Article

Educational Policy
1–42
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0895904819843593
journals.sagepub.com/home/epx
SAGE

Teacher-Level Value-Added Models on Trial: Empirical and Pragmatic Issues of Concern Across Five Court Cases

Audrey Amrein-Beardsley[1] **and Kevin Close**[1]

Abstract

Ongoing or recently completed across the United States are a series of lawsuits via which teacher plaintiffs are contesting how they are being evaluated using value-added models (VAMs) as part of states'/districts' teacher accountability systems. To investigate the empirical and pragmatic matters addressed in court, researchers conducted a case study analysis of the documents submitted for five such cases. Researchers framed analyses using measurement concepts resident within the *Standards for Educational and Psychological Testing*, given issues with (a) reliability, (b) validity, (c) bias, (d) transparency, and (e) fairness, with emphases also on (f) whether VAMs are being used to make consequential decisions using concrete (e.g., not arbitrary) evidence, and (g) whether VAMs' unintended consequences are also of legal pertinence and concern.

Keywords

accountability, education policy, educational reform, evaluation and assessment, legal issues, policy implementation, teacher quality

[1]Arizona State University, Phoenix, USA

Corresponding Author:
Audrey Amrein-Beardsley, Professor, Educational Policy and Evaluation Program, Mary Lou Fulton Teachers College, Arizona State University, P.O. Box 37100, Phoenix, AZ 85069-7100, USA.
Email: audrey.beardsley@asu.edu

Introduction

In May 2016 in New York, an 18-year veteran, fourth-grade, National Board Certified Teacher (NBCT), in an upscale suburb of Long Island, sued the state over her value-added teacher evaluation score. Although the district recognized her as having a "flawless" teaching record and being "highly regarded as an educator" (Strauss, 2014), the state used her value-added score to classify her as "ineffective." In 2014, as per the growth (or lack thereof) that her fourth-grade students registered on their large-scale standardized tests from years prior (i.e., aggregated at the teacher level), she received a score of 1 out of 20. One year prior, the evaluation system classified her as "effective" with a score of 14 out of 20 using the same test-based indicators.

Her husband, who was a lawyer, took his wife's case to the State of New York Supreme Court (i.e., the intermediate level of the state's judicial system, with the state's highest court being the State of New York Court of Appeals). They won, with the presiding judge ruling that the state's recently reformed teacher evaluation system, as primarily based on teachers' value-added estimates, was "arbitrary and capricious." The presiding judge defined "arbitrary and capricious" as actions "taken without sound basis in reason or regard to the facts" (State of New York Supreme Court, 2016, p. 11).

This teacher's case was one of approximately 15 lawsuits occurring at the time ("Teacher Evaluation Heads to the Courts," 2015), related to the nation's reformed teacher evaluation systems. Specifically, these cases were located across seven states: Florida $n = 2$, Louisiana $n = 1$, Nevada $n = 1$, New Mexico $n = 4$, New York $n = 3$, Tennessee $n = 3$, and Texas $n = 1$. Through these cases, plaintiffs were targeting the value-added indicators that were (and in some cases still are being) adopted, implemented, and used as the primary or key teacher evaluation system components across the Unites States (Doherty & Jacobs, 2015). Across cases, plaintiffs consisting of public school teachers, teacher unions, and teacher unions' local affiliates were taking school districts and state departments of education to court over these systems. Plaintiffs were arguing that the evaluation systems, as noted prior, were not only "arbitrary and capricious" (State of New York Supreme Court, 2016, p. 11), but also "irrational" (VAMboozled, 2014a), "unfair" (VAMboozled, 2014b, 2014c, 2018), and being used in violation of teachers' due process rights (VAMboozled, 2014b), again, as based in large part upon teachers' value-added estimates.

Indeed, all of the 15 teacher evaluation systems at issue in court had a value-added component, around which other evaluative indicators (e.g., observational components) revolve. The "preponderant" value-added component, as written into policy and oft-enacted in practice, sometimes carries more weight or altogether trumps other system indicators (Doherty & Jacobs, 2015). In New York, for example, the value-added indicator was to count for 50% of a teacher's overall evaluation score, but if the other indicators collected at the same time about the teachers' effectiveness (e.g., the observational component) contradicted the value-added output, the value-added output was to trump the other indicator(s), yielding 100% of a teacher's overall evaluation score (National Association of Secondary School Principals [NASSP], n.d.). In other states, at the time, it was most often the case that a state's teacher-level value-added component was to carry between 33% and 50%, "[a]ccording to the best-available research" (Bill & Melinda Gates Foundation, 2013).

Notwithstanding, plaintiffs argued that the preponderant use of value-added indicators was more egregious when high-stakes decisions were attached to value-added output. Hence, this evaluative component warrants attention and gives rise to concern in practice, in policy, and also in court.

In general, system output includes teacher-level estimates as based on value-added models (VAMs) and growth models (hereafter referred to more generally as VAMs[1]). VAMs, in the simplest of terms, classify teachers' effectiveness according to teachers' statistically measurable (and purportedly) causal impacts on their students' standardized test scores over time (Amrein-Beardsley, 2014), although there is certainly debate about the extent to which VAMs can or do separate out a teacher's impact from other classroom-level factors (see, for example, Rothstein, 2009, 2010). Nonetheless, the intent of VAMs is to help to identify teachers whose students outperform their projected levels of growth as effective or of "value-added" and teachers whose students fall short as ineffective or not of "value-added" (Sanders, 2003, 2006). As mentioned, system output also oft-includes estimates based on supervisors' observations of teachers teaching in practice, using observational rubrics to identify, document, and score teachers' desirable, nonlatent behaviors.

High-stakes decisions attached to system output include, but are not limited to, teachers' permanent files being flagged with their value-added or overall effectiveness categories (e.g., "highly effective," "effective," "ineffective," "highly ineffective") that has prevented teachers from moving teaching positions across districts (VAMboozled, 2018); the awarding or revocation of teacher licenses or tenure; salary increases, decreases, or merit pay; and probation or termination (Amrein-Beardsley, 2014). Similarly, it is this set of high-stakes consequences, predominantly attached to the value-added overobservational components of these reformed teacher evaluation systems, which are at the core of these lawsuits in court ("Teacher Evaluation Heads to the Courts," 2015).

As also explained in "Teacher Evaluation Heads to the Courts" (2015), the fundamental reasons "lots of litigation" have been filed, especially from a federal education policy level, are related to the reformed, and hence stronger, and more objective teacher accountability systems called for and popularized across the United States over the past decade. More specifically, these reformed state and district policies were initially spurred by former U.S. Secretary of Education Spellings's growth model pilots (U.S. Department of Education, 2006a, 2006b), incentivized by President Obama's Race to the Top Competition (2011; see also Duncan, 2009), and then punctuated by the No Child Left Behind (NCLB) waivers that excused states from not meeting NCLB's prior 100% student proficiency by 2014 goals should they implement reformed teacher evaluation systems (U.S. Department of Education, 2010a, 2014).

Moreover, econometricians (e.g., Chetty, Friedman, & Rockoff, 2014a, 2014b; Hanushek, 2009; T. J. Kane, 2015; Sanders, 2003, 2006), high-profile political (e.g., Duncan, 2011; Rhee, 2011) philanthropic figures (e.g., Brown, 2014; Gates, 2013; see also Bill & Melinda Gates Foundation, 2013), and nonpartisan foundations and think-tanks (e.g., Bill & Melinda Gates Foundation, National Council on Teacher Quality, The New Teacher Project [TNTP]; see, for example, Doherty & Jacobs, 2015; T. J. Kane & Staiger, 2012; TNTP, 2012; Walsh, Joseph, Lakis, & Lubell, 2017; Weisberg, Sexton, Mulhern, & Keeling, 2009), supported these reformed teacher accountability policies contributing to the spread of these policies across states and districts throughout the United States.

In sum, multiple federal acts and policies encouraged states to attach high-stakes consequences to VAM-based teacher evaluations, setting the groundwork for these cases. States with stronger consequences attached to system output were more likely to receive Race to the Top funds (e.g., Florida at US$700 million, New York at US$700 million, Tennessee at US$500 million; see also U.S. Department of Education, 2010b) as well as NCLB waivers (e.g., Florida, Louisiana, Nevada, New Mexico, New York, Tennessee), although 46 states ultimately received NCLB waivers (U.S. Department of Education, 2015). Consequently, it can be argued that because the states in which lawsuits exist are those in which legislators best complied with the aforementioned federal policy calls (see also Mathematica Policy Research, 2014), this may explain why these states' teacher evaluation systems are/were at issue across U.S. courts.

Similarly, while the federal passage of the Every Student Succeeds Act (ESSA) has since curbed such efforts, no longer requiring or incentivizing all states to engage in such high-stakes teacher evaluation policies as based in large part on VAMs, stronger teacher accountability systems remain, particularly in these states, given the substantive financial and human resources already invested (Close, Amrein-Beardsley, & Collins, 2018; Excel*in*Ed, 2017; Kraft & Gilmour, 2017).

Regardless, the way the plaintiffs framed the lawsuits led to direct and indirect impacts. Direct impacts included actual court rulings. Indirect impacts included providing examples to lawmakers for what to avoid (Close, & Amrein-Beardsley, 2018), producing headlines in popular media (Rhee, 2011), and challenging the assumptions of major players in education such as the Gates Foundation (Darville, 2017). Determining the actual impact of these lawsuits on teacher evaluation systems in the future is too complex, but one scholar, Superfine (2016), argued that the lawsuits provided an arena to attempt broad policy reform and to work through complex issues between policy makers and teachers. This study provides a close-up of that arena showing the lines of argument in several states.

Purpose of the Study

The cases across states contain not only common but also distinct lines of legal argument and dispute. In other words, plaintiffs have argued their cases in unique ways with shared and distinctive foci, again, as largely defined by states' constitutions and legislative actions, or by district policies and administrative actions in the relatively fewer district-level suits. These common and distinct cases, accordingly, are worthy of further exploration and consideration.

The purpose of this study, accordingly, was to examine five of these lawsuits to investigate and make more transparent their common and also unique, critical, empirical, and pragmatic issues as presented to the court. The secondary purpose of this study was to underscore some of the most problematic features of the education policies that have landed states and districts in court, given the legal implications at issue (see, for example, B. D. Baker, Oluwole, & Green, 2013). The purpose of this study was not, however, to examine the broader potential (see, for example, Pullin, 2013, 2014, 2015) and actual legal issues writ large (see, for example, Superfine, 2016), also given researchers in this study are not legal but rather education policy scholars. Rather, the purpose of this study was to examine five real cases, in depth, to make known the education policy and educational measurement issues being both registered and recognized in court from an academic point of view.

Correspondingly, researchers conducted a case study analysis framing the pertinent legal issues, as per the current scholarly literature on the topic and the key educational measurement criteria written into the most current *Standards for Educational and Psychological Testing* (American Educational

Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014), hereafter referred to as the *Standards* (AERA et al., 2014). Researchers conducted this case study to make more transparent and known the key educational measurement matters "at issue" across these five cases for audiences unfamiliar with these court cases and proceedings, as well as their implications for, and potential or actual impacts on, educational policy and practice writ large.

Conceptual Framework

The conceptual framework researchers used in this study consists of seven measurements issues common in implementing, interpreting, and making decisions based on VAM-based output. The researchers found and defined these issues within the most current *Standards* (AERA et al., 2014), all of which affect the viability of the evaluation systems being used.

With VAM-based output, there are empirical and, as also framed in this piece, legal issues with (a) reliability, (b) validity, (c) bias, (d) transparency, and (e) fairness. Emphases are also on (f) whether VAMs are being used to make consequential decisions using solid evidence (e.g., not arbitrary, capricious, irrational) and (g) whether VAMs' unintended consequences (e.g., teachers avoiding certain students or grade levels) are being registered and addressed. The researchers also note that other teacher evaluation tools or instruments (e.g., observations, student and parent surveys) face very similar measurement issues. In this piece, researchers focus on the issues presented to the court that directly pertained to VAMs.

Reliability

The *Standards* (AERA et al., 2014) define reliability as the degree to which test- or measurement-based scores "are consistent over repeated applications of a measurement procedure [e.g., a VAM] and hence and inferred to be dependable and consistent" (pp. 222-223) for the individuals (e.g., teachers) to whom the test- or measurement-based scores pertain. VAMs are reliable when within-group VAM estimates of teacher effectiveness are more or less consistent over time, from one year to the next, regardless of the type of students and perhaps subject areas teachers teach. Consistency over time is typically captured using "standard errors, reliability coefficients per se, generalizability coefficients, error/tolerance ratios, item response theory (IRT) information functions, or various indices of classification consistency" (AERA et al., 2014, p. 33) that help to situate and make explicit VAM estimates and their (sometimes sizable) errors. These measures are also captured to help others understand the errors that come along with VAM estimates and to better contextualize the VAM-based inferences that result.

Current research on the reliability (or intertemporal stability) of VAMs suggests that teachers classified as “effective” one year might have a 25% to 59% chance of being classified as “ineffective” the next year, or vice versa, with other permutations possible (Chiang, McCullough, Lipscomb, & Gill, 2016; Martinez, Schweig, & Goldschmidt, 2016; Schochet & Chiang, 2013; Shaw & Bovaird, 2011; Yeh, 2013). If a teacher who is classified as a “strong” teacher this year is classified as a “weak” teacher next year, and vice versa, this casts doubt on the reliability of VAMs for the purpose of identifying and making high-stakes decisions regarding teachers (Aaronson, Barrow, & Sanders, 2007; Ballou, 2005; Koedel & Betts, 2007; McCaffrey, Sass, Lockwood, & Mihaly, 2009).

Across VAMs, reliability is a hindrance, especially when unreliable measures are to be used for consequential purposes. Similarly, issues with reliability undermine validity in that reliability is a necessary or qualifying condition for validity. If scores are unreliable, stakeholders may be unable to make or support valid, authentic, and accurate inferences from scores (Brennan, 2006, 2013; M. T. Kane, 2006, 2013; Messick, 1975, 1980, 1989, 1995).

Validity

As per the *Standards* (AERA et al., 2014), validity “refers to the degree to which evidence and theory support the interpretations of test scores for [the] proposed uses of tests” (p. 11). Similarly, “[v]alidity is a unitary concept,” as measured by “the degree to which all the accumulated evidence supports the intended interpretation of [the test-based] scores for [their] proposed use[s]” (p. 14). Accordingly, one must be able to support validity arguments with quantitative or qualitative evidence that the data derived allow for accurate inferences (Brennan, 2006, 2013; M. T. Kane, 2006, 2013; Messick, 1975, 1980, 1989, 1995).

Here, for example, of great import, as well as debate, is the extent to which VAMs can demonstrate that a particular teacher *causes* a particular change in a collective group of students’ test scores over time. The “fundamental concern is that, if making causal attributions is the goal, [whether a] statistical model, however, complex . . .” can satisfy this high bar (Braun, 2005, p. 7). Indeed, it involves a series of “heroic” assumptions (Rubin, Stuart, & Zanutto, 2004) when assuming VAMs yield causal decisions about teachers’ direct impacts on their students’ learning, as measured by large-scale standardized test scores over time (see also Amrein-Beardsley, 2008; Harris, 2011; Wainer, 2004). These and other large-scale standardized test-based concerns (e.g., whether using similar tests under similar situations and conditions yield different VAM-based results; see, for example, Papay, 2011) can be captured under the concerns pertaining to validity.

Following suit, VAM researchers have delved into searching for evidence of many subareas of validity, including but not limited to (a) content-related evidence of validity—"what is to be validated is not the test . . . but the inferences derived from [the] test scores" (Messick, 1989, p. 5); (b) concurrent-related evidence of validity—"the degree of relationship between the test scores and [other] criterion scores" taken at the same time (Messick, 1989, p. 7; see also Messick, 1980); (c) predictive-related evidence of validity, whereas VAM-based estimates might be used to predict future outcomes on a related academic (M. T. Kane, 2013; see also T. J. Kane, McCaffrey, Miller, & Staiger, 2013) or nonacademic measure (e.g., lifetime earnings, pregnancy; see also Chetty et al., 2014a, 2014b); and (d) consequence-related evidence of validity—"[t]he only form of validity evidence [typically] bypassed or neglected in these traditional formulations . . . that . . . bears on the social consequences of test interpretation and use" (Messick, 1980, p. 8; see also M. T. Kane, 2013). However, while all this evidence of validity helps to support construct-related evidence of validity, in VAM research most researchers rely on gathering concurrent-related evidence of validity.

Concurrent validity assesses, for example, whether teachers who post large and small value-added gains or losses over time are the same teachers deemed effective or ineffective, respectively, over the same period using other independent quantitative and qualitative measures of teacher effectiveness (e.g., supervisors' observational scores). If all measures line up and theoretically validate one another, then confidence in them as independent measures of the same construct increases (Messick, 1975, 1980, 1989, 1995; see also Chin & Goldhaber, 2015; Hill, Kapitula, & Umland, 2011). If all indicators do not point in the same direction, something may be wrong with either or both indicators.

Regarding the validity of VAMs, many researchers have investigated whether measures of teacher value-added are substantively related to at least one other criterion of teacher effectiveness (e.g., teacher observational indicators; Grossman, Cohen, Ronfeldt, & Brown, 2014; Hill et al., 2011; T. J. Kane & Staiger, 2012; Polikoff & Porter, 2014; Wallace, Kelcey, & Ruzek, 2016). They also debate whether the concurrent-related evidence of validity that does exist is strong or substantive, with debates primarily focused upon how large in magnitude a correlation might need to be to meet what is still an arbitrary standard of strength (see, for example, T. J. Kane & Staiger, 2012). Few researchers have empirically investigated whether these indicators should align in the first place (see, for example, Chin & Goldhaber, 2015).

Bias

As per the *Standards* (AERA et al., 2014), bias pertains to the validity of the inferences that stakeholders draw from test-based scores. The *Standards* define bias as the "construct underrepresentation of construct-irrelevant components of test scores that differentially affect the performance of different groups of test takers and consequently the . . . validity of interpretations and uses of their test scores" (p. 216). Biased estimates, also known as systematic error as concerning "[t]he systematic over- or under-prediction of criterion performance" (p. 222), are observed when the said criterion performance varies for "people belonging to groups differentiated by characteristics not relevant to the criterion performance" (p. 222) of measurement.

Specific to VAMs, unpredictable characteristics of students can bias the estimates. Schools do not randomly assign teachers the students they teach, so whether their students are invariably more or less motivated, smart, knowledgeable, or capable can bias students' test-based data and teachers' test-based data when aggregated. The current research suggests that VAM-based estimates sometimes present biased results, especially when relatively homogeneous sets of students (i.e., English-language learners [ELLs], gifted and special education students, free-or-reduced-price lunch eligible students) are nonrandomly concentrated into schools, purposefully placed into classrooms, or both. Research suggests that this can happen regardless of the sophistication of the statistical controls used to block said bias (Amrein-Beardsley, 2014; B. D. Baker, 2012; E. L. Baker et al., 2010; Capitol Hill Briefing, 2011; Collins, 2014; Darling-Hammond & Haertel, 2012; P. C. Green, Baker, & Oluwole, 2012; Kappler Hewitt, 2015; Koedel, Mihaly, & Rockoff, 2015; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Newton, Darling-Hammond, Haertel, & Thomas, 2010; Rothstein & Mathis, 2013).

In perhaps the most influential study on this topic, Rothstein (2009, 2010) illustrated VAM-based bias when he found that a student's fifth-grade teacher was a better predictor of a student's fourth-grade growth than was the student's fourth-grade teacher. While others have called into question Rothstein's findings (Goldhaber & Chaplin, 2015; Guarino, Reckase, Stacy, & Wooldridge, 2014; Koedel & Betts, 2009), over the past decade researchers have empirically investigated VAM-based evidence at least 33 times in articles published in high-quality peer-reviewed journals (Lavery et al., 2019). The primary debate raging across articles concerns whether statistically controlling for potential bias by using complex statistical approaches to account for nonrandom student assignment makes bias negligible or rather "ignorable" (Rosenbaum & Rubin, 1983; see also Chetty et al., 2014a, 2014b; Koedel et al., 2015; Rothstein, 2014). While some argue that bias can be effectively controlled for if models include prior achievement, others argue that bias may still exist even if models include prior achievement *and* other available covariates. On this note, Koedel et al. (2015) most recently noted that models should include all potential biasing variables to effectively control for bias as much as possible.

Transparency

While the *Standards* do not specifically define transparency (AERA et al., 2014), transparency pertains to the use of the inferences derived via educational measurements and instruments including VAMs. Hence, for the purposes of this study, researchers define transparency as the extent to which something is accessible and understandable. In terms of VAMs, the main issue presented is that VAM-based estimates may not make sense to those receiving the estimates. Teachers and principals may not understand the models being used to evaluate their performance; hence, they are reportedly quite-to-very unlikely to use the output for formative purposes (Eckert & Dabrowski, 2010; Gabriel & Lester, 2013; Goldring et al., 2015; Graue, Delaney, & Karch, 2013). Practitioners often describe value-added data reports as confusing, not comprehensive in terms of the key concepts and objectives taught, ambiguous regarding teachers' efforts at both the student and composite levels, and often received months after students leave teachers' classrooms.

For example, teachers in Houston (home to one of the cases in this study) expressed that they are learning little about what they did effectively or might use their value-added data to improve their instruction (Collins, 2014). Teachers in North Carolina reported that they were "weakly to moderately" familiar with their value-added data (Kappler Hewitt, 2015). Eckert and Dabrowski (2010) also demonstrated that in Tennessee (home to two of the legal cases in this study) teachers maintained that there was very limited support or explanation helping teachers use their value-added data to improve upon their practice (see also Harris, 2011). Altogether, this is problematic because the main purported strength of really all VAMs is the wealth of positive diagnostic information accumulated for the said formative purposes (see, for example, Sanders, Wright, Rivers, & Leandro, 2009), although simultaneously model developers sometimes make "no apologies for the fact that [their] methods [are] too complex for most of the teachers whose jobs depended on them to understand" (Carey, 2017; see also Gabriel & Lester, 2013).

Fairness

The *Standards* define fairness as the impartiality of "test score interpretations for intended use(s) for individuals from *all* relevant subgroups" (AERA et al., 2014, emphasis added, p. 219). Issues of fairness arise when a test or test use affects some more than others in unfair or prejudiced, yet often consequential ways (Dorans & Cook, 2016).

The main issue here is that states and districts can only produce VAM-based estimates for approximately 30% to 40% of all teachers (B. D. Baker et al., 2013; Gabriel & Lester, 2013; Harris, 2011). The other 60% to 70%, which sometimes include entire campuses of teachers (e.g., early elementary and high school teachers) or teachers who do not teach the core subject areas assessed using large-scale standardized tests (e.g., mathematics and English/language arts), cannot be evaluated or held accountable using teacher-level value-added data. What VAM-based data provide, then, are measures of teacher effectiveness for only a large "handful of teachers" (B. D. Baker et al., 2013, p. 12; see also Collins, 2014; Gabriel & Lester, 2013; Harris & Herrington, 2015; Jiang, Sporte, & Luppescu, 2015; Papay, 2011). When stakeholders use these data to make consequential decisions, issues with fairness arise. Some teachers in certain grades or subject areas experience the negative or positive consequences of these VAM-based data more than their colleagues.

Consequential Use

As per Messick (1989), "[t]he only form of validity evidence [typically] bypassed or neglected in these traditional formulations is that which bears on the social consequences of test interpretation and use" (p. 8). In other words, the social and ethical consequences matter as well (M. T. Kane, 2013; Messick, 1980). The *Standards* (AERA et al., 2014) recommend ongoing evaluation of both the intended and unintended consequences of any test as an essential part of any test-based system, including those based upon VAMs.

The *Standards* (AERA et al., 2014) state that the responsibility of ongoing evaluation of social and ethical consequences should rest on the shoulders of the governmental bodies that mandate such test-based policies as they are those who are to "provide resources for a continuing program of research and for dissemination of research findings concerning both the positive and the negative effects of the testing program" (AERA, 2000; see also AERA Council, 2015). However, this rarely occurs. The burden of proof, rather, typically rests on the shoulders of VAM researchers to provide evidence about the positive and negative effects that come along with VAM use, to explain these effects to external constituencies including policy makers, and to collectively work to determine whether VAM use, given the consequences, can be rendered as acceptable and worth the financial, time, and human resource investments (see also M. T. Kane, 2013).

Intended Consequences

The primary intended consequence of VAM use is to improve teaching and help teachers (and schools/districts) become better at educating students by measuring and then holding teachers accountable for their effects on students (Burris & Welner, 2011). The stronger the consequences, the stronger the motivation leading to stronger intended effects. Secondary intended consequences included replacing the nation's antiquated teacher evaluation systems (see, for example, Weisberg et al., 2009).

Yet, in practice, research evidence supporting whether VAM use has led to these intended consequences is suspect, given the void of evidence supporting such intended effects. For improving teaching and student learning, as noted prior, VAM estimates may tell teachers, schools, and states little-to-nothing about how teachers might improve upon their instruction, or how all involved might collectively improve student learning and achievement over time (Braun, 2015; Corcoran, 2010; Goldhaber, 2015). For reforming the nation's antiquated teacher evaluation systems, recent evidence suggests that this has not occurred (Kraft & Gilmour, 2017).

Unintended Consequences

Simultaneously, stakeholders often fail to recognize VAMs' unintended consequences (AERA, 2000; see also AERA Council, 2015). Policy makers must present evidence on whether VAMs cause unintended effects and whether the said unintended effects outweigh their intended effects, all things considered. Policy makers should also contemplate the educative goals at issue (e.g., increased student learning and achievement), alongside the positive and negative implications for both the science and ethics of using VAMs in practice (Messick, 1989, 1995).

As summarized by Moore Johnson (2015), unintended consequences include, but are not limited to, (a) teachers being more likely to "literally or figuratively 'close their classroom door' and revert to working alone . . . [which] . . . affect[s] current collaboration and shared responsibility for school improvement" (p. 120); (b) teachers being "[driven] . . . away from the schools that need them most and, in the extreme, causing them to leave [or to not (re)enter] the profession" (p. 121); and (c) teachers avoiding teaching high-needs students if teachers perceive themselves to be at greater risk of teaching students who may be more likely to hinder their value-added, "seek[ing] safer [grade level, subject area, classroom, or school] assignments, where they can avoid the risk of low VAMS scores" (p. 120), all the while leaving "some of the most challenging teaching assignments . . . difficult to fill and likely . . . subject to repeated [teacher] turnover" (p. 120; see also B. D. Baker et al., 2013; Collins, 2014; Hill et al., 2011). The findings from these studies and others point to damaging unintended consequences where teachers view and react to students as "potential score increasers or score compressors; [s]uch discourse dehumanizes students and reflects a deficit mentality that pathologizes these student groups" (Kappler Hewitt, 2015, p. 32; see also Darling-Hammond, 2015; Gabriel & Lester, 2013; Harris & Herrington, 2015; Mellon, 2010).

In sum, as per the *Standards* (AERA et al., 2014), policy makers should commit to ongoing evaluation of all these issues. The American Statistical Association (ASA, 2014), the AERA Council (2015), the National Academy of Education (E. L. Baker et al., 2010), and the NASSP, n.d.), have also underscored similar calls for research within their associations' positions statements about VAMs and VAM use.

Researchers used all of these overarching issues, accordingly, to frame this case study analysis about the pragmatic and empirical issues presented to the court. Researchers set out to make this set of issues transparent to help others better understand the issues of primary dispute in each of the five court cases they selected for this study herein.

Method

For the purposes of this study, researchers conducted a case study (Campbell, 1975; Flyvbjerg, 2011; Gerring, 2004; Ragin & Becker, 2000; Thomas, 2011; VanWynsberghe & Khan, 2007) to examine the collective cases of the five lawsuits at focus. The case study approach, according to VanWynsberghe and Khan (2007), best suits research with (a) a small number of participants, (b) a focus on contextual detail, (c) a focus on nonexperimentally controlled events, (d) a well-defined case, and (e) multiple data sources. Consequently, researchers (a) examined five lawsuits, (b) positioning them within their appropriate contexts, and (c) describing the lawsuits as they exist in the real world, (d) given that it was these five cases for which researchers had full access to all exhibits and legal documents (i.e., including documents inaccessible via legal databases such as Lexis/Nexis or Westlaw) from which they could (e) conduct across- and within-case analyses.

As noted, these legal cases were similar and separate enough to permit such a case study approach, especially as the plaintiffs across each lawsuit had associated experiences and could serve as comparable instances of the same legal phenomenon (Ragin & Becker, 2000). Statistical generalizations should not be permitted given researchers' sample of convenience, although naturalistic generalizations might be warranted (Stake, 1978; Stake & Trumbull, 1982).

Researchers analyzed and used these practical experiences to help others better understand how multiple reformed teacher evaluation systems were being used and experienced within and across states, districts, and schools. (Flyvbjerg, 2011) All of this, again, was predicated upon large- and small-scale educational policies, and many policy-based, consequential decisions also at stake and of legal dispute.

Case Sample With Summaries

Researchers conducted this case study using a sample of convenience, given these five cases, again, were the five cases to which researchers had access to the full set of exhibits and legal documents (i.e., including documents inaccessible via legal databases such as Lexis/Nexis or Westlaw) that researchers needed to conduct this case study analysis. The five cases of interest included the following, as organized by the states in which each case was filed. Please note that researchers include, below, background information per case, as pulled per court documents:

1. New Mexico: *American Federation of Teachers—New Mexico and the Albuquerque Federation of Teachers (Plaintiffs) v. New Mexico Public Education Department (Defendants)*, State of New Mexico, County of Bernanillo, First Judicial District Court.
   - *Background*: The state's homegrown or locally developed VAM (see Swedian, 2014), at the time of this case being filed, comprised 50% of a teacher's overall evaluation score.
   - *Alleged violations*: Plaintiffs are arguing that teachers received poor VAM-based ratings because of flawed or incomplete data, because teachers were linked to the wrong students, or students they never taught, or subject areas they never taught as also sometimes assessed by state-level tests that did not map onto that which they taught.
   - *Status*: The court granted a preliminary injunction in December 2015, and the case was to be heading back to court in October 2017, but the case has since been postponed with it now scheduled to be heard in the spring 2019.
2. New York: *Sheri G. Lederman (Plaintiff) v. John B. King, Jr. Commissioner, New York State Education Department (Defendants)*. Supreme Court of the State of New York, County of Albany.
   - *Background*: The state's homegrown VAM (i.e., the New York growth system), at the time of this case being filed, comprised 50% of a teacher's overall evaluation score.

- ○ *Alleged violations*: The plaintiff's teacher-level value-added score and subsequent rating as "ineffective" were "arbitrary and capricious." New York's system also unfairly penalized teachers whose students consistently scored well and could not demonstrate growth upward (e.g., teachers of gifted students, as this teacher normally taught).
- ○ *Status*: The court ruled in favor of the plaintiff in May of 2016.

3. Tennessee: *Lisa Trout (Plaintiff) v. Knox County Board of Education (Defendant)*. United States District Court, Eastern District of Tennessee, Knoxville Division.
   - ○ *Background*: The state's homegrown, but also the first VAM of its kind (i.e., the Tennessee Value-Added Assessment System [TVAAS], although this VAM is now more popularly recognized and used throughout the nation as the more generalized Education Value-Added Assessment System [EVAAS]),[2] at the time of this case being filed, comprised 35% to 50% of a teacher's overall evaluation score.
   - ○ *Alleged violations*: The plaintiff was unfairly denied a merit pay bonus because of her TVAAS scores, which were supposed to have been calculated using the systemwide TVAAS scores because of her teaching position within an alternative school. In addition, 10 of the plaintiff's students were not included in her TVAAS score.
   - ○ *Status*: The court dismissed the case in February of 2016.
4. Tennessee: *Mark Taylor (Plaintiff) v. Kevin S. Huffman, William Edward Haslam and Knox County Board of Education (Defendants)*, United States District Court, Eastern District of Tennessee, Knoxville Division.
   - ○ *Background*: See the *Background* for Case 3 above.
   - ○ *Alleged violations*: The plaintiff was unfairly denied a merit pay bonus because of his TVAAS scores, based on the test scores of 22 of his 142 students. Because the plaintiff taught four upper-level science courses and one regular eighth-grade science class, and there were no state-level tests to evaluate students in the four upper-level courses, his relatively higher performing students were not included in his TVAAS score.
   - ○ *Status*: The court dismissed the case in February of 2016. **Note*: Plaintiffs in two of the three total court cases in Tennessee combined efforts before the cases were heard before the court; hence, and hereafter, researchers present the same backgrounds and the same statuses for Cases 3 and 4. Researchers also analyzed both

cases using the official legal documents accessed and retrieved from what ultimately became one case titled as follows: *Lisa Trout et al. (Plaintiffs) v. Knox County Board of Education (Defendant)*.

5. Texas: *Houston Federation of Teachers (Plaintiff) v. Houston Independent School District (Defendant)*, United States District Court, Southern District of Texas, Houston Division.
    - *Background*: The district used the aforementioned generalized EVAAS at the time of this case being filed, comprising 50% to 100% of a district teacher's overall score.
    - *Alleged violations*: Plaintiffs are arguing that EVAAS output is inaccurate; the EVAAS is unfair; that teachers are being evaluated via the EVAAS using tests that do not match the curriculum they teach; that the EVAAS system fails to control for student-level factors that affect how well teachers perform but that are outside of teachers' control (e.g., parental and student socioeconomic effects); that the EVAAS is opaque, incomprehensible (e.g., a "black box" model) and, hence, very difficult if not impossible to use to improve upon their instruction (i.e., not actionable); and that teachers' due process rights are being violated because teachers do not have adequate opportunities to change their professional practice as a result of their EVAAS reports.
    - *Status*: The court ruled in favor of plaintiffs in May of 2017 on two counts.

Data Sources

For each case, researchers analyzed official court documents including affidavits, case documents (e.g., official court filings), and rebuttals for all five cases (turned four total cases, see Nos. 3 and 4 above, see also the *Status* section of No. 4 above). Across cases, this included an average of 6.5 documents per case (calculated with four cases), with the most documents coming from the case in Houston, Texas and the fewest coming from the case in New Mexico. Otherwise, course documents included 18 affidavits (including one oral affidavit transcribed by a court reporter), five case documents, and two rebuttals. In total, 1,063 pages of legal documents made up of 336,491 words served as researchers' primary and official data sources. For each court case, the documents also varied in terms of balance on the sides of the plaintiff(s), defendant(s), or on either side or with no intent (e.g., official course filings). Across cases, there tended to be more written on the side of the plaintiffs although the reverse was true in the case in Tennessee.

Data Analyses

To analyze the written pages, researchers read through each case document coding for text, quotes, and concepts related to the elements of their a priori framework (Miles & Huberman, 1994). More specifically, using this deductive approach to coding (in which the categories were preselected from this framework), researchers read through the material in each case issue-by-issue to collectively determine and agree upon the main measurement issues for each case. Next, researchers negotiated, agreed upon, and then grouped text, quotes, and concepts by measurement element as per their a priori framework, all the while whittling down the textual material into themes by measurement construct per case. Thereafter, researchers compared the cases overall, using a constant comparative method in which "the data in hand [were] then analyzed again and compared with the new data" to improve trustworthiness (Boeije, 2002, p. 393) and to also examine how the cases overlapped or diverged (Ragin & Becker, 2000). This systematic approach to coding modeled a framework method (Gale, Heath, Cameron, Rashid, & Redwood, 2013; Ritchie, Lewis, Nicholls, & Ormston, 2013).

Findings

The following section addresses each key issue. Specifically, researchers attempt to make transparent how each of the cases addresses each issue (e.g., explicitly, or as couched in a research narrative). Accordingly, in each subsection, researchers present readers with a case- or theme-based discussion, not a step-by-step, case-by-case, analysis.

Reliability

Authors of court documents across the cases acknowledged disputes surrounding VAMs' levels of reliability. In the New York case, the alleged poor reliability of the state's VAM was probably the most significant issue acknowledged by the court. Fittingly, this case relied upon examples of individual growth scores varying "wildly" year to year, in New York and elsewhere throughout the literature (Chiang et al., 2016; Martinez et al., 2016; Schochet & Chiang, 2013; Shaw & Bovaird, 2011; Yeh, 2013).

In the Houston and Tennessee cases, case documents also presented reliability as a central issue. In the Tennessee case, plaintiffs invoked research surrounding reliability in general, but not research specific to the TVAAS, given reliability estimates were unknown at the time. In the Houston case, plaintiffs showed that the EVAAS yielded statistically large standard errors making rankings unstable. Defendants did not defend these issues of dispute despite examples of literature that invoke reliability coefficients from other statistical spaces (e.g., baseball) and argue that the "relatively" low levels of reliability observed might not be as problematic as argued (see also McCaffrey et al., 2009). While performance may be relatively inconsistent over time across a variety of domains of human behavior, from baseball to teaching, plaintiffs in Houston argued that attaching consequences to such inconsistencies was concerning.

Relatedly, in New Mexico, plaintiffs argued that the state's VAM was also unreliable because the state did not use enough years of data to calculate teachers' VAM-based estimates. While defendants acknowledged that the state's teacher evaluation model needed at least 3 years of student achievement data to be as reliable as possible, many teachers' VAM-based estimates were calculated using only 1 or 2 years. In many ways, this had to do with the state's recent implementation of its reformed evaluation system, after which the state attached consequences to estimates, regardless. Plaintiffs argued that New Mexico implemented the system too quickly and injudiciously.

Validity

The aforementioned types of evidence of validity were also presented as essential for VAM adoption, implementation, and use across these cases. In New York, one of the plaintiff's expert witnesses wrote that the New York VAM did not yield valid estimates from which valid inferences could be made about high-achieving students. This and other expert witnesses argued that ceiling effects also caused this teacher's value to decline during the year she was classified as "ineffective." No defendants countered these claims.

In all the other cases, debates considered whether the states or district validated the tests that students took for (a) measuring student growth over time and then (b) directly attributing that growth to students' individual teachers (i.e., teachers' causal effects). No defendants, across cases, showed that the tests were validated for these purposes. Relatedly, across cases, court discussions also centered around confounding variables (i.e., other factors that might cause variation in student test scores besides teachers' effects, distorting validity of inference). Accordingly, all plaintiffs' witnesses pressed the VAM position statement released by the ASA (2014), arguing that teacher influences are associated with 1% to 14% of the variance in the test scores used to estimate teachers' value-added effects and that, relatedly, VAMs measure correlation and not causation. Houston and New Mexico defendants, however, argued that policy makers could make causal claims using one set of peer-reviewed articles to press their case (i.e., Chetty et al., 2014a, 2014b).

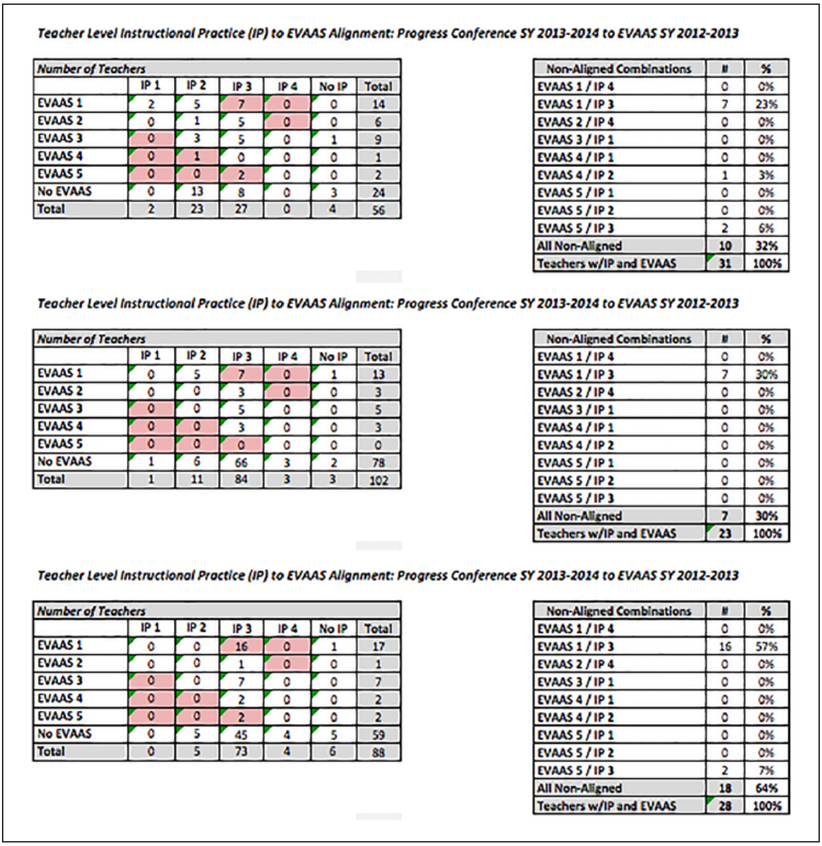
Defendants' claims included, for example, "A teacher's effectiveness has profound consequences for the achievement and earnings of students." In no case was this issue resolved, just as it has not been resolved within the academic community.

Across cases, plaintiffs and defendants concentrated upon the concurrent-related types of evidence of validity defined prior. Only in Houston did stakeholders measure correlation coefficients between teachers' EVAAS and observational estimates, with plaintiffs showing that the correlations between both measures mirrored other VAMs on the market (see also Grossman et al., 2014; Hill et al., 2011; T. J. Kane & Staiger, 2012; Polikoff & Porter, 2014; Wallace et al., 2016). That is, correlations were statistically significant yet weak for the 3 years of concern in that case (r = .30, r = .28, r = .34). Interestingly, though, the district set out to deliberately improve these correlations by building policies to encourage rating alignment, which also became a validity concern in court.

More specifically, Houston plaintiffs registered concerns about how teachers' observational scores, used alongside their EVAAS estimates, were being "artificially conflated" as written into policy. In Houston, district policy required that principals whose observational scores for individual teachers deviated too far from their EVAAS estimates were to revise the more subjective observational scores to better match them with teachers' more objective EVAAS output, to also keep observers' subjectivities in check (see Figure 1 for three schools' alignment matrices with associated descriptive statistics). One of the defendants' expert witnesses argued that both indicators should not be independent of one another, but influence each other to help offset both indicators' methodological imperfections, regardless of the implications for validity.

The same validity issue emerged in the case in Tennessee, with the entire state, as per state-level policy, encouraging similar alignment practices. Not only did the state use similar alignment matrices to make sure teachers' observational scores aligned with their individual TVAAS estimates, but also the state put into place policy indicating how principals whose scores did not satisfactorily align should receive support to enhance their alignment and reduce their subjectivities. See an "alignment of scores" instructional table used in Tennessee in Figure 2 to help facilitate this process. Defendants did not counter these claims.

Also of importance in Tennessee and Houston and unique to the TVAAS/EVAAS, statisticians also revised teachers' value-added scores retroactively as more information became available about teachers' students over time, to get at more valid inferences. The plaintiffs argued that this practice voided teachers' prior estimates and the validity of the decisions made as based on



Teacher Level Instructional Practice (IP) to EVAAS Alignment: Progress Conference SY 2013-2014 to EVAAS SY 2012-2013

Number of Teachers						
	IP 1	IP 2	IP 3	IP 4	No IP	Total
EVAAS 1	2	5	7	0	0	14
EVAAS 2	0	1	5	0	0	6
EVAAS 3	0	3	5	0	1	9
EVAAS 4	0	1	0	0	0	1
EVAAS 5	0	0	2	0	0	2
No EVAAS	0	13	8	0	3	24
Total	2	23	27	0	4	56

Non-Aligned Combinations	#	%
EVAAS 1 / IP 4	0	0%
EVAAS 1 / IP 3	7	23%
EVAAS 2 / IP 4	0	0%
EVAAS 3 / IP 1	0	0%
EVAAS 4 / IP 1	0	0%
EVAAS 4 / IP 2	1	3%
EVAAS 5 / IP 1	0	0%
EVAAS 5 / IP 2	0	0%
EVAAS 5 / IP 3	2	6%
All Non-Aligned	10	32%
Teachers w/IP and EVAAS	31	100%

Teacher Level Instructional Practice (IP) to EVAAS Alignment: Progress Conference SY 2013-2014 to EVAAS SY 2012-2013

Number of Teachers						
	IP 1	IP 2	IP 3	IP 4	No IP	Total
EVAAS 1	0	5	7	0	1	13
EVAAS 2	0	0	3	0	0	3
EVAAS 3	0	0	5	0	0	5
EVAAS 4	0	0	3	0	0	3
EVAAS 5	0	0	0	0	0	0
No EVAAS	1	6	66	3	2	78
Total	1	11	84	3	3	102

Non-Aligned Combinations	#	%
EVAAS 1 / IP 4	0	0%
EVAAS 1 / IP 3	7	30%
EVAAS 2 / IP 4	0	0%
EVAAS 3 / IP 1	0	0%
EVAAS 4 / IP 1	0	0%
EVAAS 4 / IP 2	0	0%
EVAAS 5 / IP 1	0	0%
EVAAS 5 / IP 2	0	0%
EVAAS 5 / IP 3	0	0%
All Non-Aligned	7	30%
Teachers w/IP and EVAAS	23	100%

Teacher Level Instructional Practice (IP) to EVAAS Alignment: Progress Conference SY 2013-2014 to EVAAS SY 2012-2013

Number of Teachers						
	IP 1	IP 2	IP 3	IP 4	No IP	Total
EVAAS 1	0	0	16	0	1	17
EVAAS 2	0	0	1	0	0	1
EVAAS 3	0	0	7	0	0	7
EVAAS 4	0	0	2	0	0	2
EVAAS 5	0	0	2	0	0	2
No EVAAS	0	5	45	4	5	59
Total	0	5	73	4	6	88

Non-Aligned Combinations	#	%
EVAAS 1 / IP 4	0	0%
EVAAS 1 / IP 3	16	57%
EVAAS 2 / IP 4	0	0%
EVAAS 3 / IP 1	0	0%
EVAAS 4 / IP 1	0	0%
EVAAS 4 / IP 2	0	0%
EVAAS 5 / IP 1	0	0%
EVAAS 5 / IP 2	0	0%
EVAAS 5 / IP 3	2	7%
All Non-Aligned	18	64%
Teachers w/IP and EVAAS	28	100%

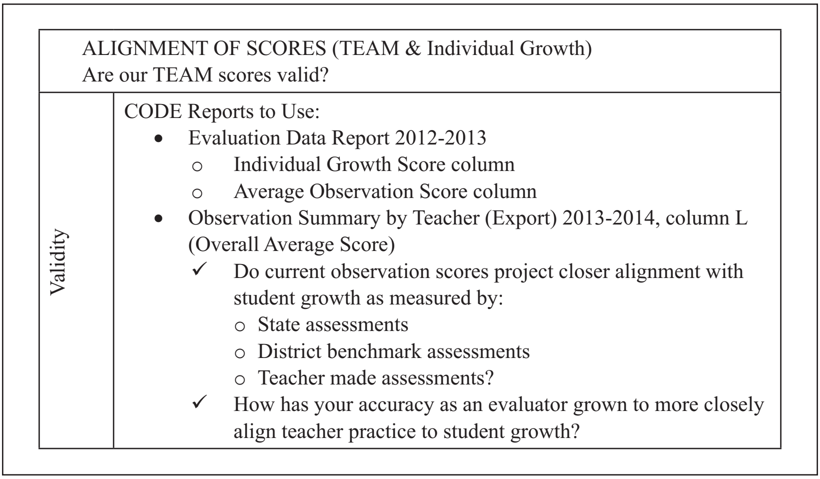
Figure 1. Three Houston schools' rating alignment matrices with associated descriptive statistics.

Note. Cells highlighted in red indicate that teachers' EVAAS and IP scores are unacceptably misaligned. EVAAS = Education Value-Added Assessment System; IP = instructional practice.

their prior estimates. This, too, has major implications for validity, whereas this shows that the inferences based on EVAAS estimates could not be taken as valid or "true" as they consistently changed in retrospect (see Ballou & Springer, 2015).

Bias

Examinations of bias also traversed cases in New Mexico, New York, and Houston. In these cases, defendants, again, cited one of the two peer-reviewed



ALIGNMENT OF SCORES (TEAM & Individual Growth) Are our TEAM scores valid?	
Validity	CODE Reports to Use: • Evaluation Data Report 2012-2013 o Individual Growth Score column o Average Observation Score column • Observation Summary by Teacher (Export) 2013-2014, column L (Overall Average Score) ✓ Do current observation scores project closer alignment with student growth as measured by: o State assessments o District benchmark assessments o Teacher made assessments? ✓ How has your accuracy as an evaluator grown to more closely align teacher practice to student growth?

Figure 2. An "alignment of scores" instructional table used in Tennessee, as per Tennessee State Board of Education (TSBE, 2012) policy (Anonymous, personal communication, October 15, 2015; recreated from a photograph of an official document).

Chetty et al. (2014a) studies, arguing that VAM-based estimates were not biased. Plaintiffs criticized the same piece pointing out that VAMs "on average" may or may not be biased, but year-to-year they may yield biased measurements when teachers teach homogeneous classes of high- or low-performing students. This study and topic turned out to be a source of disagreement that stumped the courts.

Only in Houston did expert witnesses directly assess VAM-based estimates for bias. Plaintiffs showed that the estimates derived via the EVAAS may be relatively more biased against teachers teaching racial minority, ELLs, free-or-reduced-price lunch eligible, and special education students, despite the frequent statements about the EVAAS's capacity to block and control for bias (Sanders et al., 2009). Inversely, defendants argued that plaintiffs' evidence and opinions pertaining to bias represented a minority view. They argued that teachers' value-added scores would still provide useful information to principals, even if partially biased.

The cases in Tennessee and New Mexico also addressed bias, but in terms of subject- and grade-level bias. In both cases, the evaluation systems rated English/language arts and mathematics teachers significantly differently. This same pattern applied to teachers in different grade levels in Tennessee. Both sides argued that this systematic pattern occurred for one of two reasons: (a) these subject- and grade-level teachers were simply better or worse as a whole, or (b) there was another biasing factor at play (see also Holloway-Libell, 2015).

Transparency

Authors of court documents across the cases also acknowledged disputes surrounding VAMs' levels of transparency. In the Houston case, however, transparency was probably the most noteworthy issue acknowledged by the court. Again, throughout Houston, the district used the aforementioned EVAAS to evaluate teachers—the EVAAS, which is sold at least in part to provide a "wealth of positive diagnostic information" for formative purposes (Sanders et al., 2009, p. 9). However, Sanders (the EVAAS creator) continued to make "no apologies for the fact that his methods were too complex for most of the teachers whose jobs depended on them to understand" (Carey, 2017; see also Gabriel & Lester, 2013). Hence, in Houston (as well as in Tennessee, given its use of the original TVAAS), two forces conflicted. On one hand, teachers deserved transparency so they could access, understand, and then learn from their value-added estimates to become better educators. On the other hand, EVAAS owners desired to protect their intellectual property (e.g., proprietary algorithms and source codes), albeit at the cost of millions of public taxpayers' dollars per year (e.g., US$680,000 per year in Houston).

This conflict played out in the court, accordingly. Despite a Public Information Act request submitted by the plaintiffs, EVAAS owners provided very limited access to the data used to evaluate Houston teachers, providing only general descriptions of their statistical procedures and formulas, and technical reports. They did not, however, release their decision rules, statistical "secrets," or source codes. What they did permit was access to protected information to one of the plaintiffs' expert witnesses, although this was done using a highly controlled and monitored procedure and place (i.e., demarcated by the presiding judge as "extremely restrictive access"). This eventually led to defendants filing an alleged violation of their protective order, with defendants arguing that what this witness wrote into his or her affidavit exposed some of the EVAAS's proprietary information. In the end, the presiding judge ruled that defendants "interpret[ed] the[ir] protective order too broadly," adding that this "overly broad interpretation urged by [EVAAS owners] would inhibit legitimate discussion about the lawsuit."

Perhaps, then, what became even more of an issue in court was whether teachers, without having the access that the aforementioned expert witness had to what plaintiffs argued was a "black box" model, were able to access, understand, make sense of, or replicate their scores. Hence, this became one of the chief complaints in the case, ultimately playing out in the presiding judge's final ruling on whether such secrecy should be tolerated (see forthcoming).

In New York, state law mandates that the teacher evaluation process be transparent, yet key features of the VAM also remained opaque there. For example, although state representatives argued that they made the state model transparent by posting their VAM on their Department of Education website, one expert witness argued that a practicing teacher could not possibly understand the website. Teachers in New Mexico also described a similar lack of transparency, impeding their abilities to understand and use their VAM-based scores to become better teachers.

Fairness

The discussion about fairness across the four cases related to, as mentioned, whether VAMs or VAM uses affected some teachers more than others in unfair and consequential ways; although issues with fairness were either peripheral (e.g., in the New York and Tennessee cases) or at the forefront of these cases. In the Houston case, plaintiffs argued that some teachers were being evaluated via the EVAAS using tests that did not match the curriculum they were teaching. Plaintiffs also argued that although more teachers were held accountable using the EVAAS in Houston, given the district also had large-scale standardized tests in science and social studies (i.e., allowing for more than the typically 30%-40% of teachers held accountable using VAM-based systems), because such highly consequential decisions were tied to their EVAAS output (including termination), this was still an issue with fairness.

In Tennessee and Houston, given their use of the TVAAS/EVAAS, respectively, the court also addressed how teachers who teach small classes were more likely to perform at average levels. The system put out this performance measurement not because they were actually average, but because the methods that modelers used to counteract the errors caused by small sample sizes (e.g., shrinkage methods) pulled teachers of small class sizes value-added scores toward the mean (Ballou & Springer, 2015; see also New York). In Tennessee, the sitting EVAAS director positioned this "a [necessary] statistical protection put into place so that we're not misclassifying teachers."

In New Mexico, plaintiffs charged, as well, that the state's reformed teacher evaluation system was unfair, harming some teachers over others. A list of highly respected educators (including one state senator who retired from teaching) argued these points. This state senator, more specifically, positioned as a fairness issue that "glaring errors" marred state's ratings of some teachers over others (e.g., calculus teachers being evaluated using the state's mathematics tests, as also noted in Houston), and that the state adjusted some teachers' ratings when errors were identified by teachers' districts (i.e., not the state) and the state, accordingly (albeit unfairly), adjusted those teachers' scores versus others. Defendants did not explicitly defend these points, with the exception of explanations regarding how the state was attempting to include more teachers in the state's evaluation system (e.g., via the adoption of more tests at more grade levels) and to reduce data and scoring errors in more fair and systematic ways.

Consequential Uses

As noted, the high-stakes consequences at the core of these lawsuits included teachers' permanent files being flagged; possible professional development or interventions; awarding or revocation of teacher licenses or tenure; salary increases, decreases, or the granting of merit pay; and probation or termination. Most notable across suits, though, was the case in Houston, as Houston is widely recognized for its use of the EVAAS for more consequential decision-making purposes (i.e., teacher merit pay and termination) than anywhere else in the nation (Collins, 2014; Corcoran, 2010; Harris, 2011). For example, 221 Houston teachers were terminated in 2011 based predominantly (and arguably solely) on their EVAAS scores, given teachers termination letters noted that the terminated teachers demonstrated "a significant lack of student progress attributable to the educator" or "insufficient student academic growth reflected by [their EVAAS] value-added scores." The district also terminated teachers for the same reasons in the years before and after, which in many ways brought this case forward. That is, the highly consequential nature of Houston's system is quite literally what landed this district in court. Specifically, that limited-to-no evidence existed showing that terminating these teachers (or attaching money to teachers' value-added) yielded the system's intended consequences (i.e., increasing student achievement). This point was also of great importance because, as argued by one of the defendants' expert witnesses, the EVAAS (and observational component) in Houston "[were] primarily used for feedback and targeted professional development." See Figure 3, as taken from court records, for other test-based indicators of whether the same theory of change as applied in Houston benefited students, given the intended consequences alleged.

In Tennessee, plaintiffs charged that they were denied monetary bonuses after their TVAAS estimates were miscalculated. Teachers who should have received bonuses (e.g., US$2,000) received nothing. In addition, the state also used TVAAS scores for a variety of other high-stakes decisions such as tenure decisions, advancement of teaching licenses, and also establishing

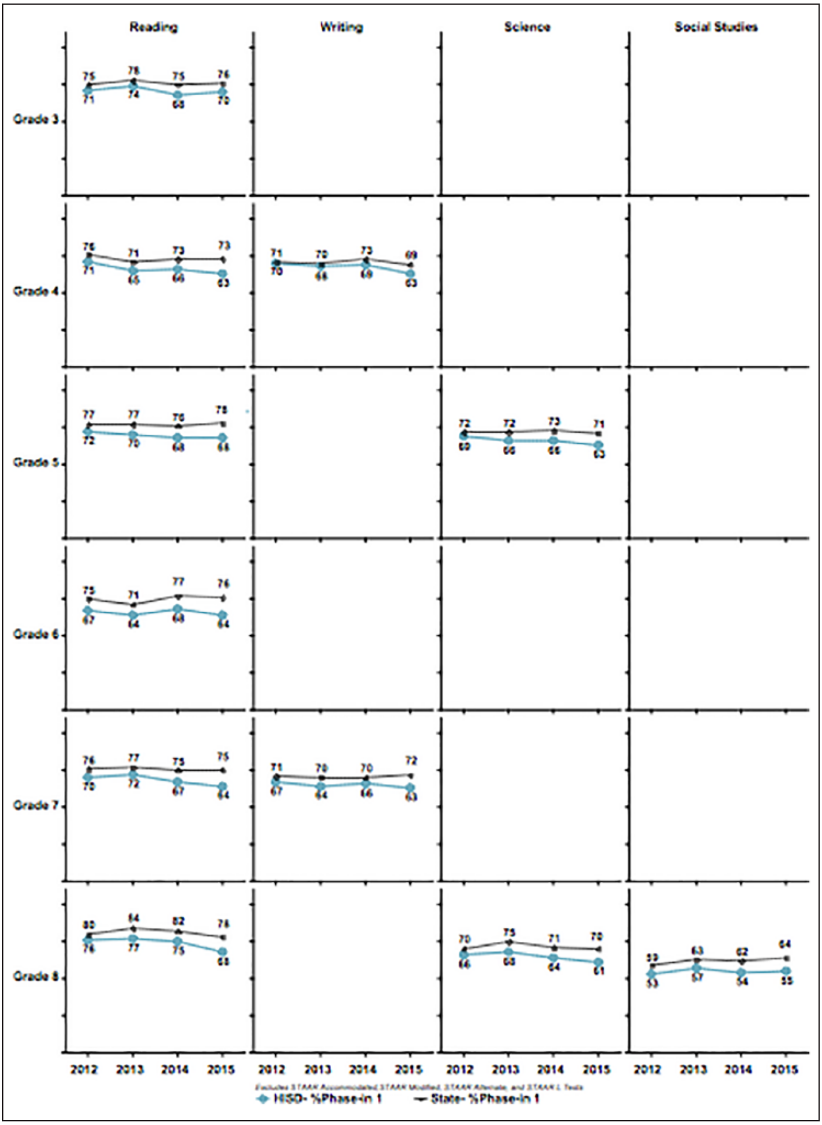


Figure 3. Houston's performance (i.e., percent met standard), as compared with the state, on Texas's large-scale standardized tests in Grades 3 to 8 in reading, writing, science, and social studies, 2012-2015 (spring administration)

"inefficiency" as a rationale for termination. In this case, however, counter-evidence supporting the state's intended consequences—gains demonstrated on the National Assessment of Educational Practice (NAEP) and touted throughout the state as evidence that the state's TVAAS-based system yielded improved student achievement (see, for example, Huffman, 2014; see also Rubinstein, 2014)—was never debated (see also forthcoming).

In New Mexico, despite a widespread understanding that 2013-2014 was to be a "hold-harmless" year so that the state could, with prudence, test and roll out its reformed teacher evaluation system, the state flagged teachers' files with their value-added or overall effectiveness categories (e.g., "highly effective," "effective," "ineffective," "highly ineffective") and, in some cases, placed teachers with low value-added scores on professional improvement plans. This problem was exacerbated, again, given the state also admitted to miscalculating many New Mexico teachers' value-added and overall evaluation scores. Because the state was rolling out such a new system, however, no other discussions in court regarding the system's intended consequences were pertinent, given no one had a chance to argue, evaluate, or assess them.

Finally, in New York, the court viewed the consequences attached to its VAM differently, positioning an "ineffective" tag as consequential in terms of public shame and loss of reputation in the professional and greater community for a teacher. Hence, court documents presented how stigma might also play a role as a high-stakes consequence, regardless of the intended consequences that were not of explicit concern in this case.

Unintended Consequences

Perhaps, some of the most interesting elements across all cases were also the unintended consequences playing out across these lawsuits (e.g., teachers working in isolation, competing, leaving the profession, and avoiding certain grade levels, subject areas, and students). However, across cases, plaintiffs mentioned these consequences as more circumstantial, anecdotal, and, consequently, of less substantive concern. For example, in New Mexico, district-level experts reported an exodus of teachers and administrators who were "fed up" with the evaluation system, with one district superintendent testifying that, after the 2013-2014 school year, the district set an all-time record for resignations and transfers. The judge heard this type of evidence, but did not officially acknowledge the evidence (e.g., in the court's official ruling), although this may not be as surprising because rules of evidence and past precedent constrain judges.

The only thing unique, perhaps, that came out of the cases in Houston and Tennessee as officially acknowledged by the court, was that teachers claimed that their administrators were changing and therefore distorting their other evaluative (i.e., observational) indicators of effectiveness. Policy required administrators to trust teachers' value-added more than their observational scores and they acted accordingly. This was registered as an unintended consequence that also violated validity, via the deliberate forcing of increased correlations between both indicators to demonstrate increased alignment and increased (manufactured) objectivity on the observational side (i.e., termed "artificial conflation," as noted prior).

Rulings

Given all of the aforementioned issues addressed in court, the court rulings were as follows. In New Mexico, a state court judge denied a motion to dismiss the lawsuit and granted a preliminary injunction, preventing the state from making any more consequential decisions about teachers throughout the state, until the state (and/or others external to the state) could evidence to the court (with evidence explicitly aligned with the abovementioned educational measurement principles) during another trial that such consequences are warranted and nonarbitrary, and that the system is legally defensible as well as "uniform and objective" as per state constitutional requirements. Again, the final hearing of this case has been postponed multiple times since its initial scheduling in October 2017 to its current scheduling in the spring 2019; however, with the recent (i.e., November 2018) election of a democratic governor who explicitly vowed to overhaul the state's teacher evaluation system, namely, in terms of *not* using teachers' students' test scores to hold New Mexico's teachers accountable for their effects, it is possible this case may ultimately be dismissed.

In New York, the court ruled in favor of the plaintiff in May 2016, with the court's conclusion being based upon the following: (a) the evidence of VAM bias against teachers at both ends of the spectrum (e.g., those with homogeneous sets of high-performing students or low-performing students), (b) the effect of the plaintiff's small class size on her growth score, (c) the unexplained swing in the plaintiff's growth score from one year to the year following, and, most tellingly, (d) the strict imposition of relativistic rating constraints forced to fit a "bell curve" by placing teachers in predetermined categories regardless of their "true" levels of effectiveness.

In Tennessee, the court dismissed the case in February 2016 before it went to trial. More specifically, the court dismissed the plaintiffs' claims. While the court was "not unsympathetic to the teachers' claims," and the court registered substantial concerns about the reliability and validity of the TVAAS, the court ruled that the state satisfied the threshold of the “rational basis test” at issue, writing that

> [w]hile the court expresses no opinion as to whether the Tennessee Legislature has enacted sound public policy, it finds that the use of TVAAS as a means to measure teacher efficacy survives minimal constitutional scrutiny. If this policy proves to be unworkable in practice, plaintiffs are not to be vindicated by judicial intervention but rather by democratic process.

In short, evaluating teachers using the TVAAS was “rationally related to a legitimate government interest.”

Finally, in Houston, the court ruled in favor of plaintiffs in May of 2017 on two counts, ruling that the plaintiffs had legitimate claims regarding (a) how EVAAS was a violation of their Fourteenth Amendment due process protections, and (b) whether the system permitted teachers to act to ensure their VAM scores were accurate as due process requires. In October 2017, the parties settled the case when the district agreed to remove from its evaluation system any scoring system, including but not limited to the EVAAS, that could not be verified.

Conclusion and Implications

The way the framing of the lawsuits affected future policy and practice is too complex to untangle exactly, but the lawsuits did present some important implications for policy and practice worth mention. First, the lawsuit framing had direct implications for future policy because the judges ruled only on what the plaintiffs argued. In particular, the New York judge validated the arguments of plaintiffs who claimed that VAMs could bias certain teachers. In addition, the New York judge presented an example of direct and blunt court intervention in education policy related to the rights of teachers (i.e., fairness).

Some rulings complicated the picture, however, showing that the arguments presented sometimes had no bearing on the legal status of policy. The judge in Tennessee complicated the precedent set by the New York judge, by setting a line for how and why a court should intervene with education policy. Although the court showed concern about reliability and validity of the VAM in question, the judge ruled that the policy should be tested through voting, not a court judgment. Legally, according to the judge’s ruling in Tennessee, the arguments made seemed to have merit, but were not legally relevant.

Less direct implications of the arguments included the public coverage of the lawsuits and the arguments made by plaintiffs. As Superfine (2016) argued, the lawsuits served as a place for teachers to be heard. Teachers, who may have had less voice or been overruled in public policy decisions at the federal and state level, now had a platform from which to make arguments about their rights. Through the lawsuits, covered by popular media (Rhee, 2011), teachers and teacher unions presented their side to issues concerning fairness and the unintended consequences of such policies. In New Mexico, issues regarding teacher evaluation played a role in gubernatorial elections, indicating the potential real-world impact of the arguments made in court and echoed by the media. Again, linking the lawsuits directly to the results of political elections is a stretch. Too many factors play a role to clearly present a link; however, the lawsuits certainly presented political points of contention or "political stumbling blocks" (Sawchuk, 2016, p. 1).

That all of these court cases fundamentally rested upon not only a variety of educational research studies (i.e., hundreds of peer-reviewed study citations across documents) but also educational measurement concepts afforded by the *Standards* (AERA et al., 2014), given all of the cases addressed in this study attended to all of these measurement concepts to some degree, is important to note, although it is also important to note that just because VAMs might not satisfy these *Standards*, VAMs may still not necessarily be considered illegitimate or illegal (e.g., in the case of Tennessee where the court ruled that, regardless of the measurement issues, using VAMs for teacher evaluation purposes was "rational"). Although the *Standards* may serve as a useful touchstone for understanding cases involving such educational measurement issues, they are not the only important touchstone for understanding VAMs in court.

Nonetheless, that these legal arguments represented hundreds of hours of work by plaintiffs and defendants, including not only lawyers but also expert witnesses representing the academic community, is also noteworthy.[3] This is true given the educational research community, as a whole, is currently engaged with how educational research might have a direct and measurable impact on policy and practice writ large (Deshpande, 1981; Fischman & Tefara, 2014; Lingard, 2013; Nutley, Walter, & Davies, 2003). See also the themes from the 2016 and 2017 AERA annual meetings that speak to this regarding "Public Scholarship" and "Knowledge to Action" (see also Welner, 2012). These also speak to how the academy is seeking to rework models for how research might affect policy and practice.

Therefore, while much discussion has transpired regarding the extent to which researchers might make their empirical findings more impactful (Furlough, 2010; M. F. Green, 2000; Shulenburger, 2005), especially in terms of policy and practice, in few legal cases prior (e.g., in desegregation law; see, for example, Erickson & Simon,1998) and of late (e.g., teacher tenure in *Vergara v. California*, as related; school choice in, for example, Prothero, 2018; and school finance in, for example, Farmer, 2017) has the education academy seen its influence affect change more than via these legal cases.

While the adversarial processes upon which courts rely may inspire debate as to whether those processes are well suited to identify reliable and valid evidence to inform court decisions, these cases demonstrate how America's judicial system uses and interprets this type of evidence and how those interpretations affect educational policy and practice as a result. See again what transpired in Houston with the Fourteenth Amendment due process ruling, in New Mexico in terms of withholding all consequences until evidence warrants their administration, and in New York in terms of the "arbitrary and capricious" ruling and subsequent definition as such actions being "taken without sound basis in reason or regard to the facts." At minimum, these rulings show how research evidence "added value" in these particular cases.

Relatedly, the ways via which plaintiffs and defendants framed these critical measurement issues is important, also in terms of how these issues directly affect teachers (i.e., the positive and negative consequences attached) and school systems (i.e., forcing administrators to manipulate numbers, losing teachers from the field). Although the effects varied to some extent, all of the cases framed such consequences as major factors to be considered. This may illustrate one major difference between academic and legal work on such teacher evaluation systems, with most academic work in this area focusing more heavily on the measurement issues, oft-regardless of how they play out from theory to practice. Although some scholars consider both measurement issues and on-the-ground consequences in their scholarly work, the court valued these issues, perhaps, more so, more holistically, and more on account of the general good (e.g., taxpayers paying for these systems using federal and state revenues).

Understanding the value of on-the-ground consequences is germane to our collective understandings about this area of research and reform, and how our conceptions of and research on both might evolve. This is especially important as many states continue to employ similar teacher evaluation systems despite the measurements at issue and the pragmatic issues of concern. This also continues despite the fact that, as per the ESSA, states are no longer required (or incentivized) to engage in these policies (see also Kraft & Gilmour, 2017). However, while the passage of ESSA no longer mandates such systems, because states' systems were enacted into state law, these systems will likely continue to remain for some time, given the financial and human resources investments already made and given the time it might take states to revise, adopt, and then implement new policies in effect. Related, current evidence suggests that most of state's current teacher evaluation policies continue to look much the same as they did prior to the passage of ESSA (Close et al., 2018; Excel*in*Ed, 2017).

As such, this is certainly not a call to discontinue such research. This is a call to continue such research, perhaps, in more universal ways. Researchers might better marry theory and practice when researching such systems if they take into consideration these systems' empirical and pragmatic issues, benefits, and challenges. It is precisely this set of research that played the most central role in these cases and, therefore, affected policy and practice. Similarly, additional investigations into what the courts valued, with investigations into what critical evidence was also missing, might better inform future research as well as future research's bearings on policy and practice.

Declaration of Conflicting Interests

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

Funding

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

Notes

1. The main differences between value-added models (VAMs) and growth models are how precisely estimates are made and whether control variables are included. Different from the typical VAM, for example, the student growth models are more simply intended to measure the growth of similarly matched students to make relativistic comparisons about student growth over time, typically without any additional statistical controls (e.g., for student background variables). Students are, rather, directly and deliberately measured against or in reference to the growth levels of their peers, which de facto controls for these other variables.
2. The Education Value-Added Assessment System (EVAAS) is advertised as "the most comprehensive reporting package of value-added metrics available in the educational market." The EVAAS also offers states, districts, and schools "precise, reliable and unbiased results that go far beyond what other simplistic [value-added] models found in the market today can provide" (SAS Institute Inc., n.d.). The EVAAS comes in different versions for different states and for different large and small school districts. For each consumer, EVAAS modelers choose one of two primary linear mixed models. These include either the preferred multivariate response model (MRM), which essentially entails a multivariate repeated-measures analysis of variance (ANOVA) approach, and the less ideal univariate response model (URM) that essentially entails a traditional

analysis of covariance (ANCOVA) approach, which resembles certain hierarchical linear model (HLM) approaches (Sanders, 2003, 2006). The better the test data available, even if taken from different types of standardized achievement tests (e.g., large-scale standardized tests that are aligned to either national or specific state standards), the better the model used and the better (i.e., more valid, reliable, and unbiased) the model estimates (see also Wright, Sanders, & Rivers, 2006; Wright, White, Sanders, & Rivers, 2010).

3. Also noteworthy in and of itself is that the sitting director overseeing all EVAAS and sales at SAS Institute Inc., John White, who has an undergraduate, master's, and PhD in statistics, admitted on the record that he is "not a psychometrician." More importantly, when he was asked whether he was "familiar with the standards for educational and psychological testing," he responded, "I'm familiar enough to know that it's usually psychometric principles that they're dealing with, but other than that, not exactly." The plaintiffs' attorney asked, so "It's something that you, as [an educational] statistician, really have to be concerned with, right?" Answer: "I think that it is more about the psychometrician developing assessments. That's my understanding of them." That the sitting director of this system, one of the most widely implemented and used systems in the nation, was unfamiliar with the *Standards* (AERA et al., 2014) is, accordingly, important to note.

References

Aaronson, D., Barrow, L., & Sanders, W. (2007). Teachers and student achievement in the Chicago Public High Schools. *Journal of Labor Economics*, *25*, 95-135. doi:10.1086/508733

American Educational Research Association. (2000). *Position statement on high-stakes testing in PreK-12 education*. Retrieved from http://www.aera.net/AboutAERA/AERARulesPolicies/AERAPolicyStatements/PositionStatementonHighStakesTesting/tabid/11083/Default.aspx

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

American Educational Research Association Council. (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 Statistical Association. (2014). *ASA statement on using value-added models for educational assessment*. Alexandria, VA: Author. Retrieved from http://www.amstat.org/policy/pdfs/asa_vam_statement.pdf

Amrein-Beardsley, A. (2008). Methodological concerns about the Education Value-Added Assessment System (EVAAS). *Educational Researcher*, *37*(2), 65-75. doi: 10.3102/0013189X08316420

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

Baker, B. D. (2012, December). It's good to be King: More misguided rhetoric on the NY State Eval System. *School Finance 101*. Retrieved from http://schoolfinance101.wordpress.com/2012/12/12/its-good-to-be-king-more-misguided-rhetoric-on-the-ny-state-eval-system/

Baker, B. D., Oluwole, J. O., & Green, P. C. (2013). The legal consequences of mandating high stakes decisions based on low quality information: Teacher evaluation in the Race-to-the-Top era. *Education Policy Analysis Archives*, *21*(5), 1-71. Retrieved from https://epaa.asu.edu/ojs/article/view/1298

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 http://www.epi.org/publications/entry/bp278

Ballou, D. (2005). Value-added assessment: Lessons from Tennessee. In R. W. Lissitz (Ed.), *Value-added models in education: Theory and application* (pp. 272-297). Maple Grove, MN: JAM Press.

Ballou, D., & Springer, M. G. (2015). Using student test scores to measure teacher performance: Some problems in the design and implementation of evaluation systems. *Educational Researcher*, *44*, 77-86. doi:10.3102/0013189X15574904

Bill & Melinda Gates Foundation. (2013, January 8). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study*. Seattle, WA. Retrieved from http://k12education.gatesfoundation.org/resource/ensuring-fair-and-reliable-measures-of-effective-teaching-culminating-findings-from-the-met-projects-three-year-study/

Boeije, H. (2002). A purposeful approach to the constant comparative method in the analysis of qualitative interviews. *Quality & Quantity*, *36*, 391-409.

Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton, NJ: Educational Testing Service. Retrieved from www.ets.org/Media/Research/pdf/PICVAM.pdf

Braun, H. I. (2015). The value in value added depends on the ecology. *Educational Researcher*, *44*, 127-131. doi:10.3102/0013189X15576341

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 1-16). Westport, CT: American Council on Education/Praeger.

Brennan, R. L. (2013). Commentary on "validating interpretations and uses of test scores." *Journal of Educational Measurement*, *50*, 74-83. doi:10.1111/jedm.12001

Brown, C. (2014, July 31). *Stephen Colbert interview with Campbell Brown: The Colbert report*. New York, NY: Comedy Central. Retrieved from http://www.cc.com/video-clips/2mpwlv/the-colbert-report-campbell-brown

Burris, C. C., & Welner, K. G. (2011). *Letter to Secretary of Education Arne Duncan concerning evaluation of teachers and principals*. Boulder, CO: National Education Policy Center. Retrieved from http://nepc.colorado.edu/publication/letter-to-Arne-Duncan

Campbell, D. (1975). Degrees of freedom and the case study. *Comparative Political Studies*, *8*, 178-185.

Capitol Hill Briefing. (2011, September 14). *Getting teacher evaluation right: A challenge for policy makers* (Research in brief). Washington, DC: Dirksen Senate Office Building.

Carey, K. (2017, May). The little-known statistician who taught us to measure teachers. *The New York Times*. Retrieved from https://www.nytimes.com/2017/05/19/upshot/the-little-known-statistician-who-transformed-education.html?_r=0

Chetty, R., Friedman, J., & Rockoff, J. (2014a). Measuring the impacts of Teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*, 2593-2632.

Chetty, R., Friedman, J., & Rockoff, J. (2014b). Measuring the impacts of Teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, *104*, 2633-2679.

Chiang, H., McCullough, M., Lipscomb, S., & Gill, B. (2016). *Can student test scores provide useful measures of school principals' performance?* Washington, DC: U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/pubs/2016002/pdf/2016002.pdf

Chin, M., & Goldhaber, D. (2015). *Exploring explanations for the "weak" relationship between value added and observation-based measures of teacher performance*. Cambridge, MA: Center for Education Policy Research at Harvard University. Retrieved from http://cepr.harvard.edu/files/cepr/files/sree2015_simulation_working_paper.pdf

Collins, C. (2014). Houston, we have a problem: Teachers find no value in the SAS Education Value-Added Assessment System (EVAAS®). *Education Policy Analysis Archives, 22*. Retrieved from http://epaa.asu.edu/ojs/article/view/1594

Corcoran, S. P. (2010). *Can teachers be evaluated by their students' test scores? Should they be? The use of value-added measures of teacher effectiveness in policy and practice*. Providence, RI: Annenberg Institute for School Reform. Retrieved from https://eric.ed.gov/?id=ED522163

Close, K., & Amrein-Beardsley, A. (2018). Learning from what doesn't work in teacher evaluation. *Phi Delta Kappan*, *100*(1), 15-19. Retrieved from http://www.kappanonline.org/learning-from-what-doesnt-work-in-teacher-evaluation/

Close, K., Amrein-Beardsley, A., & Collins, C. (2018). *State-level assessments and teacher evaluation systems after the passage of the every student succeeds act: Some steps in the right direction*. Boulder, CO: Nation Education Policy Center (NEPC). Retrieved from http://nepc.colorado.edu/publication/state-assessment

Darling-Hammond, L. (2015). Can value added add value to teacher evaluation? *Educational Researcher*, *44*, 132-137. doi:10.3102/0013189X15575346

Darling-Hammond, L., & Haertel, E. (2012, November 5). A better way to grade teachers. *Los Angeles Times*. Retrieved from https://www.latimes.com/opinion/la-xpm-2012-nov-05-la-oe-darling-teacher-evaluations-20121105-story.html

Darville, S. (2017, October 19). Gates foundation to move away from teacher evals, shifting attention to "networks" of public schools. *Chalkbeat*. Retrieved from https://www.chalkbeat.org/posts/us/2017/10/19/gates-foundation-to-move-away-from-teacher-evals-shifting-attention-to-networks-of-public-schools/

Deshpande, R. (1981). Action and enlightenment functions of research: Comparing private-and public-sector perspectives. *Knowledge*, *2*, 317-330. doi:10.1177/107554708100200302

Doherty, K. M., & Jacobs, S. (2015). *State of the states 2015: Evaluating teaching, leading and learning*. Washington, DC: National Council on Teacher Quality. Retrieved from http://www.nctq.org/dmsView/StateofStates2015

Dorans, N. J., & Cook, L. L. (2016). *Fairness in educational assessment and measurement*. New York, NY: Routledge.

Duncan, A. (2009, July 4). *The race to the top begins: Remarks by Secretary Arne Duncan*. Retrieved from https://www.ed.gov/news/speeches/race-top-begins

Duncan, A. (2011, March 9). *Winning the future with education: Responsibility, reform and results: Testimony given to the U.S. Congress*. Washington, DC. Retrieved from http://www.ed.gov/news/speeches/winning-future-education-responsibility-reform-and-results

Eckert, J. M., & Dabrowski, J. (2010, May). Should value-added measures be used for performance pay? *Phi Delta Kappan*, *91*(8), 88-92.

Erickson, R. J., & Simon, R. J. (1998). *The use of social science data in Supreme Court decisions*. Champaign: University of Illinois Press.

Excel*in*Ed. (2017). *ESSA state plans: 50-state landscape analysis*. Tallahassee, FL. Retrieved from https://www.excelined.org/wp-content/uploads/2017/12/ExcelinEd.Quality.ESSA_.50StateAnalysis.Dec072017.pdf

Farmer, L. (2017, December). In school funding court battles, there's been a winning shift. *Governing*. Retrieved from http://www.governing.com/topics/finance/gov-winning-shift-school-funding-court-battles.html

Fischman, G., & Tefara, A. (2014). If the research is not used, does it exist? *Teachers College Record*, *17570*, 1-10.

Flyvbjerg, B. (2011). Five misunderstandings about case-study research. *Qualitative Inquiry*, *12*, 219-245. doi:10.1177/1077800405284363

Furlough, M. (2010). Open access, education research, and discovery. *Teachers College Record*, *112*, 2623-2648. Retrieved from http://www.tcrecord.org/Content.asp?ContentId=15874

Gabriel, R., & Lester, J. N. (2013). Sentinels guarding the grail: Value-added measurement and the quest for education reform. *Education Policy Analysis Archives*, *21*(9), 1-30. Retrieved from http://epaa.asu.edu/ojs/article/view/1165

Gale, N. K., Heath, G., Cameron, E., Rashid, S., & Redwood, S. (2013). Using the framework method for the analysis of qualitative data in multi-disciplinary health research. *BMC Medical Research Methodology, 13*, Article 117. doi:10.1186/1471-2288-13-117

Gates, B. (2013, April 3). Bill Gates: A fairer way to evaluate teachers. *The Washington Post*. Retrieved from https://www.washingtonpost.com/opinions/

bill-gates-a-fairer-way-to-evaluate-teachers/2013/04/03/c99fd1bc-98c2-11e2-814b-063623d80a60_story.html

Gerring, J. (2004). What is a case study and what is it good for? *American Political Science Review*, *98*, 341-354. doi:10.1017/S0003055404001182

Goldhaber, D. (2015). Exploring the potential of value-added performance measures to affect the quality of the teacher workforce. *Educational Researcher*, *44*, 87-95. doi:10.3102/0013189X15574905

Goldhaber, D., & Chaplin, D. D. (2015). Assessing the "Rothstein Falsification Test": Does it really show teacher value-added models are biased? *Journal of Research on Educational Effectiveness*, *8*, 8-34. doi:10.1080/19345747.2014.978059

Goldring, E., Grissom, J. A., Rubin, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make room value added: Principals' human capital decisions and the emergence of teacher observation data. *Educational Researcher*, *44*, 96-104. doi:10.3102/0013189X15575031

Graue, M. E., Delaney, K. K., & Karch, A. S. (2013). Ecologies of education quality. *Education Policy Analysis Archives*, *21*(8), 1-36. Retrieved from http://epaa.asu.edu/ojs/article/view/1163

Green, M. F. (2000). Bridging the gap: Multiple players, multiple approaches. *New Directions for Higher Education*, *110*, 107-113.

Green, P. C., Baker, B. D., & Oluwole, J. (2012). The legal and policy implication of value-added teacher assessment policies. *Brigham Young University Education and Law Journal*, *2012*, 1-29.

Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher*, *43*, 293-303. doi:10.3102/0013189X14544542

Guarino, C. M., Reckase, M. D., Stacy, B. W., & Wooldridge, J. M. (2014). *Evaluating specification tests in the context of value-added estimation*. East Lansing: The Education Policy Center at Michigan State University.

Hanushek, E. (2009). Teacher deselection. In D. Goldhaber & J. Hannaway (Eds.), *Creating a new teaching profession* (pp. 165-180). Washington, DC: Urban Institute Press.

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

Harris, D. N., & Herrington, C. D. (2015). Editors' introduction: The use of teacher value-added measures in schools: New evidence, unanswered questions, and future prospects. *Educational Researcher*, *44*, 71-76. doi:10.3102/0013189X15576142

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, *48*, 794-831. doi:10.3102/0002831210387916

Holloway-Libell, J. (2015). *Evidence of grade and subject-level bias in value-added measures*. Teachers College Record. Retrieved from http://www.tcrecord.org/Content.asp?ContentID=17987

Huffman, K. (2014). *Kevin Huffman at TEDxNashville* [Video file]. Retrieved from https://www.youtube.com/watch?v=8IznMHnRH5c

Jiang, J. Y., Sporte, S. E., & Luppescu, S. (2015). Teacher perspectives on evaluation reform: Chicago's REACH students. *Educational Researcher*, *44*, 105-116. doi: 10.3102/0013189X15575517

Johnson, S. M. (2015). Will VAMS reinforce the walls of the egg-crate school? *Educational Researcher*, *44*, 117-126. doi:10.3102/0013189X15573351

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Washington, DC: National Council on Measurement in Education & American Council on Education.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1-73. doi:10.1111/jedm.12000

Kane, T. J. (2015, April). Teachers must look in the mirror. *New York Daily News*. Retrieved from http://www.nydailynews.com/opinion/thomas-kane-teachers-mirror-article-*1*.2172662

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment* (MET project). Bill & Melinda Gates Foundation. Retrieved from https://files.eric.ed.gov/fulltext/ED540959.pdf

Kane, T. J., & Staiger, D. (2012). *Gathering feedback on teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://k12education.gatesfoundation.org/resource/gathering-feedback-on-teaching-combining-high-quality-observations-with-student-surveys-and-achievement-gains-3/

Kappler Hewitt, K. (2015). Educator evaluation policy that incorporates EVAAS value-added measures: Undermined intentions and exacerbated inequities. *Education Policy Analysis Archives*, *23*(76), 1-49. Retrieved from http://epaa.asu.edu/ojs/article/view/1968

Koedel, C., & Betts, J. R. (2007, April). *Re-examining the role of teacher quality in the educational production function* (Working Paper No. 07-08). University of Missouri. Retrieved from https://economics.missouri.edu/working-papers/2007/wp0708_koedel.pdf

Koedel, C., & Betts, J. R. (2009, July). *Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique* (Working Paper No. 2009-01). Nashville, TN: National Center on Performance Incentives.

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

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting the Widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher*, *46*, 234-249. doi:10.3102/0013189X17718797

Lavery, M. R., Amrein-Beardsley, A., Pivovarova, M., Holloway, J., Geiger, T., & Hahs-Vaughn, D. L. (2019). *Do value-added models (VAMs) tell truth about teachers? Analyzing validity evidence from VAM scholars*. Paper presented at the Annual Meeting of the American Educational Research Association (AERA), Toronto, Canada.

Lingard, B. (2013). The impact of research on education policy in an era of evidence-based policy. *Critical Studies in Education*, *54*, 113-131. doi:10.1080/17508487.2013.781515

Martinez, J. F., Schweig, J., & Goldschmidt, P. (2016). Approaches for combining multiple measures of teacher performance: Reliability, validity, and implications for evaluation policy. *Educational Evaluation and Policy Analysis*, *38*, 738-756. doi:10.3102/0162373716666166

Mathematica Policy Research. (2014). *Alignment of state teacher evaluation policies with Race to the Top priorities*. Washington, DC. Retrieved from https://www.mathematica-mpr.com/news/alignment-of-state-teacher-evaluation-policies-with-race-to-the-top-priorities

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*, 67-101. Retrieved from www.rand.org/pubs/reprints/2005/RAND_RP1165.pdf

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy*, *4*, 572-606. doi:10.1162/edfp.2009.4.4.572

Mellon, E. (2010, January 14). HISD moves ahead on dismissal policy: In the past, teachers were rarely let go over poor performance, data show. *Houston Chronicle*. Retrieved from http://www.chron.com/disp/story.mpl/metropolitan/6816752.html

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, *30*, 955-966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*, 1012-1027.

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

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741-749.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: A sourcebook*. Beverly Hills, CA: SAGE.

National Association of Secondary School Principals. (n.d.). *Value-added measures in teacher evaluation: Position statement*. Reston, VA: NASSP Board of Directors. Retrieved from https://www.nassp.org/who-we-are/board-of-directors/position-statements/value-added-measures-in-teacher-evaluation?SSO=true

The New Teacher Project. (2012). *The irreplaceables: Understanding the real retention crisis in America's urban school*. Brooklyn, NY: Author. Retrieved from http://tntp.org/assets/documents/TNTP_Irreplaceables_2012.pdf

Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis Archives*, *18*(23), 1-27. Retrieved from http://epaa.asu.edu/ojs/article/view/810

Nutley, S., Walter, I., & Davies, H. T. (2003). From knowing to doing: A framework for understanding the evidence-into-practice agenda. *Evaluation*, *9*, 125-148. doi:10.1177/1356389003009002002

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal*, *48*, 163-193. doi:10.3102/0002831210362589

Polikoff, M. S., & Porter, A. C. (2014). Instructional alignment as a measure of teaching quality. *Educational Evaluation and Policy Analysis*, *36*, 399-416. doi:10.3102/0162373714531851

Prothero, A. (2018, May). Fate of Washington's charter school law to be decided, again, by state supreme court. *Education Week*. Retrieved from http://blogs.edweek.org/edweek/charterschoice/2018/05/fate_of_washingtons_charter_school_law_to_be_decided_again_by_state_supreme_court.html

Pullin, D. (2013). Legal issues in the use of student test scores and value-added models (VAM) to determine educational quality. *Education Policy Analysis Archives*, *21*(6), 1-27. doi:10.14507/epaa.v21n6.2013

Pullin, D. (2014). Professional testing standards in the eyes of the law. *Educational Measurement: Issues and Practice*, *33*(4), 19-21.

Pullin, D. (2015). Performance measures for teachers and teacher education: Corporate education reform opens the door to new legal issues. *Education Policy Analysis Archives*, *23*(81), 1-36. Retrieved from https://epaa.asu.edu/ojs/article/view/1980

Race to the Top Act of 2011, S. 844–112th Congress. (2011). Retrieved from http://www.govtrack.us/congress/bills/112/s844

Ragin, C. C., & Becker, H. S. (2000). Cases of "what is a case?" In C. C. Ragin & H. S Becker (Eds.), *What is a case? Exploring the foundations of social inquiry* (pp. 1-17). Cambridge, UK: Cambridge University Press.

Rhee, M. (2011, April 6). The evidence is clear: Test scores must accurately reflect students' learning. *The Huffington Post*. Retrieved from http://www.huffingtonpost.com/michelle-rhee/michelle-rhee-dc-schools_b_845286.html

Ritchie, J., Lewis, J., Nicholls, C. M., & Ormston, R. (Eds.). (2013). *Qualitative research practice: A guide for social science students and researchers*. Los Angeles, CA: SAGE.

Rosenbaum, P., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41-55. doi:10.2307/2335942

Rothstein, J. (2009). Student sorting and bias in value added estimation: Selection on observables and unobservables. *Education Finance and Policy*, *4*, 537-571. doi:10.3386/w14666

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, *125*, 175-214. doi:10.1162/qjec.2010.125.1.175

Rothstein, J. (2014). *Revisiting the impacts of teachers* (Working paper). Berkeley: University of California, Berkeley.

Rothstein, J., & Mathis, W. J. (2013, January). *Review of two culminating reports from the MET project*. Boulder, CO: National Education Policy Center. Retrieved from http://nepc.colorado.edu/thinktank/review-MET-final-2013

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*, *29*, 103-116. doi:10.3102/10769986029001103

Rubinstein, G. (2014, July). Pay no attention to the falling Tennessee reading test scores. *Gary Rubinstein's Blog*. Retrieved from https://garyrubinstein.wordpress.com/2014/07/04/pay-no-attention-to-the-falling-tennessee-reading-test-scores/

Sanders, W. L. (2003, April). *Beyond "No Child Left Behind."* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Sanders, W. L. (2006, October). *Comparisons among various educational assessment value-added models*. Paper presented at The Power of Two—National Value-Added Conference, Columbus, OH.

Sanders, W. L., Wright, S. P., Rivers, J. C., & Leandro, J. G. (2009, November). *A response to criticisms of SAS® EVAAS*. Cary, NC: SAS Institute Inc. Retrieved from http://www.sas.com/resources/asset/Response_to_Criticisms_of_SAS_EVAAS_11-13-09.pdf

SAS Institute Inc. (n.d.). *SAS® EVAAS® for K-12: Assess and predict student performance with precision and reliability*. Cary, NC: Author. Retrieved from https://www.sas.com/en_us/software/evaas.html

Sawchuk, S. (2016, January). ESSA loosens reins on teacher evaluations, qualifications. *Education Week*, *35*(15), 14-15.

Schochet, P. Z., & Chiang, H. S. (2013). 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

Shaw, L. H., & Bovaird, J. A. (2011, April). T*he impact of latent variable outcomes on value-added models of intervention efficacy*. Paper presented at the Annual Conference of the American Educational Research Association, New Orleans, LA.

Shulenburger, D. E. (2005). Public goods and open access. *New Review of Information Networking*, *11*, 3-11. doi:10.1080/13614570500268282

Stake, R. E. (1978). The case study method in social inquiry. *Educational Researcher*, *7*(2), 5-8.

Stake, R. E., & Trumbull, D. (1982). Naturalistic generalizations. *Review Journal of Philosophy and Social Science*, *7*, 1-12.

State of New York Supreme Court. (2016). *Decision, order and judgment*. Roger D. McDonough, Justice. Retrieved from http://vamboozled.com/wp-content/uploads/2016/05/ruling.pdf

Strauss, V. (2014, October). High-achieving teacher sues state over evaluation labeling her "ineffective." *The Washington Post*. Retrieved from https://www.washingtonpost.com/news/answer-sheet/wp/2014/10/31/high-achieving-teacher-sues-state-over-evaluation-labeling-her-ineffective/?utm_term=.2ff667771a76

Superfine, B. M. (2016). Judging teacher evaluation: The legal implications of high-stakes teacher evaluation policy. In K. Kappler Hewitt & A. Amrein-Beardsle (Eds.), *Student growth measures in policy and practice: Intended and unintended* *consequences of high-stakes teacher evaluations* (pp. 223-243). Basingstoke, UK: Palgrave Macmillan.

Swedian, J. (2014, September). Statistical guru for evaluations leaving PED. *Albuquerque Journal*. Retrieved from https://www.abqjournal.com/463424/statistical-guru-for-evaluations-leaving-ped.html

Teacher evaluation heads to the courts. (2015, October). *Education Week*. Retrieved from http://www.edweek.org/ew/section/multimedia/teacher-evaluation-heads-to-the-courts.html

Tennessee State Board of Education (TSBE). (2012). *Teacher and principal evaluation policy*. Nashville, TN. Retrieved from https://www.tn.gov/assets/entities/sbe/attachments/7-27-12-II_C_Teacher_and_Principal_Evaluation_Revised.pdf

Thomas, G. (2011). A typology for the case study in social science following a review of definition, discourse, and structure. *Qualitative Inquiry*, *17*, 511-521. doi:10.1177/1077800411409884

U.S. Department of Education. (2006a, November 9). *Secretary spellings approves additional growth model pilots for 2006-2007*. Retrieved from https://www.webwire.com/ViewPressRel.asp?aId=37501

U.S. Department of Education. (2006b, May 17). *Secretary spellings approves Tennessee and North Carolina growth model pilots for 2005-2006*. Retrieved from https://votesmart.org/public-statement/174269/secretary-spellings-approves-tennessee-and-north-carolina-growth-model-pilots-for-2005-2006%23.U2kVosf94a8#.XJ4ehJgza1s

U.S. Department of Education. (2010a). *A blueprint for reform: The reauthorization of the Elementary and Secondary Education Act*. Retrieved from https://www2.ed.gov/policy/elsec/leg/blueprint/blueprint.pdf

U.S. Department of Education. (2010b). *Delaware and Tennessee win first Race to the Top grants*. Retrieved from https://www.ed.gov/news/press-releases/delaware-and-tennessee-win-first-race-top-grants

U.S. Department of Education. (2014). *States granted waivers from No Child Left Behind allowed to reapply for renewal for 2014 and 2015 school years*. Washington, DC. Retrieved from http://www.ed.gov/news/press-releases/states-granted-waivers-no-child-left-behind-allowed-reapply-renewal-2014-and-2015-school-years

U.S. Department of Education. (2015). *Title I waivers*. Washington, DC. Retrieved from https://www2.ed.gov/nclb/freedom/local/flexibility/waiverletters2009/index.html

VAMboozled. (2014a, April). Another lawsuit in Tennessee. Retrieved from http://vamboozled.com/another-lawsuit-in-tennessee/

VAMboozled. (2014b, May). Breaking news: Houston teachers suing over their district's EVAAS use. Retrieved from http://vamboozled.com/breaking-news-houston-teachers-suing-over-their-districts-evaas-use/

VAMboozled. (2014c, May). Florida's VAM-based evaluation system ruled "unfair but not unconstitutional." Retrieved from http://vamboozled.com/floridas-vam-based-evaluation-system-ruled-unfair-but-not-unconstitutional/

VAMboozled. (2018, June). New Mexico teacher evaluation lawsuit updates. Retrieved from http://vamboozled.com/new-mexico-teacher-evaluation-lawsuit-updates/

VanWynsberghe, R., & Khan, S. (2007). Redefining case study. *International Journal of Qualitative Methods*, *6*, 80-94.

Wainer, H. (2004). Introduction to a special issue of the Journal of Educational and Behavioral Statistics on value-added assessment. *Journal of Educational and Behavioral Statistics*, *29*, 1-3. doi:10.3102/10769986029001001

Wallace, T. L., Kelcey, B., & Ruzek, E. (2016). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the tripod student perception survey. *American Educational Research Journal*, *53*, 1834-1868. doi:10.3102/0002831216671864

Walsh, K., Joseph, N., Lakis, K., & Lubell, S. (2017). *Running in place: How new teacher evaluations fail to live up to promises*. Washington, DC: National Council on Teacher Quality. Retrieved from http://www.nctq.org/dmsView/Final_Evaluation_Paper

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The Widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. New York, NY: The New Teacher Project. Retrieved from http://tntp.org/assets/documents/TheWidgetEffect_2nd_ed.pdf

Welner, K. G. (2012). Scholars as policy actors: Research, public discourse, and the zone of judicial constraints. *American Educational Research Journal*, *49*, 7-29. doi:10.3102/0002831211415253

Wright, S. P., Sanders, W. L., & Rivers, J. C. (2006). Measurement of academic growth of individual students toward variable and meaningful academic standards. In R. Lissitz (Ed.), *Longitudinal and value added models of student performance* (pp. 385-406). Maple Grove, MN: JAM Press.

Wright, S. P., White, J. T., Sanders, W. L., & Rivers, J. C. (2010, March 25). *SAS® EVAAS® statistical models* Cary, NC: SAS Institute Inc.

Yeh, S. S. (2013). A re-analysis of the effects of teacher replacement using value-added modeling. *Teachers College Record*, *115*(12), 1-35. Retrieved from http://www.tcrecord.org/Content.asp?ContentID=16934

Author Biographies

Audrey Amrein-Beardsley, PhD, is a Professor in the Mary Lou Fulton Teachers College at Arizona State University. Her research focuses on the use of value-added models (VAMs) in and across states before and since the passage of the Every Student Succeeds Act (ESSA). More specifically, she is conducting validation studies on multiple system components, as well as serving as an expert witness in many legal cases surrounding the (mis)use of VAM-based output.

Kevin Close is currently pursuing a PhD in the Learning, Literacies, and Technologies program at Arizona State University. His research focused on digital adaptive assessments and nation-wide teacher evaluation systems based on high-stakes tests. His interests lie in using technology to change the way we assess and measure progress.