Journal of Educational Research and Practice
2018, Volume 8, Issue 1, Pages 146–178
DOI:10.5590/JERAP.2018.08.1.12

Using the Texas Value-Added Assessment System (TxVAAS) to Improve Teacher Effectiveness: Investigating the Research-Situated "Truths" Behind TxVAAS Claims

Audrey Amrein-Beardsley
Arizona State University

Clarin Collins
Arizona State University

At the time of this study, the Texas Value-Added Assessment System (TxVAAS) was being piloted throughout Texas to hold teachers more accountable for their effects on students' achievement (i.e., teachers' value added). It is still being used by districts throughout Texas today. Using a framework informed by the *Standards for Educational and Psychological Testing*, researchers conducted a content analysis of the marketing and research-based claims asserted about the TxVAAS, to (a) examine the "truth" of each claim to (b) help others critically consume the marketing claims using (c) nonproprietary, peer-reviewed literature. Given that the more popular, and also proprietary version of the TxVAAS—the Education Value-Added Assessment System—continues to be sold and marketed to other states and districts in similar ways, researchers deemed it critical to intervene before other states and districts might blindly trust the marketing and research-based claims presented.

Keywords: *accountability testing, consequential validity, formative assessment, large-scale assessment, teacher evaluation*

Introduction

The first part of the 21st century brought several iterations of educational accountability policies, fixed upon educational reform by holding students, teachers, schools, districts, and states responsible for their impacts on student achievement over time. Given the federal government's expanded role in public education at the time (e.g., via Race to the Top, 2011), compliant states and districts spent much effort and expense to adopt and implement such policy-mandated metrics (e.g., growth or value-added models [VAMs]).

In the simplest of terms, statisticians use growth models and VAMs (hereafter referred to more generally as VAMs) to measure the predicted then actual "value" a teacher "adds" to (or detracts from) student achievement from one year to the next. Modelers typically do this by measuring student growth over time on large-scale standardized tests (e.g., as mandated by No Child Left Behind, 2001), and 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, English-language learner [ELL], special education [SPED] students) and school-level variables (e.g., class size, resources), although control variables vary by model. Notwithstanding, much debate continues to exist about the actual, intended successes such policies have had on improving educational outcomes and the extent to which pernicious, unintended consequences have resulted instead (see, for example, Amrein-Beardsley, 2014; Koretz, 2017).

Please address queries to: Audrey Amrein-Beardsley, Arizona State University. Email: audrey.beardsley@asu.edu

Because of these and many other reasons (e.g., too much instructional time focused on testing and test preparation), the federal government reduced its policy mandates with the Every Student Succeeds Act (2016). This allowed for greater local control over such policies. In April of 2016, all 50 states submitted their Every Student Succeeds Act plans to the federal government, with most states diminishing the accountability plans they recently put into place. Fewer than half still intend to measure student achievement as based on proficiency; the majority significantly lessened or removed the consequences attached to schools', teachers', or students' test-based performance; and many decreased or entirely withdrew efforts to hold teachers accountable for their effectiveness using teachers' students' achievement (e.g., via VAMs; see, for example, Close, Amrein-Beardsley, & Collins, 2018; ExcelinEd, 2017). Fifteen states moved teacher accountability decisions to local district control with regard to type of VAMs used, percentage of growth teachers are held accountable for, and consequences attached to VAMs (Close et al., 2018).

Texas, for example, was one such state which gave local control back to districts in all regards, including teacher-level accountability policies. Districts are no longer required to submit a plan for teacher evaluation or submit teacher evaluation results to the state. Despite having local control options regarding teacher evaluation, however, there is still a "recommended" teacher appraisal system endorsed by the Texas Commissioner of Education—the Texas Teacher Evaluation and Support System (T-TESS). T-TESS is currently in use in more than 1,000 of the 1,217 districts served by Texas regional education service centers (Texas Education Agency [TEA], 2017a, 2017b), representing a large majority of Texas districts (n = 1,000/1,217; 82.2%).

The Texas Teacher Evaluation and Support System (T-TESS)

As part of the T-TESS, districts are permitted to use student learning objectives, student portfolios, pre- and posttest results on district assessments, and value-added data. At the time of this study, Teachers' Value Added data were to be calculated via the state's Texas Value-Added Assessment System (TxVAAS), as formerly made available via a state-level pilot and contract with TxVAAS vendor SAS Institute, Inc.

Accordingly, the TxVAAS was of primary interest in this study. It should be noted here, though, that the pilot did not yield the results needed to warrant statewide expansion, or perhaps the cost of expansion at approximately $6 million per school year in terms of net benefit. The TxVAAS as well as the website sources that researchers analyzed within this study (n = 13) were consequently halted and taken down (L. Johnson, personal communication, January 15, 2018). More specifically, the information placed on the TEA website disappeared postpilot (n = 8*;* see TEA, n.d.; TEA & SAS, n.d.-a, n.d.-b, n.d.-c, n.d.-d, n.d.-e, n.d.-f, and n.d.-g), although the websites cited by TEA as located on a SAS website are still available (n = 5*;* see SAS, n.d.-a., n.d.-b, n.d.-c, n.d.-d, and n.d.-e).

Notwithstanding, TxVAAS is still one of the four options districts can select from to measure student growth (e.g, available to districts for 2017–2018), with student growth to account for a recommended 20% of teachers' evaluation scores (Castellano, Alvarez-Calderon, Heinchon, & Hoyer, 2017). If districts select TxVAAS, though, they must individually contract with SAS, and no state funding may be used toward their costs.

Likewise, and in addition to use in Texas, SAS is still selling its more generalized VAM, the Education Value-Added Assessment System (EVAAS) across other states and districts (e.g., value-added assessment systems in Tennessee and Pennsylvania, etc.), warranting what researchers in this study investigated and found in Texas as still pertinent and meaningful, especially as other districts and states consider adopting the EVAAS.

The Texas Value-Added Assessment System (TxVAAS)

At the time of this study, a simple Internet search of TxVAAS took one to a cosponsored website that served as the primary source of interest in this study (TEA & SAS, n.d.-e). It included information about the TxVAAS (TEA & SAS, n.d.-c) and information about what was then defined as a partnership between the TEA and SAS given their collective efforts to provide the TxVAAS to school districts across T-TESS users. The explicit goals of this partnership were to (a) encourage the increased adoption/implementation of the TxVAAS to (b) better hold Texas teachers accountable for their effectiveness, as defined by their students' growth in achievement over time (i.e., their value added).

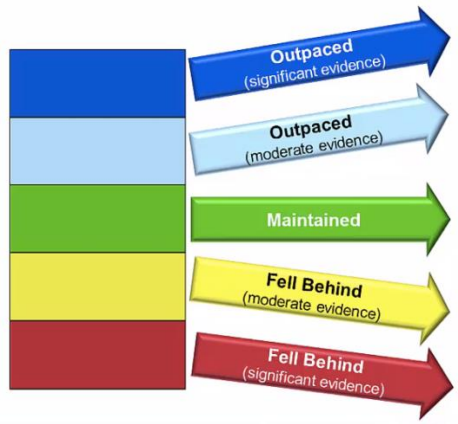
This was important in that, according to SAS, teachers impact students in many ways. TEA and SAS, fittingly, cited Chetty, Friedman, and Rockoff (2014b) in support given Chetty et al.'s well-known (albeit controversial) findings that teachers' have considerable impacts on students' future college going rates, long-term incomes, rates of pregnancy, and the like (TEA & SAS, n.d.-c, 23:15 min). Hence, TEA and SAS argued that measuring whether teachers' students significantly "outpace, maintain, or fall behind" their peers' per year was both critical and necessitous. Accordingly, TEA and SAS collectively argued that implementing and using the TxVAAS was required to do this (TEA & SAS, n.d.-c, 17 min).

In the simplest of terms, how this was to be done is illustrated in Figure 1, which captures the TxVAAS's use of students' (and then teachers') predicted and then actual value-added scores (see also TEA, n.d.). Put simply, predicted scores were what one might expect a student to score on a standardized test if the student made typical progress over 1 year (as compared to students' academically similar peers), and this was to be compared to how a student actually progressed, after the fact. Predicted scores at the teacher level, then, were what one might expect a teacher to score via his or her students' predicted aggregate scores and then actually measured to indicate the progress a teacher's students actually made.

However, beyond this figure and what was available on the aforementioned website of interest in this study (TEA & SAS, n.d.-e), consumers were directed to the SAS (n.d.-a) website for more information. Information here included how the EVAAS models work, given the more generalized EVAAS's publically accessible (i.e., nonproprietary) procedures, formulas, calculations, and the like. All else is held as proprietary.

Purpose of the Study

For purposes of this study, accordingly, researchers conducted a content analysis of all that was asserted across the 14 documents included on the aforementioned website of interest (TEA & SAS, n.d.-e). Researchers then situated that which TEA and SAS asserted across documents within the current research on teacher-level VAMing and the EVAAS, specifically. As such, the primary purposes of this study were to (a) analyze each marketing or other TEA and SAS asserted claim within these 14 documents and then to (b) determine, using the conceptual framework and research literature presented next, whether each claim was supported and to what extent. Again, researchers used the literature as pertinent to the EVAAS specifically or VAMs in general. Researchers' tertiary goal was to (c) provide nonproprietary, peer-reviewed feedback to help others critically consume the marketing and other research-based claims surrounding the TxVAAS and its broader EVAAS equivalent, as presented.



Level 5	**Level 5, Well Above Expectations:** Significant evidence that the teacher's students made more progress than the Growth Standard (the teacher's index is 2 or greater).
Level 4	**Level 4, Above Expectations:** Moderate evidence that the teacher's students made more progress than the Growth Standard (the teacher's index is equal to or greater than 1 but less than 2).
Level 3	**Level 3, At Expectations:** Evidence that the teacher's students made progress similar to the Growth Standard (the teacher's index is equal to or greater than -1 but less than 1).
Level 2	**Level 2, Below Expectations:** Moderate evidence that the teacher's students made less progress than the Growth Standard (the teacher's index is equal to or greater than -2 but less than -1).
Level 1	**Level 1, Well Below Expectations:** Significant evidence that the teacher's students made less progress than the Growth Standard (the teacher's index is less than -2).

Figure 1. *Illustration Indicating Teachers' Value Added Given Whether, on Average, Their Students Outpaced, Maintained, or Fell Behind Their Peers on the TxVAAS, After Which Teachers' Are Classified by Level (TEA & SAS, n.d.-c, 17 min, 21:15 min)*

With each of these three purposes at the heart of this research study, the ultimate research question researchers set out to answer, or rather make more transparent primarily for audiences including districts and states (current or future users of the EVAAS), was whether the claims marketed in the case of Texas with its TxVAAS could be substantiated or warranted and to what extent (e.g., generalize; Miles & Huberman, 1994). If substantiated, this would imply that those using or looking to adopt the EVAAS (or any of its derivatives) could, more or less, trust that which is marketed. If not substantiated, at minimum, current or future users should more critically consume that which is marketed about this VAM and ask for evidence to help substantiate that which is advertised and suspect. Because this particular model is more popularly known, above all other VAMs as the "black box" model (see, for example, Amrein-Bearsdley, 2008, 2014; Ballou & Springer, 2015; Eckert & Dabrowski, 2010; Gabriel & Lester, 2013), the ultimate purpose of this study was to not only address conflicts in that which is marketed, again as per the current research literature, but also address gaps in the knowledge of primarily educational leaders and practitioners who continue to adopt (and often spend taxpayer dollars on) this model, blind to the research-situated truths that surround this particular VAM.

Conceptual and Literature Lenses

In general, there are empirical issues working with VAMs with (a) reliability, (b) validity, (c) bias, (d) transparency and formative use, and (e) fairness. Also at play are (f) VAMs' intended and unintended consequences. Researchers framed their analyses using these key categories as defined by the *Standards for Educational and Psychological Testing,* hereafter referred to as the *Standards* (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014).

Reliability

According to the *Standards* (AERA et al., 2014), reliability is defined as the degree to which test-based scores "are consistent over repeated applications of a measurement procedure [e.g., a VAM] and hence and inferred to be dependable and consistent" (pp. 222–223) for the individuals (e.g., teachers) to whom the scores pertain. In terms of VAMs, reliability is quantified when VAM estimates of teacher effectiveness are consistent from year to year (intertemporal stability), regardless of the types of students and subject areas taught. This is typically captured using "standard errors, reliability coefficients per se, generalizability coefficients, error/tolerance ratios, item response theory…information functions, or various indices of classification consistency" (AERA et al., 2014, p. 33), all of which help to situate and make explicit or transparent VAM estimates and their (sometimes sizeable) errors. This is also done to help situate and inform the validity of the inferences that result.

Validity

According to 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). Correspondingly, when establishing evidence of validity one must be able to support with quantitative or qualitative evidence that accurate inferences can be drawn from the data derived (Brennan, 2006, 2013; M. T. Kane, 2006, 2013; Messick, 1980, 1989).

Following suit, VAM researchers have delved into searching for evidence of many subareas of validity, including but not limited to (a) content-related evidence of validity in that "what is to be validated is not the test…but the inferences derived from [the] test scores" (Messick, 1989, p. 5), (b) concurrent-related evidence of validity or "the degree of relationship between the test scores and [other] criterion scores" taken at the same time (Messick, 1989, p. 7; see also Messick, 1980), and (c) consequence-related evidence of validity which is"[t]he only form of validity evidence [typically] bypassed or neglected in these traditional formulations" that "bears on the [intended and unintended] social consequences of test interpretation and use" (Messick, 1980, p. 8; see also M. T. Kane, 2013). It should be noted that while all of these evidences of validity help to support construct-related evidence of validity, in this area of research most researchers rely heavily on gathering concurrent-related evidence of validity.

Here, it is necessary to assess whether teachers who post large and small value-added gains over time are those deemed effective, and inversely, using other independent measures of teacher effectiveness (e.g., supervisors' observational scores, student survey results). If all measures line up (i.e., correlate) and validate one another, confidence in them as independent measures of the same construct increases (Messick, 1980, 1989; see also Chin & Goldhaber, 2015; Hill, Kapitula, & Umlan, 2011). If all indicators do not line up, something may be wrong with either or both indicators being used (i.e., using bivariate correlations).

Bias

According to the *Standards* (AERA et al., 2014), bias is defined 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). Also known as systematic error given "[t]he systematic over- or under-prediction of criterion performance" (p. 222), biased estimates in the case of VAMs are observed when performance varies for "people [i.e., teachers] belonging [or in this case teaching]…groups [of students (e.g., impoverished, gifted, SPED, ELL students)] differentiated by characteristics not relevant to the criterion performance" (p. 222) measure.

Transparency and Formative Use

While transparency is not uniquely defined within the *Standards* (AERA et al., 2014), transparency pertains to the extent to which VAM-based output are made accessible and can be understood, consumed, and then used in (in)formative ways. Hence, researchers defined for purposes of this study transparency and formative use as the extent to which VAM-based output are accessible and capable of being understood so that output can be used by teachers to improve their effectiveness. This is imperative in that VAM estimates, once accessible, do not seem to make sense to those at their receiving ends (Eckert & Dabrowski, 2010; Gabriel & Lester, 2013; Goldring et al., 2015; Graue, Delaney, & Karch, 2013).

Fairness

According to the *Standards* (AERA et al., 2014), fairness is defined as the impartiality of "test score interpretations for intended use(s) for individuals from *all* relevant subgroups" (p. 219, emphasis added). 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. This is different than bias in that bias relates to the validity of the inference as potentially distorted by bias and fairness pertains to disparate impacts that may occur regardless of the bias present or not in any test-based measure.

The main issue here is that, currently, VAM-based estimates can be produced for approximately 30–40% of all teachers (B. D. Baker, Oluwole, & Green, 2013; Gabriel & Lester, 2013; Harris, 2011). The other 60–70%, which sometimes includes entire campuses of teachers (e.g., early elementary teachers, high school teachers, teachers of non-core-subject areas), cannot be evaluated or held accountable using value-added data (B. D. Baker et al., 2013; Gabriel & Lester, 2013; Harris, & Herrington, 2015; Jiang, Sporte, & Luppescu, 2015; Papay, 2010). When these data are used to make consequential decisions, issues with fairness become even more important, with some teachers relatively more or less likely to realize the negative or reap the positive consequences attached to VAM-based estimates.

Intended and Unintended Consequences

According to Messick (1989), "[t]he only form of validity evidence [typically] bypassed or neglected…is that which bears on the social consequences of test interpretation and use" (p. 8). Intended and unintended, social and ethical consequences matter as well (Messick, 1980; M. T. Kane, 2013). Accordingly, the *Standards* (AERA et al., 2014) indicate that ongoing evaluation of all consequences of any test use is essential with any test-based system, including those as based on VAMs.

While this is an imperative that might (or arguably should) rest on the shoulders of the governmental bodies (e.g., state boards of education and districts) 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), this rarely occurs. The burden of proof, rather, rests on the shoulders of VAM researchers, and vendors as is the case in this particular study, to provide objective and credible evidence about the positive and negative effects that come along with VAM use. They are also to explain these effects to external constituencies including policymakers and potential customers (e.g., state department of education and district leaders) and to collectively work to determine whether VAM use, given the intended and unintended consequences registered or involved, can be rendered as acceptable and worth their financial, time, and human resource investments.

Intended Consequences

In terms of VAMs' intended consequences, the primary intent behind VAM use is to improve teacher effectiveness by holding teachers accountable for their effects (Burris & Welner, 2011). The stronger the consequences, the more evident the intended effects are to be (e.g., improved teacher effectiveness and subsequent student achievement). In practice, however, whether this is realized is of debate, also due to the evidence void supporting VAMs' intended effects (Braun, 2015; Corcoran, 2010; Goldhaber, 2015).

Unintended Consequences

At the same time, VAMs' unintended consequences are often going underexplored (see also AERA Council, 2015). Here, evidence is needed to evaluate the unintended effects of VAMs and whether such unintended outweigh their intended effects, or vice versa, all things considered. To be contemplated are the educative goals at issue (e.g., increased student achievement) alongside the positive and negative implications for the science and ethics of using VAMs (Messick, 1989, M. T. Kane, 2006, 2013).

As summarized by Moore Johnson (2015), unintended consequences include but are not limited to (a) teachers being more likely to "literally or figuratively 'close their classroom door' and revert to working alone…[which]…affect[s] current collaboration and shared responsibility for school improvement" (p. 120); (b) Teachers being "[driven]…away from the schools that need them most and, in the extreme, causing them to leave [or to not (re)enter] the profession" (p. 121); and (c) Teachers avoiding teaching high-needs students if teachers perceive themselves to be at greater risk of teaching students who may be more likely to hinder teachers' value added. These teachers are those who are apparently "seek[ing] safer [grade level, subject area, classroom, or school] assignments, where they can avoid the risk of low VAMS scores" (p. 120; see also Amrein-Beardsley, 2014; B. D. Baker et al., 2013; Hill et al., 2011). These contentions are no doubt troubling, whereas teachers seem to be reacting to students as "potential score increasers or score compressors…[with such] discourse dehumaniz[ing] students and reflect[ing] a deficit mentality that pathologizes these student groups" (Kappler Hewitt, 2015, p. 32; see also Darling-Hammond, 2015; Gabriel & Lester, 2013; Harris & Herrington, 2015).

Summary

Notwithstanding, the *Standards* (AERA et al., 2014) underscore that ongoing evaluation of all of these measurement issues is essential although this has not been committed to, nor is this being diligently done. The American Statistical Association (2014), the AERA Council (2015), the National Academy of Education (E. L. Baker et al., 2010), and the National Association of Secondary School Principals (n.d.) have also underscored similar calls for such research within their position statements on VAMs and VAM use.

Accordingly, researchers used these educational measurement (and also pragmatic) issues to help frame their content analyses in this study. It is this set of issues that researchers set out to examine

in the case of Texas' TxVAAS to help others understand the issues once facing educators throughout Texas, and still facing others beyond Texas's borders (i.e., in other states and districts currently using or contemplating purchasing and using this particular VAM).

Method

Again, researchers conducted a content analysis across all 13 documents included on the aforementioned website (TEA & SAS, n.d.-e). Researchers also situated each claim resident within each online document within the current literature. Researchers did this to best determine, using the current research, whether the claims advanced were supported and to what extent.

Research Design

Because all of the 14 documents that researchers examined included only text, and no measurable or quantitative data beyond some basic descriptive statistics that likely any lay audience of practitioners might understand, researchers engaged in only qualitative analyses of the textual data included within- and across (Miles & Huberman, 1994) these 14 documents. This was appropriate given the goal was to help readers add "meaning to the descriptive [and] inferential information compiled" (p. 56) and because qualitative analyses are also often used to help others determine whether investments, such as those likely at play here, are worthwhile (i.e., whether the proclaimed benefits are worth the time, money, and human resources spent or to be spent). This is typically done while also taking into consideration the theoretical and practical or actual values of a product being sold.

Data Collection

Documents (n = 14) included those used by SAS to market its EVAAS system to states and districts (see SAS, n.d.-a., n.d.-b, n.d.-c, n.d.-d, and n.d.-e) and documents including information about VAMs in general (TEA, n.d.) and in the specific cases of the EVAAS/TxVAAS (see TEA & SAS, n.d.-a., n.d.-b, n.d.-c, n.d.-d, n.d.-e, n.d.-f, n.d.-g, n.d.-h). Documents were targeted at teachers and teacher leaders including, for example, frequently asked questions (FAQs) about the TxVAAS (e.g., about whether teachers whose students had large-scale test scores would be held accountable differently than teachers who taught students in subject areas or grade levels without large-scale standardized tests), how teachers' index and effectiveness levels would be calculated and categorized, and how teachers of subject areas not tested or tested consecutively would also be identified by teachers' value added.

All 14 documents that researchers analyzed are provided in the Appendix. Although, again, because the website sources that researchers analyzed within this study were consequently taken down (L. Johnson, personal communication, January 15, 2018). Hence, beyond the descriptions provided above, all that could be included here to quasi-represent what these documents looked like is provided here in Figure 2 (taken from SAS, n.d.-b). Likewise, because n = 8 of the documents noted with asterisks are no longer available online (see TEA, n.d.; TEA & SAS, n.d.-a, n.d.-b, n.d.-c, n.d.-d, n.d.-e, n.d.-f, and n.d.-g), any replication of this study is not possible. The n = 5 documents still available for researchers, perhaps, looking to conduct a similar study can still be found on the SAS website. Likewise, there are many other documents also ripe for similar analyses there, beyond those that were cited at the time of this study for the state of Texas (see SAS, n.d.-a., n.d.-b, n.d.-c, n.d.-d, and n.d.-e).

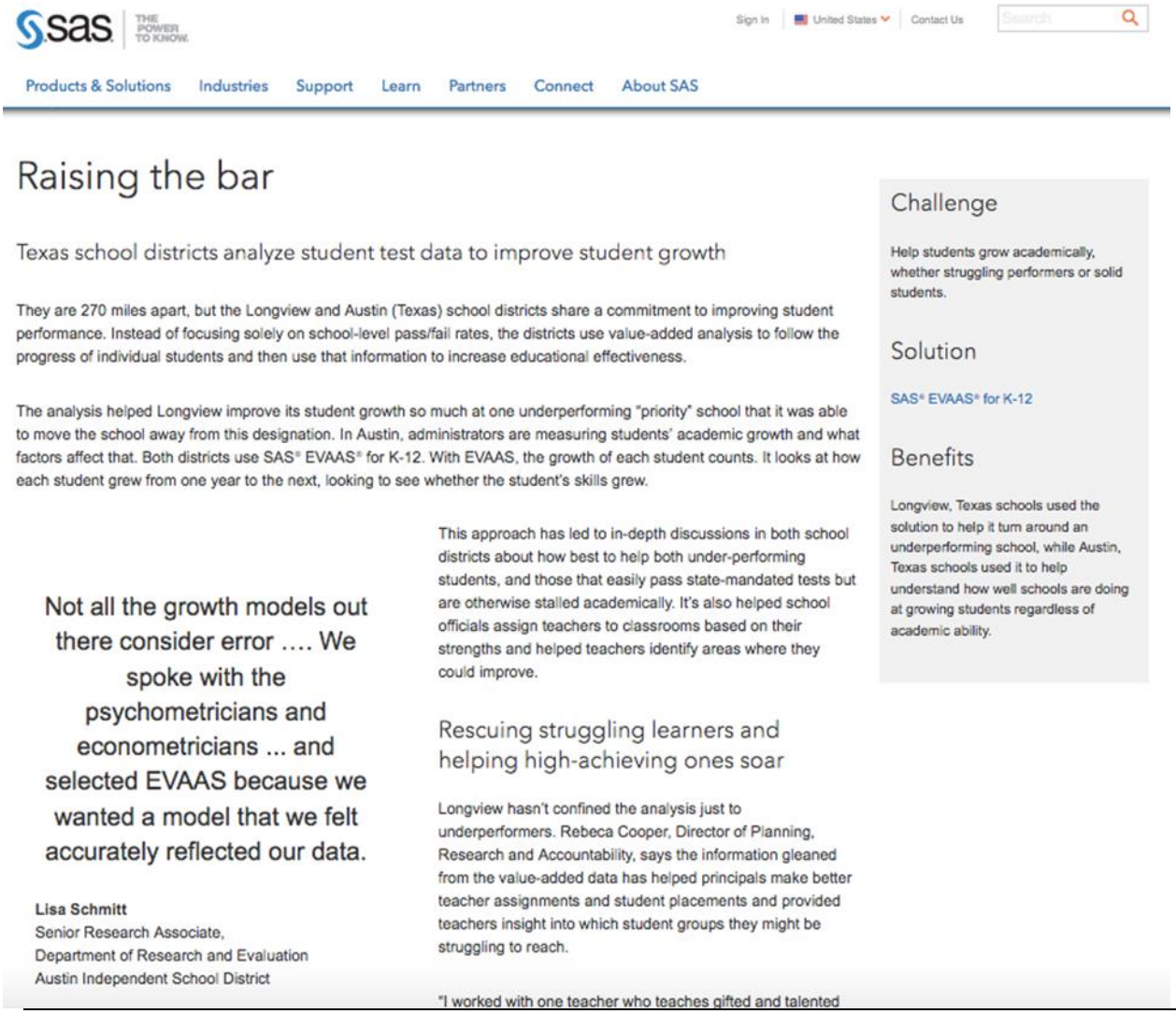


Figure 2. *Screen Shot of One of the n = 14 Documents Analyzed for Purposes of This Study (see the Full Document at SAS, n.d.-b)*

Analytical Method

To analyze the content of each of the 14 documents, researchers read through each document coding for text, quotes, and concepts related to the elements of their measurement-oriented, conceptual framework (Miles & Huberman, 1994). More specifically, researchers used a spreadsheet with tabs aligned with the descriptive concepts and codes used for purposes of analyses (e.g., reliability, validity, bias) and inserted claims resident within the documents into this spreadsheet. Once researchers organized all claims by measurement area, they collapsed and then positioned similar claims within the research literature. They organized the literature also by measurement area via a previous study in which Lavery, Holloway-Libell, Amrein-Beardsley, Pivovarova, and Hahs-Vaughn (2018) used the same framework to systematically review 470 unique sources about VAMs that were published in both traditional (i.e., academic journals, federal documents, policy institutes' technical reports) and nontraditional outlets (e.g., newspapers, newsletters, blog posts). They used these literatures, given both conceptual frameworks were the same, to assess whether the literature was in support or opposition of each of the TxVAAS claims. Researchers modeled their deductive

approach to coding on the framework method described in Gale, Heath, Cameron, Rashid, and Redwood (2013) and Ritchie, Lewis, Nicholls, and Ormston (2013).

Validity

As per Miles and Huberman (1994), researchers aligned their analyses to five key evaluative criteria in support of validity. First, researchers attempted to keep themselves and one another "relatively neutral" to ensure a certain level of "reasonable freedom" from bias (Miles & Huberman, 1994, p. 229); although, given the research they have also contributed to the contemporary literature in this area, researchers would classify themselves as not biased but research informed. Others, who may be more pro-VAM than the researchers of this study (e.g., Chetty et al., 2014a, 2014b, 2014c; T. J. Kane, 2013; T. J. Kane & Staiger, 2012) might call this bias, and vice versa. This is clearly a conundrum not specific to the contentious issues still surrounding VAMs. Second and in terms of dependability (Miles & Huberman, 1994, p. 278), researchers consulted with outside scholars, with similar expertise and a similar stance, to spot check their interpretations of a small sample of the documents included herein. Third, in terms of this study's internal validity (Miles & Huberman, 1994, p. 228), researchers consistently asked themselves, and will trust readers or consumers of this study to assess whether the findings of the study make sense (see also Stake, 1978; Stake & Trumbull, 1982). Fourth, in terms of this study's external validity, researchers made the case above and will continue to make the case below that, indeed, the findings from this study "can be transferred to other contexts, and [thus] generalizable beyond the present study" (Miles & Huberman, 1994, p. 228). For example, that one of the documents analyzed for purposes of this study (SAS, n.d.-a), thereafter removed from the Texas website of interest in this study (TEA & SAS, n.d.-e), showed up in Pennsylvania validates this point. Fifth, in terms of what this study might do for participants (e.g., "utilization"; Miles & Huberman, 1994, p. 242), this is also explained earlier but in more detail in the findings and conclusions presented next.

Findings

It is important to highlight one document—"Current Knowledge About Value-Added Modeling"—that SAS (n.d.-a) advanced as an introduction to the main website of interest. In this document, SAS wrote that over the past 20 years (although VAM research extends back 50 years; see, e.g., Hanushek, 1971) "there have been many studies, articles, and commentaries regarding value-added modeling" (p. 1). SAS (n.d.-a) cited 22 such papers. See Figure 3.

SAS® EVAAS®

sas THE POWER TO KNOW.

Current Knowledge About Value-Added Modeling

Over the past two decades, there have been many studies, articles, and commentaries regarding value-added modeling. The papers listed in this document are related to value-added modeling itself, how value-added modeling is used to measure teacher effectiveness, and how value-added modeling affects educational policies.

Introduction to Value-Added Modeling and SAS® EVAAS® Reporting

These papers provide general information about value-added modeling, as well as detailed information about the EVAAS approach.

SAS: *Comparisons Among Various Educational Assessment Value-Added Models* is an introduction to various value-added models, noting the advantages and disadvantages of each.

SAS: Key Research Findings summarizes key findings, research, and milestones by the EVAAS team over the past 30 years.

SAS: *Addressing Common Concerns about Value-Added Modeling* briefly addresses several common concerns and misconceptions of EVAAS value-added reports.

SAS: *Adjusting for Student Characteristics in Value-Added Models* uses both theoretical and empirical data to illustrate why some value-added models do not make adjustments for student characteristics.

SAS: Summary of RAND Research on Value-Added Modeling RAND has published many papers that focus on the technical aspects of value-added modeling. Because these papers are so technical and often lengthy, EVAAS compiled a summary of these papers.

The Washington Post: *Seven Misconceptions About Value-Added Measures* is an interview of Doug Harris by columnist Jay Mathews. This article addresses common concerns in a non-technical way.

Importance of Value-Added Modeling and Effective Teaching

These papers focus on the importance of value-added modeling and effective teaching, and they assess value-added modeling in the context of other means of evaluation or assessment.

The Brookings Institute: *Climate Change and Value-Added: New Evidence Requires New Thinking* was written by Tom Kane, and it makes the case that the relevant question for educators and policymakers is *how* to include growth measures in teacher evaluation rather than *whether* to include them.

The Brookings Institute: *Evaluating Teachers: The Important Role of Value-Added* was written by Steven Glazerman, Dan Goldhaber, Susanna Loeb, Stephen Raudenbush, Douglas Staiger and Grover Whitehurst. While acknowledging potential limitations of value-added modeling, they put these in context of other means of evaluation and address many common concerns.

Measures of Effective Teaching (MET) Project: *Ensuring Fair and Reliable Measures of Effective Teaching: Culminating Findings from the MET Project's Three-Year Study: Brief* was funded by the Bill and Melinda Gates Foundation, and this policy brief represents the study's final conclusions, particularly in relation to randomized assignment and composite teacher measures.

National Bureau of Economic Research (NBER): *The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood* was written by Raj Chetty, John N. Friedman, and Jonah E. Rockoff. It assesses long-term outcomes related to effective teaching.

Additional Technical Papers

These papers go into detail regarding the value-added models used in EVAAS reporting.

SAS: SAS EVAAS Statistical Models details the general statistical approaches to EVAAS models.

SAS: *Do Teacher Effect Estimates Persist When Teachers Move to Schools with Different Socioeconomic Environments* focuses on whether teacher effects persist when teachers move to schools with different socioeconomic environments (Source: Prepared for Performance Incentives: Their Growing Impact on American K-12 Education in Nashville, Tennessee on February 29, 2008.)

Knowledge Briefs

These papers present findings within the context of educational policy considerations.

National Center for Analysis of Longitudinal Data in Educational Research (CALDER): CALDER publishes many reports on value-added reporting. *Portability of Teacher Effectiveness Across Schools* which finds conclusions similar to those in the SAS paper on teacher mobility (listed above).

Carnegie Knowledge Network (CKN): CKN publishes many reports on value-added modeling, such as *Do Value-Added Models Level the Playing Field*; *How Stable are Value-Added Estimates across Years, Subjects, and Student Groups*; *and* *How Do Value-Added Indicators Compare to Other Measures of Teacher Effectiveness?*

Council of Chief State Schools Office (CCSSO): CCSSO provided a fairly non-technical introduction to a variety of value-added and growth models in its A Practitioner's Guide to Growth Models.

Battelle for Kids: *Selecting Growth Measures: A Guide for Education Leaders* provides an overview of growth measures propelled by Race to the Top legislation in many states.

Policy Papers

Five Thirty Eight: *The Science of Grading Teachers Gets High Marks* discusses the growing body of research supporting the use of value-added as a valid indicator of teaching effectiveness, even without randomized experiments with student assignment.

Education Week: *Doug Harris Crunches Critics in Value-Added Smackdown* is a blog entry by Rick Harris, and it illustrates the passionate rhetoric used by both supporters and opponents of value-added modeling.

Chetty, Friedman, and Rockoff: *Discussion of the American Statistical Association's Statement (2014) on Using Value-Added Models for Educational Assessment* is a point-by-point response to frequent mischaracterizations of the initial statement on value-added analysis by the American Educational Research Association

WestEd: *Options for Incorporating Student Academic Growth as One Measure of the Effectiveness of Teachers in Tested Grades and Subjects: A Report to the North Carolina Department of Public Instruction* was prepared for the North Carolina Department of Public Instruction to assist in the selection of a statewide value-added model. As a result of its research, WestEd recommended EVAAS models for their statistical advantages as well as policy considerations.

Figure 3. *Screen Shot of the 22 Articles Cited Within One of the n = 14 Documents Analyzed for Purposes of This Study (Document Formerly Available at SAS, n.d.-a)*

The breakdown of these 22 papers follows: Six provided general information about VAMs; although, five were published by SAS and one was a *Washington Post* interview with Harris who is "among the legion of economists" in support of VAMs (Mathews, 2011). According to Lavery et al. (2018), there have been 142 empirical studies and articles published in highly regarded academic journals (i.e., journals with the highest relative rankings and impact factors) across education, economics, and finance. Hence, this representation of 22 articles was clearly limited and hardly representative of the literature regarding value-added modeling.

Likewise, four of these 22 papers focused on the importance of VAMs as related to effective teaching, although two of these four (T. J. Kane, 2013; Glazerman et al., 2010) were produced by the Brookings Institute (i.e., an independent and nonpartisan think tank that also promotes VAMs); one was the final measures of effective teaching report produced by the Bill and Melinda Gates Foundation (2013; as directed by the abovementioned Kane who is a prominent supporter of VAMs); and the other was the aforementioned Chetty et al. (2014b) article with authors also recognized as prominent supporters of VAMs.

Otherwise, two of these 22 papers went into detail regarding the value-added models used in EVAAS reporting, both of which were SAS produced (SAS, n.d.-e.; Sanders, Wright, & Langevin, 2008). Six of the 22 papers "present[ed] findings within the context of educational policy considerations" (SAS, n.d.-a, p. 2), with one published by another VAM-contractor (i.e., the American Institute of Research; Xu, Özek, & Corritore, 2012), three published by the Carnegie Knowledge Network (see Harris, 2012; Loeb & Candelaria, 2012; McCaffrey, 2012), one published by the Council of Chief State School Officers (2013), and one published by Battelle for Kids—a SAS EVAAS partner that is also supported by the pro-VAM Gates Foundation (Battelle for Kids, 2011).

The four other papers were described as policy papers and included the *FiveThirtyEigh*t blog post celebrating the same Chetty et al. (2014b) study, as well as the aforementioned T. J. Kane's pro-value-added works (Flowers, 2015; T. J. Kane & Staiger, 2012); the "Straight Up" blog post published in *Education Week* in which the aforementioned Harris "Crunche[d] Critics in [a] Value-Added Smackdown" (Hess, 2012); another Chetty et al. piece, albeit unsolicited commentary in which they (Chetty et al., 2014a) criticize the American Statistical Association's (2014) position statement on VAMs; and a pro-VAM executive summary published by the North Carolina Department of Public Instruction (2012). This was seemingly random, or not, given SAS is located in North Carolina, which uses the EVAAS statewide.

Clearly, SAS's (n.d.-a) representation of the "many [i.e., 22] studies, articles, and commentaries regarding value-added modeling" (p. 1) was incomplete, biased, and accordingly misleading as not representative of the set of literatures capturing our current knowledge about value-added modeling. Consequently, this introductory document foreshadows researchers' findings to come.

Reliability

SAS made many claims throughout the documents they made available (TEA & SAS, n.d.-e) pertaining to the reliability of their TxVAAS (and more generalized EVAAS) model. For example, TEA and SAS noted that the TxVAAS provides "reliable measures of the growth that students [make] in each district, school, and classroom" (TEA & SAS, n.d.-d, p. 1); the TxVAAS affords "the most reliable measures of the impact the teacher [has] on students' academic growth" (TEA & SAS, n.d.-f, p. 1); and the TxVAAS yields "highly reliable teacher value-added reporting…[evidenced by a]….repeatability correlation [at] about 0.70–0.80 for three-year teacher value-added estimates" (SAS, n.d.-e, p. 2). However, across these and all other claims about TxVAAS reliability (see also SAS, n.d.-b; TEA & SAS, n.d.-f) only one untitled, unpublished, and inaccessible report (White, Wright, & Sanders, 2011) was referenced in support of this latter reliability coefficient.

In terms of the current research on reliability as pertaining to VAMs in general, issues with reliability are quite consistently evident. Teachers classified as "effective" one year often have around a 25%–59% chance of being classified as "ineffective" the next, or vice versa, with other permutations noted. Using correlation coefficients akin to the aforementioned statistics cited by SAS (n.d.-e), researchers have accordingly estimated year-to-year correlations across VAMs to be in the range of near zero (i.e., $0 \leq r \leq 0.3$) or higher (i.e., $0.3 \leq r \leq 0.5$). Hence, the best one can assert given the research is that VAM-based levels of reliability are very weak to moderate (Chiang, McCullough, Lipscomb, & Gill, 2016; Martinez, Schweig, & Goldschmidt, 2016; Schochet & Chiang, 2013; Shaw & Bovaird, 2011; Yeh, 2013). Consequently, what might be asked here is whether the TxVAAS (or EVAAS) is twice as reliable as all other VAMs, and where the actual evidence is to prove it as currently missing.

Recently, however, a related study was ordered by a Texas court that forced SAS to acquiesce their EVAAS data from the Houston Independent School District (HISD). Researchers found that correlations between teachers' EVAAS estimates across years were moderate, at best, with correlation coefficients ranging between $.41 \leq r \leq .52$ (Amrein-Beardsley & Geiger, 2018). These coefficients are similar to the general VAM correlations noted above, but they fall below the $r = 0.70$–0.80 coefficients referred to in the inaccessible White et al. (2011) reference.

Perhaps more importantly given the grander purposes of this study, what this means is that the EVAAS might not offer states, districts, and schools "*reliable*…results that go *far beyond* what other simplistic [value-added] models found in the market today can provide" (SAS, n.d.-c, emphasis added). What this also means across VAMs is that reliability is still a hindrance, thwarting the usability of teacher-level VAM estimates, especially when unreliable measures are to be used for consequential decision-making purposes.

Validity

SAS made additional claims throughout the documents on the website of interest (TEA & SAS, n.d.-g) about validity. Primarily, these included claims about the standardized tests used to measure growth in achievement over time and then used to aggregate and attribute said growth to students' teachers (i.e., value added). For example, one claim was that the tests used across EVAAS models were "reviewed and analyzed to ensure that they [were] appropriate for value-added analysis…to ensure that reliable and valid results [would be] provided to educators to make accurately informed decisions" (TEA, n.d., p. 2). TEA and SAS also noted that the TxVAAS could "accommodate tests on different scales" (e.g., from the Texas Assessment of Academic Knowledge and Skills and the State of Texas Assessments of Academic Readiness tests). That is, "[i]t is not required for the prior test scores to be on the same scale as the current test…The important thing is…how the students' place[s] in the statewide distribution [change] from one year to the next" (TEA, n.d., p. 4).

What is problematic here is that, fundamentally, VAMs require that the scales used to measure growth from one year to the next rely upon vertical, interval scales of equal units (see, e.g., M. T. Kane, 2017). To yield reliable and valid results, these scales should connect consecutive tests on the same fixed ruler, making it possible to measure growth from one year to the next across different grade-level tests. Here, a ten-point difference (e.g., between a score of 50 and 60 in the fourth grade) on one test might mean essentially the same thing as a 10-point difference (e.g., between a score of 80 and 90 in the fifth grade) on a similar test 1 year later. Then, one could more feasibly (and more accurately) measure growth upwards across tests. However, the scales of the large-scale standardized achievement test scores used across all VAMs are not vertically aligned as oft-assumed (Ballou, 2009; Braun, 2004; Doran & Fleischman, 2005; Harris, 2009a; Newton, Darling-Hammond, Haertel, & Thomas, 2010; Reckase, 2004). "[E]ven the psychometricians who are responsible for test scaling shy away from making [such an] assumption" (Harris, 2009b, p. 329).

Related, and maybe more important, is that nowhere do VAM (or EVAAS) technicians note that all VAMs, including the EVAAS, are based on the assumption that test scores can be used to accurately measure student growth using these tests, not to mention teachers' causal and autonomous impacts on student growth on these tests over time. This is an issue with content validity in that these tests have only been designed and validated to measure student achievement at one point of time, not growth over time at the student or teacher level.

Otherwise, and in terms of validity, of concern is whether teachers who post large and small value-added gains or losses over time are the same teachers deemed effective or ineffective, respectively, using other independent measures of teacher effectiveness (e.g., supervisors' observational scores). Once among the documents included on the website of interest in this study (TEA & SAS, n.d.-g) was such evidence of validity discussed, whereby teachers were explicitly instructed to consider their "local data, personal knowledge, and [data from their] classroom observation[s]" to help them better understand their TxVAAS estimates (TEA & SAS, n.d.-h, p. 2). Here, it was implied that teachers' observational data, as their other, primary, and independent measure of their effectiveness should be used as a common indicator.

Elsewhere, concurrent-related evidence of validity pertaining to the EVAAS or TxVAAS went unmentioned. This is also important to note in that concurrent-related evidence of validity and reliability, as noted prior, are likely the two most common sources of evidence present throughout the research literature in this area.

Likewise, it has been established that strong evidence of validity is still lacking across VAMs. The correlation coefficients typically yielded between varieties of VAMs and observational systems typically range between $.10 \leq r \leq .40$ (Grossman, Cohen, Ronfeldt, & Brown, 2014; Hill et al., 2011; T. J. Kane & Staiger, 2012; Polikoff & Porter, 2014; Wallace, Kelcey, & Ruzek, 2016). What is known in terms of the concurrent-related evidence of validity pertaining to the EVAAS (and by default the TxVAAS) specifically, though, comes from the study mentioned and summoned by a Texas court. Researchers also found that the correlations between Houston teachers' EVAAS and observational scores were comparable to other VAMs on the market, yielding weak coefficients ranging from $.28 \leq r \leq .34$ (Amrein-Beardsley & Geiger, 2018).

Bias

SAS made many claims throughout the same set of documents (TEA & SAS, n.d.-g) about bias, or how EVAAS (and by default TxVAAS) output was essentially free from bias. This was true, they argued, primarily due to teachers' opportunities to enter their percentages of responsibility per student using the EVAAS rostering process and regardless of whether teachers' students are "low or high achieving" (TEA, n.d., p. 2), "part of a certain socioeconomic/demographic…group" (p. 2), "identified as Special Education" (p. 3), "frequently absent" (p. 3), or "highly mobile" (p. 3).

According to these same documents, bias was also not of concern if teachers worked in teams in that teachers were permitted "to enter the percentage of instructional responsibility they had for each student…to account for scenarios where more than one instructor had responsibility for a student's learning in a given subject/grade" (TEA, n.d., p. 4). This apparently (also without cited evidence) averted the biasing effects also caused by other teachers' impacts across subject areas and grade levels.

Elsewhere, SAS claimed that there was "virtually no relationship between a student's background (demographics) and cumulative academic growth" as measured by the EVAAS (and by default the TxVAAS; SAS, n.d.-e, p. 1). The research they cited in support, however and again, was nonexistent. This also forced researchers to turn to the peer-reviewed, academic research.

Current research suggests that all VAM-based estimates are sometimes biased, especially when relatively homogeneous sets of certain student types (e.g., high achieving or gifted students, students of certain socioeconomic groups, SPED students, students who are frequently absent, students retained in grade, students who are ELLs) are nonrandomly placed or assigned into classrooms. Research suggests that bias is also evident often despite the statistical controls used to block said bias (E. L. Baker et al. 2010; Green, Baker, & Oluwole, 2012; M. T. Kane, 2017; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Newton et al., 2010; Paufler & Amrein-Beardsley, 2014; Rothstein & Mathis, 2013).

In perhaps the most well-known study on this topic, Rothstein (2009, 2010) illustrated VAM-based bias when he found that a student's fifth-grade teacher was a better predictor of a student's fourth-grade growth than was the student's fourth-grade teacher. This was due to the biasing influences of student demographic variables and standard, nonrandom student-to-classroom assignment practices. While others have since critiqued Rothstein's findings (Goldhaber & Chaplin, 2011; Guarino, Reckase, Stacy, & Wooldridge, 2014; Koedel & Betts, 2009), what is generally known in this regard is that bias exists at least some of the time. Over the past decade VAM bias has been investigated at least 33 times in articles published in 142 empirical studies and articles published in the most highly regarded academic journals (Lavery et al., 2018; see also Koedel, Mihaly, & Rockoff, 2015). Of great debate across articles is, still, whether controlling for bias by using complex statistical approaches makes such biasing effects negligible or "ignorable" (Rosenbaum & Rubin, 1983; see also Chetty et al., 2014c; Koedel et al., 2015; Rothstein, 2014).

It should also be noted that, not without controversy, the EVAAS is the VAM in which developers claim that the statistical controls and mechanisms built into it (which include students' prior test histories and typically no other variables accounting for other student-level demographics) unequivocally control for or block the bias caused by nonrandom student sorting. While this practice counters current recommendations pertinent to VAM-related bias (Koedel et al., 2015), EVAAS statisticians continue to argue that their VAM is unimpaired by such variables. Rather, allowing students to serve as their own controls (using students' testing histories in the formula) dissolves all bias of concern (Sanders & Horn, 1994, 1998; Sanders, Wright, Rivers, & Leandro, 2009; Wright, White, Sanders, & Rivers, 2010). While some others agree with this approach (Franco & Seidel, 2014), the consensus is that it is "statistically prudent to use all available information about students and the schooling environment to mitigate as much bias in value-added estimates as possible" (Koedel et al., 2015, p. 188).

What is known about EVAAS bias specifically comes from internal reports, as well as a few reports conducted externally in which researchers express concerns regarding such claims (Amrein-Beardsley, 2008, 2014; Eckert & Dabrowski, 2010; Gabriel & Lester, 2013; Kappler Hewitt, 2015; Kupermintz, 2003). What is also known in terms of EVAAS (and TxVAAS) bias was illustrated in the aforementioned study summoned by a Texas court. Here, researchers found that either Houston teachers teaching relatively smaller populations of racial minority, students receiving free or reduced lunches, students retained in grade, and SPED and ELL students were among the best teachers in the district or the EVAAS was biased against teachers who taught relatively higher proportions of these students (Amrein-Beardsley & Geiger, 2018).

Perhaps more importantly for purposes of this study here, though, was that this means the EVAAS might not offer states, districts, and schools "*unbiased* results that go *far beyond* what other simplistic [value-added] models found in the market today can provide" (SAS, n.d.-c, emphasis added). Unfortunately, however, it is likely that the population to whom the EVAAS is marketed (i.e., state policy-makers or district decision-makers) may trust this and these other marketing claims at face value without knowledge of the actual (i.e., externally vetted) research in these regards.

Transparency and Formative Use

TEA and SAS also made many claims throughout documents (TEA & SAS, n.d.-g) about the extent to which VAM estimates were accessible (i.e., transparent), after which estimates could be consumed and understood to direct the (in)formative uses of the estimates derived. Claims included but were not limited to declarations about how using the TxVAAS would "promote dialogue and continual reflection for [teachers'] professional growth" (TEA & SAS, n.d.-f, 23:30 min); allow teachers to "examin[e] [their] professional practice in the most recent school year and over time" (29:30 min); "help teachers understand each student's academic needs, set meaningful individual goals for students, and plan appropriate strategies for differentiating instruction" (TEA & SAS, n.d.-g, p. 2); "provide insight into how successful instruction was…and draw attention to areas of focus" for the future (p. 2); help teachers "identify areas of strengths and weaknesses across content areas…to positively grow all students in all subjects" (TEA & SAS, n.d.-h, p. 2); allow "educators to reflect on their practices and have a positive impact on all students' achievement" (TEA, n.d., p. 1); and help teachers "[i]dentify strengths and weaknesses of [their] curriculum or instructional practices" (p. 1; see also additional claims in TEA, n.d., p. 1; TEA & SAS, n.d.-f; TEA & SAS, n.d.-h, p. 2).

While SAS also offered teachers support systems to do this, including on-demand e-learning modules, quick video clips (e.g., to facilitate EVAAS/TxVAAS system navigation, report interpretation, and data), live question and answer seminars with TEA and EVAAS support specialists, online help built into every section of the EVAAS's web-based system, helpdesk support provided via email, and additional supporting documents (TEA & SAS, n.d.-g), whether any or all of these help teachers understand and use their EVAAS data in practice is also suspect. This is also true given the current research surrounding this particular VAM.

In general, EVAAS reports are often described by practitioners as confusing, not comprehensive in terms of the key concepts and objectives teachers teach, ambiguous in terms of teachers' efforts at both the individual student and teacher composite levels, and often received months after students leave teachers' classrooms. For example, teachers in Houston collectively expressed that they were learning little about what they did effectively or how they might use their EVAAS data to improve their instruction (Collins, 2014). Teachers reported frustration with not understanding how their scores were generated, and when they inquired about this with their principals, principals were also unable to respond to the teachers' questions. Teachers who were plaintiffs in the lawsuit also charged that their EVAAS scores were incomprehensible and, therefore, very difficult if not impossible to use to improve upon their instruction.

Elsewhere, teachers in North Carolina reported that they were "weakly to moderately" familiar with their EVAAS data (Kappler Hewitt, 2015). Approximately "22% of respondents felt that the professional development they received on EVAAS/value-added was fairly (19%) or completely (3%) sufficient" (p. 12). Teachers who were EVAAS eligible "were significantly more likely to disagree with the use of value-added for educator evaluation," with evidence also suggesting that the teachers who had more experience with the EVAAS were significantly "more skeptical" of its potential (p. 15). Likewise, Eckert and Dabrowski (2010) demonstrated that in Tennessee (home to two other EVAAS lawsuits), teachers maintained that there was very limited support or explanation helping teachers use Teachers' Value Added data to improve upon their practice (see also Harris, 2011). This is, at least in part, why the EVAAS may still be more popularly recognized as the black box VAM.

Altogether, this is problematic in that one of the main and alleged strengths of really all VAMs is the wealth of diagnostic information to be accumulated for formative purposes. In the specific case of the EVAAS (see also Sanders et al., 2009), however, promoters simultaneously make "no apologies for the fact that [their] methods [are] too complex for most of the teachers whose jobs depend on them to understand" (Carey, 2017; see also Gabriel & Lester, 2013).

Fairness

Again, the issue with fairness is that VAM-based estimates can typically be produced for approximately 30–40% of all teachers (B. D. Baker et al., 2013; Gabriel & Lester, 2013; Harris, 2011). Consequently, when these data are used to make consequential decisions, issues with fairness become even more important whereas some become more likely than others, by teacher type, to realize the positive or reap the negative consequences attached to VAM scores.

In Texas with the TxVAAS, fairness was possibly less of a concern given the subject area tests that were to be included when calculating teachers' value added scores. Teachers who were TxVAAS eligible included teachers of reading in Grades 3–8, mathematics in Grades 3–8, writing in Grade 7, science in Grades 5 and 8, social studies in Grade 8, and teachers of English I, English II, Algebra I, Biology, and U.S. History at the high school level. While this excluded many teachers from being evaluated for teachers' value added, this included more teachers than standard.

In practice, however, this was likely not to work as promoted, either. This was because the tests to be used, other than those in reading and mathematics (i.e., the tests mandated by No Child Left Behind, 2001), did not have the pretests needed in adjacent years to assess individual teachers' impacts in the subject areas discontinuously tested. Rather, for the teachers teaching subject areas without consecutive test administrations (i.e.., writing in Grade 7, science in Grades 5 and 8, social studies in Grade 8, and all of the high school subject areas), teachers were to be held accountable in what might also be perceived as unfair ways (e.g., including other teachers' residual effects; see also Briggs & Weeks, 2009; Lockwood, McCaffrey, Mariano, & Setodji, 2007; McCaffrey et al., 2004).

TEA and SAS (n.d.-a) asserted, instead and without evidence, again, that this could be done because a "student's performance on reading assessments" was "related to how the student perform[ed] on the [in this case] social studies assessment" (p. 1). Hence, they could use students' performance on prior reading test(s) as the pretest(s) required when measuring growth in achievement in said subject areas without adjacent tests (e.g., social studies in the eighth grade, the subject area in which all Texas students are still not tested before the eighth grade). Put differently, their statistical approach could "[leverage] the relationships that exist[ed] among the subjects" (p. 1) so that test scores could be taken from any subjects or grades prior, and linked to any subject area test scores later. "This average performance [would yield] a *reasonable* expectation" (p. 1, emphasis added), although this was also stated sans evidence.

Intended and Unintended Consequences

According to Messick (1989) "[t]he only form of validity evidence bypassed or neglected in these traditional formulations is that which bears on the social consequences of test interpretation and use" (p. 8). The intended and unintended social and ethical costs matter as well (Messick, 1980; M. T. Kane, 2013). Hence, the *Standards* (AERA et al., 2014) emphasized that ongoing evaluation of these is also vital with any test-based system, including those with VAMs.

Intended Consequences

In terms of VAMs' intended consequences, TEA and SAS made many (TEA & SAS, n.d.-g) in addition to those noted prior (e.g., about how TxVAAS use would improve teacher effectiveness, and consequently improve student achievement). These include but are not limited to TxVAAS's value-added and diagnostic reports offer "powerful tools for reporting [the student-achievement] component of teacher effectiveness" (TEA & SAS, n.d.-f, 23:30 min); help teachers to "look back" in terms of "evaluating [their] effectiveness" (25:45 min); and provide "a robust tool for educators to reflect on their practices [to] have a positive impact on all students' achievement" (TEA, n.d., p. 1; see also (TEA & SAS, n.d.-f, p. 1; TEA, n.d., p. 2). No evidence was cited in support of these or other

claims; however, they did drive consumers to an anecdotal series of "Success Stories" that primarily came from other districts in Texas (i.e., not from Houston, the largest school district in Texas, see forthcoming) using the TxVAAS at the time. Researchers do not review these stories herein as decidedly anecdotal, but reference them instead in the event readers might want to look into these further (Young, 2014a, 2014b, 2014c).

The fact that Houston and its teachers'/administrators' use of the EVAAS was not referenced throughout the stories given as support for TxVAAS is important to note in that Houston used the EVAAS for likely longer than the featured Texas districts put together (e.g., since around 2009). Why? Perhaps because the EVAAS evidently failed in Houston in that, after years of high-stakes use (e.g., including merit pay and the termination of teachers, including 221 teachers in 2011), it yielded the results illustrated in Figure 4 (taken from HISD, 2015b, Figure 1, p. 13) and Figure 5 (taken from HISD, 2015a, Figure 2, p. 12).

What evidently happened (or did not happen, as illustrated) in Houston is indeed relevant elsewhere, except maybe here as the TEA and SAS continued to push the TxVAAS despite the lawsuit and other allegations surrounding Houston's EVAAS use at the time (Collins, 2014).

Unintended Consequences

In terms of VAMs' unintended consequences, TEA and SAS made no claims throughout the same documents (TEA & SAS, n.d.-g) of interest in this study. However, researchers found four claims that were somewhat related. The first claim pertained to the limitations of VAMs in that they

> measure teaching effectiveness through student performance on tests, which is *only one aspect* of effective teaching…That said, recent *studies* show that value-added modeling also captures the long-term effects that teachers have on their students' future earnings, retirement savings, and educational attainment. (TEA, n.d., p. 2, emphasis added)

Only the study conducted by Chetty et al. (2014b) was cited, again, as representative of these "studies" to counter the noted "only one aspect" concern.

The second claim pertained to the black box reference mentioned prior. If VAMs seemed like a black box to a teacher,

> You [should] rest assured that the TxVAAS modeling use[s] standard, vetted statistical models to produce valid and reliable results. Non-partisan, independent experts in value-added analysis [have] replicated models similar to the TxVAAS approach *over the years*, and *their most recent findings* suggest that these models [are] among the most reliable, least biased, and most suitable for teacher value-added modeling. (TEA, n.d., p. 2, emphasis added)

One non-peer-reviewed paper presented at a national conference by McCaffrey and Lockwood (2008) was cited in support. This "*variety* of public and private sector experts" was also referenced once elsewhere (SAS, n.d.-e, p. 1, emphasis added), with the same McCaffrey and Lockwood (2008) study cited alongside another study conducted a year prior by the same authors reversed (Lockwood & McCaffrey, 2007). Rather, a small handful of researchers over the past 10–15 years have replicated approaches akin to the EVAAS (see these studies cited in Sanders et al., 2009), although they have done so without access to EVAAS's proprietary information (i.e., equations, computer source codes, decision rules, assumptions integral to permit review and replication).

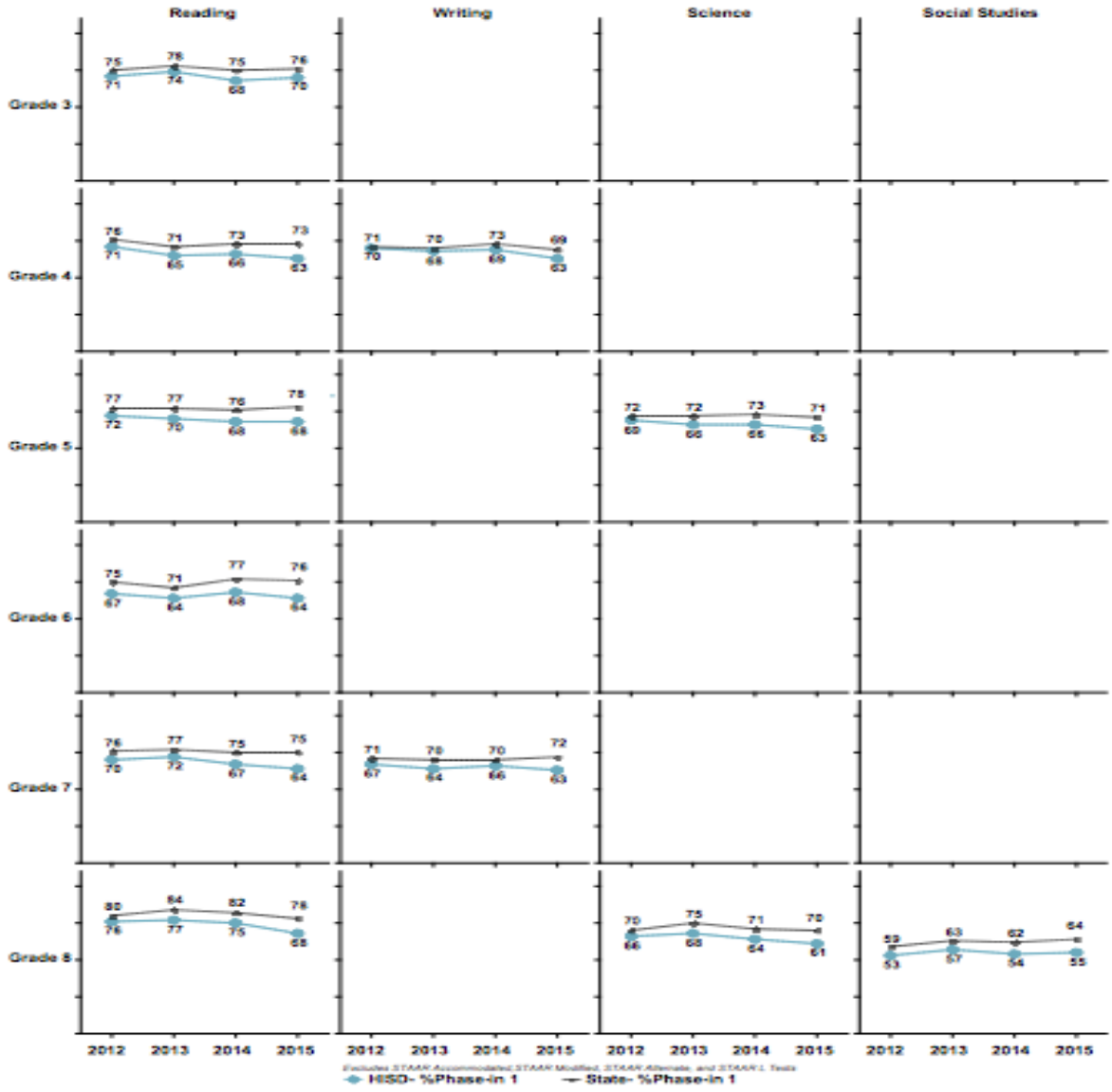


Figure 4. *Percent of Students Meeting Standards in Houston Independent School District (HISD) Versus the State of Texas, 2012–2015, on All State Tests, Combined by Subject and Grade Level, With All Students Tested*

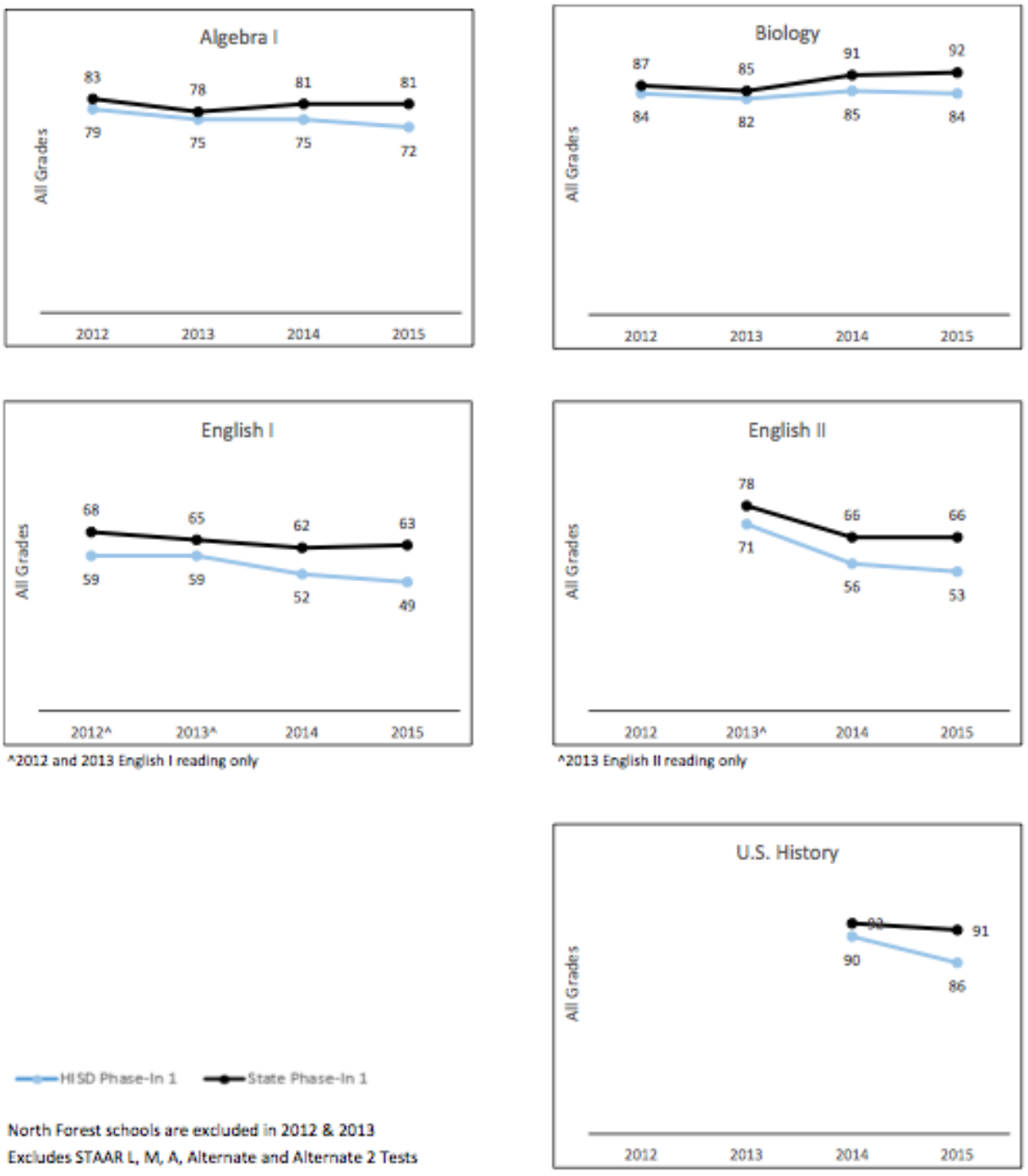


Figure 5. *Percent of Students Meeting Standards in Houston Independent School District (HISD) Versus the State of Texas, 2012–2015, on All State End-of-Course Tests, Combined by Subject and Grade Level, With All Students Tested*

One of the expert witnesses in the aforementioned Houston case was granted more access than prior, under court orders, but he was unable to replicate Houston teachers' EVAAS scores. His inability to do this supported the black box concerns of the court and contributed to the final verdict in plaintiffs' (i.e., teachers') favor (see also Amrein-Beardsley, 2018; Ballou & Springer, 2015). The judge ultimately ruled that the system was opaque and unfair to the teachers it was supposedly objectively evaluating. More specifically, the judge ruled in favor of plaintiffs in May 2017 on two counts, ruling that the plaintiffs had legitimate claims regarding (a) how EVAAS was a violation of their 14th Amendment due process protections (i.e., no state or in this case organization shall deprive any person of life, liberty, or property, without due process) and (b) whether the system permitted teachers to act to ensure their VAM scores were accurate as due process requires. In October 2017, the parties settled the case when the district agreed to remove from its evaluation system any scoring system that could not be verified.

Related, the third claim regarding unintended consequences also pertains to the ongoing criticisms surrounding this particular VAM, as oft made by VAM scholars who continue to critique the model for preventing external review and vetting. Again, others have replicated approaches akin to the EVAAS, but they have done so without access. This is where another primary point of contention exists, and where one event that occurred during the court case in Houston comes back into play.

In 2016, the American Federation of Teachers and Houston Federation of Teachers (i.e., the key plaintiffs in support of teachers in this lawsuit) published an "EVAAS Litigation Update" (National Educational Policy Center, 2016). This summarized a portion of the same expert witness's affidavit in which he concluded that teachers do not have the ability to meaningfully verify their EVAAS scores and that, at most, teachers could request information about the students for which they were held accountable. They could also read literature released by SAS (i.e., not written or released via an independent organization or set of researchers) about their EVAAS estimates and their validity.

Lawyers representing SAS, soon thereafter, charged that the "EVAAS Litigation Update" (National Educational Policy Center, 2016) violated a protective order that was put into place prior to also protect the EVAAS's computer source code during this case. Even though there was nothing in the update that disclosed the actual or any parts of the source code, SAS objected to any conclusions put forward as based on this expert witness's confidential access to the proprietary code. As such, SAS threatened HFT, its lawyers, and the expert witness with monetary sanctions.

HFT went back to court to get the court's interpretation of the protective order. The court ruled that SAS's lawyers

> interpret[ed] the protective order too broadly in this instance. [The expert witness'] opinion regarding the inability to verify or replicate a teacher's EVAAS score essentially mimic[ked] the allegations of HFT's complaint. The Update made clear that [the expert witness] confirmed this opinion after review of the source code; but it [was] not an opinion 'that could not have been made in the absence of review' of the source code. [The expert witness also] testified by affidavit that his opinion [was] not based on anything he saw in the source code, but on the *extremely restrictive access* permitted by SAS. (*Houston Federation of Teachers Local 2415 et al v. Houston Independent School District*, 2017, pp. 4–5, emphasis added)

The fourth and final claim related to unintended consequences pertains to fairness, whereby the website failed to adequately answer a FAQ concerning whether teachers whose students were tested using large-scale test scores would be held accountable differently than teachers with students without. TEA's response was that "There is [sic] different data for teachers who have value-added scores than for teachers who don't have scores. Both groups of teachers, however, will use measures

of student growth as part of their evaluation process" (TEA, n.d., p. 4). TEA did not provide detail beyond that, although one can reasonably assume that school-level growth might have been used to evaluate these Teachers' Value Added, or perhaps indicators derived via, for example, student learning objectives.

Conclusions

The purpose of this study was to (a) analyze the marketing and other research-based claims made about the EVAAS and, in the state of Texas, TxVAAS. Researchers did this to (b) determine whether the claims made by SAS (and TEA for purposes specific to Texas) were or could be supported with empirical evidence. Researchers' also desired (c) to provide nonproprietary, peer-reviewed feedback to help others critically consume the marketing and other research-based claims surrounding the TxVAAS as well as its broader and more generalized EVAAS equivalent. Understanding the breadth of research that exists on VAMs and the expanse of the empirical research related to the educational measurement issues connected to VAMs, researchers conducted this study to also take on the burden of proof for others already using or considering the use of the EVAAS writ large.

It should be noted here, though, that researchers could not go into detail per claim and counter claim given space limitations and the sheer number of studies published on these topics (see, for example, the over 100 references already cited below; see also Lavery et al., 2018, in which researchers conducted a systematic literature review of these and other published pieces). What researchers did provide, however, is a comprehensive bibliography that could be useful to readers who might be interested in reading more about VAMs or any of the measurement subareas reviewed above (i.e., reliability, validity, bias, transparency and formative use, fairness, intended and unintended consequences). It was not the researchers' intent to leave the counter-research as almost opaque as the black box researchers critique herein.

Notwithstanding, when analyzing the claims made by the TEA and SAS throughout the websites and documents used to promote the TxVAAS, researchers explained the shortcomings and arguably blatant misguidings of the documents presented with regard to the abovementioned measurement criteria, again, given the evidence provided and the actual research evidence as cited. This is significant in and of itself should current or future consumers read this piece and become equipped to question and critically consume that which is marketed versus that which is "true" as it pertains to this particular VAM (i.e., the EVAAS).

Related, it is clear that the documents provided represented a certain set of interests over others. While it is not necessarily new news that corporations stretch the truth in their marketing materials, the question here was whether EVAAS promoters exaggerated or obfuscated the truth and how, with particular instances situated in the empirical truth.

This is true in that findings also suggest that that which was marketed also did not support the interests of teachers, schools, districts, and states, in terms of advancing their collective efforts to increase educational effectiveness. While some might argue that SAS and those who are still endorsing the EVAAS across states (e.g., Battelle for Kids) might genuinely believe that they are supporting the interests of these educational stakeholders, others might argue that their efforts may be more disingenuous, with profit interests as priority. This might also be true given that which is being sold and the traces of arguably mendacious marketing and other research-based claims asserted without evidence in support. With around $6 million of taxpayer funds paid to SAS for one academic year's worth of EVAAS services in just Texas (TEA, n.d.), it is also clear not only that SAS had a vested interest in Texas but continues to have similar interests in other states and districts still looking for a "research-based" VAM.

Put differently, if one is not to critically consume that which is delivered as "truth" in this case (and maybe others), this VAM may look better than all others. After all, EVAAS superiority is how this VAM is oft-marketed as "the most comprehensive reporting package of value-added metrics available in the educational market" (SAS, n.d.-c). If, however, policymakers, leaders, are practitioners are to read even just one article like this, perhaps they can exercise better judgment and, at minimum, make much more deliberate and informed decisions.

References

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

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*, 448–452. doi:10.3102/0013189X15618385

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

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

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

Amrein-Beardsley, A., & Geiger, T. J. (2018). *Using test scores to evaluate and hold school teachers accountable in the U.S.: One state's value-added model (VAM) on trial.* Manuscript under review.

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*, 1–71. Retrieved from http://epaa.asu.edu/ojs/article/view/1298

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., … Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers*. Washington, DC: Economic Policy Institute. Retrieved from http://www.epi.org/publications/entry/bp278

Ballou, D. (2009). Test scaling and value-added measurement. *Education Finance and Policy, 4*, 351–383. doi:10.1162/edfp.2009.4.4.351

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

Battelle for Kids. (2011). *Selecting growth measures: A guide for education leaders*. Seattle, WA: Bill and Melinda Gates Foundation. Retrieved from static.battelleforkids.org/documents/edgrowth/Selecting_Growth_measures.pdf

Bill and Melinda Gates Foundation. (2013, January 8). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study*. Seattle, WA: Author. Retrieved from http://www.gatesfoundation.org/press-releases/Pages/MET-Announcment.aspx

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

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

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

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

Briggs, D. C., & Weeks, J. P. (2009). The sensitivity of value-added modeling to the creation of a vertical scale score. *Education Finance and Policy, 4*, 384–414. doi:10.1162/edfp.2009.4.4.384

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

Carey, K. (2017). 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

Castellano, P., Alvarez-Calderon, A., Heinchon, S., & Hoyer, S. L. (2017). *T-TESS student growth measures overview* [PowerPoint slides]. Retrieved from https://www.neisd.net/cms/lib/TX02215002/Centricity/Domain/267/Meeting%203/Student%20Growth%20Measures%20Overivew%20DEIC%20Jan%202017.pdf

Chetty, R., Friedman, J., & Rockoff, J. (2014a). Discussion of the American Statistical Association's Statement (2014) on using value-added models for educational assessment. *Statistics and Public Policy, 1*, 111–113. doi:10.1080/2330443X.2014.955227

Chetty, R., Friedman, J., & Rockoff, J. (2014b). Measuring the impact of teachers, I: Teacher value-added and student outcomes in adulthood. *American Economic Review, 104*, 2593–2632. doi:10.3386/w19424

Chetty, R., Friedman, J., & Rockoff, J. (2014c). Measuring the impact of teachers, II: Evaluating bias in teacher value-added estimates. *American Economic Review, 104*, 2593–2632. doi:10.3386/w19424

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

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

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

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*, 1–42. doi: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://www.annenberginstitute.org/publications/can-teachers-be-evaluated-their-students%E2%80%99-test-scores-should-they-be-use-value-added-me

Council of Chief State School Officers. (2013). *A practitioner's guide to growth models*. Washington, DC. Retrieved from http://www.ccsso.org/documents/2013growthmodels.pdf

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

Doran, H. C., & Fleischman, S. (2005, November). Challenges of value-added assessment. *Assessment to Promote Learning, 63*, 85–87.

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

Every Student Succeeds Act 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

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

Flowers, A. (2015). The science of grading teachers gets high marks. *FiveThirtyEight*. Retrieved from https://fivethirtyeight.com/features/the-science-of-grading-teachers-gets-high-marks/

Franco, M. S., & Seidel, K. (2014). Evidence for the need to more closely examine school effects in value-added modeling and related accountability policies. *Education and Urban Society, 46*, 30–58. doi:10.1177/0013124511432306

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

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

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G. (2010). *Evaluating teachers: The important role of value-added*. Washington, DC: The Brookings Institution. Retrieved from https://www.brookings.edu/wp-content/uploads/2016/06/1117_evaluating_teachers.pdf

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

Goldhaber, D., & Chaplin, D. (2011, October 31). *Assessing the "Rothstein test:" Does it really show teacher value-added models are biased?* Seattle, WA: Center for Education Data and Research. Retrieved from http://www.mathematica-mpr.com/publications/pdfs/education/rothstein_wp.pdf

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

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

Green, P. C., Baker, B. D., & Oluwole, J. (2012). Legal and policy implications of value-added teacher assessment policies. *The Brigham Young University Education and Law Journal, 2012*, 2.

Retrieved from https://digitalcommons.law.byu.edu/cgi/viewcontent.cgi?article=1306&context=elj

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

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

Hanushek, E. A. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *The American Economic Review, 61*, 280–288.

Harris, D. N. (2009a). Teacher value-added: Don't end the search before it starts. *Journal of Policy Analysis and Management, 28*, 693–700. doi:10.1002/pam.20464

Harris, D. N. (2009b). Would accountability based on teacher value added be smart policy? An evaluation of the statistical properties and policy alternatives. *Education Finance and Policy, 4*, 319–350. doi: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. (2012). *How do value-added indicators compare to other measures of teacher effectiveness?* Retrieved from http://www.carnegieknowledgenetwork.org/briefs/value-added/value-added-other-measures/

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

Hess, R. (2012). Doug Harris crunches critics in value-added smackdown. *Education Week*. Retrieved from http://blogs.edweek.org/edweek/rick_hess_straight_up/2012/02/doug_harris_crunches_critics_in_value-added_smackdown.html

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

Houston Federation of Teachers Local 2415 et al v. Houston Independent School District. Civil Action H-14-1189 (S.D. Tex. 2017)

Houston Independent School District (HISD). (2015a). *State of Texas Assessments of Academic Readiness (STAAR) end-of-course results, Spring 2015*. Retrieved from http://www.houstonisd.org/Page/69852

Houston Independent School District (HISD). (2015b). *State of Texas Assessments of Academic Readiness (STAAR) performance, Grades 3-8, Spring 2015*. Retrieved from http://www.houstonisd.org/site/default.aspx?PageType=14&DomainID=8269&PageID=63696&ModuleInstanceID=96404&ViewID=1e008a8a-8e8a-4ca0-9472-a8f4a723a4a7&IsMoreExpandedView=True

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

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

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

Kane, M. T. (2017). *Measurement error and bias in value-added models* (ETS RR–17-25). Princeton, NJ: Educational Testing Services. doi:10.1002/ets2.12153

Kane, T. J. (2013). *Climate change and value-added: New evidence requires new thinking*. Washington, DC: The Brookings Institution. Retrieved from https://www.brookings.edu/research/climate-change-and-value-added-new-evidence-requires-new-thinking/

Kane, T., J., & Staiger, D. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Seattle, WA: Bill and Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Practioner_Brief.pdf

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

Koedel, C., & Betts, J. R. (2009). *Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique* [Working paper 2009-01]. Nashville, TN: National Center on Performance Incentives.

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

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

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value-Added Assessment System. *Educational Evaluation and Policy Analysis, 25*, 287–298. doi:10.3102/01623737025003287

Lavery, M. R., Holloway-Libell, J., Amrein-Beardsley, A., Pivovarova, M., & Hahs-Vaughn, D. L. (2018). *Evaluating the validity evidence surrounding the use of value-added models to evaluate teachers: A systematic review*. Manuscript under review.

Lockwood, J. R., & McCaffrey, D. F. (2007). Controlling for individual heterogeneity in longitudinal models, with applications to student achievement. *Electronic Journal of Statistics, 1*, 223–252. doi:10.1214/07-ejs057

Lockwood, J. R., McCaffrey, D. F., Mariano, L. T., & Setodji, C. (2007). Bayesian methods for scalable multivariate value-added assessment. *Journal of Educational and Behavioral Statistics, 32*, 125–150. doi:10.3102/1076998606298039

Loeb, S., & Candelaria, C. A. (2012). *How stable are value-added estimates across years, subjects, and student groups?* Retrieved from http://www.carnegieknowledgenetwork.org/briefs/value-added/value-added-stability/

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

Mathews, J. (2011). Seven misconceptions about value-added measures. *The Washington Post*. Retrieved from https://www.washingtonpost.com/blogs/class-struggle/post/seven-

misconceptions-about-value-added-measures/2011/04/19/AFbUNV7D_blog.html?utm_term=.127dcfd928a8

McCaffrey, D. F. (2012). *Do value-added methods level the playing field for teachers?* Retrieved from http://www.carnegieknowledgenetwork.org/briefs/value-added/level-playing-field/

McCaffrey, D. F., & Lockwood, J. R. (2008). *Value-added models: Analytic issues*. Paper presented at the Workshop on Value-Added Modeling sponsored by the National Research Council and the National Academy of Education, Board on Testing and Accountability, Washington, DC.

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

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.

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

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

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

National Educational Policy Center. (2016). *Houston lawsuit update, with summary of expert witnesses' findings about the EVAAS*. Boulder, CO. Retrieved from http://nepc.colorado.edu/blog/houston-lawsuit

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*, 1–27. Retrieved from http://epaa.asu.edu/ojs/article/view/810

No Child Left Behind Act of 2001, Pub. L. No. 107-110, § 115 Stat. 1425. (2001). Retrieved from http://www.ed.gov/legislation/ESEA02/

North Carolina Department of Public Instruction. (2012). *Executive summary*. Retrieved from http://www.ncpublicschools.org/docs/sbe-archives/meetings/2012/02/tcp/02tcp04.pdf

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. doi: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, 51*, 328–362. doi:10.3102/0002831213508299

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

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

Reckase, M. D. (2004). The real world is more complicated than we would like. *Journal of Educational and Behavioral Statistics, 29*, 117–120. doi:10.3102/10769986029001117

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

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

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

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

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

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

Sanders, W. L., & Horn, S. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*, 299–311. doi:10.1007/bf00973726

Sanders, W. L., & Horn, S. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*, 247–256.

Sanders, W. L., Wright, P., & Langevin, W. E. (2008). *Do teacher effect estimates persist when teachers move to schools with different socioeconomic environments?* Nashville, TN: National Center on Performance Incentives. Retrieved from https://my.vanderbilt.edu/performanceincentives/files/2012/10/200820_Sanders_TeacherEffectEstimates1.pdf

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. (SAS). (n.d.-a). *Current knowledge about value-added modeling*. Cary, NC: Author. Retrieved July 20, 2014, from https://tea.sas.com/support/CurrentKnowledgeAboutValueAddedModeling.pdf

SAS Institute Inc. (SAS). (n.d.-b). *Raising the bar: Texas school districts analyze student test data to improve student growth*. Cary, NC: Author. Retrieved July 20, 2014, from https://www.sas.com/en_us/customers/texas-public-schools.html

SAS Institute Inc. (SAS). (n.d.-c). *SAS EVAAS for K–12: Assess and predict student performance with precision and reliability*. Cary, NC: Author. Retrieved July 20, 2014, from https://www.sas.com/en_us/software/evaas.html#

SAS Institute Inc. (SAS). (n.d.-d). *SAS EVAAS for K–12: Statistical models. Cary, NC*: Author. Retrieved July 20, 2014, from https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/sas-evaas-k12-statistical-models-107411.pdf

SAS Institute Inc. (SAS). (n.d.-e). *SAS EVAAS: Key research findings*. Cary, NC: Author. Retrieved July 20, 2014, from https://tea.sas.com/support/KeyResearchFindings.pdf

Schochet, P. Z., & Chiang, H. S. (2013). What are error rates for classifying teacher and school performance using value-added models? *Journal of Educational and Behavioral Statistics, 38*, 142–171. doi:10.3102/1076998611432174

Shaw, L. H., & Bovaird, J. A. (2011, April). *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, New Orleans, LA

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

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

Texas Education Agency (TEA). (n.d.). *Common questions about value-added modeling*. Retrieved July 20, 2017, from https://tea.sas.com/support/TEA_CommonQuestionsAboutValueAddedModeling.pdf

Texas Education Agency (TEA). (2017a). *Agreement affirms key components of teacher evaluations in Texas*. Retrieved from https://tea.texas.gov/About_TEA/News_and_Multimedia/Press_Releases/2017/Agreement_Affirms_Key_Components_of_Teacher_Evaluations_in_Texas/

Texas Education Agency (TEA). (2017b). *Texas school directory, 2017–2018*. Retrieved from http://tea4avholly.tea.state.tx.us/TEA.AskTED.TSD/TSDfiles/tsd2018/tagged/texas_school_directory_2017-18_tagged.pdf

Texas Education Agency (TEA) & SAS Institute Inc. (SAS). (n.d.-a). *Measuring growth in a subject not previously tested*. Retrieved July 20, 2017, from https://tea.sas.com/support/MeasuringGrowthInSubjectNotPreviouslyTested.pdf

Texas Education Agency (TEA) & SAS Institute Inc. (SAS). (n.d.-b). *SAS EVAAS for K–12 statistical models*. Retrieved July 20, 2017, from https://tea.sas.com/support/EVAASStatisticalModels.pdf

Texas Education Agency (TEA) & SAS Institute Inc. (SAS). (n.d.-c). *TxVAAS: An introduction to SAS EVAAS for Texas*. Retrieved July 20, 2017, from https://tea.sas.com/videos/TEA/Intro_to_TxVAAS.mp4

Texas Education Agency (TEA) & SAS Institute Inc. (SAS). (n.d.-d). *TxVAAS: Quick reference guide to predicted scores*. Retrieved July 20, 2017, from https://tea.sas.com/support/PredictedScores.pdf

Texas Education Agency (TEA) & SAS Institute Inc. (SAS). (n.d.-e). *TxVAAS: Texas Education Agency*. Retrieved July 20, 2017, from https://tea.sas.com/

Texas Education Agency (TEA) & SAS Institute Inc. (SAS). (n.d.-f). *TxVAAS: The teacher index and effectiveness levels*. Retrieved July 20, 2017, from https://tea.sas.com/support/TeacherIndexAndEffectivenessLevels.pdf

Texas Education Agency (TEA) & SAS Institute Inc. (SAS). (n.d.-g). *TxVAAS: Using student projections to differentiate instruction*. Retrieved July 20, 2017, from https://tea.sas.com/support/UsingStudentProjections.pdf

Texas Education Agency (TEA) & SAS Institute Inc. (SAS). (n.d.-h). *TxVAAS: Using teacher reports to support instructional improvement*. Retrieved July 20, 2017, from https://tea.sas.com/support/UsingTeacherReports.pdf

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

White, J. T., Wright, S. P., & Sanders, W. L. (2011). [untitled]. Unpublished report.

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

Xu, Z., Özek, U., & Corritore, M. (2012). *Portability of teacher effectiveness across school settings*. Washington, DC: American Institute of Research. Retrieved from http://www.caldercenter.org/sites/default/files/wp77.pdf

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

Young, N. (2014a). *Teacher effectiveness culture shifts in Lubbock ISD schools – Part 1: The teachers* [Blog post]. Retrieved from http://blogs.sas.com/content/statelocalgov/2014/06/04/teacher-effectiveness-culture-shifts-in-lubbock-isd-schools-part-1-the-teachers/

Young, N. (2014b). *Teacher effectiveness culture shifts in Lubbock ISD schools – Part 2: The principals* [Blog post]. Retrieved from http://blogs.sas.com/content/statelocalgov/2014/06/09/teacher-effectiveness-culture-shifts-in-lubbock-isd-schools-part-2-the-principals/

Young, N. (2014c). *Teacher effectiveness culture shifts in Lubbock ISD schools – Part 3: The superintendent* [Blog post]. Retrieved from http://blogs.sas.com/content/statelocalgov/2014/06/10/teacher-effectiveness-culture-shifts-in-lubbock-isd-schools-part-3-the-superintendent/

[*Appendix follows*]

Appendix

The 14 Documents Analyzed for This Study

SAS Institute Inc. (SAS). (n.d.-a). *Current knowledge about value-added modeling*. Cary, NC. Retrieved July 20, 2014, from https://tea.sas.com/support/CurrentKnowledgeAboutValueAddedModeling.pdf

SAS Institute Inc. (SAS). (n.d.-b). *Raising the bar: Texas school districts analyze student test data to improve student growth*. Cary, NC. Retrieved July 20, 2014, from https://www.sas.com/en_us/customers/texas-public-schools.html

SAS Institute Inc. (SAS). (n.d.-c). *SAS EVAAS for K–12: Assess and predict student performance with precision and reliability*. Cary, NC. Retrieved July 20, 2014, from https://www.sas.com/en_us/software/evaas.html#

SAS Institute Inc. (SAS). (n.d.-d). *SAS EVAAS for K–12: Statistical models*. Cary, NC. Retrieved July 20, 2014, from https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/sas-evaas-k12-statistical-models-107411.pdf

SAS Institute Inc. (SAS). (n.d.-e). *SAS EVAAS: Key research findings*. Cary, NC. Retrieved July 20, 2014, from https://tea.sas.com/support/KeyResearchFindings.pdf

Texas Education Agency (TEA). (n.d.). *Common questions about value-added modeling*. Retrieved July 20, 2017, from https://tea.sas.com/support/TEA_CommonQuestionsAboutValueAddedModeling.pdf

Texas Education Agency (TEA) & SAS Institute Inc. (SAS). (n.d.-a). *Measuring growth in a subject not previously tested*. Retrieved July 20, 2017 from https://tea.sas.com/support/MeasuringGrowthInSubjectNotPreviouslyTested.pdf

Texas Education Agency (TEA) & SAS Institute Inc. (SAS). (n.d.-b). *SAS EVAAS for K–12 statistical models*. Retrieved July 20, 2017, from https://tea.sas.com/support/EVAASStatisticalModels.pdf

Texas Education Agency (TEA) & SAS Institute Inc. (SAS). (n.d.-c). *TxVAAS: An introduction to SAS EVAAS for Texas*. Retrieved July 20, 2017 from https://tea.sas.com/videos/TEA/Intro_to_TxVAAS.mp4

Texas Education Agency (TEA) & SAS Institute Inc. (SAS). (n.d.-d). *TxVAAS: Quick reference guide to predicted scores*. Retrieved July 20, 2017, from https://tea.sas.com/support/PredictedScores.pdf

Texas Education Agency (TEA) & SAS Institute Inc. (SAS). (n.d.-e). *TxVAAS: Texas Education Agency*. Retrieved July 20, 2017, from https://tea.sas.com/

Texas Education Agency (TEA) & SAS Institute Inc. (SAS). (n.d.-f). *TxVAAS: The teacher index and effectiveness levels*. Retrieved July 20, 2017, from https://tea.sas.com/support/TeacherIndexAndEffectivenessLevels.pdf

Texas Education Agency (TEA) & SAS Institute Inc. (SAS). (n.d.-g). *TxVAAS: Using student projections to differentiate instruction*. Retrieved July 20, 2017, from https://tea.sas.com/support/UsingStudentProjections.pdf

Texas Education Agency (TEA) & SAS Institute Inc. (SAS). (n.d.-h). *TxVAAS: Using teacher reports to support instructional improvement*. Retrieved July 20, 2017, from https://tea.sas.com/support/UsingTeacherReports.pdf

The ***Journal of Educational Research and Practice*** provides a forum for studies and dialogue that allows readers to better develop social change in the field of education and learning. Journal content may focus on educational issues of all ages and in all settings. It also presents peer-reviewed commentaries, book reviews, interviews of prominent individuals, and additional content. The objectives: We publish research and related content that examines current relevant educational issues and processes aimed at presenting readers with knowledge and showing how that knowledge can be used to impact social change in educational or learning environments. Additional content provides an opportunity for scholarly and professional dialogue regarding that content's usefulness in expanding the body of scholarly knowledge and increasing readers' effectiveness as educators. The journal also focuses on facilitating the activities of both researcher-practitioners and practitioner-researchers, providing optimal opportunities for interdisciplinary and collaborative thought through blogging and other communications.

Walden University Publishing: http://www.publishing.waldenu.edu