Educational Assessment, Evaluation and Accountability (2020) 32:187–235
https://doi.org/10.1007/s11092-020-09324-w

Using test scores to evaluate and hold school teachers accountable in New Mexico

Tray J. Geiger[1] · Audrey Amrein-Beardsley[1] · Jessica Holloway[2]

Received: 9 September 2019 / Accepted: 20 May 2020 / Published online: 12 June 2020

Abstract

For this study, researchers critically reviewed documents pertaining to the highest profile of the 15 teacher evaluation lawsuits that occurred throughout the U.S. as pertaining to the use of student test scores to evaluate teachers. In New Mexico, teacher plaintiffs contested how they were being evaluated and held accountable using a homegrown value-added model (VAM) to hold them accountable for their students' test scores. Researchers examined court documents using six key measurement concepts (i.e., reliability, validity [i.e., convergent-related evidence], potential for bias, fairness, transparency, and consequential validity) defined by the *Standards for Educational and Psychological Testing* and found evidence of issues within both the court documents as well as the statistical analyses researchers conducted on the first three measurement concepts (i.e., reliability, validity [i.e., convergent-related evidence], and potential for bias).

Keywords Bias · Convergent-related evidence · Educational policy · Reliability · Teacher accountability · Teacher evaluation · Validity · Value-added models

1 Introduction

Summarized in an article published in the United States (U.S.), in its national popular press outlet Education Week (2015), as of 2015 there were 15 lawsuits throughout the

✉ Audrey Amrein-Beardsley
audrey.beardsley@asu.edu

Tray J. Geiger
tjgeiger@asu.edu

Jessica Holloway
jessica.holloway@deakin.edu.au

[1] Arizona State University, Tempe, AZ, USA

[2] Australian Research Council, Research for Educational Impact (REDI) Centre, Deakin University, Geelong, Australia

U.S. in which teacher plaintiffs were contesting how they were being evaluated and held accountable using their students' standardized test scores. These 15 cases were located within seven states: Florida ($n = 2$), Louisiana ($n = 1$), Nevada ($n = 1$), New Mexico ($n = 4$), New York ($n = 3$), Tennessee ($n = 3$), and Texas ($n = 1$). Teacher plaintiffs across cases were contesting the high-stakes consequences attached to their alleged impacts on their students' test scores over time, including but not limited to merit-pay in Florida, Louisiana, and Tennessee; teacher tenure decisions in Louisiana; teacher termination in Houston, Texas, and Nevada; and other "unfair penalties" in New York. To measure teachers' impacts on their students' achievement over time, the education metric of choice, and of issue across cases, was the value-added model (VAM).

In the simplest of terms, VAMs and growth models (hereafter referred to more generally as VAMs[1]) help to statistically measure and then classify teachers' levels of effectiveness according to teachers' purportedly causal impacts on their students' achievement over time. VAM modelers typically calculate teacher effects by measuring student growth over time on standardized tests (e.g., the tests mandated throughout the U.S. by the federal No Child Left Behind [NCLB] Act 2001), and then aggregating this growth at the teacher-level while statistically controlling for confounding variables such as students' prior test scores and other student-level (e.g., free-and-reduced lunch [FRL], English language learner [ELL], special education [SPED]) and school-level variables (e.g., class size, school resources), although control variables vary by model. Teachers whose students collectively outperform students' projected levels of growth (i.e., typically estimated 1 year prior) are to be identified as teachers of "added value," and teachers whose students fall short are to be identified as teachers of the inverse (e.g., teachers not of "added value").

Given the statistical sophistication VAMs were to bring to the objective evaluation of teachers' effects, and stronger accountability policies and initiatives in the name of U.S. educational reform (see, for example, Collins 2014; Eckert and Dabrowski 2010; Kappler Hewitt 2015), VAMs were incentivized by former U.S. President Obama's Race to the Top Competition (2011). Via Race to the Top, the U.S. government offered states $4.35 billion in federal support (upon which all U.S. states historically rely), on the condition that states would attach teachers' evaluations to students' test scores using a VAM-based system. States' VAM implementations were also underscored via the U.S. congressional authorization of states' NCLB waivers (U.S. Department of Education 2014), excusing states who implemented a VAM-based teacher evaluation system from not needing to meet NCLB's former goal that required that 100% of students across states to reach 100% proficiency by 2014 (U.S. Department of Education 2010). Consequently, as a result of both of these federal policy initiatives, almost all 50 states plus Washington D.C. constructed or adopted and then implemented a VAM-based teacher evaluation system by 2014 (Paufler and Amrein-Beardsley 2014).

[1] The main differences between VAMs and growth models are how precisely estimates are made and whether control variables are included. Different than 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.

While the federal government has since reduced the strictness of its aforementioned policy-based footholds, namely through the passage of its Every Student Succeeds Act (ESSA 2016) which reduced federal mandates surrounding VAMs, it is worth noting that many states and school districts continue to use VAMs for high-stakes employment decisions. More specifically, while the federal passage of ESSA allowed for greater local control over states' and school districts' teacher evaluation systems and no longer required states and school districts to rely on VAMs as a measure meant to "meaningfully differentiate [teacher] performance… including as a significant factor, data on student growth [in achievement over time] for all students" (U.S. Department of Education 2012), current evidence indicates that many states and school districts still continue to use VAMs state- and district-wide (Close et at. 2020; Ross and Walsh 2019). For example, 12 states still allow or encourage districts to make teacher termination decisions as solely or primarily based on their VAM data, and 23 states allow or encourage such decisions as per a combination of districts' VAM and other evaluative (e.g., observational) data (Close et al. 2020). Given the human and financial capital that states had already invested in their use, it stands to reason that policy inertia is providing for the continued use of VAMs, also in the New Mexico case at hand.

Likewise, the theory of change at the time (and still ongoing; see Koretz 2017) is that by objectively holding teachers accountable for that which they did or did not do effectively, in terms of the "value" they did or did not "add" to their students' achievement over time, teachers would be incentivized to teach more effectively and students would consequently learn and achieve more. If high stakes were attached to teachers' students' test output (e.g., teacher pay, tenure, termination), teachers would take their teaching and their students' learning and achievement more seriously. This, along with the data that VAMs were to also provide to help teachers improve upon their practice (i.e., the formative functions of VAMs), would ultimately help improve achievement throughout the U.S., especially in the disadvantaged schools most in need of educational reform, all of which would help the U.S. reclaim its global superiority (see, for example, Weisberg et al. 2009).

It should be noted here, though, that this theory of change, along with the use of VAMs to help satisfy it, was not isolated to the U.S. In his 2011 book *Finnish Lessons*, for example, Sahlberg coined the *Global Educational Reform Movement* (GERM) acronym which captured other countries' (e.g., Australia, England, Korea, Japan) similar policy movements to accomplish similar goals and outcomes. In short, GERM "radically altered education sectors throughout the world with an agenda of evidence-based policy based on the [same] school effectiveness paradigm…combin[ing] the centralised formulation of objectives and standards, and [the] monitoring of data, with the decentralisation to schools concerning decisions around how they seek to meet standards and maximise performance" (p. 5).

Likewise, while the U.S. was leading other nations in terms of its policy-backed and funded initiatives as based on VAMs, other countries (e.g., Chile, Ecuador, Denmark, England, Sweden) continue to entertain similar policy ideas. Put differently, "in the U.S. the use of VAM as a policy instrument to evaluate schools and teachers has been taken *exceptionally far* [emphasis added] in the last 5 years," while "most other high-income countries remain [relatively more] cautious towards the use of VAM[s]" (Sørensen 2016, p. 1). Although with the support of global bodies such as the Organisation for Economic Co-operation and Development (OECD), VAMs continue to be adopted worldwide for purposes and uses similar to those in the U.S. (see also Araujo et al. 2016).

Notwithstanding, because the U.S. took VAMs and VAM-based educational reform policies "exceptionally far" (Sørensen 2016, p. 1), what we have learned from states' uses of VAMs can and should be understood by others across nations considering whether to adopt or implement VAMs for similar purposes. In the U.S., because VAMs were also literally that which landed the above states (and districts within states) in court, this manuscript should be of interest to others both nationally and internationally. It is prudent that others within and beyond U.S. borders pay attention to that which happened within and across these cases, and what eventually led to the "lots of litigation" filed throughout U.S. courts in these regards by 2015 (Education Week 2015).

2 Purpose of the study

Given Education Week (2015) presents a broad, case-by-case description of the aforementioned 15 cases, for this study researchers purposefully selected one of these 15 cases to help others: (1) better understand the measurement and pragmatic issues at play, in some way or another, across cases in that the issues generalize across cases, and likely states, especially given their ongoing use of VAMs (Amrein-Beardsley and Geiger 2019; Close et al. 2020); and (2) better understand these issues in context, in the case of the largest and arguably most high-profile, controversial, and consequential cases of the set. The case in point occurred in New Mexico, with consequences to be attached to teachers' VAM scores including but not limited to the flagging of teachers' professional files if determined to be not of "added value," which ultimately prevented teachers from moving teaching positions within the state given their official "ineffective" classifications. Also at issue were teacher termination policies attached to New Mexico teachers' VAM scores.

3 The case of New Mexico

During the 2013–2014 through 2015–2016 school years, New Mexico's teacher evaluation system, the NMTEACH Educator Effectiveness System (EES), was seen by many as one of the toughest across the U.S. (Burgess 2017; Kraft and Gilmour 2017; see also Amrein-Beardsley and Geiger 2019). Not only was student growth data the preponderant criterion informing a teacher's overall evaluation score, but the overall distributions of teacher effectiveness ratings across the state were nearly normal for these 3 years. That is, the majority of teachers were rated as effective; fewer and nearly equal proportions of teachers were rated as slightly worse and slightly better than effective, respectively; and even fewer and nearly equal proportions of teachers were rated as much worse and much better than effective, respectively (see Fig. 1).

Unlike other states (e.g., North Carolina, Tennessee) that contracted with companies (e.g., SAS Institute Inc. 2019) that sold VAMs (e.g., the EVAAS), the New Mexico VAM was a homegrown model created by Pete Goldschmidt (see Martinez et al. 2016; see also Reiss 2017). Per teacher, a separate VAM score was calculated per subject, per grade, and per standardized assessment. Each teacher's overall VAM score was a

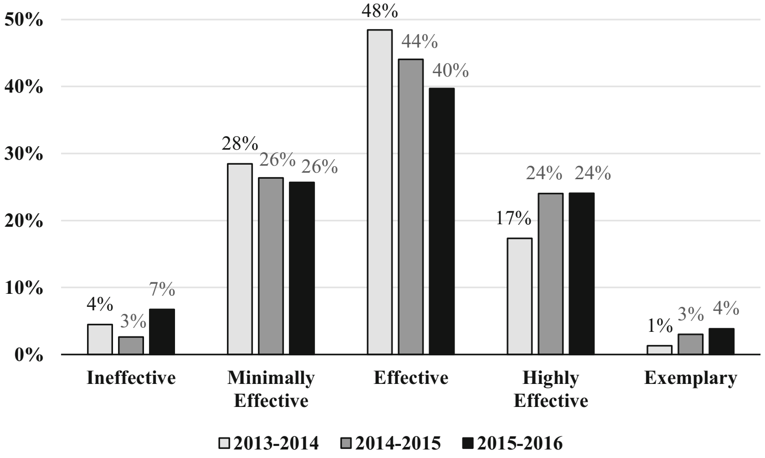

Fig. 1 Distribution of teachers' overall evaluation ratings from the 2013–2014 to 2015–2016 school years

weighted average of all individual VAM scores. The statistical model used to generate each individual VAM score was supposed to control for whether a course had been identified as an "intervention course," the student's grade level (if a course contained students from multiple grades), and the proportion of time a student had spent with each specific teacher (New Mexico Public Education Department [NMPED] 2016; see pp. 14–22 for full model specifics and formulas). A VAM-eligible teacher (i.e., one who taught in a tested subject and grade) had up to half of their overall evaluation based on their VAM score, with the remaining half based on a combination of classroom observations, student surveys, and attendance (NMPED 2016) (see Table 1).

In comparison, other states' evaluation systems either did not rely on student growth data nearly as much, and/or teachers had a much higher likelihood of receiving an evaluation of effective or better. For example, during the 2015–2016 school year, only one third of states in the U.S. (including New Mexico) required that student growth data be the "preponderant criterion" in teachers' overall evaluation scores (Doherty and Jacobs 2015). Further, New Mexico was only one of two states where over 1% of all

Table 1 Components of teachers' evaluation scores from the 2013–2014 to 2015–2016 school years

	Teacher in tested subjects/grades		Teacher not in tested subjects/grades
Years of student achievement data	3+	1–2	0
Evaluation component			
VAS data	50%	25%	–
Classroom observations	25%	40%	50%
PPP	15%	25%	40%
Student/family surveys	5%	5%	5%
Teacher attendance	5%	5%	5%
Total	100%	100%	100%

Adapted from NMPED (2016, p. 6)

teachers received an overall evaluation score in the lowest possible category (Kraft and Gilmour 2017).

In the New Mexico case, which was titled *American Federation of Teachers – New Mexico and the Albuquerque Federation of Teachers (Plaintiffs) v. New Mexico Public Education Department (Defendants)* and was being tried and heard in the state's First Judicial District Court, the primary issue was the use of the state's homegrown VAM (see Swedien 2014), specifically to account for up to 50% of every VAM-eligible teacher's annual evaluation. The specific violations contested were that (1) New Mexico teachers received poor VAM-based ratings because of flawed and incomplete student-level data (e.g., teachers were linked to the wrong students, students they never taught, subject areas they never taught, or using tests that did not map onto that which they taught); and (2) the aforementioned consequential decisions to be attached to all VAM-eligible teachers' VAM-based evaluation scores (e.g., flagging files and teacher termination decisions) were arbitrary and not legally defensible, as per the education profession's *Standards for Educational and Psychological Testing* (American Educational Research Association (AERA), American Psychological Association (APA),, and National Council on Measurement in Education (NCME) 2014), hereafter referred to as the *Standards*.

4 Conceptual framework

In order to investigate the empirical and pragmatic matters addressed in this court case, researchers conducted a case study analysis of all documents, exhibits, and data submitted for this case. Researchers framed their analyses using the key measurement concepts resident within the *Standards* (AERA et al. 2014), more specifically given issues with: (1) reliability, (2) validity (i.e., convergent-related evidence), (3) bias, (4) fairness, and (5) transparency, with emphases also on (6) consequential validity, as per (6a) whether VAMs are being used to make consequential decisions using concrete (e.g., not arbitrary) evidence and (6b) whether VAMs' unintended consequences are also of legal pertinence and concern. Researchers define and describe each of these areas of measurement and pragmatic concern next, also as per the current research literature per concept.

4.1 Reliability

As per the *Standards* (AERA et al. 2014), reliability is defined 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" (p. 222–223) for the individuals (e.g., teachers) to whom the test- or measurement-based scores pertain. In terms of VAMs, reliability (also known as intertemporal stability; see, for example, McCaffrey et al. 2009) should be observed when VAM estimates of teacher effectiveness are more or less consistent over time, from 1 year to the next, regardless of the type of students and perhaps subject areas teachers teach. This 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 both situate and make explicit VAM estimates and their (sometimes sizeable) errors.

Reporting on reliability is also done to make transparent the sometimes sizeable errors that come along with VAM estimates, to better contextualize the VAM-based inferences that result. This is especially critical when VAM outputs are to be attached to high-stakes consequences upon which the stability of these measures over time also rely, because without adequate reliability, valid interpretations and uses are difficult to defend.

What is becoming increasingly evident across research studies in this area is that VAMs are often unreliable, unstable, and sometimes "notoriously" unhinged (Ballou and Springer 2015, p. 78). While "proponents of VAM[s] are quick to point out that any statistical calculation has error of one kind or another" (Gabriel and Lester 2013, p. 4; see also Harris 2011), the errors prevalent across VAMs are consistently large enough to warrant caution, especially before high-stakes consequences are attached to VAM output.

More pragmatically, researchers have found that the possibility of teachers being misclassified (i.e., classified as adding value 1 year and then not adding value 1 to 2 years later) can range from 25% to as high as 59% (Martinez et al. 2016; Schochet and Chiang 2013; Yeh 2013). While reliability can be increased with 3 years of data, there still exists at least a 25% chance that teachers may be misclassified. Additionally, after including 3 years of data, the strength that more data add to VAM reliability plateaus (Brophy 1973; Cody et al. 2010; Glazerman and Potamites 2011; Goldschmidt et al. 2012; Harris 2011; Ishii and Rivkin 2009; Sanders as cited in Gabriel and Lester 2013).

What this means in practice, for example, with a correlation of $r = 0.40$ and $R^2 = 0.16$, which is an optimistic r and R^2 given the current research, is illustrated in Table 2 (adapted with permission and corrections from Raudenbush and Jean 2012).

Illustrated is that out of every 1000 teachers, 750 teachers would be identified correctly and 250 teachers would not. That is, one in four teachers would be falsely identified as either being worse or better than they were originally classified.

Of concern here is the prevalence of false positive or false discovery errors (i.e., type I errors), whereas an ineffective teacher is falsely identified as effective. However, the inverse is equally likely, defined as false negative or false nondiscovery errors (i.e., type II errors), whereas an ineffective teacher might go unnoticed instead. Regardless of which type of error is worse, in that "[f]alsely identifying [effective] teachers as being

Table 2 Correct and incorrect interpretations (possibly leading to correct and incorrect decisions) if $r = 0.4$ and $R^2 = 0.16$

	Teachers truly below the 25th percentile	Teachers truly above the 25th percentile	Total
Teachers estimated to be below the 25th percentile	80 (*teachers correctly identified as below the 25th percentile*)	80 (*teachers falsely identified as below the 25th percentile*)	160
Teachers estimated to be above the 25th percentile	170 (*teachers falsely identified as above the 25th percentile*)	670 (*teachers correctly identified as above the 25th percentile*)	840
Total	250	750	1000

below a threshold poses risk to teachers but failing to identify [ineffective] teachers who are truly ineffective poses risks to students" (Raudenbush and Jean 2012), there still exists one in four teachers who are likely to be falsely identified. This is clearly problematic, especially when consequential decisions are to be tied to VAM output (see also Briggs and Domingue 2011; Chester 2003; Glazerman et al. 2011; Guarino et al. 2012; Harris 2011; Rothstein 2010; Shaw and Bovaird 2011; Yeh 2013).

4.2 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). Likewise, "[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). As per Kane (2017):

> If *validity* is to be used to support a score interpretation, validation would require an analysis of the plausibility of that interpretation. If validity is to be used to support score uses, validation would require an analysis of the appropriateness of the proposed uses, and therefore, would require an analysis of the consequences of the uses. In each case, the evidence need for validation would depend on the specific claims being made. (p. 198)

When establishing evidence of validity, or explicating validity (Kane 2017), accordingly, one must be able to support with evidence that accurate inferences can be drawn from the data being used for whatever inferential purposes are at play (see also Cronbach and Meehl 1955; Kane 2006, 2013).

While there is a "multiplicity of validity vocabularies" (Markus 2016, p. 252), however, as well as multiple forms of validity evidences (e.g., content-related, criterion-related, construct-related, consequential-related) that are used to both accommodate and differentiate validations of test score inferences and justifications of test use (Cizek 2016), most often examined in this area of research is convergent-related evidence of validity (i.e., "the degree of relationship between the test scores and [other] criterion scores" taken at the same time; Messick 1989, p. 7). While current conceptions of validity have evolved well beyond capturing any specific evidences or instances of validity (e.g., convergent-related evidence of validity), especially in isolation of other evidences of validity (e.g., that should be used to capture a more holistic interpretation of validity), it is important to note, again, that researchers examining VAM-related evidences of validity have consistently and disproportionately focused on convergent-related evidences of validity in this area of research. While arguably overly simplistic and reductionistic (see, for example, Newton and Shaw 2016), because these evidences are what exist in this area of research, so much so that other types of validity are rarely mentioned or discussed (e.g., consequential-related evidences of validity), and given the purpose of this study revolves around the court documents pertaining to the New Mexico case in which only these evidences were presented to the court, only evidences of convergent-related validity are discussed and examined in this study.

In New Mexico, accordingly, convergent-related evidences of validity were used to assess the extent to which measures of similar constructs concurred or converged. The construct at issue here was teacher effectiveness, and evidences of convergent-related evidence of validity were positioned to the court as necessary to assess, for example, whether teachers who posted large and small value-added gains or losses over time were the same teachers deemed effective or ineffective, respectively, using other measures of teacher effectiveness (e.g., observational scores, student survey scores) collected at the same time.

In terms of the VAM literature writ large, convergent-related evidences of validity as per the current research suggest that VAM estimates of teacher effectiveness do not strongly correlate with the other measures typically used to measure the teacher effectiveness construct (i.e., observational scores). While some argue that the measures other than VAMs are at fault given their imperfections, others argue that all of the measures, including VAMs, are at fault because they are all in and of themselves imperfect and flawed. Likewise, while some also argue that knowing there is a low correlation between any VAM and any set of observational scores tells us nothing about whether either one, neither, or both measures are useful, others argue that should these indicators be mapped onto a general construct called teaching effectiveness, they should correlate. Should high-stakes decisions be attached to output from either or multiple measures, these correlations must be higher before consequences can be defended.

If the large-scale standardized achievement test scores that contribute to VAM estimates and the other measures typically used to measure teacher effectiveness were all reliable and valid measures of the teaching effectiveness construct, effective teachers would rate well, more or less continuously, from 1 year to the next, across indicators. Conversely, ineffective teachers would rate poorly, more or less consistently, from 1 year to the next, across indicators used. However, this does not occur in reality, except in slight magnitudes, whereby the correlations being observed among both mathematics and English/language arts value-added estimates and teacher observational or student survey indicators are low to moderate[2] in size (Sloat et al. 2018; Grossman et al. 2014; Harris 2011; Hill et al. 2011; see also Koedel et al. 2015). These correlations are also akin to those observed via the renowned Bill & Melinda Gates Foundation's Measures of Effective Teaching (MET) studies in which researchers searched for and assessed the same evidences of convergent-related validity (Cantrell and Kane 2013; Kane and Staiger 2012; see also Polikoff and Porter 2014; Rothstein and Mathis 2013).

4.3 Bias

As per the *Standards* (AERA et al. 2014), bias pertains to the validity of the inferences to be drawn 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 pertaining to "[t]he systematic over- or under-prediction of criterion

[2] As per teachers' contract statuses, HISD teachers new to the district are put on probationary contracts for 1 year if they have more than 5 years of prior teaching experience, or they are put on probationary contracts for 3 years if they are new teachers. Otherwise, teachers are on full contract.

performance" (p. 222), are observed when 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, since schools do not randomly assign teachers the students they teach (Paufler and Amrein-Beardsley 2014), whether teachers' students are invariably more or less motivated, smart, knowledgeable, capable, and the like can bias students' test-based data and teachers' test-based data once aggregated. Bias is subsequently observed when VAM estimates of teacher effectiveness correlate with student background variables other than the indicators of interest (i.e., student achievement). If VAM-based estimates are highly correlated to biasing factors, it becomes impossible to make valid inferences about the causes of student achievement growth and teachers' impacts on said growth, given the factors that bias such indicators and ultimately distort their interpretations (Messick 1989; see also Haladyna and Downing 2004).

Many VAM researchers agree that estimates of teachers who teach smaller classes; disproportionate percentages of gifted students, ELLs, special education (SPED) students, and students who receive free-or-reduced (FRL) lunches; and students retained in grade are adversely impacted by bias (Ballou and Springer 2015; McCaffrey et al. 2009; Newton et al. 2010; Rothstein 2009, 2010, 2017). In perhaps the most influential study on this topic, Rothstein (2009, 2010) illustrated VAM-based bias when he found that students' 5th grade teachers were better predictors of students' 4th grade growth than were the students' 4th grade teachers. While others have certainly called into question Rothstein's work (see, for example, Goldhaber and Chaplin 2015; Guarino et al. 2014; Koedel and Betts 2009), what is known in this regard is that over the past decade VAM-based evidence of bias has been investigated at least 33 times in articles published in top peer-reviewed journals (Lavery et al. 2019). Demonstrated across these articles is that bias is still of great debate, as is whether statistically controlling for bias by using complex statistical approaches to account for nonrandom student assignment makes such biasing effects negligible or "ignorable" (Rosenbaum and Rubin 1983; see also Chetty et al. 2014; Koedel et al. 2015; Rothstein 2017).

4.4 Fairness

The *Standards* define fairness as the impartiality of "test score interpretations for intended use(s) for individuals from *all* [emphasis added] relevant subgroups" (AERA et al. 2014, p. 219). Issues of fairness arise when a test or test-based inference or use impacts some more than others in unfair or prejudiced, yet often consequential ways. Given the recent emphases on issues of fairness, accordingly, the *Standards* have prompted an expanded book focusing on such issues (Dorans and Cook 2016).

The main issue here is that states and districts can only produce VAM-based estimates for approximately 30–40% of all teachers (Baker et al. 2013; Gabriel and Lester 2013; Harris 2011). The other 60–70%, which sometimes includes 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 relatively small handful of teachers (Harris and

Herrington 2015; Jiang et al. 2015; Papay 2010). When stakeholders use these data to make consequential decisions, issues with fairness become even more important, whereas some teachers are more likely to realize the negative or positive consequences attached to VAM-based data, simply given the grades and subject areas they teach.

4.5 Transparency

While the *Standards* (AERA et al. 2014) do not explicitly define transparency, this concept pertains to the openness, understandability, and eventual use of the inferences derived via educational measurements and instruments including VAMs. For the purposes of this study, researchers define transparency as the extent to which something is quite simply accessible (e.g., attainable) and then comprehensible and usable.

In terms of VAMs, the main issue here is that most VAM-based estimates do not seem to make sense to those at the receiving end. Teachers and principals do not seem to understand the models being used to evaluate teachers; hence, they are reportedly unlikely to use VAM output for the formative purposes for which VAMs are also intended (Eckert and Dabrowski 2010; Gabriel and Lester 2013; Goldring et al. 2015; Graue et al. 2013; Kelly and Downey 2010). Rather, practitioners describe VAM reports as inaccessible, confusing, not comprehensive in terms of the concepts and objectives teachers teach, ambiguous in terms of teachers' efforts at individual student and composite levels, and oft-received months after students leave teachers' classrooms.

For example, teachers in Houston (home to another one of the legal cases highlighted in Education Week 2015) collectively expressed that they are learning little about what they did effectively or how they 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, p. 11). Eckert and Dabrowski (2010) also demonstrated that in Tennessee (home to two other legal cases cited in Education Week 2015), teachers maintained that there was no to very limited support in helping teachers understand or use their value-added data to improve upon their practice (see also Harris 2011). Altogether, this is problematic in that one of the main purported strengths of nearly all VAMs is the wealth of diagnostic information accumulated for formative purposes (see also Sanders et al. 2009), though at the same time, 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 and Lester 2013).

4.6 Consequential validity

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 (Messick 1980; Kane 2013). 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 policymakers, 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 Kane 2013).

Intended consequences As noted prior, 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 and Welner 2011). The stronger the consequences, the stronger the motivation leading to stronger intended effects. Secondary intended consequences included replacing and improving upon 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 also 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 and Gilmour 2017).

Unintended consequences Simultaneously, stakeholders often fail to recognize VAMs' unintended consequences (AERA 2000; see also AERA Council 2015). Policymakers must present evidence on whether VAMs cause unintended effects and whether the said unintended effects outweigh their intended effects, all things considered. Policymakers 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: (1) 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); (2) 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 (3) 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 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 and Lester 2013; Harris and Herrington 2015).

In sum, as per the *Standards* (AERA et al. 2014), ongoing evaluation of all these measurement issues as pertaining to VAMs and VAM use is essential, although this is not being committed to, nor committedly done. The American Statistical Association (ASA 2014), the AERA Council (2015), and the National Academy of Education (Baker et al. 2010) have underscored similar calls for research within their associations' positions statements about VAMs (see also Harris and Herrington 2015).

Notwithstanding, researchers used the above to frame this case study analysis, again, as primarily related to the measurement and pragmatic issues presented to the court in this case in New Mexico. It is this set of issues that researchers set out to make more transparent, to help others throughout the U.S. and internationally better understand the issues of primary dispute in this high-profile case, and also across cases (see Education Week 2015), all of which surround states' (or districts') high-stakes uses of teachers' value-added estimates and these measurement areas of concern. It is important to note again, however, that while researchers examined court documents using these six key measurement concepts (i.e., reliability, validity [convergent-related evidence], bias, fairness, transparency, and consequential validity) and found evidences of issues across all areas of interest, researchers conducted actual analyses using data as pertinent to only the first three (i.e., reliability, validity [i.e., convergent-related evidence], and bias).

5 Methods

Researchers conducted a case study analysis (Campbell 1975; Flyvbjerg 2011; Gerring 2004; Ragin and Becker 2000; Thomas 2011; VanWynsberghe and Khan 2007) to examine the legal documents in this case including official complaints, exhibits, affidavits, and other court documents including court rulings. The case study approach, according to VanWynsberghe and Khan (2007), best suits this type of study given researchers' (1) nonrepresentative sample of participants (i.e., from the state of New Mexico), (2) emphases on contextual detail (i.e., official court documents as situated within New Mexico), (3) focus on nonexperimentally controlled events (i.e., a teacher-evaluation system implemented and lived in practice), (4) well-defined parameters (i.e., the state of New Mexico's teacher evaluation system), and (5) multiple data sources (i.e., legal documents including complaints, exhibits, affidavits, etc.).

Researchers' primary intent was to determine what the issues of the case were, as aligned to the afore-described conceptual framework (i.e., reliability, validity [i.e., convergent-related evidence], bias, fairness, transparency, and consequential validity). Researchers' intent was to help others understand how this teacher evaluation system was being used and played out in the real world (Flyvbjerg 2011). Again, this teacher evaluation system was predicated upon one state's large-scale educational policies, with high-stakes consequential decisions also at stake and legal dispute.

However, generalizations may not be permitted given researchers' sample of convenience (i.e., all documents taken from one state that may or may not generalize to other states' teacher evaluation systems), although naturalistic generalizations (e.g., across other states with similar teacher evaluation systems) might certainly be warranted on a case-by-case basis (Stake 1978; Stake and Trumbull 1982).

Researchers also calculated and assessed New Mexico's teacher evaluation systems' actual levels of reliability, validity (i.e., convergent-related evidence), and potential for bias (as also aligned with researchers' afore-described conceptual framework). Researchers did this statistically, with their primary intent to determine whether New Mexico's teacher evaluation data were indeed reliable, valid, and biased (or unbiased), in isolation and in comparison, to other states' or districts' teacher evaluation systems as based on the current research literature. Researchers' intent was not to test the state-level VAM (e.g., by running the state data through another VAM to assess the state's homegrown VAM). Rather, researchers' intent was to assess the actual VAM output as used throughout New Mexico and situate findings within the current literature capturing what we know about other VAMs in terms of their levels of reliability, validity (i.e., convergent-related evidence), and potential for bias (or a lack thereof) to determine how this system was functioning, relatively and perhaps well enough to make and then legally defend the consequential decisions being attached to New Mexico teachers' VAM estimates.

5.1 Data collection

As mentioned, researchers examined the official complaints, exhibits, affidavits, and other court documents of legal pertinence in this case. More specifically, researchers examined 26 exhibits (i.e., exhibits A–Z) that included the following: exhibits A–E included documents describing the key components of New Mexico's teacher evaluation system, as well as the VAM upon which the system was based (i.e., as developed by the state's former Assistant Secretary for Assessment and Accountability and value-added modeler, Pete Goldschmidt; see also Swedien 2014), although it should be mentioned that nowhere across exhibits and court documents is the actual model equation illustrated or explained (exhibit D was the closest, but only included information about general VAMs and VAM equation approximations and assumptions). Otherwise, no information was made available to the court to describe the model, or the model's technical properties or merits. Apparently, "post implementation/use statistics from the [New Mexico] model…[are]…managed by the [New Mexico] Public Education Department." Those external to that system are not privy to data or results (T. Hand, personal communication, March 17, 2017). This also speaks to the model's potential issues regarding transparency, as defined prior (see also forthcoming).

Exhibits F–I included memorandums released to all New Mexico school superintendents and human resource officers to supplement, help explain, or respond to questions given the information included in exhibits A–E. Exhibits J–N included technical information about the student-level tests that were used to calculate teachers' value-added estimates (e.g., the state of New Mexico's Standards Based Assessments [SBAs], End of Course assessments [EOCs], Dynamic Indicators of Basic Early Literacy Skills [DIBELS]). Exhibits O–S included information about the state's observational system, modified[3] from Charlotte Danielson's Framework for Teaching (Danielson Group, n.d.), and the state's homegrown student survey system, both of which were considered to be the state's other "multiple measures" (along with teacher attendance[4]) used to help evaluate New Mexico's teachers. Exhibit T included the state's press release regarding the teacher evaluation scores from academic year 2014–2015. Lastly, exhibits U–Z included the results from analyses of teachers' evaluation scores conducted by a New Mexico Legislative Education Study Committee that was created "to compile general perceptions, issues, and concerns into a summary report, which [was] to be provided to the [state]." Related, also included within exhibits U–Z was a series of letters submitted by teachers concerned about the scores they received, and a series of tables describing the errors that the state had allegedly made when calculating teachers' value-added estimates. These errors included but were not limited to teachers being held accountable for test scores of students they did not teach, teachers' students whose scores were missing, teachers being listed as teaching grade levels or subject areas different than those they actually taught, teachers being held accountable using test data for subject areas they did not teach, and the like. Researchers also examined two affidavits, both of which were submitted by the expert witness working on behalf of the plaintiffs in this case (totaling 113 pages of text), and all other relevant court documents including plaintiff and defendant witness disclosures, retention agreements, deposition notices, official complaints, case documents, and court rulings.

It should also be noted here, though, that in many cases researchers also sought out supplementary resources to help them explore particular questions they had while reviewing the documents officially submitted for this case. However, they did not find any additional documents of value in terms of helping to explain, for example, the state's VAM, the technical properties of the tests used by the states beyond that which was included in exhibits J–N, user guides or other technical information about the observational and survey systems used in the state, and the like. While problematic, also for members of the profession or public looking for this type of information that many might argue should be made more publicly available, this also verifies that that which was reviewed for this study as taken directly from the case was all that really spoke to the issues at hand, again, in court as well as in practice.

For researchers' statistical analyses of New Mexico's teacher evaluation systems' levels of reliability, validity (i.e., convergent-related evidence), and potential for bias, researchers collected: (1) teacher-level VAM-based estimates for all New Mexico teachers, as calculated by the state; (2) classroom observation scores with (2a) scores from Danielson's

[3] The New Mexico's modified Danielson model consists of four domains (as does the Danielson model): "Domain 1: Planning and Preparation" (which is the same as Danielson), "Domain 2: Creating an Environment for Learning" (which is "Classroom Environment" as per Danielson), Domain 3: "Teaching for Learning" (which is "Instruction" as per Danielson), and "Domain 4: Professionalism" (which is "Professional Responsibilities" as per Danielson). Domains 1 and 4 when combined yield New Mexico's Planning, Preparation, and Professionalism (PPP) dimension. It is uncertain how the state adjusted the Danielson model for observational purposes, or whether the state had permission to do so from the Danielson Group (n.d.).

[4] In terms of teacher attendance, the state's default teacher attendance cut scores were based on days missed as follows: 0–2 days missed = Exemplary, 3–5 days missed = Highly Effective, 6–10 days missed = Effective, 11–13 days missed = Minimally Effective, and 14+ days missed = Ineffective. However, some districts did not include teacher attendance data for various reasons (e.g., "because absences are often attributed to the Family and Medical Leave Act, bereavement, jury duty, military leave, religious leave, professional development, and coaching") making system fidelity and fairness, again, suspect.

modified domains 2 and 3 (i.e., as related to student "learning") weighted differentially than (2b) scores from Danielson's modified domains 1 and 4 (i.e., as related to Planning, Preparation, and Professionalism [PPP]); and (3) student survey scores. Researchers used the latter indicators to assess whether teachers' VAM scores converged (i.e., convergent-related evidence of validity) with teachers' observational and survey scores.

Population and subsample The initial data files provided by the state for this lawsuit included 26,966 unique teachers across three academic years: 2013–2014, 2014–2015, and 2015–2016. Of the 26,966 educators, 97.0% ($n = 26{,}160$) were certified teachers. For purposes of this analysis, researchers restricted this sample to include certified teachers who had VAM and observation data for all three academic years, as is standard and recommended practice (Brophy 1973; Cody et al. 2010; Glazerman and Potamites 2011; Goldschmidt et al. 2012; Harris 2011; Ishii and Rivkin 2009; Sanders as cited in Gabriel and Lester 2013). Researchers did this while also taking into consideration that these two measures carry the majority of weight in the state's teacher evaluation system (i.e., these two measures are of most evaluative value as weighted, although weights were not used in these analyses). This resulted in the final sample including 7777 teachers, which was 28.8% of the full dataset or 29.7% of all certified teachers. That the final sample included approximately 30% of the teachers included in the main data files is also important to note as directly related to issues of fairness, which were also of concern to the court in this case (see also forthcoming).

5.2 Data analyses

For case study purposes, researchers analyzed all of the written text included within the pages of the documents described prior. Researchers read through each document coding for text, quotes, and concepts related to the elements of their a priori framework (Miles and Huberman 1994) aligned with the *Standards* (AERA et al. 2014). Using this deductive approach to coding (in which the categories were preselected from this framework), researchers grouped text, quotes, and concepts by element in the framework per document. This systematic approach to coding also modeled the framework method (Gale et al. 2013; Ritchie et al. 2013).

In terms of the statistical analyses researchers conducted in order to assess the New Mexico teacher evaluation system's levels of reliability, validity (i.e., convergent-related evidence), and potential for bias, researchers engaged in the following methods of data analyses per area of interest. In terms of reliability, researchers investigated the distribution of teachers' VAM estimates per teacher over time, as well as the correlations among scores over time. For comparative purposes, researchers calculated the correlations among teachers' observational, PPP, and student survey scores over the same period of time. Researchers conducted chi-square (χ^2) tests to determine if teachers' ratings along the four variables' score distributions and degrees of score variation significantly differed from year to year. To do this, they calculated quintiles of each measure's scores to determine what percentages of teachers moved among quintiles (chosen given the state of New Mexico classifies teachers as per their evaluation output using five effectiveness ratings) from 1 year to the next.

In terms of convergent-related evidence of validity, researchers investigated the (cor)relationships between the same indicators (i.e., teachers' VAM-based estimates, observational scores, PPP scores, and student survey scores). Researchers analyzed correlations among all four variables via calculations of Pearson's *r* coefficients between each pair of variables for each year. To determine if the differences between bivariate correlations from year to year were significant, researchers used Fisher's *Z* tests (Dunn and Clark 1969, 1971).

In terms of potential for bias (or the lack thereof), researchers compared the scores of all four measures per multiple teacher- and school-level subgroups (see Appendix 1 for a full list of the demographics used). Researchers calculated descriptive statistics for teachers' scores by these teacher- and school-level subgroups, while also analyzing statistically significant differences using *t* tests or fixed effects analyses of variance (ANOVA). It should also be noted that researchers did not have access to more nuanced or granular data, such as criterion-specific scores for each observation domain or individual item scores for the student surveys, that would have allowed for a more sophisticated analysis of potential bias.

6 Findings

6.1 Reliability

Across the documents analyzed for the case study section of this study, the concept of reliability was noted once and only peripherally. Only in exhibit E did the state include information about how teachers' students' test scores would be used to calculate teachers' value-added given how many years of VAM-based data teachers had. The more value-added data the teacher had, the more weight (i.e., at least 50%) the teacher's value-added scores were to carry. While written in exhibit E was the goal to have 3 years of data per teacher, also written was that teachers with less than 3 years of data (but no less than 1 year of data) would be held accountable for their value-added, or lack thereof.

Again, according to the literature, reliability can be increased with 3 years of data, although after including 3 years of data the strength more data adds to efforts to increase reliability plateaus. As such, having a minimum of 3 years of data is now widely accepted as standard practice in order to get the most reliable VAM estimates possible. What New Mexico had made official, then, contradicts field standards in terms of guaranteeing all evaluated teachers a 3-year minimum. This is especially important when consequential decisions are at play (Brophy 1973; Cody et al. 2010; Glazerman and Potamites 2011; Goldschmidt et al. 2012; Harris 2011; Ishii and Rivkin 2009; Sanders as cited in Gabriel and Lester 2013).

In terms of researchers' calculations of the state's actual levels of reliability, using only those teachers for whom 3 years of value-added data were available, researchers found that a plurality of teachers' VAM-based quintile rankings were the same from year to year (i.e., 31.6% [$n = 2455/7771$] from 2013–2014 to 2014–2015 and 31.6% [$n = 2454/7766$] from 2014–2015 to 2015–2016) or differed by one quintile (i.e., 40.6% [$n = 3157/7771$] from 2013–2014 to 2014–2015 and 39.4% [$n = 3060/7766$] from 2014–2015 to 2015–2016). However, many teachers also received dissimilar quintile rankings over the same period of time, with over 25% of teachers with scores that differed by two or more quintiles year to year (i.e., 27.8% [$n = 2159/7771$] from 2013–2014 to 2014–2015; 29.0% [$n = 2252/7766$] from 2014–2015 to 2015–2016).

To help illustrate what this looked like in practice, researchers generated a figure illustrating New Mexico teachers' VAM ratings from year one (2013–2014) to year three (2015–2016) to illustrate how teachers' value-added scores fluctuated over time. Evidenced in Fig. 2 are the percentages of teachers who got a given rating in 1 year (i.e., illustrated on the left) and the same teachers' subsequent ratings 2 years later (i.e., illustrated on the right).

Visible in Fig. 2 is that 43.1% of teachers who scored in the top quintile as per their VAM ratings in 2013–2014 remained in the same quintile (e.g., highly effective) 2 years later (2015–2016); 24.0% of these same teachers dropped one quintile (e.g., from highly effective to effective); 14.4% dropped two quintiles (e.g., from highly effective to average); 10.9% dropped three quintiles (e.g., from highly effective to ineffective); and 7.5% dropped four quintiles (e.g., from highly ineffective to highly ineffective). Inversely, 17.71% of teachers who scored in the bottom quintile as per their VAM ratings in 2013–2014 remained in the same quintile 2 years later (e.g., highly ineffective); 19.5% of these same teachers moved up by one quintile (e.g., from highly ineffective to ineffective); 19.9% moved up by two quintiles (e.g., from highly ineffective to average); 21.0% moved up by three quintiles (e.g., from highly ineffective to effective); and 17.7% moved up by four quintiles (e.g., from highly ineffective to highly effective). See also all other permutations illustrated.

These results make sense when situated in the current literature, as also noted prior, whereas teachers classified as "effective" 1 year typically have a 25% to 59% chance of being classified as "ineffective" 1 or 2 years later (Martinez et al. 2016; Schochet and

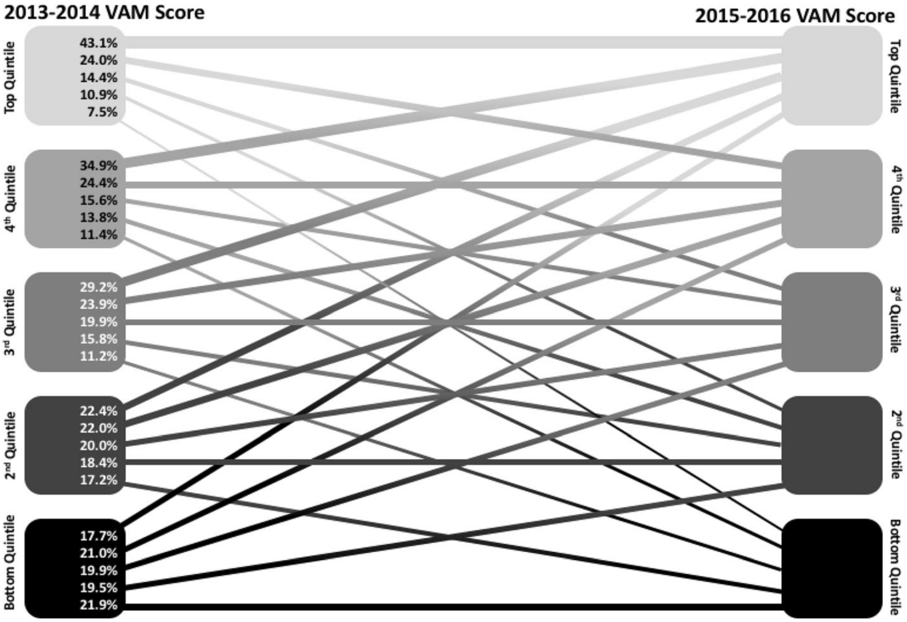

Fig. 2 Distributions of VAM-based ratings of New Mexico teachers with 3 years of VAM data across quintiles from 2013–2014 to 2015–2016

Chiang 2013; Yeh 2013). In New Mexico, 18.4% of teachers deemed "effective" (e.g., highly effective or effective) in the first year were deemed "ineffective" (e.g., ineffective or highly ineffective) 2 years later, and 38.7% of teachers deemed "ineffective" (e.g., ineffective or highly ineffective) in the first year were deemed "effective" (e.g., highly effective or effective) 2 years later. This yields an average movement, akin to the afore-cited literature, of 29% or approximately one out of every three teachers jumping effectiveness ratings within 3 years.

In comparison to the other measures used to evaluate New Mexico teachers, teachers' observation scores appeared to be the most stable or reliable over time, again, as based on similar reliability estimates of teachers' observational, PPP, and student survey scores. See Appendix 2 for these scores' levels of reliability over time.

6.2 Validity

Across the documents analyzed for the case study part of this study, the concept of validity (i.e., convergent-related evidence of validity) was noted nowhere. Validity, more broadly speaking, was mentioned once in terms of content-related evidence of validity, or whether test scores can be used to make inferences about student achievement, over time, as well as teachers' impacts on student achievement over time. In exhibit N, the state noted that districts might opt to use tests in addition to those that were the state-approved tests, but districts must evidence to the state that each test pass a content review during which "a panel of content experts and skilled item writers should evaluate the quality of newly-written items" included in the tests to be used for teacher evaluation purposes, after which the district would be required to submit a report to the state describing results. Here, and again elsewhere across documents, the state stops at the most basic level of validity, encouraging or relying seemingly only on logic, face validity, and nothing more empirical. Related, across exhibits J–N that included technical information about the tests used to calculate teachers' value-added estimates for all of New Mexico's VAM-eligible teachers, no empirical information was presented to evidence that the states' tests could or should be used to evaluate student achievement over time or teachers' impacts on student achievement over time. Rather, standard statistics (e.g., internal consistency statistics, test–retest and split-half reliability indicators, proportion-correct [*p*] values) were used to evidence that the tests were reliable and valid for purposes of measuring student achievement at one point in time. It seems that because many of these tests are already in place across states (e.g., as mandated via the federal government's former NCLB Act 2001), these tests are being used more out of convenience (and cost savings) for their newly but not yet validated tasks and uses.[5]

[5] For example, most VAMs require that the scales that are used to measure growth from 1 year to the next can be appropriately positioned upon vertical, interval scales of equal units. These scales should connect consecutive tests on the same fixed ruler, so-to-speak, making it possible to measure growth from 1 year to the next across different grade-level tests. Here, for example, a ten-point difference (e.g., between a score of 50 and 60 in fourth grade) on one test should mean the same thing as a ten-point difference (e.g., between a score of 80 and 90 in fifth grade) on a similar test 1 year later. However, the scales of all large-scale standardized achievement test scores used in all current value-added systems do not even come close to being vertically aligned, as so often assumed (Baker et al. 2010; Ballou 2004; Braun 2004; Briggs and Betebenner 2009; Ho et al. 2009; Newton et al. 2010; Papay 2010; Sanders et al. 2009). "[E]ven the psychometricians who are responsible for test scaling shy away from making [such an] assumption" (Harris 2009, p. 329).

Otherwise, and in terms of researchers' calculations of the state's VAM-based levels of convergent-related evidence of validity as aligned with current practice in this area of research, researchers found that correlations among all measures used in New Mexico to evaluate its teachers were weak to very weak.[6] This was true across all years, with the exception of the correlations between teachers' observation and PPP scores, which were strong[7] (see Table 3).

The strong correlations observed between teachers' observation and PPP scores, however, make sense given teachers' observation and PPP scores came from within the same observational instrument (i.e., as modified from the Danielson's Framework [Danielson Group, n.d.]). If anything, such strong correlations might suggest that separating out teachers' observational scores (i.e., from modified domains 2 and 3) from teachers' PPP scores (i.e., from modified domains 1 and 4) is not defensible or warranted, given such high or strong correlation coefficients often suggest a universal but not divisible or detachable factor structure. Rather, this may suggest that the factor structure pragmatically posited and used throughout New Mexico may not empirically hold (Sloat 2015; see also Sloat et al. 2017; Polat and Cepik 2015).

Notwithstanding, illustrated in Table 3 is that correlations were the weakest between teachers' VAM-based estimates and student survey scores across years, ranging between the statistically significant yet very weak[8] levels of $r = 0.031$ and 0.135. Conversely, other than the strong relationships observed between teachers' observation and PPP scores, again as taken from within the same instrument, correlations were stronger (albeit still weak[9]) between teachers' observation and student survey scores across years, ranging from $r = 0.211$ to 0.235.

Perhaps most importantly, as also situated within the current literature, the correlations between teachers' VAM and observation scores ranged from $r = 0.153$ to $r = 0.210$. Lower correlations were observed between teachers' VAM and PPP scores ranging from $r = 0.128$ to $r = 0.189$. To help illustrate what the higher of the two VAM and observational correlations looked like in practice (i.e., taking the observational score from domains 2 and 3 which mapped onto student learning), researchers generated Fig. 3 to illustrate the distributions of New Mexico teachers' VAM and observation ratings across quintiles for all 3 years of focus (i.e., 2013–2014, 2014–2015, and 2015–2016).

Demonstrated is that regardless of which year is chosen, of the three sets of VAM and observational data displayed, New Mexico's teachers' data are more or less randomly distributed. Visible is that, on average (with the average taken across the 3 years of data visualized) 30.4% of teachers who scored in the top quintile as per their VAM scores also landed in the top quintile as per their observation scores (i.e., with both indicators suggesting that these teachers were highly effective); 26.6% of these same teachers dropped one quintile from the top VAM quintile to the second highest observation quintile (e.g., from highly effective to effective); 18.4% dropped two quintiles (e.g., from highly effective to average); 16.1% dropped three quintiles (e.g.,

[6] Interpreting r: $0.8 \le r \le 1.0$ = a very strong correlation; $0.6 \le r \le 0.8$ = a strong correlation; $0.4 \le r \le 0.6$ = a moderate correlation; $0.2 \le r \le 0.4$ = a weak correlation; and $0.0 \le r \le 0.2$ = a very weak correlation, if any at all (Merrigan and Huston 2008).

[7] Ibid.

[8] Ibid.

[9] Ibid.

Table 3 Correlations among measures, per year

Measure	2013–2014	2014–2015	2015–2016
VAM & observation	0.153***	0.187***	0.210***
VAM & PPP	0.128***	0.154***	0.189***
VAM & student survey	0.031*	0.063**	0.135***
Observation & PPP	0.771***	0.774***	0.788***
Observation & student survey	0.211***	0.219***	0.235***
PPP & student survey	0.196***	0.175***	0.202***

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

from highly effective to ineffective); and 8.6% dropped four quintiles (e.g., from highly ineffective to highly ineffective). Inversely, 20.9% of teachers who scored in the bottom quintile as per their VAM scores also landed in the bottom quintile as per their observation scores (i.e., with both indicators suggesting that these teachers were highly ineffective); 23.8% of these same teachers moved up by one quintile from the bottom VAM quintiles to the second lowest observation quintile (e.g., from highly ineffective to ineffective); 19.1% moved up by two quintiles (e.g., from highly ineffective to average); 21.0% moved up by three quintiles (e.g., from highly ineffective to effective); and 15.2% moved up by four quintiles (e.g., from highly ineffective to highly effective). See also all other permutations illustrated.

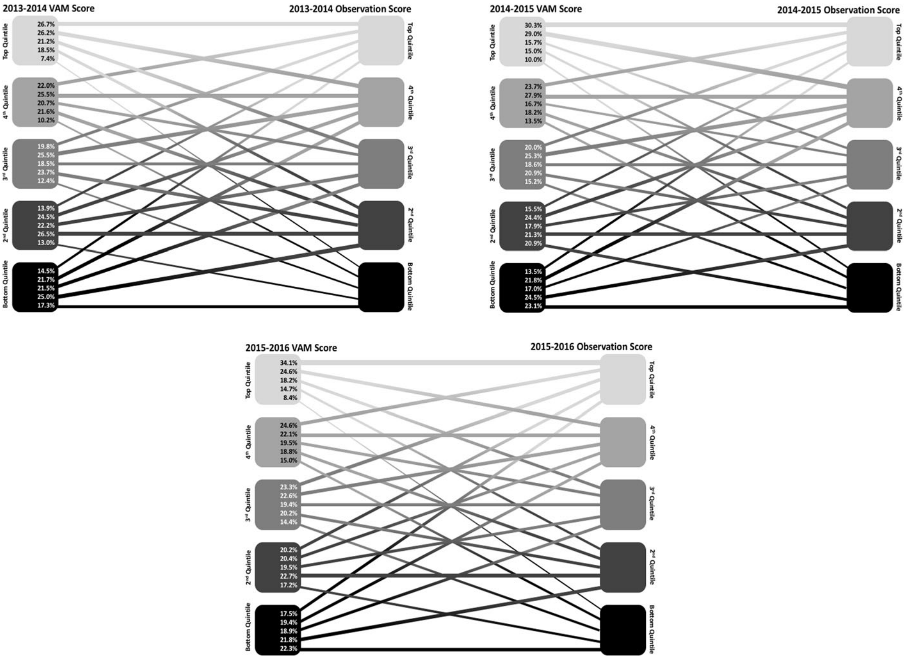

Fig. 3 Distributions of New Mexico teachers' VAM and observation ratings across quintiles (2013–2014, 2014–2015, and 2015–2016)

These results make sense as per the current literature, as also noted prior, given the correlations between teachers' VAM scores in general. Not only are the correlations for New Mexico teachers very weak,[10] they are also relatively very weak as situated within the literature. The literature has been saturated with evidence that correlations between multiple VAMs and observational scores typically range from $0.30 \leq r \leq 0.50$ (see, for example, Grossman et al. 2014; Hill et al. 2011; Kane and Staiger 2012; Polikoff and Porter 2014; Wallace et al. 2016). It can be concluded, then, that New Mexico's correlations, with regard to convergent-related evidence of validity, are very weak also in comparison to other VAMs.

6.3 Bias

Across the documents analyzed for the case study section of this study, the concept of bias was noted twice. First, in one of the memorandums (i.e., exhibit I of exhibits F–I) released to all New Mexico school superintendents and human resource officers, the state noted that the state's VAM is one of the few that does not include covariates to control for or block student demographic variables that may bias teachers' value-added scores (e.g., socioeconomic status, race, ELL status). The memo did not offer anything further, however, by disclosing or discussing the extent to which not controlling for such variables may yield more or less biased VAM output.

Elsewhere in exhibit N, as also related to the above comments about the tests that districts might use in addition to the state-level tests required, the state also encouraged districts to review the tests they might use for bias within the tests themselves. However, the state again stopped short of requiring empirical evidence (e.g., analyses of bias in test score outcomes by student demographics) and required, instead, "bias reviews" by "panel[s] of subgroup representatives who are competent in [the] content area [who] can conduct an acceptable bias evaluation." The state suggested one analytical technique to do this, that of "Differential Item Functioning (DIF) [which] is perhaps the most widely-recognized bias examination method;" although, the state also noted that "the [state would] accept," but not require such evidence, as this "require[d] highly specialized expertise to perform."

Otherwise, and as previously explained, potentially biased estimates of teacher effectiveness are observed when performance varies for different subgroups of teachers, often even despite the sophistication of statistical controls put into place to control for or block such bias (see also Paufler and Amrein-Beardsley 2014; Baker et al. 2010; Collins 2014; Kappler Hewitt 2015; McCaffrey et al. 2004; Michelmore and Dynarski 2016; Newton et al. 2010; Rothstein and Mathis 2013). In fact, current practice suggests that all such demographics should be included in any VAM just to be increasingly certain that the biasing effects of such variables are negated (Koedel et al. 2015). With that being said, researchers found indicators of bias throughout the data examined at the teacher (e.g., gender, race/ethnicity) and school levels (e.g., proportion of ELL students, proportion of minority students) likely (or at least in part) as a result. Researchers also found that the potential for bias existed across the measures of teacher effectiveness used in New Mexico, or that bias was not just limited to the state's VAM.

[10] Ibid.

6.3.1 Teacher-level differences

Researchers found evidence of possible teacher-level bias across all 3 years, all four of New Mexico's measures of teacher effectiveness, and as per the five teacher-level subgroups that they analyzed (see Appendix 3, Tables 6, 7, 8, and 9).

Gender Compared to female teachers, male teachers had significantly higher VAM scores in 2013–2014 ($t = 3.471$, $p = 0.001$), but significantly lower VAM scores in 2015–2016 ($t = 7.150$, $p < 0.001$). VAM-based bias did not hold by gender. However, female teachers had significantly greater observation scores, PPP scores, and student survey scores compared to male teachers for each of the 3 years; hence, if anything, these three measures (i.e., not including VAM estimates across years) might have been biased in favor of female teachers (see also Bailey et al. 2016; Steinberg and Garrett 2016; Whitehurst et al. 2014), although some might argue that female teachers were better than their male colleagues.

Race/ethnicity While there were no significant differences between Caucasian and non-Caucasian teachers' VAMs (with non-Caucasian teachers defined as Asian (< 1%), African American, Hispanic, and Native American as per the state's classifications), there were statistically significant differences between the two groups of teachers on the other three measures. Caucasian teachers had significantly higher observation and PPP scores than non-Caucasian teachers for each of the 3 years. Hence, it could be concluded that Caucasian teachers may be perceived as better teachers than non-Caucasians given these instruments or the scorers observing teachers in practice may be biased against some versus other teacher types by race (see also Bailey et al. 2016; Steinberg and Garrett 2016; Whitehurst et al. 2014). Inversely, non-Caucasian teachers had higher student survey scores than Caucasian teachers for all 3 years, with 2014–2015 ($t = 4.258$, $p < 0.001$) and 2015–2016 ($t = 5.607$, $p < 0.001$) being statistically significantly different. Again, this could be due to bias (i.e., that teachers' students' had differing perspectives of their teachers' qualities by race), given standard issues with surveys (e.g., low response rates that might distort validity; Nunnally 1978), or given non-Caucasian teachers may have indeed been better teachers than their Caucasian colleagues.

Years of experience Overall, teachers with fewer years of experience had VAM estimates that were significantly lower than teachers with more years of experience. Similar patterns were observed for teachers' observation scores and PPP scores, with teachers with the least amount of experience routinely earning scores that were significantly lower than their more experienced counterparts across each of the 3 years. This could mean, as also in line with common sense as well as the current research (Darling-Hammond 2010), that teachers with more experience are typically better teachers. These findings might support the validity of teachers' VAMs and observational scores in this regard. Survey scores did not follow this pattern, however, which might be due to issues with survey research also noted prior.

Grades taught In 2013–2014, elementary school teachers had significantly lower VAM scores than middle school teachers and high school teachers ($F = 149.465$,

$p < 0.001$), and high school teachers had significantly higher VAM estimates than both elementary teachers and middle school teachers in 2014–2015 ($F = 6.789$, $p = 0.001$). This might mean, in the simplest of terms, that elementary school teachers might be worse than teachers in high school, or that VAM estimates and the ways they are calculated (e.g., using different tests at different levels) might be biased against teachers of younger students. There were no other significant differences observed across teachers' observation and PPP scores based on grades taught, although significant differences did exist for survey scores. High school teachers had the lowest scores as based on their student survey data, and elementary school teachers had the highest scores in 2013–2014 ($F = 60.929$, $p < 0.001$) and 2014–2015 ($F = 178.239$, $p < 0.001$).

Subject taught Overall, teachers who taught ELL or SPED classes had lower VAM estimates across all 3 years than those who did not teach such classes. Those that were significant were for ELL teachers who had significantly lower VAMs than non-ELL teachers in 2014–2015 ($t = 2.001$, $p = 0.046$), and SPED teachers who had significantly lower VAMs than non-SPED teachers in 2014–2015 ($t = 2.248$, $p = 0.025$) and 2015–2016 ($t = 7.354$, $p < 0.001$). Contrariwise, teachers who taught gifted classes had significantly higher VAMs than nongifted teachers in 2013–2014 ($t = 4.724$, $p < 0.001$) and 2014–2015 ($t = 3.147$, $p = 0.002$). This runs counter to the research evidencing that gifted students often prevent gifted teachers from displaying growth given ceiling effects (Cole et al. 2011; Kelly and Monczunski 2007; Koedel and Betts 2007; Linn and Haug 2002; Wright et al. 1997).

Otherwise, patterns similar to those mentioned above were also observed for ELL, SPED, and gifted teachers on their observation and PPP scores. Consistently across all years, non-ELL, non-SPED, and gifted teachers had significantly better observation and PPP scores than their ELL, SPED, and nongifted counterparts (see also Bailey et al. 2016; Steinberg and Garrett 2016; Whitehurst et al. 2014). The one exception to this pattern was for ELL teachers in 2015–2016, as while their PPP scores were higher than those of non-ELL teachers, the difference was not significant. Regarding student survey scores, ELL teachers had significantly higher scores than non-ELL teachers for all 3 years, and SPED teachers had significantly higher scores than non-SPED teachers in 2015–2016. There were no significant differences between gifted and nongifted teachers' survey scores.

6.3.2 School-level differences

Researchers found evidence of the potential for school-level bias across all 3 years, all four of New Mexico's measures of teacher effectiveness, and as per the six school-level subgroups that researchers analyzed (see Appendix 3, Tables 10, 11, 12, and 13).

6.3.3 Total enrollment

Teachers in schools with low enrollments (i.e., enrollment less than the sample median; hereafter referred to as low enrollment schools) had significantly higher VAMs in 2013–2014 ($t = 2.428$, $p = 0.015$) and 2014–2015 ($t = 5.017$, $p < 0.001$) than teachers in high enrollment schools, although this was reversed in 2015–2016 where teachers in high enrollment schools had significantly greater VAMs ($t = 4.300$, $p < 0.001$). Observation scores significantly differed only in 2015–2016, where teachers in low enrollment schools scored significantly higher than teachers in high enrollment schools ($t = 2.888$, $p = 0.004$). Teachers in low enrollment schools had significantly lower PPP scores in 2013–2014 ($t = 5.386$, $p < 0.001$), but significantly higher PPP scores in year 2015–2016 ($t = 2.807$, $p = 0.005$). Only on teachers' student survey scores did researchers observe a consistent pattern. Teachers in low enrollment schools had significantly higher survey scores for all the 3 years, although concerns about the student survey measures noted prior likely also come into play here.

SPED student population Teachers in schools with low populations of SPED students (i.e., hereafter referred to as low SPED schools) consistently had significantly greater VAMs, observation scores, and PPP scores than teachers in high SPED schools across all 3 years. This suggests that teachers in low SPED teachers are as a group better or that VAM estimates might be biased against teachers teaching in high SPED schools, preventing them from demonstrating comparable growth. This pattern was reversed for teachers' student survey scores, however, as teachers in low SPED schools had lower survey scores than teachers in high SPED schools. This was trued across the board, although the only significant difference was in 2015–2016 ($t = 3.003$, $p = 0.003$).

ELL student population Similar to teachers in low SPED schools, teachers in low ELL schools consistently had significantly greater VAMs, observation scores, and PPP scores than teachers at high ELL schools for all 3 years. Teachers in low ELL schools also had lower survey scores for each of the 3 years, although the only statistically significant difference was, again, in 2015–2016 ($t = 9.131$, $p < 0.001$). This would suggest that teachers in low ELL schools are as a group better, or that VAM estimates might be biased against teachers teaching in high ELL schools, preventing them from demonstrating comparable growth.

FRL student population Similar to teachers in low ELL and low SPED schools, teachers in low FRL schools consistently had significantly greater VAM, observation, and PPP scores than teachers at high FRL schools for all 3 years. Again, this would suggest that teachers in low FRL schools are as a group better, or that VAM estimates might be biased against teachers teaching in such schools, preventing them from demonstrating comparable growth. Teachers in high FRL schools had significantly higher survey scores for each of the 3 years.

Gifted student population Related to the discussion about gifted teachers prior, teachers in schools with higher proportions of gifted students (hereafter referred to as high gifted schools) had significantly greater VAM scores as a whole than teachers at low gifted schools in 2013–2014 ($t = 2.980$, $p = 0.003$) and 2014–2015 ($t = 6.075$, $p < 0.001$). This pattern was the same for observation and PPP scores; hence, this would, again, suggest that teachers in high gifted schools are as a group better, or that VAM estimates might be biased against teachers teaching in schools with fewer gifted students, preventing them from demonstrating comparable growth. Teachers in low gifted schools, however and perhaps not surprisingly at this point, had significantly greater survey scores than teachers in high gifted schools for each of the 3 years.

Minority student population Lastly, and in line with teachers in low SPED, ELL, and FRL schools, teachers in schools with lower populations of non-Caucasian students (hereafter referred to as low minority schools) consistently had significantly higher VAMs than teachers in high minority schools for all 3 years. Again, this was the case for teachers' observation and PPP scores, again suggesting that teachers in low minority schools are as a group better, or that VAM estimates might be biased against teachers teaching in schools with more minority students, preventing them from demonstrating comparable growth. Survey scores were slightly more varied based on minority student population, with the only significant difference between scores in 2014–2015, where teachers in low minority schools had significantly higher survey scores than teachers in high minority schools ($t = 2.902$, $p = 0.004$).

6.4 Fairness

Across the documents analyzed for the case study section of this study, the concept of fairness was noted multiple times. In terms of the state's general teacher evaluation model, the state made it explicit across multiple documents that in order to include more teachers in the state's teacher evaluation system, the state developed three different evaluation models to be more inclusive in its efforts to evaluate three different teacher types, as per the grades and subject areas they taught. These teacher types included group a, group b, and group C teachers. Group A teachers taught grades (e.g., grades 3–11) and subject areas that were tested using the state's SBAs (e.g., in mathematics, English/language arts, science, and social studies), and were, therefore, the teachers who were eligible for teacher-level VAM scores (i.e., VAM-eligible). These teachers' students' test scores were to count for at least 50% of these teachers' overall evaluation scores, alongside their other evaluation data. However, in some cases, VAM data were used in isolation of the other data. For example, some plaintiff teachers' files were flagged solely given their VAM scores, thus suggesting that these teachers were ineffective. Regardless of these teachers' other data, when the state used the VAM score to flag these (and other) teachers' files, this meant that the VAM carried 100% of the evaluative weight.

Group B and group C teachers were VAM-ineligible teachers because they taught students in nontested subject areas or nontested grade levels. More specifically, group B teachers included all physical education, music, art, foreign language, etc. teachers; grade 9 and grade 12 mathematics and English/language arts teachers; and all high school science [e.g., biology, chemistry, physics] and social studies teachers [e.g., world history, geography, economics]. Group C teachers taught students in grades kindergarten through grade 2.

While at face value this might seem fair, this still leaves group A teachers as the only ones who are VAM-eligible. Given that VAM scores are the scores to which the state ties its most important decisions (e.g., flagging teachers' files and teacher termination decisions), this reinforces New Mexico's issues with fairness in that, as also situated in the current research, still only approximately 30–40% of all teachers in the state are VAM- as well as consequence-eligible (Baker et al. 2013; Gabriel and Lester 2013; Harris 2011).

Perhaps not surprisingly, when researchers conducted their statistical analyses of New Mexico's VAM's levels of reliability, validity (i.e., convergent-related evidence), and bias (or the lack thereof) described prior, they analyzed a final sample including 7777 teachers. As mentioned, this final sample also included 28.8% of the full dataset and 29.7% of all certified teachers, and this proportion held constant across the 3 years of data analyzed. That the final sample omitted approximately 70% of the state's teachers is also important to note as directly related to the national statistics cited. This yields evidence of issues with fairness, and this yields evidence that New Mexico is no different than most if not all other states and districts, at least throughout the U.S. Indeed, some teachers (i.e., 30% in New Mexico and 30% in states and districts elsewhere) are more likely to realize the negative or positive consequences attached to value-added output, simply given the types of students they teach by grade and subject area. Inversely, 70% or so of New Mexico teachers (and 70% of teachers in states and districts elsewhere) are VAM and, hence, accountability and consequence immune.

6.5 Transparency

Across the documents analyzed for the case study section of this study, transparency was not mentioned or exhibited as valued. Nowhere across the exhibits was the actual VAM equation listed, much less explained. Only in exhibit D were the purposes behind using a VAM in New Mexico listed (e.g., improved student performance, ridding New Mexico schools of teachers who are "free riders[11]"), and all points listed were listed without references in support. In fact, the only studies cited within exhibit D were one coauthored by New Mexico VAM developer Goldschmidt (Goldhaber et al. 2013; cited in exhibit D as Goldschmidt 2011) and three citations authored by others who are relatively well-known for their pro-VAM stances (Koedel and Betts 2007; Rivkin et al. 2005; Rockoff 2004). Also listed in exhibit D was a very rudimentary definition of a VAM as a measure of "teacher [*sic*] contribution to student learning" with emphases on "mak[ing] this [*sic*] a causal estimate." This was quite simply the most thorough offering across all case documents to help others (e.g., New Mexico's teachers and principals) understand New Mexico's VAM.

Perhaps most interesting in terms of exhibit D, though, was that at the end of the document anybody with additional questions was directed as follows: "If [anyone has] further questions about [his/her] teacher evaluation, [(s)he is to] contact [his/her] school principal or [his/her] school district testing director." Ironically, when researchers contacted the testing director of the second largest school district in New Mexico about some basic questions (e.g., where the VAM equation could be found, whether there was evidence of reliability in a technical or other report somewhere), (s)he had little to share about the state's VAM besides referring researchers back to the same exhibit D (e.g., T. Hand, personal communication, March 17, 2017). That might serve as one indicator of how far one might get if following such directions when trying to find out more about the state's VAM or value-added teacher evaluation system. Beyond this, only other simple details about how New Mexico's VAM was to be used to evaluate teachers using VAM-eligible teachers' students' test scores were offered.

[11] While "free riders" are not defined in exhibit D, researchers assume this means that this term refers to teachers who obtain something without comparable effort.

Relatively much more information was offered about the state's observational system (e.g., in exhibits O and P), again, as modified or "[b]ased largely on the Danielson Framework" (Danielson Group, n.d.). In exhibit O, for example, the state made claims the researchers found reasonable about this observational system in the sense that it might provide "critical and meaningful," "regular and purposeful," and "timely, targeted, and actionable" feedback. All of these claims, despite the lack of citations in support, are likely "true" as such observational systems, while imperfect, are often touted for being instructionally useful, if not the most instructionally useful evaluation measures currently in play. In exhibit P, the actual rubric was provided to teachers so that they would know precisely on what they would be evaluated when being observed. Researchers saw this as a direct evidence of transparency, as per *The Standards* (AERA et al. 2014).

Perhaps it is no wonder why teachers and principals in exhibits U–V charged that the state's teacher evaluation system, and primarily its VAM as the key component, did not help teachers and principals in their efforts to support increased student learning and achievement. Teachers noted that this was the case primarily because the system was confusing, "rushed," constructed without educator input, not transparent which also prohibited application and use, and that no training or professional development had been offered to help them understand the VAM or how to use VAM output. This all had to do with this system's VAM and its transparency (or lack thereof).

6.6 Consequential validity

Across the documents analyzed for the case study section of this study, researchers identified the concept of consequential validity being present across multiple documents. Across numerous informational and training documents published by the NMPED (e.g., exhibits A–D, T), the state cited incorporating (and heavily weighting) VAM data as a way to improve the accuracy and "overall predictive power" of the NMTEACH system, which would consequently result in the improved identification of effective teachers. Across exhibits A–D and T, researchers noted that the concept of consequential validity was always discussed by the state in a positive light, in that it consistently positioned the incorporation of VAM data as a way to ultimately improve educational outcomes for its students, and more so than other traditional measures of teacher evaluation (e.g., teaching experience, graduate degrees).

Further, although the 2013–2014 school year was supposed to be a "hold harmless" year for teachers so the NMPED could carefully roll out and test its new NMTEACH system, the state flagged teachers' files regardless with their value-added or overall effectiveness categories. In some cases, teachers with low VAM scores were placed on professional improvement plans, even after the state realized and admitted that many teachers' VAM scores (and therefore overall scores) were miscalculated.

Compared to state-issued documents, researchers found that the concept of consequent validity was discussed in a much different light by school administrators and teachers (i.e., in exhibits U–V). Many school administrators expressed feelings of frustration and uncertainty around the new NMTEACH system and, specifically, the incorporation of VAM data. Overall, administrators indicated that the lack of transparency (as previously discussed) and schools and districts rushing to implement the system (to meet NMPED and state policy deadlines) resulted in gross confusion, among administrators themselves and their teachers. This confusion was further exacerbated when administrators realized the extent of missing or incorrect teacher-level data. For example, in exhibit U, administrators indicated that teachers in their schools or districts were improperly coded (e.g., as group B instead of group A), were linked to the wrong students, had missing student test score data, and/or had missing or late summative evaluations. These errors resulted in administrators feeling angry and frustrated, and several cited greatly reduced morale and buy-in from teachers regarding the evaluation process writ large.

As evidenced in exhibit V, many teachers experienced unintended consequences from the new NMTEACH system as a whole, as well as the incorporation of VAM-based data. Echoing administrators' sentiments, teachers consistently indicated that they felt angry and frustrated about being linked to the wrong students and/or wrong test data, and hopeless and confused about how to improve their instructional methods as they could not understand why they received such low VAM scores. Additional unintended consequences included teachers being denied promotion due to faulty student test data, having to pay hundreds of dollars to fight such data inaccuracies, and being penalized for taking time off to attend to documented health concerns. District-level leaders also 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.

7 Discussion

The purpose of this study was to critically review all documents pertaining to what is arguably still the highest profile of the 15 cases (Education Week 2015) in which teacher plaintiffs throughout the U.S. were contesting how they were being evaluated and held accountable using VAMs. Researchers conducted this study, accordingly, to help both national and international educationists better understand the measurement and pragmatic issues at play and better understand these issues in multiple high-stakes contexts.

While authors of other VAM-based studies have offered a good amount of information about how these systems might work in theory and practice, these studies have been conducted primarily in the U.S. (see, for example, the aforementioned MET studies, which to date are still considered the most invested studies in this area of research at $45 million; Cantrell and Kane 2013; Kane and Staiger 2012). The majority of VAM-based studies that followed Race to the Top (2011) mostly did the same as well, in terms of researchers primarily assessing the reliability, validity, and potential for bias of such systems (see these studies systematically reviewed in Author(s), revised and resubmitted).

Perhaps more importantly, though, researchers of the MET studies as well as most of these others alluded to above did not look at how VAM-based systems actually played out in practice, especially when high-stakes consequences (e.g., termination, denial of tenure, merit pay) were attached to system output, like they were in all 15 of the aforementioned lawsuits (Education Week 2015). In addition, the MET studies were completely empirical, with students randomly assigned to teachers, with no real stakes or decisions attached to teacher-level output. In this case, not only did a state (i.e., New Mexico) develop what researchers would classify as a MET-aligned teacher evaluation system (i.e., the NMTEACH), state leaders implemented it in the real world (i.e., without random assignment, which makes sense given schools' current placement practices; see Sloat et al. 2014), and they did this very quickly (and injudiciously, as alleged), without allowing time for the construction of appropriate datasets and rostering systems; without proper piloting, validation, and revising; without collecting or considering teacher or administrator feedback anywhere throughout the process; and the like.

One thing New Mexico did not do, and other states (or districts) certainly should do, would be to follow the *Standards* in this area, given the firm recommendations of AERA et al. (2014) of states'/districts' ongoing evaluations of both the intended and unintended consequences of any test as an essential part of any test-based system, including those based upon VAMs. More specifically, the authors of the *Standards* write that

> "[E]vidence of the validity of a given interpretation of test scores for a specified use is a necessary condition for the justifiable use of the test. Where sufficient evidence of validity exists, the decision as to whether to actually administer a particular test generally takes additional considerations into account…[as well as]…consideration of negative consequences of test use, and weighing of any negative consequences against the positive consequences of tests use." (AERA et al. 2014, p. 11; see also AERA 2000; AERA Council 2015)

While, in general, actual examinations of consequential evidences rarely occur, this was certainly the case in New Mexico. Burdens of proof, instead, rested on the shoulders of plaintiffs and plaintiffs' expert witnesses to provide evidence about the positive and negative effects that come along with high-stakes VAM use.

Notwithstanding, in the U.S., VAM use (and, perhaps, abuse) is still pertinent, even post the passage of ESSA (2016), which reduced federal mandates surrounding VAMs, as states are still using VAMs, namely, to make teacher termination decisions as solely or primarily based on teachers' VAM data, or encourage such decisions as per a combination of districts' VAM and other evaluative (e.g., observational) data (Cantrell and Kane 2013; Kane and Staiger 2012). Globally, the teaching profession and, specifically, teacher evaluation continues to become datafied (Anderson 2008; Grek 2009; Selwyn 2015), and similar VAM-related initiatives remain ongoing in other industrialized nations (e.g., in Australia, England, Korea, Japan). While most countries have shown greater restraint than the U.S. in terms of directly linking student test scores to teacher performance, such test scores are increasingly becoming a key priority in teacher evaluation systems (Smith and Kubacka 2017).

With that being said, in this study researchers found that, in terms of reliability, New Mexico's teachers "jumped around" in ways similar to all other teachers being evaluated using VAMs, whereas teachers classified as "effective" 1 year had about a 30% chance of being classified as "ineffective" 2 years later, and vice versa (see also Martinez et al. 2016; Schochet and Chiang 2013; Yeh 2013). In addition, of all measures used to evaluate New Mexico teachers, teachers' VAM scores were also the least reliable or consistent over the same period of time.

In terms of convergent-related evidence of validity, researchers found that the correlations between teachers' VAM and observation and PPP scores ranged from $r = 0.128$ to $r = 0.210$. The literature has been saturated with evidence of such correlations ranging between $0.30 \leq r \leq 0.50$ (Grossman et al. 2014; Hill et al. 2011; Kane and Staiger 2012; Polikoff and Porter 2014; Wallace et al. 2016). Hence, the evidence suggests that New Mexico and its VAM is actually performing much worse in terms of its convergent-related evidence of validity as compared to other VAMs.

In terms of potential for bias, researchers found evidence of possible teacher- and school-level bias across all 3 years, all four of New Mexico's measures of teacher effectiveness, and as per the five teacher-level and six school-level subgroups that they analyzed. Most notably at the teacher level, no consistent significant differences in VAM scores were observed by teacher gender or race/ethnicity, but laudable differences were observed by teachers' years of experience (i.e., potentially validating VAM estimates), and potential for bias was present via teachers' grade and subject areas taught (see also Holloway-Libell 2015). Indicators of bias were also consistently observed across the observational and PP measures used in New Mexico. Most notably at the school level, no consistent VAM bias was observed by schools' enrollment numbers, but teachers in low SPED, low ELL, low FRL, and high gifted schools consistently had significantly greater VAMs, observation scores, and PPP scores across all 3 years.

Whether all of the teachers noted by subgroup above were in fact better or worse than other teachers teaching different proportions or populations of students, though, is certainly something of interest, and also of methodological and pragmatic concern. While some might argue that the by-school or by-teacher-type findings above simply suggest that better teachers are resident within certain types of schools, others might argue that the types of students who "better" teachers teach are simply more likely to demonstrate growth, compared to the other types or proportions of students that "worse" teachers teach (i.e., as nonrandomly assigned into their classrooms). Hence, the real questions pertinent to these findings are (1) whether students, regardless of the type of school, have equal opportunities to demonstrate growth, and (2) whether students' teachers, regardless of the types of schools in which they teach, have equal opportunities (i.e., a "level the playing field") to demonstrate growth once scores are aggregated.

Also, of note is that more or less, across analyses was that the scores derived via student surveys yielded scores most often "opposite" of what was observed across the other measures included within this particular teacher evaluation system. Again, this could very well be due to the measurement issues pertaining to survey research noted above, as well as in general (e.g., reliability, validity [i.e., convergent-related evidence], and bias, as often related to inadequate response rates), especially when evaluating teachers in public schools (Nunnally 1978).

In terms of fairness, researchers found that approximately 30% in New Mexico (and 30% in states and districts elsewhere) were more likely to realize the negative or positive consequences attached to value-added output, simply given the types of students they taught by grade and subject area. Inversely, 70% or so of New Mexico teachers (and 70% in states and districts elsewhere) were VAM and, hence, accountability and consequence ineligible.

In terms of transparency, researchers found that there was essentially little-to-none. If researchers were hard pressed to find out information about the VAM used throughout New Mexico, via thorough examinations of all court documents submitted in what is likely the highest profile lawsuit ongoing across the U.S. in these regards, it goes without saying that teachers and principals would be hard pressed to do the same. Likewise, it is also important to note that for purposes of this study, and researchers' understandings of the VAM in use throughout New Mexico, researchers also searched for additional information, especially about the state's VAM. These efforts also yielded no results that were of more "added value" than that which was already included in the documents submitted as part of this particular case.

Finally, in terms of consequential validity, perhaps most important was that regardless of the fact that the 2013–2014 school year was supposed to be a "hold harmless" year, the state flagged teachers' files regardless with their value-added or overall effectiveness categories (e.g., teachers placed on professional improvement plans and teachers being denied teaching positions, or rather making horizontal moves within the state given their flagged files), even after the state realized and admitted that many teachers' VAM scores (and therefore overall scores) were miscalculated. These are the very set of consequences that landed the state in court. Once the case was heard, however, other issues related to primarily unintended consequences came to bear (e.g., frustrations, as well as threats to validity caused by gross data errors; general feelings of confusion, frustration, anger, hopelessness, and fear; lack of transparency, which lead to lack of formative use and subsequent improvement; reduced teacher morale and buy-in, as well as administrator buy-in; and teacher and administrator resignations, transfers, and departures).

8 Implications

Again, these findings are of importance not only for others who are still grappling with the use of VAMs to evaluated and "better" hold teachers accountable for that which they do or do not do well, these findings are also of importance should others continue to adopt, implement, or even consider the adoption of VAMs for similar goals and purposes (e.g., as also often endorsed by the OECD). As alluded to prior, while VAM output might be statistically sophisticated, this study's findings can serve as illustrative evidence for what can occur if states or districts continue to value complex and complicated data and metrics (e.g., VAMs) over professional judgment and pragmatic concerns. While VAM output and similar data can potentially provide valuable information about teachers (Lingard 2011), provided technical errors are kept to a minimum, such "objectivity" should never be prioritized at the expense of the professional judgments and contextual factors that affect measurement output.

In addition, it is important to remind others that the *Standards* (AERA et al. 2014) exist and should be used for their intended purposes and reasons. As relevant to this study and teacher evaluation more broadly, the *Standards* should be used to help others develop (and evaluate) test-based systems, such as the VAM-based systems still being used to evaluate teachers, post-ESSA (2016; see Close et al 2020), and hold them accountable for their effects. These standards, again, pertain to reliability (*Standards* 2.0, 2.2, 2.4, 2.7, 2.15, 2.19), validity (*Standards* 1.0, 1.1, 1.3, 1.4, 1.25), bias (*Standards* 3.0, 3.2, 3.6, 3.7, 3.16), and use, as well as consequential use in practice (*Standards* 13.2, 13.3, 13.4, 13.5, 13.6, 13.8, 13.9).

While this study focused on one U.S. state's teacher evaluation system, it can be used as an example for understanding the logics and instruments that have permeated much of global education governance more broadly. The U.S. serves as a particularly interesting case in this regard due to its steadfast approach to using statistical measurement to assess teacher quality and hold teachers accountable, as evidenced throughout the past several decades of U.S. policy history.

As the collection, analysis, reporting, and acting upon "objective" data surrounding most aspects of education becomes increasingly normal practice across education systems, and as the "trust in numbers" (Ozga 2016; Porter 1996; see also Amrein-Beardsley and Collins 2012) continues to be solidified as a meaningful practice, it is worth seeing how different countries navigate and fare in this "data deluge" (Anderson 2008; Grek 2009; Selwyn 2015). While the logic of value-added measurement has been extensively critiqued—philosophically, pragmatically, and empirically (Grek and Ozga 2010; Lingard et al. 2013; Mathis 2011; Nichols and Berliner 2007; Timar and Maxwell-Jolly 2012; see also Denby 2012; Mathews 2013; Pauken 2013)—the legal dimension of teacher evaluation, and especially that associated with VAMs, is still a relatively new area of study (see Paige 2016). This is a cautionary tale for educational systems and policymakers, in the U.S. and abroad, as they consider options for VAM-based policy or other test-based accountability policies and practices more generally.

9 Limitations and further research

With that being said, this study also had its limitations. First and foremost, researchers only investigated and analyzed one state's teacher evaluation system and its related policies. As previously discussed, while results from this study are likely relevant across multiple geographical, institutional, and policy contexts, readers should be careful to not blindly generalize findings. Additionally, researchers were limited in the analyses they could conduct given the quality and structure of the teacher evaluation data they were provided by the NMPED.

Finally, as also described prior, the ways in which VAM researchers have approached evidences of validity, in reductionistic versus more holistic ways as per current validity theory, is also a limitation of this study. Again, the only validity evidences that researchers reviewed, analyzed, and presented herein were consistent with the concepts related to capturing convergent evidences of validity, independent of and as not in line with current conceptualizations of validity (see, for example, Cizek 2015; Markus 2016; Kane 2013, 2016; Newton and Shaw 2016; see also AERA et al. 2014). It is not that researchers have a limited view of validity, although this could certainly be true given that researchers of this study are also engrained in the current literature surrounding VAMs in which convergent-related validity evidence is valued above and beyond other evidences (e.g., consequence-related evidences of validity), it is that in this area of research, convergent evidence of validity is what is being disproportionally valued, examined, and reported in the literature, as unfortunate and overly simplistic as the case may be. With that being said, one area of future research, especially in order to capture VAM-related validity (perhaps, with a capital V), would be for validity theorists and experts to analyze all validity evidences surrounding VAMs, perhaps using a framework similar to the one researchers used herein but also more advanced and in line with current theory and practice, to ultimately better situate and help others better understand the uses and inferences drawn from VAMs.

A related extension of this research would be to conceptualize or build a better teacher evaluation system, or system options, whether reliant upon or even inclusive of a VAM component, that aligns with current validity theory and/or *The Standards* (AERA et al. 2014) themselves. This would be especially important should teachers within or beyond U.S. borders continue to be held accountable using high-stakes consequential decisions, as tied to whatever teacher evaluation measurement tools might be used. Developed at minimum, perhaps, could be a user-friendly set of teacher evaluation guidelines, so as to introduce those often tasked with such design obligations with the key measurement language, definitions, criteria, and requirements pertinent to any such measurement and evaluation system.

10 Conclusions

Perhaps, then, it is appropriate to close with some of the recommendations advanced by the ASA—not just America's but the world's largest community of statisticians. In their 2014 "ASA Statement on Using Value-Added Models [VAMs] for Educational Assessment," the ASA wrote, as pertinent in New Mexico and states and nations elsewhere, that "Estimates from VAMs should always be accompanied by measures of precision…These limitations are particularly relevant if VAMs are used for high-stakes purposes" (p. 1). This relates to reliability as studied herein. The ASA also asserted that "VAMs are generally based on standardized test scores, and do not directly measure potential teacher contributions toward [these and] other student outcomes" (p. 2). This was also discussed prior in terms of validity, with the point being that never have the test scores being used across states for said purposes been validated to measure teachers' contributions to student achievement over time. In addition, whether the correlations between VAM and other teacher evaluation indicators yield adequate convergent-related evidence of validity should also be of interest.

Related, "VAMs typically measure correlation, not causation: Effects – positive or negative –attributed to a teacher may actually be caused by other factors that are not captured in the model" (ASA 2014, p. 2). Related, "[m]ost VAM studies find that teachers account for about 1% to 14% of the variability in test scores" (p. 2), which means that these other out-of-school variables indeed have a potential for a biasing impact. This, of course, pertains to bias as also discussed and detailed with evidence.

With regard to fairness, the ASA noted that "[a]ttaching too much importance to a single item of quantitative information is counter-productive—in fact, it can be detrimental to the goal of improving quality" (p. 5). In this case, the growth measure counted for at least 50% of teachers' overall evaluation scores, and the weight of the VAM measure could carry up to 100% of the weight when it trumped teachers' other evaluative measures if other measures contradicted the VAM. This, as per the ASA, is also inappropriate and unfair. With regard to transparency, the ASA noted that "[w]hen used appropriately, VAMs may provide quantitative information that is relevant for improving education processes" (p. 5), but a primary pre-condition for this is that the VAM and VAM output are accessible, understandable, and able "to provide meaningful information [to facilitate] a teacher's ability to promote student learning" (p. 5).

In terms of the lawsuit under investigation, as it pertains to the measurement concepts investigated in this study, a New Mexico state judge ultimately 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 were warranted and nonarbitrary. The state was also to present to the court that the system was legally defensible as well as "uniform and objective" as per state constitutional requirements. The trial was on hold for years in that the state never presented to the court this evidence; hence, the preliminary injunction held until the state of New Mexico's new governor was sworn in (2019), after which she signed an Executive Order for the entire state's teacher evaluation system to be amended. The state's teacher evaluation system was to no longer involve any value-added data to evaluate New Mexico's teachers.

Appendix 1. Demographic variables

Table 4 Demographics used in the analyses

Teacher gender	Male, female
Teacher ethnicity	Asian, African American, Caucasian, Hispanic, Native American
Teacher ethnicity	White, Non-White
Teacher years of experience*	0–2 years, 3–8 years, 9–15 years, 16+ years
Grade level**	Elementary, middle, or high school
ELL teacher	Yes, no
SPED teacher	Yes, no
Gifted teacher	Yes, no
Student enrollment	Low (4–502 students)
	High (503–2389 students)
Student SPED school population	Low (0.00–13.99% of all students)
	High (14.00–100.00% of all students)
Student ELL school population***	Low (0.00–10.69% of all students)
	High (10.70–85.20% of all students)
Student FRL school population	Low (0.00–74.49% of all students)
	High (74.50–100.00% of all students)
Student gifted school population***	Low (0.00–3.49% of all students)
	High (3.50–44.44% of all students)
Student minority school population***	Low (5.60–78.29% of all students)
	High (78.30–100.00% of all students)

*These categories were calculated by determining quartiles as the year of experience variable in the initial dataset provided by the state was continuous

**This data was only available for 2013–2014 and 2014–2015; thus, grade level analyses were not conducted for the 2015–2016 school year

***The low end of the "low" range and the high end of the "high" range represents the lowest and highest proportions in the dataset, respectively; as such, not all low ends of the range are 0.00% and not all high ends of the range are 100.00%

Appendix 2. Deviations in scores over time

Table 5 Teachers' variations in score quintiles over time

	VAM scores				Observation scores			
	Year 1 to year 2		Year 2 to year 3		Year 1 to year 2		Year 2 to year 3	
	%	*n*	%	*n*	%	*n*	%	*n*
No variation	31.6	2455	31.6	2454	39.0	3031	41.3	3210
One quintile variation	40.6	3157	39.4	3060	39.7	3087	39.7	3084
Two quintile variation	20.1	1564	19.9	1547	15.8	1230	13.6	1061
Three quintile variation	6.0	465	7.3	570	4.7	366	4.4	343
Four quintile variation	1.7	130	1.7	135	0.8	63	1.0	79
Total	100.0	7771	100.0	7766	100.0	7777	100.0	7777
	PPP scores				Student survey scores			
	Year 1 to year 2		Year 2 to year 3		Year 1 to year 2		Year 2 to year 3	
	%	*n*	%	*n*	%	*n*	%	*n*
No variation	37.2	2298	41.2	2369	33.5	778	41.5	894
One quintile variation	39.9	2468	36.1	2079	38.3	889	39.1	844
Two quintile variation	15.6	964	15.4	889	16.2	375	13.8	297
Three quintile variation	6.3	389	6.0	343	9.0	208	5.0	108
Four quintile variation	1.0	63	1.3	76	3.1	71	0.7	14
Total	100.0	6182	100.0	5756	100.0	2321	100.0	2157

Appendix 3. Means of teacher effectiveness measures, per teacher and school subgroups

Table 6 VAM means, per teacher subgroup

	Year 1	Year 2	Year 3
Gender			
Male	27.69	37.42	42.67
Female	25.94	38.24	47.29
Race/ethnicity			
Asian	27.25	36.18	45.53
African American	23.71	34.96	42.24
Caucasian	26.64	38.30	46.56
Hispanic	26.18	38.12	45.64
Native American	22.06	31.63	46.75
Non-Caucasian	25.94	37.65	45.60
Years of experience			
0–2 years	22.82	31.37	40.52
3–8 years	24.24	35.03	45.51
9–15 years	27.42	40.39	47.56
16+ years	28.02	46.26	47.27
Grade level			
Elementary school	21.91	37.27	N/A
Middle school	29.54	37.75	N/A
High school	30.87	39.48	N/A
ELL teachers			
Yes	24.67	35.48	44.88
No	26.41	38.10	46.20
SPED teachers			
Yes	25.86	36.20	40.62
No	26.40	38.19	46.73
Gifted teachers			
Yes	32.66	43.06	43.63
No	26.21	37.89	46.24

Teacher grade level assignment was not provided for year 3, so disaggregating scores by grades taught was not possible

Table 7 Observation score means, per teacher subgroup

	Year 1	Year 2	Year 3
Gender			
Male	0.663	0.682	0.709
Female	0.679	0.708	0.738
Race/ethnicity			
Asian	0.640	0.665	0.691
African American	0.658	0.686	0.715
Caucasian	0.680	0.707	0.734
Hispanic	0.670	0.697	0.729
Native American	0.634	0.651	0.389
Non-Caucasian	0.667	0.693	0.726
Years of experience			
0–2 years	0.662	0.682	0.712
3–8 years	0.669	0.700	0.730
9–15 years	0.678	0.704	0.735
16+ years	0.682	0.707	0.733
Grade level			
Elementary school	0.677	0.705	N/A
Middle school	0.671	0.697	N/A
High school	0.676	0.701	N/A
ELL teachers			
Yes	0.657	0.680	0.710
No	0.676	0.702	0.732
SPED teachers			
Yes	0.658	0.680	0.709
No	0.677	0.704	0.733
Gifted teachers			
Yes	0.708	0.751	0.764
No	0.674	0.700	0.730

Teacher grade level assignment was not provided for year 3, so disaggregating scores by grades taught was not possible

Table 8 PPP score means, per teacher subgroup

	Year 1	Year 2	Year 3
Gender			
Male	0.676	0.708	0.713
Female	0.700	0.744	0.757
Race/ethnicity			
Asian	0.673	0.714	0.704
African American	0.673	0.719	0.726
Caucasian	0.702	0.743	0.753
Hispanic	0.685	0.726	0.740
Native American	0.647	0.682	0.689
Non-Caucasian	0.682	0.723	0.736
Years of experience			
0–2 years	0.677	0.707	0.727
3–8 years	0.689	0.732	0.745
9–15 years	0.697	0.738	0.751
16+ years	0.701	0.745	0.748
Grade level			
Elementary school	0.693	0.734	N/A
Middle school	0.692	0.738	N/A
High school	0.698	0.734	N/A
ELL teachers			
Yes	0.667	0.712	0.737
No	0.695	0.736	0.746
SPED teachers			
Yes	0.679	0.715	0.726
No	0.696	0.737	0.748
Gifted teachers			
Yes	0.728	0.777	0.786
No	0.693	0.734	0.745

Teacher grade level assignment was not provided for year 3, so disaggregating scores by grades taught was not possible

Table 9 Survey score means, per teacher subgroup

	Year 1	Year 2	Year 3
Gender			
Male	0.763	0.793	0.792
Female	0.784	0.824	0.816
Race/ethnicity			
Asian	0.732	0.747	0.766
African American	0.760	0.790	0.798
Caucasian	0.777	0.807	0.803
Hispanic	0.786	0.831	0.822
Native American	0.767	0.802	0.805
Non-Caucasian	0.782	0.825	0.819
Years of experience			
0–2 years	0.770	0.815	0.803
3–8 years	0.788	0.828	0.819
9–15 years	0.780	0.820	0.809
16+ years	0.776	0.805	0.806
Grade level			
Elementary school	0.803	0.863	N/A
Middle school	0.768	0.795	N/A
High school	0.756	0.781	N/A
ELL teachers			
Yes	0.803	0.860	0.863
No	0.778	0.813	0.808
SPED teachers			
Yes	0.772	0.825	0.829
No	0.780	0.815	0.808
Gifted teachers			
Yes	0.773	0.789	0.795
No	0.779	0.816	0.810

Teacher grade level assignment was not provided for year 3, so disaggregating scores by grades taught was not possible

Table 10 VAM means, per school subgroup

	Year 1	Year 2	Year 3
Total enrollment			
High	26.81	39.30	44.96
Low	25.82	36.73	47.36
Student SPED population			
High	25.04	37.06	45.38
Low	27.57	38.99	46.89
Student ELL population			
High	24.30	36.77	44.29
Low	28.19	39.22	47.85
Student FRL population			
High	25.10	36.03	43.68
Low	27.47	39.93	48.45
Student gifted population			
High	26.90	39.50	46.05
Low	25.67	36.39	46.25
Student minority population			
High	25.09	36.18	42.33
Low	27.49	39.79	49.73

Table 11 Observation score means, per school subgroup

	Year 1	Year 2	Year 3
Total enrollment			
High	0.677	0.701	0.727
Low	0.673	0.702	0.734
Student SPED population			
High	0.671	0.697	0.728
Low	0.678	0.706	0.734
Student ELL population			
High	0.663	0.690	0.721
Low	0.686	0.712	0.740
Student FRL population			
High	0.661	0.690	0.722
Low	0.688	0.713	0.739
Student gifted population			
High	0.679	0.706	0.732
Low	0.670	0.697	0.729
Student minority population			
High	0.660	0.686	0.718
Low	0.689	0.717	0.743

Table 12 PPP score means, per school subgroup

	Year 1	Year 2	Year 3
Total enrollment			
High	0.700	0.733	0.742
Low	0.688	0.738	0.750
Student SPED population			
High	0.691	0.729	0.742
Low	0.697	0.742	0.749
Student ELL population			
High	0.680	0.719	0.735
Low	0.707	0.750	0.756
Student FRL population			
High	0.675	0.721	0.735
Low	0.712	0.749	0.755
Student gifted population			
High	0.702	0.738	0.747
Low	0.685	0.732	0.744
Student minority population			
High	0.678	0.718	0.729
Low	0.709	0.752	0.761

Table 13 Survey score means, per school subgroup

	Year 1	Year 2	Year 3
Total enrollment			
High	0.770	0.806	0.794
Low	0.790	0.824	0.825
Student SPED population			
High	0.781	0.820	0.814
Low	0.779	0.813	0.806
Student ELL population			
High	0.780	0.819	0.823
Low	0.779	0.812	0.798
Student FRL population			
High	0.786	0.817	0.820
Low	0.774	0.815	0.801
Student gifted population			
High	0.775	0.793	0.791
Low	0.784	0.830	0.829
Student minority population			
High	0.777	0.810	0.812
Low	0.782	0.822	0.807

References

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

American Educational Research Association (AERA) Council. (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. *Educational Researcher, 44*(8), 448–452. https://doi.org/10.3102/0013189X15618385 Retrieved from http://edr.sagepub.com/content/early/2015/11/10/0013189X15618385.full.pdf+htmls.

American Statistical Association (ASA). (2014). *ASA statement on using value-added models for educational assessment*. Alexandria, VA: ASA Retrieved from https://www.amstat.org/asa/files/pdfs/POL-ASAVAM-Statement.pdf.

Amrein-Beardsley, A., & Collins, C. (2012). The SAS Education Value-Added Assessment System (SAS® EVAAS®) in the Houston Independent School District (HISD): Intended and unintended consequences. *Education Policy Analysis Archives, 20*(12). https://doi.org/10.14507/epaa.v20n12.2012.

Amrein-Beardsley, A., & Geiger, T. J. (2019). Potential sources of invalidity when using teacher value-added and principal observational estimates: Artificial inflation, deflation, and conflation. *Educational Assessment, Evaluation and Accountability, 31*(4), 465–493. https://doi.org/10.1007/s11092-019-09311-w.

Anderson, C. (2008). The end of theory: the data deluge makes the scientific method obsolete. Wired. Retrieved from https://www.wired.com/2008/06/pb-theory/.

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., & Schady, N. (2016). Teacher quality and learning outcomes in kindergarten. *The Quarterly Journal of Economics, 131*(3), 1415–1453. https://doi.org/10.1093/qje/qjw016.

Bailey, J., Bocala, C., Shakman, K., & Zweig, J. (2016). Teacher demographics and evaluation: a descriptive study in a large urban district. Washington DC: U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/edlabs/regions/northeast/pdf/REL_2017189.pdf

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R. J., & 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.

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). https://doi.org/10.14507/epaa.v21n5.2013.

Ballou, D. (2004). Rejoinder. *Journal of Educational and Behavioral Statistics, 29*(1), 131–134. https://doi.org/10.3102/10769986029001131.

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*(2), 77–86. https://doi.org/10.3102/0013189X15574904.

Braun, H. I. (2004). *Value-added modeling: what does due diligence require?* Princeton, NJ: Educational Testing Service.

Braun, H. (2015). The value in value-added depends on the ecology. *Educational Researcher, 44*(2), 127–131. https://doi.org/10.3102/0013189X15576341.

Briggs, D. C. & Betebenner, D. (2009). *Is growth in student achievement scale dependent?* Paper presented at the annual meeting of the National Council for Measurement in Education (NCME), San Diego, CA.

Briggs, D. & Domingue, B. (2011). Due diligence and the evaluation of teachers: a review of the value-added analysis underlying the effectiveness rankings of Los Angeles Unified School District Teachers by the *Los Angeles Times*. Boulder, CO: National Education Policy Center (NEPC). http://nepc.colorado.edu/files/NEPC-LAT-VAM-2PP.pdf

Brophy, J. (1973). Stability of teacher effectiveness. *American Educational Research Journal, 10*(3), 245–252. https://doi.org/10.2307/1161888.

Burgess, K. (2017, July 6). Expert: NM teacher evals are toughest in the nation. *The Albuquerque Journal.* Retrieved from https://www.abqjournal.com/1029370/expert-nm-teacher-evals-toughest-in-us.html

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 (NEPC). 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*(2), 178–185. https://doi.org/10.1177/001041407500800204.

Cantrell, S., & Kane, T. J. (2013). Ensuring fair and reliable measures of effective teaching: culminating findings from the MET project's three-year study. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from https://k12education.gatesfoundation.org/resource/ensuring-fair-and-reliable-measures-of-effective-teaching-culminating-findings-from-the-met-projects-three-year-study/.

Carey, K. (2017 19). 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

Chester, M. D. (2003). Multiple measures and high-stakes decisions: a framework for combining measures. *Educational Measurement: Issues and Practice, 22*(2), 32–41. https://doi.org/10.1111/j.1745-3992.2003.tb00126.x.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: evaluating bias in teacher value-added estimates. *American Economic Review, 104*(9), 2593–2632. https://doi.org/10.3386/w19424.

Cizek, G. J. (2016). Validating test score meaning and defending test score use: different aims, different methods. *Assessment in Education: Principles, Policy & Practice, 23*(2), 212–225. https://doi.org/10.1080/0969594X.2015.1063479.

Close, K., Amrein-Beardsley, A., & Collins, C. (2020). Putting teacher evaluation systems on the map: An overview of states' teacher evaluation systems post-Every Student Succeeds Act. *Education Policy Analysis Archives, 28*(1), 1–58. https://doi.org/10.14507/epaa.28.5252.

Cody, C. A., McFarland, J., Moore, J. E., & Preston, J. (2010). The evolution of growth models. Raleigh, NC: Public Schools of North Carolina. Retrieved from http://www.dpi.state.nc.us/docs/intern-research/reports/growth.pdf.

Cole, R., Haimson, J., Perez-Johnson, I., & May, H. (2011). *Variability in pretest-posttest correlation coefficients by student achievement level*. Washington, D.C.: U.S. Department of Education Retrieved from https://ies.ed.gov/ncee/pubs/20114033/pdf/20114033.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*(98). doi: https://doi.org/10.14507/epaa.v22.1594.

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 http://annenberginstitute.org/publication/can-teachers-be-evaluated-their-students%E2%80%99-test-scores-should-they-be-use-value-added-mea.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957.

Darling-Hammond, L. (2010). *The flat world and education: how America's commitment to equity will determine our future*. New York, NY: Teachers College Press.

Darling-Hammond, L. (2015). Can value-added add value to teacher evaluation? *Educational Researcher, 44*(2), 132–137. https://doi.org/10.3102/0013189X15575346.

Denby, D. (2012). Public defender: Diane Ravitch takes on a movement. The New Yorker, Annals of Education. Retrieved from https://www.newyorker.com/magazine/2012/11/19/public-defender.

Doherty, K. M., & Jacobs, S. (2015). *State of the states 2015: Evaluating teaching, leading and learning*. Washington DC: National Council on Teacher Quality (NCTQ) 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.

Dunn, O. J., & Clark, V. A. (1969). Correlation coefficients measured on the same individuals. *Journal of the American Statistical Association, 64*(325), 366–377. https://doi.org/10.2307/228374625.

Dunn, O. J., & Clark, V. A. (1971). Comparison of tests of the equality of dependent correlation coefficients. *Journal of the American Statistical Association, 66*(336), 904–908. https://doi.org/10.2307/2284252.

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

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

Every Student Succeeds Act (ESSA) of 2016, Pub. L. No. 114–95, § 129 Stat. 1802. (2016). Retrieved from https://www.gpo.gov/fdsys/pkg/BILLS-114s1177enr/pdf/BILLS-114s1177enr.pdf

Flyvbjerg, B. (2011). Five misunderstandings about case-study research. *Qualitative Inquiry, 12*(2), 219–245. https://doi.org/10.1177/1077800405284363.

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). https://doi.org/10.14507/v21n9.2013.

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*(1), 117. https://doi.org/10.1186/1471-2288-13-117.

Gerring, J. (2004). What is a case study and what is it good for? *The American Political Science Review, 98*(2), 341–354. https://doi.org/10.1017/S0003055404001182.

Glazerman, S. M., & Potamites, L. (2011). False performance gains: a critique of successive cohort indicators. Mathematica Policy Research. Retrieved from https://www.mathematica.org/~/media/publications/PDFs/education/false_perf.pdf.

Glazerman, S., Goldhaber, D., Loeb, S., Raudenbush, S., Staiger, D. O., & Whitehurst, G. J. (2011). *Passing muster: evaluating teacher evaluation systems*. Washington, D.C.: The Brookings Institution www.brookings.edu/reports/2011/0426_evaluating_teachers.aspx.

Goldhaber, D. (2015). Exploring the potential of value-added performance measures to affect the quality of the teacher workforce. *Educational Researcher, 44*(2), 87–95. https://doi.org/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*(1), 8–34. https://doi.org/10.1080/19345747.2014.978059.

Goldhaber, D. D., Goldschmidt, P., & Tseng, F. (2013). Teacher value-added at the high-school level: different models, different answers? *Educational Evaluation and Policy Analysis, 35*(2), 220–236. https://doi.org/10.3102/0162373712466938.

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*(2), 96–104. https://doi.org/10.3102/0013189X15575031.

Goldschmidt, P., Choi, K., & Beaudoin, J. B. (2012). Growth model comparison study: practical implications of alternative models for evaluating school performance. Technical issues in large-scale assessment state collaborative on assessment and student standards. Council of Chief State School Officers. Retrieved from https://files.eric.ed.gov/fulltext/ED542761.pdf.

Graue, M. E., Delaney, K. K., & Karch, A. S. (2013). Ecologies of education quality. *Education Policy Analysis Archives, 21*(8). https://doi.org/10.14507/epaa.v21n8.2013.

Grek, S. (2009). Governing by numbers: the PISA ‘effect’ in Europe. *Journal of Education Policy, 24*(1), 23–37. https://doi.org/10.1080/02680930802412669.

Grek, S., & Ozga, J. (2010). Re-inventing public education: the new role of knowledge in education policy making. *Public Policy and Administration, 25*(3), 271–288. https://doi.org/10.1177/0952076709356870.

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*(6), 293–303. https://doi.org/10.3102/0013189X14544542.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2012). Can value-added measures of teacher education performance be trusted? East Lansing, MI: The Education Policy Center at Michigan State University. Retrieved from http://education.msu.edu/epc/library/documents/WP18Guarino-Reckase-Wooldridge-2012-Can-Value-Added-Measures-of-Teacher-Performance-Be-T_000.pdf

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, MI: The Education Policy Center at Michigan State University.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17–27. https://doi.org/10.1111/j.1745-3992.2004.tb00149.x.

Harris, D. N. (2009). Would accountability based on teacher value added be smart policy? An evaluation of the statistical properties and policy alternatives. *Education Finance and Policy, 4*(4), 319–350. https://doi.org/10.1162/edfp.2009.4.4.319.

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*(2), 71–76. https://doi.org/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*(3), 794–831. https://doi.org/10.3102/0002831210387916.

Ho, A. D., Lewis, D. M., & Farris, J. L. (2009). The dependence of growth-model results on proficiency cut scores. *Educational Measurement: Issues and Practice, 28*(4), 15–26. https://doi.org/10.1111/j.1745-3992.2009.00159.x.

Holloway-Libell, J. (2015). Evidence of grade and subject-level bias in value-added measures. *Teachers College Record*, 117.

Ishii, J., & Rivkin, S. G. (2009). Impediments to the estimation of teacher value added. *Education Finance and Policy, 4*(4), 520–536. https://doi.org/10.1162/edfp.2009.4.4.520.

Jiang, J. Y., Sporte, S. E., & Luppescu, S. (2015). Teacher perspectives on evaluation reform: Chicago's REACH students. *Educational Researcher, 44*(2), 105–116. https://doi.org/10.3102/0013189X15575517.

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

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

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice, 23*(2), 198–211. https://doi.org/10.1080/0969594X.2015.1060192.

Kane, M. T. (2017). *Measurement error and bias in value-added models*. Princeton: Educational Testing Service (ETS) Research Report Series. https://doi.org/10.1002/ets2.12153 Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/ets2.12153/full.

Kane, T. J., & Staiger, D. (2012). Gathering feedback for teaching: combining high-quality observations with student surveys and achievement gains. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from https://files.eric.ed.gov/fulltext/ED540960.pdf.

Kappler Hewitt, K. (2015). Educator evaluation policy that incorporates EVAAS value-added measures: undermined intentions and exacerbated inequities. *Education Policy Analysis Archives, 23*(76). doi: https://doi.org/10.14507/epaa.v23.1968.

Kelly, A., & Downey, C. (2010). Value-added measures for schools in England: looking inside the "black box" of complex metrics. *Educational Assessment, Evaluation and Accountability, 22*(3), 181–198. https://doi.org/10.1007/s11092-010-9100-4.

Kelly, S., & Monczunski, L. (2007). Overcoming the volatility in school-level gain scores: a new approach to identifying value added with cross-sectional data. *Educational Researcher, 36*(5), 279–287. https://doi.org/10.3102/0013189X07306557.

Koedel, C., & Betts, J. R. (2007). *Re-examining the role of teacher quality in the educational production function* (working paper no. 2007-03). Nashville, TN: National Center on Performance Initiatives.

Koedel, C., & Betts, J. R. (2009). *Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique*. Nashville, TN: National Center on Performance Incentives. Retrieved from. https://doi.org/10.1162/EDFP_a_00027?journalCode=edfp.

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

Koretz, D. (2017). *The testing charade: pretending to make schools better*. Chicago, IL: University of Chicago Press.

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting the widget effect: teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher, 46*(5), 234–249. https://doi.org/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. Annual Meeting of the American Educational Research Association (AERA), Toronto, Canada. (Presidential Session)

Lingard, B. (2011). Policy as numbers: ac/counting for educational research. *The Australian Educational Researcher, 38*(4), 355–382. https://doi.org/10.1007/s13384-011-0041-9.

Lingard, B., Martino, W., & Rezai-Rashti, G. (2013). Testing regimes, accountabilities and education policy: commensurate global and national developments. *Journal of Education Policy, 28*(5), 539–556. https://doi.org/10.1080/02680939.2013.820042.

Linn, R. L., & Haug, C. (2002). Stability of school-building accountability scores and gains. *Educational Evaluation and Policy Analysis, 24*, 29–36. https://doi.org/10.3102/01623737024001029.

Markus, K. A. (2016). Alternative vocabularies in the test validity literature. *Assessment in Education: Principles, Policy & Practice, 23*(2), 252–267. https://doi.org/10.1080/0969594X.2015.1060191.

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*(4), 738–756. https://doi.org/10.3102/0162373716666166.

Mathis, W. J. (2011). NEPC review: *Florida formula for student achievement: Lessons for the nation*. Boulder, CO: National Education Policy Center. Retrieved from https://nepc.colorado.edu/thinktank/review-florida-formula.

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*(1), 67–101.

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

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 and 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.

Michelmore, K., & Dynarski, S. (2016). The gap within the gap: using longitudinal data to understand income differences in student achievement. Cambridge, MA: National Bureau of Economic Research (NBER). Retrieved from http://www.nber.org/papers/w22474.

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

Moore Johnson, S. (2015). Will VAMS reinforce the walls of the egg-crate school? *Educational Researcher, 44*(2), 117–126. https://doi.org/10.3102/0013189X15573351.

New Mexico Public Education Department. (2016). *NMTEACH technical guide. Business rules and calculations. 2015-2016*. Santa Fe, NM: Author.

Newton, P. E., & Shaw, S. D. (2016). Disagreement over the best way to use the word 'validity' and options for reaching consensus. *Assessment in Education: Principles, Policy & Practice, 23*(2), 178–197. https://doi.org/10.1080/0969594X.2015.1037241.

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). doi: https://doi.org/10.14507/epaa.v18n23.2010.

Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage: how high-stakes testing corrupts America's schools*. Cambridge, MA: Harvard Education Press.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Paige, M. A. (2016). *Building a better teacher: understanding value-added models in the law of teacher evaluation*. Lanham, MD: Rowman & Littlefield.

Papay, J. P. (2010). Different tests, different answers: the stability of teacher value-added estimates across outcome measures. *American Educational Research Journal., 48*, 163–193. https://doi.org/10.3102/0002831210362589.

Paufler, N. A., & Amrein-Beardsley, A. (2014). The random assignment of students into elementary classrooms: Implications for value-added analyses and interpretations. *American Educational Research Journal (AERJ), 51*(2), 328–362. https://doi.org/10.3102/0002831213508299.

Pauken, T. (2013). Texas vs. No Child Left Behind. The American Conservative. Retrieved from https://www.theamericanconservative.com/articles/texas-vs-no-child-left-behind/.

Polat, N., & Cepik, S. (2015). An exploratory factor analysis of the Sheltered Instruction Observation Protocol as an evaluation tool to measure teaching effectiveness. *TESOL Quarterly, 50*(4), 817–843. https://doi.org/10.1002/tesq.248.

Polikoff, M. S., & Porter, A. C. (2014). Instructional alignment as a measure of teaching quality. *Educational Evaluation and Policy Analysis, 36*(4), 399–416. https://doi.org/10.3102/0162373714531851.

Porter, T. M. (1996). *Trust in numbers: the pursuit of objectivity in science and public life*. Princeton, NJ: Princeton University Press.

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: The Press Syndicate of The University of Cambridge.

Raudenbush, S. W. & Jean, M. (2012). How should educators interpret value-added scores? Stanford, CA: Carnegie Knowledge Network. Retrieved from http://www.carnegieknowledgenetwork.org/briefs/value-added/interpreting-value-added/.

Reiss, R. (2017). A vindication of the criticism of New Mexico Public Education Department's teacher evaluation system. *The Beacon, XX*(1), 2-4. Retrieved from http://www.cese.org/wp-content/uploads/2017/05/2017-05-Beacon.pdf.

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.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417–458. https://doi.org/10.1111/j.1468-0262.2005.00584.x.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: evidence from panel data. *The American Economic Review, 94*(2), 247–252. https://doi.org/10.1257/0002828041302244.

Rosenbaum, P., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55. https://doi.org/10.2307/2335942.

Ross, E. & Walsh, K. (2019). *State of the states 2019: teacher and principal evaluation policy.* Washington, DC: National Council on Teacher Quality (NCTQ). Retrieved from https://www.nctq.org/pages/State-of-the-States-2019:-Teacher-and-Principal-Evaluation-Policy#footnote-15.

Rothstein, J. (2009). *Student sorting and bias in value-added estimation: selection on observables and unobservables.* Cambridge, MA: National Bureau of Economic Research (NBER). Retrieved from http://www.nber.org/papers/w14666.pdf.

Rothstein, J. (2010). Teacher quality in educational production: tracking, decay, and student achievement. *Quarterly Journal of Economics, 125*(1), 175–214. https://doi.org/10.1162/qjec.2010.125.1.175.

Rothstein, J. (2017). *Revisiting the impacts of teachers (working paper)*. Berkeley, CA: University of California, Berkeley Retrieved from https://eml.berkeley.edu/~jrothst/CFR/rothstein_cfr_workingpaper_jan2017.pdf.

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

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. (2019). SAS EVAAS for K-12. Retrieved from http://www.sas.com/en_us/industry/k-12-education/evaas.html.

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. https://doi.org/10.3102/1076998611432174.

Selwyn, N. (2015). Data entry: towards the critical study of digital data and education. *Learning, Media, and Technology, 40*(1), 64–82. https://doi.org/10.1080/17439884.2014.921628.

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

Sloat, E. F. (2015). *Examining the validity of a state policy-directed framework for evaluating teacher instructional quality: informing policy, impacting practice* (Unpublished doctoral dissertation). Arizona State University, Tempe, AZ.

Sloat, E., Amrein-Beardsley, A., & Sabo, K. E. (2017). Examining the factor structure underlying the TAP System for Teacher and Student Advancement. *AERA Open, 3*(4), 1–18. https://doi.org/10.1177/2332858417735526.

Sloat, E., Amrein-Beardsley, A., & Holloway, J. (2018). Different teacher-level effectiveness estimates, different results: Inter-model concordance across six generalized value-added models (VAMs). *Educational Assessment, Evaluation and Accountability, 30*(4), 367–397. https://doi.org/10.1007/s11092-018-9283-7.

Smith, W. C., & Kubacka, K. (2017). The emphasis of student test scores in teacher appraisal systems. *Educational Policy Analysis Archives, 25*(86). doi: https://doi.org/10.14507/epaa.25.2889.

Sørensen, T. B. (2016). *Value-added measurement or modelling (VAM)*. Brussels: Education International Retrieved from http://download.ei-ie.org/Docs/WebDepot/2016_EI_VAM_EN_final_Web.pdf.

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.

Steinberg, M. P., & Garrett, R. (2016). Classroom composition and measured teacher performance: what do teacher observation scores really measure? *Educational Evaluation and Policy Analysis, 38*(2), 293–317. https://doi.org/10.3102/0162373715616249.

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

Thomas, G. (2011). A typology for the case study in social science following a review of definition, discourse, and structure. *Qualitative Inquiry, 17*(6), 511–521. https://doi.org/10.1177/1077800411409884.

Timar, T. B., & Maxwell-Jolly, J. (Eds.). (2012). *Narrowing the achievement gap: perspectives and strategies for challenging times*. Cambridge, MA: Harvard Education Press.

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

U.S. Department of Education. (2012). Elementary and Secondary Education Act (ESEA) flexibility. Washington, D.C.: Retrieved from https://www.ed.gov/esea/flexibility

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 D.C. Retrieved from http://www.ed.gov/news/press-releases/states-granted-waivers-no-child-left-behind-allowed-reapply-renewal-2014-and-2015-school-years.

VanWynsberghe, R., & Khan, S. (2007). Redefining case study. *International Journal of Qualitative Methods, 6*(2), 80–94.

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*(6), 1834–1868. https://doi.org/10.3102/0002831216671864.

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 (TNTP). Retrieved from http://tntp.org/assets/documents/TheWidgetEffect_2nd_ed.pdf.

Whitehurst, G. J., Chingos, M. M., & Lindquist, K. M. (2014). *Evaluating teachers with classroom observations: lessons learned in four districts*. Washington, DC: Brookings Institution Retrieved from https://www.brookings.edu/wp-content/uploads/2016/06/Evaluating-Teachers-with-Classroom-Observations.pdf.

Wright, P., Horn, S., & Sanders, W. L. (1997). Teachers and classroom heterogeneity: their effects on educational outcomes. *Journal of Personnel Evaluation in Education, 11*(1), 57–67.

Yeh, S. S. (2013). A re-analysis of the effects of teacher replacement using value-added modeling. *Teachers College Record, 115*(12), 1–35.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.