Value-Added Models (VAMs): Caveat Emptor

Margarita Pivovarova, Audrey Amrein-Beardsley, and Jennifer Broatch

ABSTRACT
Value-added models (VAMs) are being used in education to link the contribution of individual teachers and schools to students’ learning. The use of VAMs has been surrounded by controversy and high-profile public debates. On April 8, 2014 the American Statistical Association (ASA) released a statement on VAMs related to their use in education practice. In this article, we lay out the discussion of the main points raised in the ASA statement within the large amount of scholarly literature published over the past decade in statistical, education, and economics journals. We identify the issues that are critical for the understanding of the VAMs’ strengths and weaknesses, and related consequences of their use for high-stakes decision-making. We conclude that the cautionary points raised in the ASA statement are supported by the findings in the existing research that, with a few exceptions, challenges the assumptions underlying the use of VAMs and demonstrates the issues that should be taken into consideration when using VAMs for consequential decisions.

1. Introduction
Value-added models (VAMs) were introduced in education with the hope that they might help to objectively measure the amount of “value” that a teacher “adds” to (or detracts from) student learning and achievement from one school year to the next. Statistically, the data on students’ standardized test scores are fitted to measure the growth in student achievement over time while controlling for students’ prior testing histories, and sometimes controlling for student-level socio-demographic characteristics (e.g., race, gender, ethnicity, levels of poverty, students’ levels of English language proficiency, special education status). Additional student-level variables such as attendance, suspension, retention records, and classroom and school-level variables are often included, if available. The value-added estimate is then the teacher fixed effect not explained by other terms in the model, or the residual that remains when all possible controls are included. The resulting value-added estimates are used in teacher evaluation systems as one of the few indicators of teacher quality, often to determine high-stakes decisions. (A basic specification of the value-added model includes prior year test scores (\(A_{i,t-1}\)), student (\(X_{it}\)), and classroom (\(Z_{jt}\)) characteristics, and a teacher fixed effect (\(\mu\)), where \(i\) indicates individual student, \(j\) indicates classroom, and \(t\) denotes time period or grade, and is represented by the following equation: 
\[
A_{ijt} = \beta_0 A_{i,t-1} + \beta_1 X_{it} + \beta_2 Z_{jt} + \mu + ijt.
\]
Student growth percentiles are estimated using a nonparametric quantile regression where student achievement in one period is modeled as a conditional density function of prior years’ achievement. Specifically, the following model is commonly estimated: 
\[
Q_{jt} (\gamma | A_{i,t-1}, \ldots, A_1) = \sum_{j=1}^{t-1} \sum_{j=1}^{m} \psi_m (A_j) p_m (\gamma),
\]
where \(Q_{jt}\) is a student’s place in the test score distribution in current period, \(t\) is a number of previous test scores, \(m\) is a number of polynomials to smooth the nonlinearity of the conditional function \(\gamma\).

While they are usually used in conjunction with other common but more subjective indicators of teacher quality such as observational scores and student or parent survey responses, VAM estimates have a larger weight in teacher evaluation compared to other indicators of teachers’ effectiveness, and thus raise more concerns.

Provided that the use of value-added estimates by schools and school districts is primarily associated with accountability and its consequences, VAMs have generated much controversy. Debate, dispute, and disagreement have accompanied the use of VAMs in education, as well as the large-scale educational policies (e.g., Race to the Top, Elementary and Secondary Education Act) advancing and incentivizing VAMs for such educational uses. To date, 44 states throughout the U.S. plus D.C. have adopted and have at least begun implementing VAMs to evaluate and, in many cases, make consequential decisions about teachers in their states (Collins and Amrein-Beardsley 2014). More recently, however, an “opt out” movement against standardized tests has gained strength across a number of states with New York likely to eliminate the role of testing in teacher evaluation (Taylor 2015).

On April 8, 2014 the American Statistical Association (ASA) released a statement on using VAMs in education (ASA 2014). In a short, accessible, and easy-to-understand statement, ASA pointed out a number of issues with VAMs that should be taken into account when using VAMs in educational contexts.
especially for high-stakes decision-making purposes (e.g., teacher merit pay, tenure, termination). Among the major issues the ASA highlighted were concerns about: (a) the high complexity of the statistical models at the foundation of VAMs, which impedes their interpretation and pragmatic use; (b) the lack of the accompanied measures of precision (e.g., confidence intervals, standard errors of measurement); (c) the strong assumptions underlying the interpretation of VAM output (e.g., that statistical controls are sophisticated enough to account for the nonrandom assignment of students into classrooms); (d) the sensitivity of VAMs to the statistical specification used to generate the output; (e) concerns about the causal interpretations of VAM output; (f) the use of standardized test scores for measuring teacher value-added when test scores are meant only to measure student not teacher achievement; and (g) the use of VAM output to make consequential decisions. The ASA also emphasized the need to use VAMs in conjunction with other measures of teacher effectiveness, and only for descriptive inference and information purposes, and not for consequential decisions.

Given the highly sensitive public policy nature of the topic, and the immediate response to the ASA statement from researchers working on the topic (Chetty, Friedman, and Rockoff 2014a; Pivovarova, Broatch, and Amrein-Beardsley 2014) as well as the important consequences of the debate for thousands of teachers and America’s public education system in general, we deemed it important to elaborate on the points made by the ASA in its recent statement. We support our views with evidence from the scholarly research conducted on VAMs over the past decade. We build our arguments using the literature that represents the current state of knowledge among academic professionals and experts about VAMs and their potential for use in teacher evaluation. (As opposed to the selective literature upon which Chetty et al. 2014a based their response to the ASA statement, we draw on the more comprehensive set of articles published across fields and using various methodological approaches. Chetty et al. (2014b) only cite 13 references to critique the ASA’s statement, one of which was the actual statement itself, leaving 12 external citations in total. Of these 12 external citations, three are references to their two forthcoming studies and a replication of these studies’ methods; three have thus far been published in peer-reviewed journals.)

This essay, however, is not a comprehensive review of the literature published on VAM. Rather, we present a current state of what is known about VAMs and their use in K–12 settings, with emphases on the outstanding issues and concerns. More specifically, we point out what we see as the five main issues of contention as identified in the ASA statement and discussed in the literature. We also broadly discuss the technical issues and concerns about the statistical properties of VAMs, while paying particular attention to the concerns that imply that the use of VAMs for holding teachers responsible for students’ test score outcomes may be misleading, counterproductive. We also contrast the ASA (2014) points with the critical response by Chetty et al. (2014a) and position both the ASAs statement and Chetty et al’s response within the current academic literature. We believe that no meaningful discussion of VAMs can be devoid of the spectrum of scholarly opinion and research conducted on the topic. In our discussion, we draw on the literature published in statistical, education, and economics journals alike.

2. **Point 1: Correlation Versus Causation**

The first major point of discussion focuses on whether the users of VAMs are aware that “VAMs typically measure correlation, not causation.” According to the ASA, “effects—positive or negative—attributed to a teacher may actually be caused by other factors that are not captured in the model” (p. 2). However, the inferences from VAM-derived indicators of teacher quality are typically based upon their causal capacities as assumed.

This point of critique is very important as it has major policy implications. The most optimistic estimates imply an additional 3-month learning gain for a teacher at the 84th percentile of the effectiveness distribution as compared to an average teacher (Hanushek and Rivkin 2010). Chetty, Friedman, and Rockoff (2014b) made the highly publicized claim that replacing a teacher in the bottom 5% by an average teacher would increase college attendance by 2.2% and earning at age 28 by 1.3%. It is then inferred that there is a causal relationship between teachers and students’ achievement. Many educational policies based on teacher rankings via VAMs hinge on the assumption that these models include all relevant causal factors, yet policymakers may not be aware that this critical assumption does not hold for the great majority of such models. Lohr (2012, 2015) showed that it was necessary for the top 2% of teachers as ranked by their value-added scores to be excluded from the final model in order to achieve the highly publicized result. Further, Johnson (2015) questioned the legitimacy of looking at value-added estimates calculated from student achievement on exams at a time where the examinations were not supported by any strong incentives to do one’s best.

Another concern relates to the proportion of variance in students’ test score gain that is explained by a teacher. This teacher-explained share of overall variability in standardized test score gains is estimated as low as 3% (Rivkin, Hanushek, and Kain 2005) with the most recent published estimates being in the range of 1%–14% (cited in ASA 2014; see also Goldhaber, Brewer, and Anderson 1999). According to ASAs estimates this means that between 86% and 99% of the variance in student's standardized test scores can be explained by factors not influenced by the teacher.

In a 2004 special issue of the Journal of Educational and Behavioral Statistics devoted to VAMs, Rubin, Stuart, and Zanutto (2004) wrote an important article that included the argument clearly expressing the concern about the lack of causal foundation (see also Wainer 2004, 2011). In this piece, Rubin, Stuart, and Zanutto (2004) presented the issue within a potential outcomes framework, or what is known in statistical inference literature as the Rubin causal model (see also Rubin 1978; Rosenbaum and Rubin 1983; Holland 1986). Using this framework, the authors demonstrated on purely statistical grounds that value-added estimates cannot be considered causal unless a set of strong or “heroic” assumptions are agreed upon and imposed. Moreover, authors of four additional studies within the

---

1 As Rubin et al. (2004) described the set of those assumptions, “If it is impossible to obtain overlapping covariate distributions using matching or sub-classifications, the conclusion should be that reliable causal inferences cannot be drawn from existing data without relying on explicitly stated, and typically heroic, assumptions.”
same special issue also collectively note how it is likely impossible to distinguish between two effects in particular: the effects of teaching practice, which are of interest to policy-makers and are so-called policy-malleable parameters, and the effects of school and home contexts that are rarely observable and tangible, and which therefore distort valid inferences (Kupermintz 2003; Raudenbush 2004; Reckase 2004; Rubin, Stuart, and Zanutto 2004). More recently, Castellano, Rabe-Hesketh, and Skrondal (2013) reinforced the point by showing that no matter how statistically precise the model is and how many factors are being controlled for, it is not possible to distinguish between two types of effects: the policy parameter that captures school practice and instruction and influences that are outside of a teacher’s control including peer and neighborhood effects (see also Coleman et al. 1996; Berliner 2013, 2014; Good 2014).

Accordingly, “anyone familiar with education will realize that this [is] …fairly unrealistic” (Rubin, Stuart, and Zanutto 2004, p. 108). “This” refers to the assumption made in the quasi-experimental or experimental2 studies to infer the causal impact of teachers on students’ outcomes (Kane et al. 2013; Chetty, Friedman, and Rockoff 2014b, 2014c). The major limitation of that assumption is that even if students are randomly assigned to teachers within a school (or vice versa, teachers are assigned to students in a random fashion), no statistical procedure can account for factors that influence student achievement outside of the randomization level, such as school or neighborhood. As has been pointed out numerous times by researchers (Amrein-Beardsley 2008; Braun 2008; Betenbener 2009b; Reardon and Raudenbush 2009; Baker et al. 2010; Briggs and Domingue 2011; Harris 2011; Scherrera 2011; Koedel, Mihaly, and Rockoff 2015) and also emphasized by the ASA (2014), random assignment of students to a teacher or a school is rare if ever the case in educational practice (Paufler and Amrein-Beardsley 2014), and even if random assignment was used, such an experiment could not address environmental confounders. Below we discuss the non-random assignment of students to teachers and its consequences for bias in VAM estimates.

One of the approaches in the literature to identify the above-mentioned bias in value-added estimates is to perform a falsification test replacing a given year teacher assignment with the following year assignment (also known as Rothstein’s falsification test, Rothstein 2010). However, while this test would indicate the presence of the bias, it would not be able to accurately quantify it. Many contributions to the literature have argued that the test does not provide a measure of the size of the bias (Goldhaber and Chaplin 2015; Chetty, Friedman, and Rockoff 2016). The only model-free way to access the causal impact of the difference between two teachers (which requires little bias in VAMs) is through random assignment, though one can approximate this with some quasi-experimental design. The most well-known to-date quasi-experimental approach rests on comparisons of teachers who switch schools, with estimates of the bias derived via this strategy ranging from a low of 2.6% in quasi-experimental settings (Chetty et al. 2014a) to a high of 50% in some cases in purely experimental studies with random assignment (Kane et al. 2013).

The second issue related to the correlation versus causation issue mentioned prior is seen through Chetty et al.’s (2014a) critique of the ASA statement: “Many of the important concerns about VAM raised by the ASA have been addressed in [four] recent experimental and quasi-experimental studies” (p. 2). Chetty et al. (2014a) argued that these experiments and quasi-experiments have already solved the “causation versus correlation” issue. This statement however contradicts other research evidence how the nonrandom assignment of students into classrooms constrains VAM users’ capacities to make causal claims (Rothstein 2009, 2010, 2014; Paufler and Amrein-Beardsley 2014).

Research shows that teachers with the same intrinsic qualities might get different value-added scores, simply because of the student populations they are assigned to teach (Hill, Kaplan, and Umland 2011). This is a serious issue when judging the validity of the inferences derived via VAMs, especially when it comes to social programs that high-stakes decisions are dependent on them (Rothstein 2009; Baker et al. 2010; Newton et al. 2010; Jones, Buzick, and Turkcan 2013). Collectively, these findings imply that students’ socio-economic status (SES), other student background factors (e.g., students English language proficiency and special education status), and the nonrandom assignment of students into classes (and teachers into classrooms) form the basis for potential biased estimates of teacher effectiveness. So far, none of the studies, regardless of whether a quasi-experimental design or otherwise has been used, has demonstrated that all these outside but critical factors can be properly accounted or controlled for (see, e.g., Rothstein 2014; Chetty et al. 2014a, 2014b, 2014c).

For example, the authors of the Measures of Effective Teaching (MET) study sponsored by the Bill and Melinda Gates Foundation, which is one of the experimental studies mentioned in Chetty et al. (2014a) stated that “we cannot say whether the [VAM] measures perform as well when comparing the average effectiveness of teachers in different schools…given the obvious difficulties in randomly assigning teachers or students to different schools” (Kane et al. 2013, p. 38). Additionally, VAM-estimates have been shown to be biased for teachers who teach more homogenous sets of students, primarily with lower levels of prior achievement, despite the levels of sophistication in the statistical controls used across VAMs (Newton et al. 2010; Wright 2010; Guarino et al. 2012; Goldhaber, Walch, and Gable 2012; Ehler et al. 2014). In the reevaluation of the Los Angeles Times teacher rankings, Durso (2011) demonstrated that the teacher effect derived from 1 year of the data reflected and was highly related to the composition of students in the incoming class. More recently, Loeb, Soland, and Fox (2014) investigated the effectiveness of teachers for different groups of students and found that while in general teachers who are effective at teaching English language learners (ELL) are also effective with non-ELLS, there are factors that make some of the teachers more effective with ELLs than with non-ELLS as judged by their value-added (see also Newton et al. 2010).

In an article that greatly clarified that point, Raudenbush (2004) wrote about how without randomized experiments we may never be able to ascertain the desired assessments of

---

2 Experimental studies are referred to as studies that employ random assignment of treatment and control units such as teachers to students. Quasi experimental design in the context of this essay and in relation to VAMs is understood as a naturally created plausibly random variation in treatment and control units such as teachers switching schools.
teacher effectiveness, or make valid inferences about school and teacher effects, even with the complex set of controls put into place to help control for the bias that comes about due to the issues caused by nonrandom assignment (see also Rothstein 2009, 2010, 2014; Corcoran 2010; Newton et al. 2010; Goldhaber, Walch, and Gabele 2012; Guarino et al. 2012; Paufler and Amrein-Beardsley 2014). Even in experimental settings, it is still not possible to distinguish between the effects of school practice, which is of interest to policy-makers, and the effects of school and home context, which are neither separate nor discrete. There are many factors at the student, classroom, school, home, and neighborhood levels that confound causal estimates.

For instance, consider moving disadvantaged students from a poor performing school with less effective teachers to an excellent school in an affluent neighborhood with a mixture of teacher effectiveness. Would we be able to separate out the effects of an increase in teacher effectiveness, the change in the school, the introduction of a new peer group, or the change in the parental or neighborhood influence? Clearly, there are many factors beyond the control of researchers that confound estimates.

3. **Point 2: Low Correlations With Observational Scores**

ASA emphasized in its (2014) statement that “attaching too much importance to a single item of quantitative information is counterproductive—in fact, it can be detrimental to the goal of improving quality.” This is rooted in W. Edwards Deming’s (1994) ideal system of education where “pupils from toddlers on up through the university take joy in learning, free from fear of grades and gold stars,” and “teachers take joy in their work, free from fear in ranking” (p. 62; see also Lohr 2012, 2014, 2015).

Indeed, low correlations between the estimates derived from VAMs and observational scores where the same teachers are observed is another issue of great concern. The recent availability of large datasets, including the Bill & Melinda Gates Foundation’s MET dataset that include both teacher value-added and other measures of teacher effectiveness or quality (e.g., observational and student survey data), provides the best opportunity to assess the consistency between VAM-based assessments and those from direct observations. If measures of teacher quality derived from standardized test scores (value-added estimates) and indicators of teacher effectiveness generated through other means (observations, surveys, etc.) are consistent with each other then we should observe high level of correlation between them. This is not the case, however, according to estimates illustrated in at least four recent studies (Hill, Kapitula, and Umland 2011; Grossman et al. 2014; Harris, Ingle, and Rutledge 2014; Polikoff and Porter 2014) where findings imply low levels of consistency among such measures—negligible to low correlations, all below $r = 0.4$, even with highly trained and monitored observers. Related, the estimates of the correlations between mathematics and English/language arts value-added and either teacher observational scores or student surveys of teacher quality have been found to be in the range of $0.15 \leq r \leq 0.55$ (Harris 2009; Rothstein 2009; Hill, Kapitula, and Umland 2011; Rothstein and Mathis 2013). These low to moderate correlations between value-added and other, non-test-based indicators imply that teachers who are ranked high (low) performing by one measure have a high chance to be ranked in the low (high) by another, and when the two measures are combined, the resulting measure is no longer informative.

Value-added measures derived from different assessments including large-scale tests and Stanford Achievement tests (SAT-9) have also been found to have low correlations with teacher observational scores (Hill, Kapitula, and Umland 2011; Grossman et al. 2014). The largest correlation between another measure of teacher instructional practice—alignment with test standards for math and English/language arts, and value-added estimates has been found to be 0.3 in the MET data (Polikoff and Porter 2014).

There might be at least three reasons for such low correlations: (1) the scores derived via different indicators are capturing different, weakly related aspects of teacher quality, (2) at least one of the measures is capturing teacher quality with a great amount of noise or error, or (3) at least one measure is failing to properly estimate teacher's inputs or rather impacts on student learning and achievement. Obviously, given the current state of knowledge in the area, more research is needed to understand the low agreement between traditional (e.g., observations, principal ratings) and newer indicators of teacher effectiveness such as value-added estimates.

Currently, states continue to “play” with different weighting schemes to incorporate at least three measures of teacher performance (for instance, 40% of teacher overall rating being based on VAMs with the remaining 60% accounted for by observational scores, principal evaluations, and/or student/parent surveys). When a composite measure is developed, a consideration should also be given to the amount of variation in each of these measures. Thus, observational scores typically vary much less than the VAM estimates for the same teachers, that is, the distribution of scores is tighter relative to value-added. As a result, independent of the weighting scheme, the differences in rating between teachers to a great extent are disproportionally determined by their value-added scores. Such emphasis on value-added estimates in teacher evaluation schemes raises concerns in the light of relatively low correlations between these and other measures (not test-scores based) of teacher effectiveness.

In a typical educational setting, administrators and practitioners value multiple measures to measure teacher effectiveness. This practice is also aligned with the current Standards for Educational and Psychological Testing (2014) jointly written by the American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME).

4. **Point 3: Issues With Large-Scale Standardized Test Scores**

Another issue concerns using standardized large-scale test scores for estimating teachers’ value-added. As noted by ASA (2014), “VAMs are generally based on [large-scale] standardized test scores, and do not directly measure potential teacher contributions toward other student outcomes.”

In VAMs, standardized tests are the most common measure of student achievement as their use simplifies and makes
efficient the statistical modeling process. Concerns have been raised with using standardized test scores as the primary index to measure teacher effectiveness. Here the two separate points should be distinguished. The first one is about the validity and reliability of standardized test scores as a measure of student achievement, and this point is not a subject of concern in this essay. The second point is whether student test scores are reliable and valid measures of teacher effectiveness, even given the most sophisticated statistical models (Baker et al. 2010). Results of test scores via value-added output are sensitive to previous year’s instruction, school and home resources and conditions, home and out-of-school experiences, summer learning gain and decay, and more (Kane and Steiger 2002; Baker et al. 2010; Schochet and Chiang 2010, 2013; Neal 2013; Berliner 2014; Konstanlopoulos 2014). Comparison of test scores across subjects as well as across widely separated grades is problematic. The reason for that is that the amount of learning represented by 10 points in math is not the same as 10 points in reading. Conversion to percentiles does not address this comparison issue either, since the distance between the same percentiles in math and reading, for instance, might represent different knowledge gains. As noted by a reviewer, this is the same as to compare whether Newton was a better physicist than Mozart was a composer.

Additionally, not all subjects and grades are tested (Baker et al. 2010; Robelen 2012; Gill, Bruch, and Booker 2013, 2014; see also Harris 2011; Richardson 2012). States only have to test students in grades 3–8 and once in high school, annually in mathematics and English/language arts, and less frequently than annually for science. This explicitly excludes from evaluation all teachers who teach other subject areas. Hence, only 20%–30% of teachers have student growth data when state standardized tests are incorporated into all VAM calculations (Harris 2011; Gabriel and Lester 2013; Whitehurst 2013).

Accordingly, VAMs based on large-scale standardized tests measure only a small fraction of a teacher’s contribution to student’s cognitive and noncognitive progress. In their position statement, ASA authors (2014) state that the standardized test scores used in VAMs should not be the only outcomes of interest for policy makers and stakeholders.

A number of separate points relate to standardized tests in general and making comparisons across teachers who teach different subject areas. For instance, the same score gain for the same student on the math test compared to the language or history test might (and is likely to) reflect different actual learning gain. In a number of studies, researchers raised concerns about the sensitivity of the estimates to tested content, especially when using different tests as the key indicators to calculate teacher value-added (Reckase 2004; Papay 2011; Grossman et al. 2014). Reckase (2004) demonstrated that if value-added is different across academic domains and each year’s tests measure different skills, then comparisons across years will not yield the desired (i.e., unbiased) estimates. Also, the vertical scaling (a psychometric procedure to bring the tests of different content difficulty to the common scale to make student performance on these tests comparable across grades) of test scores that is essential to estimate gains across all VAMs have been shown to affect the reliability and validity of teacher and school value-added as well (Martineau 2006; Briggs and Domingue 2013), yet there are some models that do not require vertical scaling (Lockwood et al. 2007; Broatch and Lohr 2012; Isenberg and Walsh 2014).

Value-added estimates also exhibit low correlations between estimates when different outcomes for the same teacher are used (Broatch and Lohr 2012). In perhaps one of the most well-known studies, Papay (2011) found that such estimates widely ranged for the same teachers using different tests (e.g., 0.2 ≤ r ≤ 0.6), and this finding held true when the same student populations were tested and the same test developers developed the tests.

5. Point 4: Model Specification

In their position statement on VAMs, ASA (2014) also expressed concerns about the sensitivity of VAM estimates to model specifications. Specifically, ASA wrote that, “under some conditions, VAM scores and rankings can change substantially when a different model or test is used, and a thorough analysis should be undertaken to evaluate the sensitivity of estimates to different models.”

The sensitivity of teacher value-added estimates to different specifications of the model has captured critical attention of most researchers working on the topic (Ballou, Sanders, and Wright 2004; McCaffrey et al. 2004, 2009; Tekwe et al. 2004; Briggs and Weeks 2011; Lefgren and Sims 2012; Briggs and Domingue 2013; Goldhaber, Goldschmidt, and Tseng 2013; Schochet and Chiang 2013). Thus, researchers estimated correlation between models that include and exclude classroom-level variables, school fixed effects, student fixed and random effects, student fixed effects and student characteristics, as well as models that include prior year test score or same year but different subject test scores, and, finally, between value-added model and student growth percentiles (Ballou, Sanders, and Wright 2004; Lockwood et al. 2007; Briggs and Domingue 2011; Lefgren and Sims 2012). Overall, the correlations between these different specifications range from 0.25 to 0.96 (Raudenbush and Jean 2012; Goldhaber and Hansen 2013). The use of each of these model specifications can be rationalized based on its statistical properties, but the policy question is whether these models yield consistent estimates of effectiveness for the same teacher.

For instance, comparison of teacher effectiveness estimates derived from four specifications of the model (student growth percentiles, VAM with student characteristics only, with addition of classroom variables, and with addition of school fixed effects) using the same data yielded correlations ranging from 0.48 to 0.99 for different subjects’ teachers (Goldhaber, Walch, and Gabele 2012). While these correlations are of moderate size, when teachers are being classified into quintiles or quartiles of effectiveness, this would lead to changes in their ratings. Thus, at least 14% of teachers would be reclassified and face different accountability consequences if a student growth percentile model was used in place of the conventional VAM (Walsh and Isenberg 2015).

As one of the real-world examples of how the divergence in effectiveness rating due to model specification could lead to (un)intended consequences, consider the data on teacher ranking published by the Los Angeles Times. When taken at face
value, teacher ranking provided parents with information about relative performance of teachers (11,500 teachers in a searchable database) and allowed parents to strategically select schools with good teachers. Upon further evaluation by Briggs and Domingue (2011) and then by Durso (2011), these 1 year estimates of teacher effectiveness have been shown to be model-dependent and, thus, ill-suited for comparison among teachers: for instance, only 46% of reading teachers remained in the same effectiveness category with 60% of math teachers staying within the same rank when two different models are used to estimate teacher value-added. In yet another study, researchers have shown that even with 3 years of prior data 25% to 50% of teachers are still at risk of being misclassified into the most and least effective categories (Schochet and Chiang 2013).

We should not see such dramatic jumps across years in value-added estimates for any teacher teaching the same subject area and grade level where he/she is supposedly (in)effective. Hence, the very weak to moderate stability of estimates observed over time across the same teachers serves as a signal that VAMs are not consistently measuring the parameter of interest well (see also Kelly and Monczunski 2007; Papay 2011; Winters and Cowen 2013; Loeb, Soland, and Fox 2014).

To prevent costly mistakes in year-to-year teacher effectiveness ratings (Sawchuck 2015), some of the school districts have recently used sophisticated models that account for a large number of factors such as student characteristics, classroom and grade characteristics, poverty status, transition of students across schools, and relationship between achievement and student characteristics (see, for instance, Isenberg and Hock 2012; Isenberg and Walsh 2014 for DC). However, the more complex models impose strong demands on data quality: at least 3 years of data for an individual teacher (e.g., misclassification of teachers based on the value-added effectiveness estimates can be as high as 35% with 1 year of data and 25% with 3 years of data; Schochet and Chiang 2011, 2013); large samples of students taught by the teachers; and low incidence of missing data.

6. Point 5: Compared to What?

While not explicitly addressed in the ASA statement (2014), a question about the merits of VAMs could be framed as a comparison to its alternatives. Given the absence of hard evidence in favor of one or another measure of teacher performance, the discussion, then, has evolved around whether we should or should not let the perfect be the enemy of the good. Because the VAMs and the estimates derived from them are more easily scrutinized by quantitative analyses, and thus their rates of error more often numerically illustrated and understood, information about their flaws becomes more transparent and grounded in both the methodological but also the policy debates surrounding their weaknesses (Hansen and Goldhaber 2015).

The precision of the value-added estimates of teachers remains the major issue researchers are grappling with. In other words, how much error in teacher ranking are we ready to tolerate when making consequential decisions: for instance, merit pay or contract termination. Obviously, such decisions require a high level of certainty measured by the variability in the teacher value-added estimates: the larger the standard error, the certainty with which high-stakes decisions are made declines, and vice versa. While different specifications of VAMs produce relatively similar estimates in terms of ranking, and also are comparable to ranking derived from student growth percentile models (Goldhaber, Walch, and Gabele 2012; Goldhaber and Hansen 2013), it is not clear what the benchmark for the acceptable error is. Thus, oftentimes, the error rates in VAMs estimates are compared to stability estimates of performance in other high-skilled jobs (see, e.g., Hanges, Schneider, and Niles 1990; Peterson 2010; Good 2014). On the other hand, we could also compare error rates in VAM-based estimates to those of alternative measures of teacher effectiveness such as observational scores, principal ratings, and student surveys. As noted in the position statement of the American Educational Research Association on VAMs (AERA 2015), while most agree that VAMs might be superior to conventional status models, there are still real dangers of the unintended and often negative consequences associated with using VAM estimates for teacher evaluation purposes. The many current and pending lawsuits throughout the nation as a result of arguably unfair teacher-level decisions based on value-added estimates is only one of the examples of such costly dangers (see, e.g., Sawchuck 2015).

7. Conclusions

In this article, we discussed key points raised by the ASA in its (2014) position statement on VAMs countered by Chetty et al. (2014a). We did this to draw the attention of researchers, policymakers, and educators alike to the multi-faceted nature of VAMs, as well as to all of the existing concerns about their use, in particular, for high-stakes decision-making. VAMs are the only alternatives that currently exist on the market for deriving a measure of teacher contribution from student test scores. Given that fact, issues with VAMs should be brought to the forefront of the debate so that states and school boards might make more informed decisions and choices about their appropriate use in their respective teacher evaluation systems. The current state of knowledge is such that only randomization would allow us to rank teachers and make any causal claims about the teacher impact on student achievement. Without randomization, models that are used to assess teacher effectiveness need to include the contribution of all causal factors, but they are not known, and what the proper functional form is also not known. Even use of techniques such as propensity scores cannot address such problems. These deficiencies bias the estimates of teacher effectiveness. Another important concern that remains unresolved but extensively studied is the low agreement among value-added estimates and other traditional indicators of teacher effectiveness, such as observational scores and survey responses.

Related to this point, and to the use of VAMs to rank teachers, is that VAMs also yield low reliability, or intertemporal stability given their estimates. The low reliability will likely lead to misclassification errors in teachers’ ratings and to the undesired and unintended consequences, which are and may continue to bring on lawsuits (see, e.g., Sawchuck 2015) as well as the loss of highly qualified teachers and the retention of less qualified teachers.
We believe in the necessity to continue the debate about the use of VAMs in education, especially given the variety of opinions on the topic and the impact research-informed opinions have on actual policy. This debate can be eloquently summarized by John Tukey: “The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.”

References

American Educational Research Association (AERA) (2015), “AERA Statement on Use of Value-Added Models (VAM) for the Evaluation of Educators and Educator Preparation Programs,” available at http://ed.sagepub.com/content/early/2015/11/11/001318915618385. full.pdf+html. [6]

American Statistical Association (ASA) (2014), “ASA Statement on Using Value-Added Models for Educational Assessment,” available at http://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf. [1,2,3,4,5,6]

Amrein-Beardsley, A. (2008), “Methodological Concerns About the Education Value-Added Assessment System (EVAAS),” Educational Researcher 37, 65–75. [3]

Baker, E., Barton, P., Darling-Hammond, L., Haertel, E., Ladd, H., Linn, R., Ravitch, D., Rothstein, R., Shavelson, R., and Shepard, L. (2010), “Problems With the Use of Student Test Scores to Evaluate Teachers,” Washington, DC: Economic Policy Institute. Available at http://www.epi.org/publications/entry/bp278 [3,5]

Ballou, D., Sanders, W. L., and Wright, P. (2004), “Controlling for Student Background in Value-Added Assessment of Teachers,” Journal of Educational and Behavioral Statistics, 29, 37–65. [5]

Berliner, D. C. (2013), “Effects of Inequality and Poverty vs. Teachers and Schooling on America’s Youth,” Teachers College Record, 115. Available at http://www.tcrecord.org/Content.aspx?ContentID=16889 [3]

—— (2014), “Exogenous Variables and Value-Added Assessments: A Fatal Flaw,” Teachers College Record, 116. Available at http://www.tcrecord.org/Content.aspx?ContentID=17293 [5]

Bettenbener, D. W. (2009b), “Norm- and Criterion-Referenced Student Growth,” Education Measurement: Issues and Practice, 28, 42–51. [3]

Braun, H. I. (2008), “Vicissitudes of the Validators,” in Presentation made at the 2008 Reedy Interactive Lecture Series, Portsmouth, NH. Available at http://www.cde.state.co.us/cdedocs/OPP/HenryBraunLectureReidy2008ppt [3]

Briggs, D., and Domingue, B. (2011), Due Diligence and the Evaluation of Teachers: A Review of the Value-Added Analysis Underlying the Effectiveness Rankings of Los Angeles Unified School District Teachers by the Los Angeles Times, Boulder, CO: National Education Policy Center. Available at http://nepc.colorado.edu/publication/due-diligence. [3,5]

—— (2013), “The Gains From Vertical Scaling,” Journal of Educational and Behavioral Statistics, 38, 551–576. [5]

Briggs, D., and Weeks, J. (2011), “The Persistence of School-Level Value-Added,” Journal of Educational and Behavioral Statistics, 36, 616–637. [5]

Broatch, J., and Lohr, S. (2012), “Multidimensional Assessment of Value Added by Teachers to Real-World Outcomes,” Journal of Educational and Behavioral Statistics, 37, 256–277. [5]

Castellano, K., Rabe-Hesketh, S., and Skrondal, A. (2013), “Composition, Context, and Endogeneity in School and Teacher Comparisons,” Journal of Educational and Behavioral Statistics, 39, 333–367. [3]

Chetty, R., Friedman, J. N., and Rockoff, J. E. (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. Available at http://amstat.tandfonline.com/doi/pdf/10.1080/2330443X.2014.955227. [2,3,6]

——(2014b), “Measuring the Impact of Teachers: Evaluating Bias in Teacher Value-Added Estimates,” American Economic Review, 104, 2593–2632. [2,3]

—— (2014c), “Measuring the Impact of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood,” American Economic Review, 104, 2633–2679. [3]

—— (2016), “Using Prior Test Scores to Access the Validity of Value-Added Models,” Paper presented at the ASSA meetings 2016, San Francisco, CA. [3]

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, F., Mood, A. M., Weinstein, F. D., and York, R. L. (1966), Equality of Educational Opportunity, Washington, DC: U.S. Government Printing Office. [3]

Collins, C., and Amrein-Beardsley, A. (2014), “Putting Growth and Value-Added Models on the Map: A National Overview,” Teachers College Record, 16. Available at http://www.tcrecord.org/Content.asp?ContentId=17291. [1]

Corcoran, S. (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, Educational Policy for Action Series. Providence, RI: Annenberg Institute for School Reform at Brown University. Available at http://files.eric.ed.gov/fulltext/EDES22163.pdf. [4]

Deming, W. E. (1994), The New Economics: For Industry, Government, Education, Cambridge, MA: Massachusetts Institute of Technology (MIT) Center for Advanced Educational Services. [4]

Durso, C. S. (2011), An Analysis of the Use and Validity of Test-Based Teacher Evaluations Reported by the Los Angeles Times, Boulan, CO: National Education Policy Center. Available at http://nepc.colorado.edu/publication/analysis-la-times-2011. [3,6]

Ehert, M., Koeckel, C., Parson, E., and Podgursky, M. J. (2014), “The Sensitivity of Value-Added Estimates to Specification Adjustments: Evidence From School- and Teacher-Level Models in Missouri,” Statistics and Public Policy, 1, 19–27. Available at http://amstat.tandfonline.com/pdf/doi/10.1080/2330443X.2013.856152 [3]

Gabriel, R., and Lester, J. N. (2013), “Sentinels Guarding the Grail: Value-Added Measurement and the Quest for Education Reform,” Education Policy Analysis Archives, 21, 1–30. Available at http://epaa.asu.edu/ojs/article/view/1165. [5]

Gill, B., Bruch, J., and Booker, K. (2013), Using Alternative Student Growth Measures for Evaluating Teacher Performance: What The Literature Says, Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. [5]

Gill, B., English, B., Furgesson, J., and McCullough, M. (2014). Alternative Student Growth Measures for Teacher Evaluation: Profiles of Early-Adopting Districts, Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. [5]

Goldhaber, D., Brewer, D., and Anderson, D. (1999), “A Three-Way Error Components Analysis of Educational Productivity,” Education Economics, 7, 199–208. [2]

Goldhaber, D., and Chaplin, D. D. (2015), “Assessing the Rothstein Falsification Test: Does it Really Show Teacher Value-Added Models are Biased?” Journal of Research on Educational Effectiveness, 8, 8–34. [3]

Goldhaber, D., and Hansen, M. (2013), “Is it Just a Bad Class? Assessing the Long-Term Stability of Estimated Teacher Performance,” Economica, 80, 589–612. [5,6]

Goldhaber, D., Walch, J., and Gabele, B. (2012), “Does the Model Matter? Exploring the Relationships Between Different Student Achievement-Based Teacher Assessments,” Statistics and Public Policy, 1, 28–39. [3,5,6]

Goldhaber, D. D., Goldschmidt, P., and Tseng, F. (2013), “Teacher Value-Added at the High-School Level: Different Models, Different Answers?” Educational Evaluation and Policy Analysis, 35, 220–236. [5]

Good, T. L. (2014), “What Do We Know About How Teachers Influence Student Performance on Standardized Tests: And Why Do We Know So Little About Other Student Outcomes,” Teachers College Record, 116, 1–41. [3,6]

Grossman, P. C., Cohen, J., Ronfeldt, M., and 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. [4,5]

Guarino, C. M., Maxfield, M., Reckase, M. D., Thompson, P., and Wooldridge, J. M. (2012), An Evaluation of Empirical Bayes’ Estimation of Value-Added Teacher Performance Measures, East Lansing,
Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty Pivovarova, M., Broatch, J., and Amrein-Beardsley, A. (2014), “Chetty
