Measuring Teaching Using Value-Added Modeling : The Imperfect Panacea

Jimmy Scherrer

NASSP Bulletin 2011 95: 122 originally published online 25 May 2011
DOI: 10.1177/0192636511410052

NASSP Bulletin
95(2) 122–140
© 2011 NASSP
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0192636511410052
http://bul.sagepub.com
SAGE

Measuring Teaching Using Value-Added Modeling: The Imperfect Panacea

Jimmy Scherrer[1]

Abstract

The use of value-added modeling (VAM) in school accountability is expanding. However, trying to decide how to embrace VAM can be rather nettlesome. Some experts claim it is "too unreliable," causes "more harm than good," and has "a big margin for error," while other experts assert VAM is "imperfect, but useful" and provides "valuable feedback." This article attempts to parse these statements by exploring the underlying statistical assumptions of VAM, the reliability of VAM's estimates, and the validity of the inferences commonly made based on the estimates of VAM. It then goes on to discuss the perverse incentives, unintended consequences, and gaming that might accompany the misuse of VAM. The article concludes that while, in many cases, VAM may be preferable to other commonly used measurement modes, it should never be used as the sole indicator of teacher effectiveness. Rather, it should just be a piece of a larger accountability system.

Keywords

value-added modeling, VAM, accountability, measurement, measuring teaching, teacher effects, high-stakes assessment

Introduction

Developing an accountability system for our education system that includes measuring teaching effectiveness is an interest of many, and for good reasons. An estimated $1.1 trillion was spent nationwide on education for the school year 2009-2010[1]

[1]The University of Pittsburgh, Pittsburgh, PA, USA

Corresponding Author:
Jimmy Scherrer, The University of Pittsburgh, 830 Learning Research and Development Center,
3939 O'Hara Street, Pittsburgh, PA 15260, USA
Email: scherrer@pitt.edu

(U.S. Department of Education, 2010a). Yet, for the first time in history, it has been reported that the Nation's younger generation is less well educated than its parents (National Academy of Science, 2010). Although such statements are difficult to parse, taxpayers have the right to be concerned about what is occurring in public schools.

Any effort to create a quality teaching force should inarguably include a system that holds teachers accountable; teachers do indeed have a large effect on student outcomes (Aaronson, Barrow, & Sander, 2007; Hanushek & Rivkin, 2006; Koedel & Betts, 2005). However, for nearly half a century, economists, sociologists, and educators have discussed the association between economic disadvantage and student achievement (e.g., Bowles, Gintis, & Groves, 2005; Coleman et al., 1966). Richard Rothstein (2004, p. 14) points out that over this time, "no analyst has been able to attribute less than two-thirds of the variation in achievement among schools to the family characteristics of their students." Such a truism leads policy makers and educators to declare current accountability schemes unfair: *If teachers cannot control two thirds of the variation in student outcomes, how can they be held accountable for it?*

Many believe they have found the answer: value-added modeling (VAM). By "controlling" for factors outside of what teachers can influence (e.g., prior achievement, socioeconomic status), VAM isolates teacher "effects," making for a fairer comparison between teachers (when compared with other currently used accountability approaches).

Indeed, the U.S. Department of Education (2010b) reports that more schools would have made AYP if value-added measures were used instead of standard "status models." To nobody's surprise, these "corrections" were mostly made in high-poverty schools where "on-track" growth was occurring.[2] Moreover, there were schools that met AYP using standard "status models" that did *not* meet appropriate growth measures, perhaps suggesting that some more advantaged schools are not meeting appropriate growth standards yet are "off the radar" because of high achievement scores. Measurement experts argue,

> NCLB and federal accountability will never reach their full potential and may even be substantially counterproductive if school performance continues to be measured by the percentage of students meeting proficiency. Such measures largely reward schools for who they teach rather than how well they teach. (Harris, 2009, pp. 345-346)

Hence, the desire for VAM.

This article examines how measuring teachers using value-added modeling may indeed be preferable to other commonly used measurement modes yet still fall short of a panacea and should never be used as the sole indicator of teacher effectiveness. The article begins by reviewing commonly used education measurement models.

Common Measurement Models

School-based accountability programs refer to "systems that use measures of student outcomes—primarily student achievement as measured by test scores—to hold schools accountable for improving the performance of their students" (Ladd, 2007, p. 1). This section reviews four measurement models commonly used in school-based accountability programs.

Status

Status models attempt to measure student performance at one particular time (a snapshot) that is usually compared with a target. These are useful when a teacher wants to know, for example, "What percent of my students are currently labeled as proficient?" The focus of No Child Left Behind (NCLB) is a status model that measures percent "proficient" and compares that measurement to a set target. The major limitation to such models is their lack of reporting growth.

Cohort to Cohort

Cohort-to-cohort models attempt to compare the status of two (possibly more) cohorts' performance at a particular time (i.e., comparing two snapshots). These are useful when a teacher wants to know, for example, "How did my class do on the science test compared to last year's class?" Since cohort-to-cohort models are a type of status model, they too do not measure growth.

Growth

Growth models attempt to measure achievement by tracking scores of the same students over time to determine progress. Growth models are often reported as *gain scores*. These are useful when a teacher wants to know, for example, "How much progress have my students made in their writing scores?" Growth models improve on status models by taking into account where students began (i.e., a baseline score). However, they implicitly, and erroneously, assume that all achievement levels ought to be expected to gain at the same rate and that the teacher (or school/program) is solely responsible for the growth.

Value Added

VAM attempts to measure a fraction of achievement growth over time that can be attributed to a particular teacher (or school/program). These are useful when a teacher wants to know, for example, "How did *my* contributions to student improvement compare to the average teacher?" VAM improves on growth models by controlling for factors known to be related to student achievement (e.g., socioeconomic disadvantages). The next section will discuss in more detail the advantages of VAM (when compared with other common measurement models).

Advantages of Value-Added Modeling

The concept "value-added modeling" has great plasticity. Different models vary from one another, and not surprisingly, outcomes may differ depending on the model chosen. However, a few features are common to just about all currently used value-added models: VAM takes into account family and community factors that contribute to achievement, determines were a student "begins" the school year, and measures his or her "growth" during one academic year. Then, taking the average "growth" of all students (usually over a multiyear period) that are educated by the same individual, a value-added score is assigned to a particular teacher. In theory, this approach holds a teacher accountable for only the expected learning gains for any particular child. This is a huge advantage of VAM when compared with the "status" and "growth" models discussed above and enables fairer comparisons between teachers (cf. Gordon, Kane, & Staiger, 2006).

In theory, VAM claims any deviation from the "expected learning" gain for a particular student (either positive or negative) is caused by the teacher. Thus, the results from VAM can be used as estimations of counterfactual quantities—one can estimate how students of Teacher A would have done if they had been taught by Teacher B. The ability to use statistical modeling to make such claims is enticing to the educational community whose hands are tied when it comes to true experimental designs inside of schools. However, like most statistical models, the validity of the inferences made based on value-added scores depend on a list of assumptions.

Assumptions of VAM

A closer look at the assumptions of VAM uncovers a gross departure from reality. This section of the article interprets VAM in the context of the validity of its underlying assumptions as laid out by Douglas Harris (2009) and will describe how in most normal school situations these assumptions are violated. To be clear, the intent of this section is *not* to build a case against using VAM; rather, it is to begin building a case that using VAM (like other current measures of performances) as the sole indicator of "effectiveness" is folly of the highest order.

> *VAM Assumption 1:* School administration and teamwork among teachers do not have a significant impact on student achievement.

Effective schooling requires more than "effective" teachers (Newmann, Rutter, & Smith, 1989). Indeed, the school environment (created largely by administration and the faculty as a whole) has a tremendous influence on the engagement and performance of students (Bryk & Schneider, 2003; Marks & Printy, 2003). John Easton (2008), director of the Institute of Education Sciences (IES), has identified "Instructional Leadership" as one of the *Five Fundamentals for School Success*.[3] Strong leadership has direct effects on the organization of the school (Leithwood & Jantzi, 2000), and leadership responsibilities have been shown to have a significant effect on academic achievement gains (see Marzano, Waters, & McNulty, 2005).

Furthermore, reports suggest responsive administration increases a school's sense of community (Newmann, Rutter, & Smith, 1989). This sense of community, for example, is essential to enact successful teacher inquiry groups advocated by the U.S. Department of Education (2000). Indeed, researchers have suggested that teachers left in isolated classrooms with little opportunity for collaboration are unlikely to improve instruction on their own (e.g., Elmore, 2004; Goldenberg, 2004; Little, 1982; Stigler & Hiebert, 1999). There is little, if any, evidence to suggest otherwise.

Simply stated, teachers' opportunities to affect student achievement are shaped by their administrators and peers.[4] Therefore, *VAM Assumption 1* is violated.

VAM Assumption 2: Controlling for previous achievement levels is sufficient to account for the impact of past school resources.

Controlling for previous achievement is a hallmark of commonly used value-added models. However, measurements of "previous achievement" are artifacts of the assessments used to create them. There is no assessment that can capture all of an individual's "previous achievement." Consider the following examples.

Imagine a seventh-grade teacher who works in a district that uses value-added scores for high-stakes decisions. Such a climate encourages the seventh-grade teacher to become a *rational teacher*—she or he learns to game the VAM and only focuses on the narrow range of low-level skills that will be on the test whose results will be used to measure her or his "effectiveness." For sure, this approach can produce high-test results, but since the material was not taught for understanding, decay will occur. That is, students will forget the "skills" they were taught since they do not understand their conceptual basis (see, e.g., Koretz, Linn, Dunbar, & Shepard, 1991). The end-of-the-year test may yield high scores, but the fragile knowledge will not transfer to eighth-grade material (Gick & Holyoak, 1983; Loewenstein, Thompson, & Gentner, 1999; Novick, 1988). For the subsequent eighth-grade teacher, there is a big difference between a student who enters his or her classroom knowing how to think versus a student who enters his or her classroom ready to execute algorithms. The latter type of student puts the eighth-grade teacher at a disadvantage.

Comparing low-level state tests to higher-level international tests can further illustrate this argument. For example, Baker et al. (2010) point out that from 2000 and 2006 state and local test scores were rising while PISA (an international exam known for assessing complex skills) scores were dropping (see Baldi et al., 2007). This is referred to as "score inflation," and it can also be seen when comparing scores from low-level state tests to the scores on higher-level NAEP tests (Jacob, 2007; Koretz &

Barron, 1998). Teaching to the test may indeed help raise test scores but will fail to develop understanding.

Moreover, any knowledge obtained through a family trip to the local museum or additional support from an older sibling on homework will be attributed to the classroom teacher (if tested on the exam). Therefore, teachers who have students whose families go to the museum, zoo, local library, or have family members who assist with class work will have an advantage over teachers who have students that do not have such families. Simply stated, all the knowledge a student obtains over the course of the year is not necessarily a result of what her or his classroom teacher did. Since additional opportunities are not randomly distributed (e.g., advantaged students tend to make more trips to the museum), there are limitations to "controlling for previous achievement."[5]

Furthermore, summer decay—the amount of "achievement" a child loses over the summer months (while not in school)—will distort value-added measures. Since students are generally tested in spring, due to summer decay, the knowledge they bring with them in the fall may not be representative of the knowledge displayed on the spring test. Statistically speaking, decay would be less of a concern if it had the same effect on all students. However, research has long reported on the "summer learning gap," where, on average, disadvantaged students endure summer decay while, on average, their more advantaged peers experience out-of-school learning (e.g., Alexander, Entwisle, & Olson, 2007; Cooper, Nye, Charlton, Lindsay, & Greathouse, 1996; Downey, von Hippel, & Broh, 2004; Heyns, 1978). Using reading proficiency as an example, disadvantaged children, on average, lose more than a month in reading achievement over the summer while their more advantaged peers actually gain in reading achievement over the summer (Cooper et al., 1996). The subsequent teacher for the disadvantaged student will be blamed for this summer decay while the teacher of the advantaged student will be credited for the gain.[6]

While attempting to control for previous achievement in VAM is certainly an improvement over other commonly used measurement models, the assumption that this feature can precisely report what a child knows and can "bring to the table" is foolish. Inasmuch as what any teacher can do with a student is largely affected by what the student's previous teachers did with him,[7] as well as the impact of his or her home environment, *VAM Assumption 2* is violated.

> *VAM Assumption 3:* Students' contributions to their own achievement can be measured with student fixed effects[8] that account for the nonrandom assignment of students to teachers.

Most neighborhoods in the United States are strikingly homogenous, and the schools that house the children in each neighborhood are as well. Simply stated, children are not randomly assigned to schools. Advantaged students usually find themselves with other advantaged students, and disadvantaged students find themselves with other disadvantaged students.[9]

These nonrandom living arrangements introduce many factors (e.g., families with the most resources tend to move to school districts with the best teachers, better teachers gravitate toward schools with more resources, differences in social capital) that all complicate the measurement of the value a teacher adds to achievement. Most VAM attempts to "control" for these differences and allow for comparisons between "like" populations. But precision in matching is extremely difficult. For example, qualification for free/reduced lunch often labels a child as having low socioeconomic status. Beyond this single variable, it is difficult to obtain (on a large scale) information about a child's family that would tell more about her or his disadvantagedness (e.g., parents' education level). Disadvantaged is a relative term, and not all disadvantaged children are equally disadvantaged.[10] Qualifying for free/reduced lunch is not the same as being homeless. Failure to account adequately for family differences distorts measures of teacher quality (Ishii & Rivkin, 2009; Meyer, 1997).

Furthermore, even if we were able to build models (and obtain data) that would account for family differences (thereby addressing the issue of nonrandom assignment of children to schools), we still have the problem of nonrandom assignment of children within the same school to classrooms. That is, on average, students within a school are not randomly assigned to teachers. For example, skilled principals often try to pair certain students with certain teachers that can meet their needs: they often create a "gifted" cluster, put English Language Learners (ELL) with teachers who have skills to support their language needs, match certain discipline problems with teachers who have good classroom management, and so on. Teacher effects can only be identified if assignment of students satisfies strong exclusion restrictions, which they do not.

To illustrate, Jesse Rothstein (2010b) developed a falsification test for widely used value-added specifications. He found that fifth-grade teachers had a significant effect on fourth-grade gains. Such a bizarre finding indeed violates the exclusion restrictions for value-added modeling (i.e., the scores cannot be simply reporting teacher effectiveness). *How can future teachers influence students' past achievement?* The answer is that they cannot; therefore, this finding suggests that students are systematically grouped in some manner,[11] thereby violating *VAM Assumption 3* and leading Rothstein to warn, "VAM-based estimates are likely to be misleading about teachers' causal effects" (p. 210).

> *VAM Assumption 4:* A one-point increase in test scores represents the same amount of learning regardless of the students' initial level of achievement or the test year.

With regard to test scores: Is a 10-point increase from a score of 20 to a score of 30 the same as a 10-point increase from a score of 70 to a score of 80? How about a 10-point increase from a score of 90 to a score of 100? Is it harder for a teacher to increase a score of 20 to a score of 30 than it is to increase a score of 90 to a score of 100? Is it easier? How about 10-point gains across content? That is, is a 10-point gain in reading comprehension the same as a 10-point gain in conventions? What about grade levels? Is a 10-point gain in fifth grade the same as a 10-point gain in ninth grade? Which is harder to achieve? Moreover, was a 10-point gain on a test administered 5 years ago the same as a 10-point gain on a current test?

These are psychometric questions that have a lot to do with the domain of behavior that is sampled on any given test, and they are not easy questions to answer. VAM assumes all 10-point gains are equal: A score increase from 20 to 30 is the same as an increase from 90 to 100, a 10-point increase in comprehension is the same as 10-point increase in conventions, a 10-point gain in fifth grade is the same as a 10-point gain in ninth grade, a 10-point gain on a test given 5 years ago is the same as a 10-point gain on a current test.[12] Most psychometricians would hesitate to make such an assertion. For certain, all these 10-point gains cannot mean the same thing, and some of them are surely harder to attain than others; therefore, *VAM Assumption 4* is violated.

VAM Assumption 5: Teachers are equally effective with all types of students.

There are certain teachers who are "born to teach first grade," yet would struggle in fifth grade, for example, and vice versa. If a value-added score is obtained using data over a 5-year period for a teacher who has taught fifth grade all five of those years, the score is really just reporting the teacher's "effectiveness" of teaching fifth grade.[13] It does not ensure a particular teacher would have the same success (or lack thereof) in another grade. Furthermore, if a teacher is assigned more than one grade level over the 5-year span and receives a low value-added score, does the low score reflect the teacher's ineffectiveness of teaching, or is it just one of those grades in which the teacher is ineffective (thereby "bringing down" her or his high score in the other grade)?

Not only does *VAM Assumption 5* imply that teachers' effectiveness does not depend on the grades they teach, it also implies that all teachers are equally effective at teaching all students within a grade level. However, in reality, this is not the case. Some teachers may be amazing at teaching high-achieving students but struggle helping low achieving students "grow." Some may do wonders with English Only (EO) students, but find they have a hard time relating to ELL. Some prefer teaching students of high socioeconomic status, while others would not dream of stepping into such classrooms. There are also studies showing teachers are more effective with students of their own race or ethnicity (Dee, 2004; Hanushek, Kain, O'Brien, & Rivkin, 2005).

Since we have already established that students are not randomly assigned to teachers, the ubiquitous knowledge among practitioners that some teachers are better at teaching a certain type of student (e.g., second graders, gifted, ELL) than are other teachers violates *VAM Assumption 5* (cf. Lockwood & McCaffrey, 2009).[14]

Reliability and Validity Issues

The other major statistical/psychometric issues that are of interest with any type of measurement are the reliability of the estimates and the validity of the inferences that are made about the results. A closer examination of VAM uncovers problems with both.

Reliability

One reason VAM should not be used in isolation when making high-stakes decisions is the weak stability of its estimates. In other words, the models are unreliable. For example, a study by Koedel and Betts (2005) that used value-added modeling to identify effective teachers reported only 35% of teachers in the top quintile during one academic year remained in the top quintile the following year. In fact, "significant fractions of teachers moved up or down by two quintiles or more" when the sample of years were shifted (p. 22). In other words, from year to year, the value-added model was identifying different teachers as the most effective. In reality, 65% of a school's best teachers do not get worse from one year to the next. This nonpersistence of performance implies nonpersistent sources of error and raises grave reliability issues.[15] Other studies have reported similar conclusions regarding the weak stability of estimated teacher effects (e.g., Aaronson, Barrow, & Sander, 2007; McCaffrey, Sass, Lockwood, & Mihaly, 2009).

Schochet and Chiang (2010) conducted another study that illustrates the unreliability of VAM. Using common VAM techniques on data that cover a 3-year period, their results suggest that more than one in every four average-performing teachers (26%) would be identified for special treatment (i.e., labeled as "above average" or "below average"). The same proportion of high-performing teachers as well as low-performing teachers would also be "overlooked" (i.e., classified as average). Even if 10 years of data were obtained, more than 1 out of every 10 teachers (12%) would still be misidentified (see Schochet & Chiang, 2010).[16]

Validity[17]

Questions of validity arise when one starts to make inferences based on value-added scores. To begin with, labeling a teacher as "effective" based on mathematics and English Language Arts (ELA) scores is invalid. There is a lot more to effective teaching than mathematics and ELA scores can report. Even if the wording of the inferences were changed to "the teacher's effectiveness at teaching mathematics and ELA," one ought to question if the tests used to create the value-added score actually measure what they claim to measure. Absent from a lot of the current value-added debates is the quality of the data being used to formulate the scores. The data inputted into most currently used value-added models come from end-of-the-year standardized tests. These tests primary focus on basic, procedural skills and not the high-level thinking and reasoning recommended by current learning theories (Rothman, Slattery, Vranek, & Resnick, 2002; R. Rothstein, Jacobsen, & Wilder, 2008). If tests continually sample low-level items and thin slices of a domain, even the most sophisticated statistical models in the world will not be able to capture important goals of education. Simply stated, "Value-added estimates are based on a set of test scores that reflect a narrower set of educational goals than most parent and educators have for their students" (Braun, Chudowsky, & Koenig, 2010, p. 30). Lauren Resnick (2006) refers to these estimates as "fool's gold," and asserts that if we do not sharpen what is tested, the nation will soon wake up and realize "we will have more and more of the least valuable coin of the realm, while the high levels of achievement we meant to create (through accountability testing) will increasingly elude us" (p. 36). For sure, it is not valid (or fair) to tell parents their child's teacher is "effective" based on such low-hanging fruit.

Moreover, because most VAM use multiple measures, the data that will be used to show "growth" ought to be obtained from vertically scaled assessments. That is, there should be coherence across the sets of knowledge and skills measured at each grade. For example, when calculating a value-added score for a fifth-grade teacher using the results from her or his end-of-the-year fifth-grade test, one usually will use the results of her or his students' end-of-the-year fourth-grade test as a baseline (to control for prior achievement). If the knowledge and skills measured in fourth and fifth grades are completely orthogonal, then discussion of growth has little meaning and any inference about "value added" is invalid.[18]

Furthermore, value-added scores, like other common measures of achievement, are artifacts of the assessment used to produce them. These measurements often do not include a low enough floor to capture true growth. That is, assessments often only test grade-level standards. A low achieving student may indeed have made much growth, yet still not be at the grade-level standard (therefore not captured on the assessment). Stated differently, "no growth" could simply mean "*no growth on what was measured*."

Finally, when two teachers have different value-added scores, we want to be able to say that the difference reflects the "effectiveness-gap" of the two teachers. However, some studies report 30% to 50% of the variation in measured teacher performance is due to sampling error from "noise" in student test scores (McCaffrey et al., 2009). This is an issue of validity. If we cannot be sure the differences between teachers are due to "effectiveness" or sampling error, our inferences are invalid (and quite useless). Indeed, "judgments made about teacher performance based on value-added may be incorrect fairly often" (Harris, 2009, p. 334).

Unintended Consequences: Becoming a Rational Teacher

So far this article has discussed how the underlying assumptions of VAM are violated in real school settings, how the models are unreliable, and how the inferences generally made from the scores are invalid, all of which support the notion that value-added scores should not be used in isolation when making high-stakes decisions (e.g., judging a teacher's "effectiveness"). This next section switches focus away from the actual models to the unintended consequences, perverse incentives, and gaming that accompany high-stakes accountability systems. Similar to previous sections, this section is not intended to discourage the use of VAM. Rather, it attempts to continue building the case that scores obtained through VAM should not be used as the sole indicator of effectiveness. This section discusses issues associated with *Campbell's Law*—the more weight you put on a single quantitative outcome, the more subject the processes that created it will be to corruption and, eventually, the more it will distort the true phenomena that you are interested in (see Adams, Heywood, & Rothstein, 2009). In education, *Campbell's Law* may indeed change practitioner's behavior, but often in unintended ways (see, e.g., Ladd & Zelli, 2002).

Narrowing of the Curriculum

Like other common measurement models, VAM uses results from standardized mathematics and ELA tests to create value-added scores. By placing so much emphasis on mathematics and ELA test scores, a rational teacher may well realize her or his energies spent in other areas are not appreciated. It is not at all surprising to read accounts of science, history, health, art, and music disappearing from the curriculum (see R. Rothstein et al., 2008).

There is also a narrowing occurring within mathematics and ELA. Since standardized tests primarily focus on basic, procedural skills and not the high level conceptual understanding that is called for in standards documents and recommended by current learning theories (Rothman et al., 2002), students are now immersed in impoverished learning environments of test-prep and memorization of basic facts that will appear on the tests. Indeed, teachers often report spending more than a month on test preparation (Haney, 2000) and spending several hours a week drilling students on practice exams (McNeil & Valenzuela, 2000). This environment is especially harmful for disadvantaged students who depend so deeply on schools and will increase the education gap between disadvantaged students and their more advantaged peers (see Anyon, 1997, 2005; Keiser, 2005; Lipman, 2004). For sure, the negative effects on opportunities to learn caused by high-stakes, low-level tests are becoming clear (Koretz & Hamilton, 2006).[19]

Ignoring the Ones Who Need It Most

Most commonly used VAM incorporates previous years' assessment data for each student as a way to "control" for prior achievement. If a student does not have the correct prior achievement data, the student cannot be included in the analysis. As a result, most VAM exclude mobile students.[20] Therefore, a rational teacher may ask, "Why even bother teaching those students?" Using value-added scores for high-stakes decisions provides incentives to spend disproportionately more time with students who have a "complete data set" and less time with those students who are not going to "count" toward the value-added score.[21] This is extremely unfortunate, since a large majority of mobile students are those who are in the dire need of extra support[22] (see R. Rothstein, 2004).

The End of Viewing Students as Our Students

VAM distributes teachers onto a normative scale. That means teachers are compared with each other, and not some criteria. With this type of distribution, there exist a limited number of "above-average" slots. Stated differently, no matter how much teachers improve, half of them will always be "below average." This creates a highly competitive situation. Proponents of high-stakes accountability will assert this type of a highly competitive environment is exactly what will produce results. One must, however, consider the likely consequences of pinning teachers against each other. For example, consider a sixth-grade teacher who, out of the kindness of his or her heart, holds an after-school math club. The teacher opens the door to all sixth graders (and interested parents) who might benefit from additional tutoring. Using VAM (at the teacher level), any learning that might occur during this math club will be attributed to the students' homeroom (i.e., regular) teacher and not the teacher who is conducting the math club. This increase in the homeroom teacher's value-added score may position him or her above the math club teacher on the normative scale (even though he or she was responsible for the learning). Therefore, a rational teacher will restrict who can attend (thereby restricting who can benefit from) the math club. Teachers have no incentive to help students in the school who are not their own, insofar as any knowledge the students may gain will be attributed to someone else. In fact, teachers have all the incentive not to help since an incorrect attribution of knowledge can seriously affect their relative position on the normative distribution.

A similar situation can be illustrated by considering a school that incorporates team teaching: Imagine a school that has two fifth-grade teachers, Raquel and Erin. Raquel does an amazing job teaching mathematics. Erin does an amazing job teaching ELA. The teachers recognize their colleague's strength and decide to "team teach." Raquel will teach her own students mathematics first thing in the morning, and Erin will teach her own kids ELA first thing in the morning. Then, they will switch; Raquel will teach Erin's students mathematics while Erin teaches Raquel's students ELA. At face value, this seems to be a wonderful idea. However, suppose Raquel becomes a rational teacher and begins to worry about her own value-added score. Raquel will quickly realize that any gains in mathematics that she can make with Erin's students would be attributed to Erin, resulting in a potential increase of Erin's value-added score. As an artifact of VAM distributing teachers on a normative scale, any increase in Erin's score has the potential to position Raquel lower on the distribution. This artifact gives Raquel the perverse incentive to alter her instruction when teaching Erin's students.[23]

Discussion

The use of VAM in school accountability is expanding (Schochet, & Chiang, 2010; U.S. Department of Education, 2009). However, trying to decide how to embrace VAM can be rather nettlesome. Some experts claim it is "too unreliable" (Darling-Hammond, 2010), causes "more harm than good" (J. Rothstein, 2010a), and has "a big margin for error" (Ravitch, 2010), while other experts assert VAM is "imperfect, but useful" (Winters, 2010) and provides "valuable feedback" (Wilkins, 2010). It was the intention of this article to illustrate that, depending on how VAM is used, all these statements are true.

Indeed, in many cases, VAM is better than the alternatives (Harris, 2009, Meyer, 1997). It is, however, not the panacea that accountability enthusiasts have been looking for. Educators, statisticians, psychometricians, and economists all question the appropriateness and ability of statistical models alone to rank teachers accurately for high-stakes decisions (e.g., Baker et al., 2010; Braun et al., 2010; Hanushek & Rivkin, 2010; Ishii & Rivkin, 2009; Koedel & Betts, 2005; Linn, 2000; McCaffrey, Koretz, Lockwood, & Hamilton, 2004; Raudenbush, 2004; R. Rothstein et al., 2008; Schochet & Chiang, 2010). In addition to measurement issues (e.g., reliability and validity of estimates), using a single quantified variable for high-stakes decisions can distort goals (Adams et al., 2009; Ladd & Walsh, 2002), discourage good teachers and administrators from working in schools serving disadvantaged students (Ladd & Walsh, 2002), and lead to downright cheating (Jacob & Levitt, 2003). For these reasons, as well as the overall volatility in test scores (Kane & Staiger, 2002), value-added measures should be just a piece of a larger accountability system.

Declaration of Conflicting Interests

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

Funding

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

Notes

1. This is the combined total of money spent at the local, state, and federal levels.
2. Indeed, "status models are not well designed to promote an equity agenda because they inevitably favor the schools with the most advantaged students" (Ladd, 2007, p. 6).
3. Ironically, VAM was used to reach these conclusions.
4. VAM best captures the combined effect of schools and teachers; using currently available accountability data, it is rather vexing to separate the two (see Raudenbush, 2004).
5. Any "pull-out" program could also result in an incorrect attribution of learning. If students are pulled out of the classroom (e.g., additional reading support), any knowledge gleaned during that time would be attributed to the classroom teacher. Teachers who work at schools with less resources and personnel (i.e., less "pull-out" type programs) are thus at a disadvantage when compared with teachers who have such programs.
6. Another way to look at this is to consider two students, one advantaged and one disadvantaged, who score the same on the spring assessment. Although the spring scores are identical, due to summer decay, the disadvantaged child, on average, will enter the subsequent year behind her or his more advantaged peer. Therefore, teachers of disadvantaged students

are at a disadvantage in analyses of spring-to-spring test gains. One solution could be to propose having a test in the fall when students first return. Then, students' end-of-the-year assessment can be compared the beginning of the year and summer decay would not be an issue. While this may seem like a reasonable remedy, two points needs to be considered: (a) a test in the fall would force schools to devote even more time to testing and (b) a fall-to-spring gain score is incredibly easy to game. That is, fall-to-spring comparisons give teachers the perverse incentive to ensure that their students utterly bomb the fall exam as to make their fall-to-spring gain appear impressive.

7. This is a serious concern in districts with high student mobility.
8. Fixed effects can be thought of as a child's innate ability.
9. New segregation patterns in the United States have been referred to as "triple segregation": where segregation is not just about color, but also about poverty and linguistic isolation (Orfield & Lee, 2006).
10. Likewise, not all advantaged students are equally advantaged.
11. This is further evidence that illustrates an administrator's ability to shape a teacher's opportunities to affect student achievement (see *VAM Assumption 1*).
12. At a micro level, *VAM Assumption 4* is asserting that every question on every test is of equal difficulty and measures the same amount of a domain.
13. Based on most commonly used VAM, this score is actually only reporting the teacher's "effectiveness" at teaching fifth-grade English Language Arts and Mathematics. Since being a good fifth-grade teacher involves a lot more than teaching English Language Arts and Mathematics, it is rather unsettling to label a teacher as effective or ineffective based on such a narrow measure. This is further discussed later in the article.
14. It is important to note here that there is also evidence that a teacher's effectiveness varies within a subject (e.g., Lockwood et al., 2007). For instance, a closer examination of scores could yield that a teacher is effective at teaching "procedures" but not "problem solving." Such findings have VAM experts warning, "Caution is needed when interpreting estimated teacher effects because there is potential for teacher performance to depend on the skills that are measured by the achievement tests" (Lockwood et al., 2007. p. 56).
15. One could argue this is more of a validity issue. However, the point being made in this paragraph is that there is no consistency in the measurement results. If scores on measurements are so inconsistent as to be essentially random numbers, the measurements ought to be thought of as unreliable. For sure, this conclusion of unreliability also affects the validity of any inferences made based on the measurements.
16. Since the authors ignored nonrandom sorting of students, their results are most likely conservative.
17. This section discusses issues of validity above and beyond violating the assumptions of VAM already discussed.
18. The new Common Core State Standards might facilitate in vertically aligning assessments (see Center for K-12 Assessment & Performance Management at ETS, 2010). Although, it should be noted that work done by Schmidt, Houang, and McKnight (2005) suggest that vertically linking assessments may lead to more emphasis on the knowledge that is common across grades, thereby leading to a bias in any estimation of teacher effectiveness.

19. See also R. Rothstein et al. (2008), Appendix 4, for actual teacher accounts of goal distortion.
20. Simply defined, a *mobile* student is one who changes schools (midyear or summer). In contrast, a *stable* student in one who does not switch schools.
21. NCLB introduced a similar phenomenon. NCLB judges schools by the number of students whose tests scores are above the "proficient" cutoff. This provides incentives for schools to focus on students that are scoring around the proficiency cutoff (commonly referred to as the "bubble students") and ignore those who are far above or far below the cutoff.
22. This perverse incentive should also be a concern for parents of stable students. Even if a child is stable (i.e., not mobile), if she or he misses the end-of-the-year assessment (maybe due to illness or vacation), then she or he becomes a member of the "incomplete data set" and her or his teacher has incentives to spend a disproportionate amount of energy with other students.
23. Or, perhaps, they both become rational teachers, realize what the other has the incentive to do, and stop team teaching. In either situation, the victims are children.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*, 95-135.

Adams, S. J., Heywood, J. S., & Rothstein, R. (2009). *Teachers, performance pay, and accountability: What education should learn from other sectors.* Washington, DC: Economic Policy Institute.

Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2007). Lasting consequences of the summer learning gap. *American Sociological Review*, *72*, 167-180.

Anyon, J. (1997). *Ghetto schooling: A political economy of urban educational reform.* New York, NY: Teachers College Press.

Anyon, J. (2005). *Radical possibilities: Public policy, urban education, and a new social movement.* New York, NY: Routledge.

Baker, E., Barton, P., Darling-Hammond, L., Haertel, E., Ladd, H., Linn, R., . . . Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers* (Briefing Paper No. 278). Washington, DC: Economic Policy Institute.

Baldi, S., Jin, Y., Skemer, M., Green, P., Herget, D., & Xie, H. (2007). *Highlights from PISA 2006: Performance of U.S. 15-year-old students in science and mathematics literacy in an international context*. Washington, DC: National Center for Education Statistics.

Bowles, S., Gintis, H., & Groves, M. (2005). *Unequal chances: Family background and economic success.* Princeton, NJ: Princeton University Press.

Braun, H., Chudowsky, N., & Koenig, J. (Eds.). (2010). *Getting value out of value-added: Report of a workshop*. Washington, DC: National Academies Press.

Bryk, A. S., & Schneider, B. (2003). Trust in schools: A core resource for school reform. *Educational Leadership*, *60*(6), 40-45.

Center for K-12 Assessment & Performance Management at ETS. (2010). *Developing an internationally comparable balanced assessment system that supports high-quality learning*. Education Testing Services. Retrieved from http://www.k12center.org/publications.html

Coleman, J., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F., & York, R. (1966). *Equality and educational opportunity*. Washington, DC: U.S. Government Printing Office.

Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research*, *66*, 227-268.

Darling-Hammond, L. (2010, September 7). Too unreliable. *The New York Times*. Retrieved from http://www.nytimes.com/roomfordebate/2010/09/06/assessing-a-teachers-value/value-added-assessment-is-too-unreliable-to-be-useful

Dee, T. S. (2004). Teachers, race, and student achievement in a randomized experiment. *Review of Economics and Statistics*, *86*, 195-210.

Downey, D. B., von Hippel, P. T., & Broh, B. A. (2004). Are schools the great equalizer? Cognitive inequality during the summer months and the school year. *American Sociological Review*, *69*, 613-635.

Easton, J. (2008, November). *Goals and aims of value-added modeling: A Chicago perspective*. Paper presented at the National Academies Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Accountability, Washington, DC. Retrieved from http://www7.nationalacademies.org/bota/VAM_Workshop_Agenda.html

Elmore, R. F. (2004). *School reform from the inside out*. Cambridge, MA: Harvard Education Press.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*, 1-38.

Goldenberg, C. N. (2004). *Successful school change: Creating settings to improve teaching and learning.* New York, NY: Teachers College Press.

Gordon, R., Kane, T. J., & Staiger, D. O. (2006). *Identifying effective teachers using performance on the job* (Discussion Paper No. 2006-01). Washington, DC: Brookings Institution Press.

Haney, W. (2000). The myth of the Texas miracle in education. *Education Policy Analysis Archives*, *8*(41). Retrieved from http://epaa.asu.edu/epaa/v8n41/

Hanushek, E. A., Kain, J. F., O'Brien, D. M., & Rivkin, S. G. (2005, February). *The market for teacher quality* (NBER Working Paper No. 11154). Cambridge, MA: National Bureau of Economic Research.

Hanushek, E. A., & Rivkin, S. G. (2006). Teacher quality. In E. Hanushek & F. Welch (Eds.), *Handbook of the economics of education* (Vol. 2, pp. 1051-1078). Amsterdam, Netherlands: North-Holland.

Hanushek, E. A., & Rivkin, S. G. (2010). *Using value-added measures of teacher quality* (CALDER Brief No. 9).Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

Harris, D. N. (2009). Would accountability based on teacher value added be smart policy? An examination of the statistical properties and policy alternatives. *Education Finance and Policy*, *4*, 319-350.

Harris, D. N., & Sass, T. R. (2009). *What makes for a good teacher and who can tell?* (CALDER Working Paper No. 30). Washington, DC: National Center Analysis of Longitudinal Data in Education Research.

Heyns, B. (1978). *Summer learning and the effects of schooling.* San Diego, CA: Academic Press.

Ishii, J., & Rivkin, S. G. (2009). Impediments to the estimation of teacher value added. *Education Finance and Policy*, *4*, 520-536.

Jacob, B. (2007, January). *Test-based accountability and student achievement: An investigation of differential performance on NAEP and state assessments*. (NBER Working Paper No. 12817). Cambridge, MA: National Bureau of Economic Research.

Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics*, *118*, 843-877.

Kane, T. J., & Staiger, D. O. (2002). Volatility in school test scores: Implications for test-based accountability systems. In D. Ravitch (Ed.), *Brookings papers on education policy* (pp. 235-283). Washington, DC: Brookings Institution.

Keiser, D. (2005). Learners not widgets: Teacher education for social justice during transformational times. In N. Michelli & D. Keiser (Eds.), *Teacher education for democracy and social justice* (pp. 31-55). New York, NY: Routledge.

Koedel, C., & Betts, J. R. (2005). *Re-examining the role of teacher quality in the educational production function* (Working Paper No. 07-08). Columbia: University of Missouri Department of Economics.

Koretz, D. M., & Barron, S. I. (1998). *The validity of gains in scores on the Kentucky Instructional Results Information System (KIRIS).* Santa Monica, CA: RAND.

Koretz, D. M., & Hamilton, L. S. (2006). Testing for accountability in K-12. In R. Brennen (Ed.), *Educational measurement* (pp. 531-578). Westport, CT: American Council on Education.

Koretz, D. M., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991, April). *The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Ladd, H. F. (2007). *Holding schools accountable revisited.* Spencer Foundation Lecture in Education Policy and Management. Retrieved from http://www.appam.org/awards/pdf/2007Spencer-Ladd.pdf

Ladd, H. F., & Walsh, R. P. (2002). Implementing value-added measures of school effectiveness: getting the incentives right. *Economics of Education review*, *21*, 1-17.

Ladd, H. F., & Zelli, A. (2002). School-based accountability in North Carolina: The responses of school principals. *Educational Administration Quarterly*, *38*, 494-529.

Leithwood, K., & Jantzi, D. (2000). The effects of transformational leadership on organizational conditions and student engagement with school. *Journal of Educational Administration*, *38*, 112-129.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, *29*(2), 4-16.

Lipman, P. (2004). *High stakes education: Inequality, globalization, and urban school reform*. New York, NY: Routledge.

Little, J. W. (1982). Norms of collegiality and experimentation: Workplace conditions of school success. *American Educational Research Journal*, *19*, 325-340.

Lockwood, J. R., & McCaffrey, D. (2009). Exploring student-teacher interactions in longitudinal achievement data. *Education Finance and Policy*, *4*, 439-467.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V. N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, *44*, 47-67.

Loewenstein, J., Thompson, L., & Gentner, D. (1999). Analogical encoding facilitates knowledge transfer in negotiation. *Psychonomic Bulletin and Review*, *6*, 586-597.

Marks, H. M., & Printy, S. M. (2003). Principal leadership and school performance: An integration of transformational and instructional leadership. *Educational Administration Quarterly*, *39*, 370-397.

Marzano, R., Waters, T., & McNulty, B. (2005). *School leadership that works: From research to results*. Alexandria, VA: Association for Supervision and Curriculum Development.

McCaffrey, D., Koretz, D., Lockwood, J. R., & Hamilton, L. S. (2004). *The promise and peril of using value-added modeling to measure teacher effectiveness* (Research Brief). Santa Monica, CA: RAND.

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

McNeil, L., & Valenzuela, A. (2000). *The harmful impact of the TAAS system of testing in Texas: Beneath the accountability rhetoric*. Cambridge, MA: Harvard Civil Rights Project.

Meyer, R. H. (1997). Value-added indicators of school performance: A primer. *Economics of Education Review*, *16*, 283-301.

National Academy of Sciences. (2010). *Rising above the gathering storm, revisited: Approaching category 5.* Washington, DC: National Academies Press.

Newmann, F. M., Rutter, R. A., & Smith, M. S. (1989). Organizational factors that affect school sense of efficacy, community, and expectations. *Sociology of Education*, *62*, 221-238.

Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 510-520.

Orfield, G., & Lee, C. (2006). *Racial transformation and the changing nature of segregation.* Cambridge, MA: Civil Rights Project, Harvard University.

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, *29*, 121-129.

Ravitch, D. (2010, September, 7). A big margin for error. *The New York Times*. Retrieved from http://www.nytimes.com/roomfordebate/2010/09/06/assessing-a-teachers-value/assessing-teachers-by-student-scores-is-too-error-riden-to-be-effective?scp=1&sq=a%20big%20margin%20of%20error&st=cse

Resnick, L. B. (2006). Making accountability really count. *Educational Measurement: Issues and Practice*, *25*, 33-37.

Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). *Benchmarking and alignment of standards and testing* (CSE Technical Report 566). Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles.

Rothstein, J. (2010a, September 7). More harm than good. *The New York Times*. Retrieved from http://www.nytimes.com/roomfordebate/2010/09/06/assessing-a-teachers-value/dont-be-too-quick-to-embrace-value-added-assessments

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

Rothstein, R. (2004). *Class and schools: Using social, economic, and educational reform to close the black-white achievement gap*. New York, NY: Teachers College Press.

Rothstein, R., Jacobsen, R., & Wilder, T. (2008). *Grading education: Getting accountability right.* New York, NY: Teachers College Press.

Schmidt, H., Houang, R., & McKnight, C. C. (2005). Value-added research: Right idea but wrong solution. In R. Lissitz (Ed.), *Value added models in education: Theory and applications* (pp. 272-297). Maple Grove, MN: JAM Press.

Schochet, P. Z., & Chiang, H. S. (2010). *Error rates in measuring teacher and school performance based on student test score gains.* Washington, DC: U.S. Department of Education.

Stigler, J. W., & Hiebert, J. (1999). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom.* New York, NY: Free Press.

U.S. Department of Education. (2000). *Before it's too late: A report to the nation from the National Commission on Mathematics and Science Teaching for the 21st Century*. Washington, DC: Author.

U.S. Department of Education. (2009). *Race to the top fund executive summary: Notice of proposed priorities, requirements, definitions, and selection criteria*. Washington, DC: Author.

U.S. Department of Education. (2010a). *The federal role in education*. Retrieved from http://www2.ed.gov/about/overview/fed/role.html

U.S. Department of Education. (2010b). *Interim report on the evaluation of the growth model pilot project*. Washington, DC: Author.

Wilkins, A. (2010, September 7). Valuable feedback. *The New York Times*. Retrieved from http://www.nytimes.com/roomfordebate/2010/09/06/assessing-a-teachers-value/valuable-feedback-for-teachers

Winters, M. (2010, September 7). Imperfect but useful. *The New York Times*. Retrieved September 17, 2010, from http://www.nytimes.com/roomfordebate/2010/09/06/assessing-a-teachers-value/value-added-assessments-are-imperfect-but-useful?scp=1&sq=imperfect%20but%20useful&st=cse

Bio

Jimmy Scherrer is part of the Learning PolicyCenter at the Learning Research and Development Center, University of Pittsburgh, USA. His work lies at the intersection of instruction and policy. Current projects include *Improving Teacher Accountability Systems and Examining the Classroom Discourse Patterns of Effective Teachers*.