related to programme improvement assumes that TPP effect estimates supply useful and relevant information. In either case, these embedded assumptions need to be critically examined.

Do TPP effect estimates provide accurate, useful, and relevant information?

Any testing policy that uses test scores for programme accountability relies on human judgement in deciding what data are used and how those data are analysed and reported. There are several, distinct decisions that must be made when using value-added assessment (Henry, Kershaw, Zulli, & Smith, 2012). First, the selection of teachers, students, subjects, and years of data must be made. For example, when selecting teachers to include in the model (e.g. minimum number of teacher graduates used; first year teachers only versus teachers within their first five years of teaching), the results can vary greatly. Also, a decision must be made regarding the methods for estimating teachers' effects on student test score gains. For example, if a two-stage VAM is used that allows teacher training effects to decay over time versus a one-stage VAM that does not, some teachers may erroneously be deemed more effective than others. Finally, a decision must be made about the way in which TPP effect estimates are reported for public consumption. Some states, for instance, report in quintiles while others report in continuous values (Lincove et al., 2014). Reporting decisions can greatly affect public interpretation of TPP effectiveness. In essence, there are evaluative judgments when using value-added assessment to estimate TPP effectiveness. Decisions related to selection, methods, and reporting can (and do) affect the accuracy of information used in accountability practices because different choices lead to very different policy implications (Lincove et al., 2014).

Another assumption is that TPP effect estimates supply useful and relevant information for programme improvement purposes. For example, TPPs are looking for information

that is useful to and relevant for data-informed programme improvement; however, TPP effect estimates do not provide feedback on programme elements, thereby giving no explicit guidance to direct programme improvement (Floden, 2012; Mihaly et al., 2013; Plecki et al., 2012).

In dimension 1 there was no clear evidence that it is technically possible to estimate value-added effects for TPP accountability purposes. In this dimension, the evidence also suggests that making more accurate decisions about TPP effectiveness using student test scores does not result from this testing policy. This severely limits the utility, propriety, and credibility of value-added assessment.

Dimension 3: is it possible to achieve secondary impacts through the testing policy implementation?

A final question that evaluates the technical quality of the testing policy relates to the secondary impacts of the test scores: '*Is it possible to achieve a range of secondary impacts through the testing policy implementation?*' Secondary impacts go beyond the stated purpose of the policy and can include a variety of intended impacts. In evaluating value-added assessment, for example, increasing the ability of principals to select better teachers when hiring could be one intended secondary impact. However, there are inherent assumptions within the claim that TPP effect estimates can provide auxiliary benefits. For example, do school districts and states currently lack adequate information for secondary decision-making? Would value-added estimates provide critical information that school administrators and states could use to improve the overall quality of teacher candidates and the teacher workforce? These assumptions are explored below.

Is there a current lack of adequate information and provision of critical information?

One assumption is that school districts and states do not currently have adequate information to improve the overall quality of teacher candidates and the teacher workforce without using value-added assessment. With regard to school districts, principals currently base hiring decisions on a combination of factors such as education, certification, experience, personal characteristics, and recommendations. Some school districts also require job candidates to teach a lesson in order to evaluate their teaching practices. However, it is by no means clear that school districts have either inadequate information that is impeding their ability to hire high-quality teachers or that TPP effect estimates are even relevant to a particular job candidate. In some ways this information may be more confusing than clarifying. For example, if a TPP graduate is interviewing for a job and the principal looks up his/her TPP rating based on other graduates' K-12 student test scores, the principal then enters into a situation where they have to figure out how to weigh the relative worth of that information versus the information supplied by the teacher candidate themselves in their resume, interview, and letters of recommendation. Furthermore, if a teacher candidate is denied a job (or even an interview) because of their programme's poor ranking, the evidence suggests that that decision would be in error as most of the variation in teacher effectiveness occurs *within* rather than *between* programmes (Koedel et al., 2012).

In synthesising across the first three dimensions, there is limited evidence that supports the technical possibility of (1) estimating TPP effects, (2) making more accurate decisions about TPP effectiveness, or (3) achieving a range of secondary impacts by implementing

the testing policy. Given the limited technical support of value-added assessment, one may decide that the testing policy should be discarded. According to Newton and Shaw's (2014) framework, however, an evaluation of the testing policy's social acceptability and appropriateness should also be conducted in order to arrive at an overall judgement.

Socially acceptable? Evaluating the social appropriateness of testing policy

In addition to considering technical possibilities, social acceptability should also be considered because a high-quality mechanism for measuring TPP effectiveness may be technically possible, but it may still not be socially appropriate. Social acceptability includes the appraisal of the intended and unintended social consequences of the applied testing policy (Messick, 1989).[5] Therefore, when considering social acceptability there is not one right or wrong answer. Instead, speculation and claims based on probable consequences can and should be considered. Although social value cannot be theorised independently of technical quality, an evaluation of a testing policy cannot ignore potential social implications, whether they are intended or not.

And yet an issue arises in examining the social acceptability of value-added assessment because currently there is limited empirical research on social impacts related to the testing policy. The fact that there is very limited research that evaluates the social acceptability of value-added assessment policies in practice requires a broader view of related policies. Policies do not operate in isolation. They are often related to, and built upon, other policies and there are often several iterations of a policy in a variety of contexts. This is especially true of value-added assessment. The same critiques of teacher evaluation policies that use VAMs could also be applied here. Also, in this paper, we use another approach to investigating the social acceptability of this testing policy; we highlight the social critiques made in the value-added assessment literature using three evaluative criteria. Although there may be various considerations when evaluating a testing policy, we highlight three central questions related to feasibility, harm, and fairness.

Dimension 4: is it feasible to measure the desired attribute using the test?

The longitudinal data systems necessary to support value-added assessment are complicated and expensive. Background demographic information for each student, as well as each student's test results, must be linked with individual teachers. Some states, incentivised in part by *Race to the Top* funds, are developing longitudinal data management and data collection systems, but many other states do not yet have the capacity to link students to individual teachers to even calculate value-added estimates (Kukla-Acevedo et al., 2009; Webber et al., 2014). For example, one state conducted a pilot project to assess the feasibility of statewide TPP evaluations and found, for a variety of reasons including data challenges, that it was not feasible to assess TPPs using K-12 student test scores (Kukla-Acevedo et al., 2009). Additionally, some TPPs simply do not produce enough graduates in the tested grades and subjects to allow estimates to be calculated because of the minimum sample size needed. As a result, the use of K-12 student test scores to calculate TPP effect estimates is emerging (Gansle et al., 2015), but not yet feasible at the present time in many states.

Dimension 5: is it harmful to make decisions using test scores?

A discussion of harm when considering the consequences of a testing policy may highlight ethical notions of 'doing good to others' and 'doing no harm' (Evans, Caines, & Thompson, 2016). For example, with regard to value-added assessment, what consequences may occur if the information provided to the public, potential employers and policy-makers about TPP effectiveness is so dependent upon choices made in the policy context? What harm might result to TPPs and their graduates if programmes are misclassified as low performing when they are not? In other words, would individual teachers and programmes experience 'harm' as a result of incorrect, misleading, and unreliable information on their effectiveness?

The technical evaluation of this testing policy revealed how different model specifications can lead to very different classification ratings (Koedel et al., 2012; Lincove et al., 2014; Mihaly et al., 2013). This can harm TPPs because decisions about cut scores and performance levels impact how many TPPs are labelled as low performing. For example, in one study that investigated the sensitivity of TPP effect estimates to cut score decisions, seventeen TPPs were identified as low performing in one scenario, but using an alternative scenario only four TPPs were low performing – a shift in classification ratings for 20% of the programmes in the study. Additionally, in some states, alternative certification programmes (programmes in which programme completers have been teachers of record for 1–3 years before they complete their programmes) are compared to university-based TPPs (Gansle et al., 2011). However, because full-time teaching has a significant influence on teacher effectiveness (Plecki et al., 2012), it should come as no surprise that in states that compare alternative programmes directly to university-based TPPs, alternative programmes generally produce higher programme effect estimates. Harm may result to programmes from such unfair comparisons.

Therefore, not only is value-added assessment currently not socially appropriate because it is not feasible to implement (dimension 4), but also because the variability in TPP effect estimates (based on human decisions) can affect the acceptability of the testing policy (dimension 5). For example, decisions about cut scores and performance levels, the inclusion of alternatively certified teachers in the sample, and other decisions related to model specification creates a situation in which it may be impossible to 'do good' and 'do no harm' at the same time.

Dimension 6: is it fair to achieve secondary impacts through the testing policy implementation?

Fairness is a ubiquitous term that encompasses a number of concerns and can be examined from multiple disciplinary perspectives (Caines, 2013; Evans, 2015). Fairness is a fundamental issue in testing and includes the ways in which test scores are reported and used, as well as the consequences of test use (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014; Caines, Bridglall, & Chatterji, 2014). However, fairness is hard to quantify and difficult to unequivocally prove (or disprove). It is also difficult to prove that secondary decisions related to a testing policy are unfair because the relationship between the use of test scores and the decision are indirect.

Given these complex relationships, it is important to examine the potential unintended secondary impacts that may result from secondary decisions that rely on TPP rankings or effect estimates. Findings from this policy's technical evaluation suggest a few key unintended secondary impacts. First, there is the suggestion that principals can actually *use* the TPP rankings or effect estimates supplied from value-added assessment data to inform their selection of 'better' teachers when hiring. However, using TPP rankings or effect estimates for hiring decisions may be unfair because there are highly effective teachers (as measured by student test scores) that graduate from every TPP if more variability occurs within programmes rather than between programmes (Koedel et al., 2012). Additionally, if TPPs are held accountable through public postings of their ranking and compared to other TPPs in the state (or nationally), admissions to certain TPPs would likely diminish. Given the desire to achieve secondary impacts alongside questions related to fairness, implementation of this testing policy may be problematic.

In synthesising across the dimensions of the testing policy evaluation framework, it becomes clear that there are critical unanswered questions about the technical quality and social acceptability of value-added assessment. There is no clear evidence that value-added estimates are accurate or stable over time, which compromises their use in high-stakes decisions (dimensions 1–3). Furthermore, there are concerns about the feasibility, potential harm, and underlying fairness related to the use of K-12 student test scores to assess TPP quality (dimensions 4–6). Stakeholders on both sides must weigh all claims and evidence in order to derive an overall judgement about implementing (or continuing to implement) the testing policy.

Overall judgement

Although examining individual dimensions related to technical quality and social acceptability of a testing policy is necessary, it is only the initial step. Where do decision-makers and stakeholders go from here? In other words, how should one determine an overall judgement regarding this testing policy? One approach is to synthesise across all the dimensions holistically in order to derive an overall judgement.

Oftentimes, real-world, discrete choices must be made regarding a testing policy (Newton & Shaw, 2014). Questions may arise regarding the propriety, feasibility, utility, and accuracy of the testing policy that may give decision-makers doubts and cause a re-examination of implementation plans. These scenarios may call for an overall evaluative judgement that must be binary – yes (implement the testing policy) or no (do not implement or continue to implement the testing policy). In order to make this overall evaluative judgement, the technical and consequential evidence is synthesised and an argument is constructed that concludes it is either acceptable or not acceptable to use VAMs to calculate TPP effect estimates.

There is also another approach to determining an overall judgement regarding this testing policy. There may be a range of different arguments about the quality and value of a testing policy from different stakeholder perspectives (Newton & Shaw, 2014). For example, TPPs may have a different perspective on the quality and acceptability of value-added assessment in comparison to federal or state policy-makers. Therefore, a continuous approach allows for divergent perspectives along a spectrum for both technical quality and social acceptability. Because no testing policy will ever be perfect in an absolute sense, conceptualising quality

and acceptability as points on a continuum highlight differences in underlying values, assumptions, and perspectives that shape overall judgments.

Additionally, overall judgments are often contextualised within the proposed purpose and use of the testing policy. For example, a testing policy used for a low-stakes purpose (e.g. programme improvement) is very different than a testing policy used for a high-stakes purpose (e.g. programme accountability). Although use can be considered within other dimensions within the framework (Newton & Shaw, 2014), we structure the overall judgments around perspective and use given the context of value-added assessment and its current applications within TPP accountability.

Policy-maker perspective

Based on the premise that teachers have the single largest impact on student achievement than any other school-based factor, identifying effective teachers has become the key to large-scale education reform in the United States (Bill & Melinda Gates Foundation, 2013; Gordon, Kane, & Staiger, 2006; Rivkin, Hanushek, & Kain, 2005). At the same time, there is a general dissatisfaction with the perceived ineffectiveness and quality of some teachers and TPPs (Levine, 2006; USDOE, 2011), as well as teacher evaluation methods (Weisberg et al., 2009). Therefore, policy-makers may attempt to address these concerns by examining the ways in which teachers are prepared, especially through holding TPPs accountable for the 'success' of their graduates in positively influencing student achievement on large-scale standardised tests (Hamel & Merz, 2005; USDOE, 2011).

State and federal policy-makers who experience pressure to improve student achievement outcomes may be less concerned about the absolute technical accuracy and precision of TPP effect estimates. Instead, policy-makers may have more pragmatic concerns about the efficiency of value-added assessment as a means for achieving certain ends. Due to the limited empirical research available, policy-makers may judge the quality and value of value-added assessment in terms of the policy's overall likelihood that it will positively impact TPP programme quality and K-12 student learning outcomes, especially in comparison to current measures. As a result, some policy-makers may argue for value-added assessment simply because it is the only outcome measure of TPP quality that actually uses student achievement data. Other outcome measures such as pass rates on certification exams, surveys of programme graduates and/or principals, and attrition/retention rates do not include K-12 student outcomes. And yet, this still leaves open the extent to which value-added assessment is used for high-stakes purposes (e.g. accreditation, public disclosure, and federal/state accountability) or low-stakes purposes (e.g. programme improvement).

Acceptability of low-stakes use

Because some studies have found that it is possible to differentiate between TPPs based on value-added effect estimates, even if those effects may be quite small (Boyd, 2006; Goldhaber et al., 2013; Koedel et al., 2012; Plecki et al., 2012), policy-makers may view value-added assessment as having adequate technical quality for the intended purpose of acting as a signal that something may be problematic. In other words, policy-makers may believe it is acceptable to implement (or continue implementing) value-added assessment because the measure may signal possible concerns around TPP quality that then could be put into the hands of TPPs to investigate further. In this case, results would not be used for high-stakes

purposes because policy-makers may be concerned about how decisions made in the policy context could lead to the misclassification of some TPPs as ineffective when they are actually effective (e.g. Lincove et al., 2014). Instead, results would be provided only to TPPs who would then act on the information, as they deem appropriate. In making this overall judgement, policy-makers would support the use of value-added assessment as just one piece of evidence that TPPs could use to help assess the quality of their own programmes.

Acceptability of high-stakes use

Other policy-makers may arrive at a different overall judgement after weighing the relative risks of adopting this policy for high-stakes purposes. For example, protecting TPPs from unjust harm (i.e. misclassification) is important, but so is protecting schools and students from TPPs who year after year inadequately prepare teachers. One might argue that a student's right to equal educational opportunity is at stake if ineffective TPPs are allowed to continue churning out ineffective graduates who may cause 'harm' to students. In this case, policy-makers may argue for value-added assessment because there is a high social value and acceptability in prioritising the needs of students above the needs of institutions. This may be preferable because it allows the risk of harm to potentially affect institutions more than students.

Additionally, policy-makers may view value-added assessment as one of multiple measures used to evaluate TPP quality with each measure exhibiting limitations. As a result, they may argue that as long as data collected from other measures (e.g. attrition/retention rates, observational protocols, and/or satisfaction surveys of graduates and employers) all point in the same general direction, then the high-stakes use of these measures is acceptable and appropriate.

TPP perspective

In many instances, TPPs are going to have a different perspective from most federal or state policy-makers on the use of value-added assessment for accountability purposes (Cody, 2014; Falk, 2014; University of Georgia College of Education, 2015). The reputation and future existence of TPPs may be at stake in the assessment of TPP effectiveness using VAMs to incorporate K-12 test scores. TPPs, therefore, may not only be concerned about the technical quality of the measure, but also the unintended consequences of the testing policy. TPPs may also have more theoretical concerns about the ability of value-added assessment to fulfil its intended purpose.

At the same time, TPPs may recognise that there is public concern about the quality and preparedness of the educators they recommend for state certification and licensure (Darling-Hammond, 2010). In response to this public concern, and in response to federal and state policy-makers who are advocating for value-added assessment, TPPs may be 'caught in a vise' (Ginsberg & Kingston, 2014, p. 4). On the one hand, TPPs desire to strengthen internal review processes in order to provide the public and policy-makers with evidence as to the quality of their programmes (Wineburg, 2006). And yet, on the other hand, TPPs may totally disagree with the external measures proposed (such as value-added assessment) because of concerns over misuse and error (e.g. University of Georgia College of Education, 2015). In other words, there may be a fundamental disagreement about how best to measure the quality of TPPs. The degree to which TPPs would argue this testing policy is acceptable to

implement (or continue implementing) depends upon the stakes attached to the policy and the relative weighting of VAMs in the overall assessment of TPP quality.

Acceptability of low-stakes use

Most of the criticisms of value-added assessment by TPPs relate to the high-stakes use of the testing policy rather than the low-stakes use (Cody, 2014; Falk, 2014; Shoffner, 2014; University of Georgia College of Education, 2015). TPPs are often concerned about the ability of an accountability policy to accurately portray the quality of a TPP and also provide useful information for programme improvement (Zeichner, 2011). For example, the risk of misclassifying TPPs as ineffective when they are actually effective may pose concerns around accuracy (Koedel et al., 2012; Mihaly et al., 2013). Additionally, there are significant social concerns around fairness and harm, particularly because models are highly sensitive to decisions made in a policy and accountability context (Lincove et al., 2014). Finally, even though TPP effect estimates could be used by TPPs as one piece of evidence that helps assess the quality of their programme, the information provided by the measure is not very helpful. The results are simply not fine-grained enough to provide meaningful information that TPPs can use to restructure or identify specific aspects of their programmes that need to be changed (Zeichner, 2011).

Acceptability of high-stakes use

Because of the likely unintended consequences that would result from the high-stakes use of such a low-quality mechanism, TPPs would likely judge this testing policy as having high negative value. For example, because the preponderance of evidence suggests that (1) there is small to no variation between TPPs using value-added assessment; (2) any variation that exists is potentially biased by three significant nonrandom factors; and (3) there is more variation within programmes than between programmes, concerns about the technical quality and social acceptability of this testing policy have surfaced from both within (University of Georgia College of Education, 2015) and outside (Blaine, 2014) teacher preparation. Additionally, because many professional associations have already issued statements that warn against the use of VAMs for making high-stakes decisions about teachers (ASA, 2014; National Association of Secondary School Principals [NASSP], 2014), why would it be acceptable to take one more step and use VAMs to make high-stakes decisions about institutions that prepare those teachers, especially without further research on their appropriate use (AERA, 2015; Cody, 2014)? In the end, TPPs may think it is simply not socially acceptable to use VAMs to incorporate K-12 test scores into the decision-making process about TPP effectiveness.

Conclusion

Given the intensified focus on teacher quality as a lever for increasing student achievement, teacher education as a field is facing increased scrutiny and higher standards worldwide. We chose the United States as a context in which to examine policies that flow from this scrutiny because policies that incorporate the use of VAMs to assess the effectiveness of TPPs are currently enacted at the federal and state level. Also, although in other contexts there may be other variables that violate the assumptions that undergird value-added assessment (such as extensive out-of-school coaching and cram schools), in the United States there

is an expected relationship between TPP quality and K-12 student achievement outcomes regardless of confounding factors. To avoid misinformed policy judgments, a critical evaluation is necessary. In the context of value-added assessment of TPPs, however, research that may inform policy decisions is extremely limited. In response, we conduct this evaluation to examine the technical quality and social acceptability of value-added assessment to aid in better policy-related decision-making.

When considering value-added assessment, there should be an evaluative process that is comprehensive, unbiased, and context-sensitive, as well as based on the goal or purpose of the testing policy. Additionally, the complicated nature of decision-making requires flexibility and thoughtfulness in evaluating testing policies that reflect 'top down' mandates, especially when they are an extension of a previously contentious testing policy.

In this paper, we demonstrate how one could use a testing policy evaluation framework to examine the use of VAMs in estimating TPP effectiveness. For example, our first research question (and the first three dimensions of the testing policy evaluation framework) relates to the technical quality of value-added assessment. We examine the research literature to determine the extent to which it is possible to measure a TPP's effect on K-12 student learning outcomes. We also evaluated if it is possible to make more accurate decisions about TPP quality using K-12 student test scores. We found that in general the research literature does not support the argument that TPP effects exist or that they can be isolated from other factors that may bias estimates. We also found that because value-added estimates of TPP effects are not completely accurate or unbiased it is difficult, if not impossible, to base decisions about programme quality or teacher candidate quality on value-added estimates.

Our second research question focuses on the social acceptability of the intended uses and interpretations of value-added estimates for TPP accountability purposes. We found that there are many unanswered questions about the feasibility, unintended harm, and overarching fairness of implementing (or continuing to implement) value-added assessment policies. However, various stakeholders from policy-makers to TPP administrators still must evaluate the claims and evidence about value-added assessment and derive their own overall judgement about acceptability of policy implementation.

We hope that this example may provide various stakeholders with the means to evaluate the technical and consequential evidence presented about any particular testing policy. By looking at the same evidence from various perspectives, we also hope to highlight the way in which assumptions and arguments are often forwarded within a policy context. For example, TPPs could use this type of framework to challenge prevailing notions about the relationship between student achievement and TPP effectiveness. Also, policy-makers who attempt to address public concerns about TPP quality could use a critical evaluation framework to counter the consequential use of controversial testing policies. Likewise, assessment professionals' writ large could use this testing policy evaluation framework when examining the claims, assumptions, and arguments related to a particular policy.

In thinking about next steps, researchers could continue to explore the methodological problems with value-added assessment including year-to-year stability of TPP effect estimates and the relative weight that value-added assessment should play in TPP evaluations. There are many more questions, however, to explore beyond simply the statistical properties and technical quality of value-added assessment. Various stakeholders especially may have questions that relate to the impacts of currently implemented value-added assessment policies that future research could address. For example, what are the impacts on the

relationships between teacher educators, programme graduates, and principals when K-12 student test scores are used to measure TPP effectiveness? What are the impacts on teaching and learning when programme graduates know their TPPs ranking or effect estimate depends at least in part on their ability to increase their students' test scores? What are the impacts on TPPs who must decide every year who to admit into the programme when they know their future effectiveness will reside in the effectiveness of the programme graduate? While there is no need to start from scratch in answering these questions, there is definitely a need to empirically investigate the intended and unintended consequences of VAMs in estimating TPP effectiveness. Until researchers have time to catch up to these types of testing policy reforms, it is critical that policy-makers proceed with caution.

Notes

1. Throughout this paper, TPPs are defined as state-approved programmes or courses of study that lead to an initial teaching credential (U.S. Department of Education [USDOE], 2013).
2. There is extensive literature available on value-added modelling and related critiques. This is, however, outside the scope of our paper. We refer readers to the reference list in Amrein-Beardsley (2014).
3. In this paper, we define TPP effectiveness as the outcomes of TPP quality with K-12 student learning as the primary focus. Although this is our focus, the evaluation of TPPs can be conducted from other related angles: positive impact on teacher candidates and positive impact on teacher workforce. We refer readers to (Brabeck et al., 2014; Cochran-Smith, 2005; Feuer, Floden, Chudowsky, & Ahn, 2013; National Research Council, 2010; Wilson & Youngs, 2005).
4. Hereafter, when we refer to 'this testing policy' we are referring to the use of VAMs to incorporate K-12 student standardised test scores to estimate TPP effectiveness.
5. In line with Newton and Shaw's (2014) framework for the evaluation of testing policy, we use the term 'social acceptability' rather than 'social consequences'. However, social acceptability includes an appraisal of the (un)intended consequences of a testing policy.

Acknowledgement

We would like to thank Todd DeMitchell, Charlie DePascale, Doug Gagnon, Chad Gotch, and Emilie Reagan for their substantive feedback on earlier drafts.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Carla M. Evans is a PhD candidate in the Education Department at the University of New Hampshire. Her research focuses on the impacts and implementation of assessment and accountability policies on teaching and learning. She is interested in three policy areas at the forefront of education assessment reform: innovative assessment and accountability systems, performance-based assessments, and teacher/teacher preparation programme effectiveness initiatives.

Jade Caines Lee is an assistant professor in the Education Department at the University of New Hampshire. Prior to her arrival at UNH, she was a K-12 public school teacher for nearly a decade and a researcher with numerous organisations including state agencies, foundations, and evaluation firms. Her research areas include assessment development, the application of validity/evaluation

frameworks to education assessments, and the evaluation of interventions that improve teaching and learning for underrepresented student groups.

References

American Educational Research Association. (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. *Educational Researcher, 44*, 448–452. doi:10.3102/0013189X15618385

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Education Research Association and National Academy of Education.

American Statistical Association. (2014). *Statement on using value-added models for educational assessment*. Alexandria, VA: American Statistical Association. Retrieved from www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf

Amrein-Beardsley, A. (2014). *Rethinking value-added models in education: Critical perspectives on tests and assessment-based accountability*. New York, NY: Routledge.

Amrein-Beardsley, A., Barnett, J. H., & Ganesh, T. G. (2013). Seven legitimate apprehensions about evaluating teacher education programs and seven "beyond excuses" imperatives. *Teachers College Record, 115*(December), 1–34.

Baker, E. L. (2013). The chimera of validity. *Teachers College Record, 115*, 1–26.

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*. Washington, DC: Economic Policy Institute. Retrieved from www.epi.org/publication/bp278/

Berliner, D. C. (2014). Exogenous variables and value-added assessments: A fatal flaw. *Teachers College Record, 116*(1), 1–31.

Bill & Melinda Gates Foundation. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study. Policy and practice brief*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/reports.php

Blaine, S. (2014). Sarah Blaine: How can we stop Arne's zany plan to grade ed school by the scores of students of their graduates?. Retrieved December 8, 2014, from http://wp.me/p2odLa-9bJ

Borsboom, D. (2015). Zen and the art of validity theory. *Assessment in Education: Principles, Policy & Practice, 23*, 415–421. doi:10.1080/0969594X.2015.1073479

Boyd, D. J. (2006). How changes in entry requirements alter the teacher workforce and affect student achievement. *Education Finance and Policy, 1*, 176–216.

Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis, 31*, 416–440. doi:10.3102/0162373709353129

Brabeck, M. M., Dwyer, C. A., Geisinger, K. F., Marx, R. W., Noell, G. H., Pianta, R. C., & Worrell, F. C. (2014). *Assessing and evaluating teacher preparation programs* (APA task force report). Washington, DC: American Psychological Association.

Caines, J. (2013). *"Doing no harm:" Expanding the notion of fairness through the evaluation of validity frameworks*. Poster presented at the annual meeting of the American Educational Research Association, San Fransisco, CA.

Caines, J., Bridglall, B. L., & Chatterji, M. (2014). Understanding validity and fairness issues in high-stakes individual testing situations. *Quality Assurance in Education, 22*, 5–18.

Center for Teacher Quality. (2007). *Teacher preparation program evaluation based on K-12 student learning and performance assessments by school principals*. Long Beach, CA: California State University Center for Teacher Quality.

Center on Education Policy. (2011). *More to do, but less capacity to do it: States' progress in implementing the recovery act education reforms*. Washington, DC: Center on Education Policy.

Chiang, Y., Cole, C., Delandshere, R., Kunzman, R., Guarino, C., Rutkowski, D., Rutkowski, L., Svetina, D., Yuan, X., & Zhou, Y. (2011). *Using value added models to evaluate teacher preparation programs (White paper)*. Bloomington, IN: Indiana University.

Cizek, G. J. (2015). Validating test score meaning and defending test score use: Different aims, different methods. *Assessment in Education: Principles, Policy & Practice, 23*, 212–225. doi:10.1080/0969594X.2015.1063479

Cochran-Smith, M. (2005). 2005 presidential address: The new teacher education: For better or for worse? *Educational Researcher, 34*, 3–17.

Cochran-Smith, M. (2013). Introduction: The politics of policy in teacher education: International perspectives. *The Educational Forum, 77*, 3–4. doi:10.1080/00131725.2013.739013

Cochran-Smith, M., Piazza, P., & Power, C. (2013). The politics of accountability: Assessing teacher education in the United States. *The Educational Forum, 77*, 6–27. doi:10.1080/00131725.2013.739015

Cochran-Smith, M., Stern, R., Sanchez, J. G., Miller, A., Keefe, E. S., Fernandez, M. B., Change, W., Carney, M. C., Burton, S., & Baker, M. (2016). *Holding teacher preparation accountable: A review of claims and evidence*. Boulder, CO: National Education Policy Center.

Cody, A. (2014). Teacher education leaders speak out: Kevin Kumashiro on teacher preparation, edTPA and reform. Retrieved October 3, 2015, from http://blogs.edweek.org/teachers/living-in-dialogue/2014/05/teacher_education_leaders_spea.html

Coggshall, J. G., Bivona, L., & Reschly, D. J. (2012). *Evaluating the effectiveness of teacher preparation programs for support and accountability*. Washington, DC: National Comprehensive Center for Teacher Quality.

Collins, C., & Amrein-Beardsley, A. (2014). Putting growth and value-added models on the map: A national overview. *Teachers College Record, 116*(1), 1–9.

Conway, P. F. (2013). Cultural flashpoint: The politics of teacher education reform in Ireland. *The Educational Forum, 77*, 51–72. doi:10.1080/00131725.2013.739021

Council for the Accreditation of Educator Preparation. (2013). *CAEP accreditation standards*. Washington, DC: Author. Retrieved from http://caepnet.org/accreditation/standards/

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

Darling-Hammond, L. (2010). Teacher education and the American future. *Journal of Teacher Education, 61*, 35–47. doi:10.1177/0022487109348024

Diez, M. E. (2010). It is complicated: Unpacking the flow of teacher education's impact on student learning. *Journal of Teacher Education, 61*, 441–450. doi:10.1177/0022487110372927

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2016). Selecting growth measures for use in school evaluation systems: Should proportionality matter? *Educational Policy, 30*, 456–500. doi:10.1177/0895904814557593

Ell, F., & Grudnoff, L. (2013). The politics of responsibility: Teacher education and "persistent underachievement" in New Zealand. *The Educational Forum, 77*, 73–86. doi:10.1080/00131725.2013.739023

Evans, C. M. (2015, June). The missing framework: A case for utilizing ethics to evaluate the fairness of educator evaluation systems [Commentary]. *Teachers College Record*. Retrieved from http://www.tcrecord.org

Evans, C. M., Caines, J., & Thompson, W. C. (2016). First, do no harm?: A framework for ethical decision-making in teacher evaluation. In K. K. Hewitt & A. Amrein-Beardsley (Eds.), *Student growth measures in policy and practice: Intended and unintended consequences of high-stakes teacher evaluations* (pp. 169–188). New York, NY: Palgrave Macmillan.

Falk, B. (2014). Memo to Barack Obama: The folks who educate teachers would like to share some lessons on positive accountability with you. Retrieved October 3, 2015, from http://hechingerreport.org/memo-barack-obama-folks-educate-teachers-like-share-lessons-positive-accountability/

Feuer, M. J., Floden, R. E., Chudowsky, N., & Ahn, J. (2013). *Evaluation of teacher preparation programs: Purposes, methods and policy options*. Washington, DC: National Academy of Education.

Floden, R. E. (2012). Teacher value added as a measure of program quality: Interpret with caution. *Journal of Teacher Education, 63*, 356–360. doi:10.1177/0022487112454175

Furlong, J. (2013). Globalisation, neoliberalism, and the reform of teacher education in England. *The Educational Forum, 77*, 28–50. doi:10.1080/00131725.2013.739017

Gansle, K. A., Burns, J. M., & Noell, G. H. (2011). *Value added assessment of teacher preparation programs in Louisiana: 2007–08 to 2009–10 (Overview of 2010–11 results)*. Baton Rouge, LA: Louisiana's Teacher Quality Initiative.

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

Gansle, K. A., Noell, G. H., Grandstaff-Beckers, G., Stringer, A., Roberts, N., & Burns, J. M. (2015). Value-added assessment of teacher preparation: Implications for special education. *Intervention in School and Clinic, 51*, 106–111. doi:10.1177/1053451215579267

Ginsberg, R., & Kingston, N. (2014). Caught in a vise: The challenges facing teacher preparation in an era of accountability. *Teachers College Record, 116*(1), 1–48.

Goldhaber, D., Liddle, S., & Theobald, R. (2013). The gateway to the profession: Assessing teacher preparation programs based on student achievement. *Economics of Education Review, 34*, 29–44. doi:10.1016/j.econedurev.2013.01.011

Gordon, R., Kane, T. J., & Staiger, D. O. (2006). *Identifying effective teachers using performance on the job* (White Paper 2006–01). Washington, DC: The Brookings Institution.

Hamel, F. L., & Merz, C. (2005). Reframing accountability: A preservice program wrestles with mandated reform. *Journal of Teacher Education, 56*, 157–167. doi:10.1177/0022487105274458

Henry, G. T., Bastian, K. C., Fortner, C. K., Kershaw, D. C., Purtell, K. M., Thompson, C. L., & Zulli, R. A. (2014). Teacher preparation policies and their effects on student achievement. *Education Finance and Policy, 9*, 264–303. doi:10.1162/EDFP_a_00134

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

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 July 2011 Carolina Institute for public policy*. Chapel Hill, NC: Carolina Institute for Public Policy.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger Publishers.

Kane, M. T. (2015). Explicating validity. *Assessment in Education: Principles, Policy & Practice, 23*, 198–211. doi:10.1080/0969594X.2015.1060192

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

Koedel, C., & Betts, J. R. (2009). *Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique*. Cambridge, MA: National Bureau of Economic Research.

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review, 47*, 180–195. doi:10.1016/j.econedurev.2015.01.006

Koedel, C., Parson, E., Podgursky, M., & Ehlert, M. (2012). *Teacher preparation programs and teacher quality: Are there real differences across programs?* (Working Paper No. WP 12-04). Columbia, MO: University of Missouri.

Kukla-Acevedo, S., Streams, M., & Toma, E. F. (2009). *Evaluation of teacher preparation programs: A reality show in Kentucky* (IFIR Working Paper No. 2009-09). Lexington, KY: Institute for Federalism and Intergovernmental Relations.

Lavigne, A. L. (2014). Exploring the intended and unintended consequences of high-stakes teacher evaluation on schools, teachers, and students. *Teachers College Record, 116*(1), 1–29.

Levine, A. (2006). *Educating school teachers*. Washington, DC: The Education Schools Project.

Lincove, J. A., Osborne, C., Dillon, A., & Mills, N. (2014). The politics and statistics of value-added modeling for accountability of teacher preparation programs. *Journal of Teacher Education, 65*, 24–38. doi:10.1177/0022487113504108

Markus, K. A. (2015). Alternative vocabularies in the test validity literature. *Assessment in Education: Principles, Policy & Practice, 23*, 252–267. doi:10.1080/0969594X.2015.1060191

Mason, P. L. (2010). *Examining FAMU's supply of teachers: A value-added analysis of college of preparation on pupil academic achievement* (MPRA Paper No. 27904). Munich Personal RePEc Archive. Retrieved from http://mpra.ub.uni-muenchen.de/27904/

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: American Council on Education and Macmillan Publishing Company.

Meyer, R., Pyatigorsky, M., & Rice, A. (2014). *Evaluation of educators and educator preparation programs: Models and systems in theory and practice* (WCER Working Paper No. 2014-6). University of Wisconsin-Madison, Wisconsin Center for Education Research. Retrieved from http://www.wcer.wise.edu/publications/workingPapers/papers.php

Mihaly, K., McCaffrey, D., Sass, T. R., & Lockwood, J. R. (2013). Where you come from or where you go? Distinguishing between school quality and the effectiveness of teacher preparation program graduates. *Education Finance and Policy, 8*, 459–493. doi: 10.1162/EDFP_a_00110

Moss, P. A. (2015). Shifting the focus of validity for test use. *Assessment in Education: Principles, Policy & Practice, 23*, 236–251. doi:10.1080/0969594X.2015.1072085

National Association of Secondary School Principals. (2014). *Position statement: Value-added measures in teacher education*. Reston, VA: National Association of Secondary School Principals.

National Research Council. (2010). *Preparing teachers: Building evidence for sound policy*. Washington, DC: National Academies Press. doi:10.1126/science.316.5829.1279

Newton, P. E., & Shaw, S. D. (2014). *Validity in educational & psychological assessment*. Cambridge: Cambridge Assessment.

Newton, P. E., & Shaw, S. D. (2015). Disagreement over the best way to use the word "validity" and options for reaching consensus. *Assessment in Education: Principles, Policy & Practice, 23*, 178–197. doi:10.1080/0969594X.2015.1037241

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Education Policy Analysis Archives, 18*, 1–27.

Noell, G. H., Gansle, K. A., Patt, R. M., & Schafer, M. J. (2009). *Value added assessment of teacher preparation in Louisiana: 2005–2006 to 2007–2008*. Louisiana Board of Regents. Retrieved from http://regents.louisiana.gov/value-added-teacher-preparation-program-assessment-model/

Osborne, C. (2012). *The Texas report: Education preparation programs' influence on student achievement*. Austin, TX: Project on Educator Effectiveness & Quality.

Patterson, K. M., & Bastian, K. C. (2014). *UNC teacher quality research: Teacher portals effectiveness report*. Chapel Hill, NC: Education Policy Initiative at Carolina.

Paufler, N. A., & Amrein-Beardsley, A. (2013). The random assignment of students into elementary classrooms: Implications for value-added analyses and interpretations. *American Educational Research Journal, 51*, 328–362. doi:10.3102/0002831213508299

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*, 318–334. doi:10.1177/0022487112447110

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*, 417–458.

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

Sass, T. R. (2011). *Certification requirements and teacher quality: A comparison of alternative routes to teaching* (CALDER Working Paper No. 64). Washington, DC: CALDER, American Insitutes for Research.

Sawchuk, S. (2012). "Value added" proves beneficial to teacher prep. *Education Week, 31*, 1–20.

Schochet, P. Z., & Chiang, H. S. (2012). What are error rates for classifying teacher and school performance using value-added models? *Journal of Educational and Behavioral Statistics, 38*, 142–171. doi:10.3102/1076998611432174

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*, 5–8, 13, 24.

Shoffner, M. (2014). We're all mad here: A response to the US Department of Education's proposed regulations for teacher preparation from the chair of the conference on English education

(CEE). Retrieved October 3, 2015, from http://www.ncte.org/library/NCTEFiles/Groups/CEE/ChrResponse_DOE_regs_dec2014.pdf

Sireci, S. G. (2015). On the validity of useless tests. *Assessment in Education: Principles, Policy & Practice, 23*, 226–235. doi:10.1080/0969594X.2015.1072084

Tatto, M. T., Savage, C., Liao, W., Marshall, S. L., Goldblatt, P., & Contreras, L. M. (2016). The emergence of high-stakes accountability policies in teacher preparation: An examination of the U.S. Department of Education's proposed regulations. *Education Policy Analysis Archives, 24*(21). doi:10.14507/epaa.24.2322

Tennessee Higher Education Commission. (2014). *2014 report card on the effectiveness of teacher training programs*. Nashville, TN: Author. Retrieved from http://www.tn.gov/thec/Divisions/AcademicAffairs/rttt/report_card/2014/report_card/14report_card.shtml

U.S. Department of Education. (2009). *Race to the Top program executive summary*. Washington, DC: Author.

U.S. Department of Education. (2011). *Our future, our teachers: The Obama administration's plan for teacher education reform and improvement*. Washington, DC: Author.

U.S. Department of Education. (2012). *The federal role in education*. Nashville, TN: Author. Retrieved October 24, 2015, from http://www2.ed.gov/about/overview/fed/role.html

U.S. Department of Education. (2013). *Preparing and credentialing the nation's teachers: The secretary's ninth report on teacher quality*. Washington, DC: Author. Retrieved from https://title2.ed.gov/TitleIIReport13.pdf

U.S. Department of Education. (2014). *Race to the top*. Washington, DC: Author.

University of Georgia College of Education. (2015). Faculty and students at UGA college of education speak out against federal proposals to regulate teacher preparation. Retrieved October 3, 2015, from https://engagedintellectual.wordpress.com/2015/02/07/faculty-and-students-at-uga-college-of-education-speak-out-against-federal-proposals-to-regulate-teacher-preparation/

Webber, A., Troppe, P., Milanowski, A., Gutmann, B., Reisner, E., & Goertz, M. (2014). *State implementation of reforms promoted under the recovery act: A report from charting the progress of education reform: An evaluation of the recovery act's role*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U. S. Department of Education.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness* (2nd ed.). Brooklyn, NY: The New Teacher Project (TNTP). Retrieved from http://tntp.org/ideas-and-innovations/view/the-widget-effect

Wilson, S. M., & Youngs, P. (2005). Research on accountability processes in teacher education. In M. Cochran-Smith & K. Zeichner (Eds.), *Studying teacher education: The report of the AERA panel on research and teacher education* (pp. 591–643). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Wineburg, M. S. (2006). Evidence in teacher preparation: Establishing a framework for accountability. *Journal of Teacher Education, 57*, 51–64. doi:10.1177/0022487105284475

Zeichner, K. M. (2011). Assessing state and federal policies to evaluate the quality of teacher preparation programs. In P. M. Earley, D. G. Imig, & N. M. Michelli (Eds.), *Teacher education policy in the United States: Issues and tensions in an era of evolving expectations* (pp. 76–105). Florence, KY: Routledge.