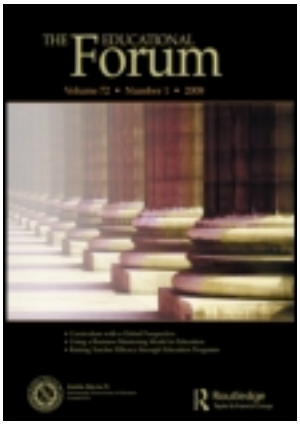

The Educational Forum

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/utef20

Reformers, Batting Averages, and Malpractice: The Case for Caution in Value-Added Use

Daniel Gleason [a]

[a] Illinois Mathematics and Science Academy , Aurora , Illinois , USA

Published online: 19 Mar 2014.

To cite this article: Daniel Gleason (2014) Reformers, Batting Averages, and Malpractice: The Case for Caution in Value-Added Use, The Educational Forum, 78:2, 128-141, DOI: 10.1080/00131725.2013.878427

To link to this article: http://dx.doi.org/10.1080/00131725.2013.878427

The Educational Forum, 78: 128–141, 2014

ISSN: 0013-1725 print/1938-8098 online
DOI: 10.1080/00131725.2013.878427

THE EDUCATIONAL Forum

Reformers, Batting Averages, and Malpractice: The Case for Caution in Value-Added Use

Daniel Gleason
Illinois Mathematics and Science Academy, Aurora, Illinois, USA

Abstract
The essay considers two analogies that help to reveal the limitations of value-added modeling: the first, a comparison with batting averages, shows that the model's reliability is quite limited even though year-to-year correlation figures may seem impressive; the second, a comparison between medical malpractice and so-called educational malpractice, suggests that strict accountability measures within education are out of line with legal precedent.

Key words: *education reform, legal precedent, statistical analysis, value-added modeling.*

The Rise of Value-Added Modeling

Value-added models emerged out of the accountability and data-driven reform movement that became dominant toward the end of the 20th century, after the shifts in education paradigms and national policies of the 1980s. Sahlberg (2011) termed the shift the Global Education Reform Movement, a movement marked by outcome-based (rather than teacher input-based) analysis, core subject focus, corporate modeling, and high-stakes accountability (pp. 100–101). Three ideas spurred the movement forward: one, constructivist pedagogy that shifted educational priorities from teaching to learning; two, the Education for All movement, which insisted that all students be given rich opportunities to learn; and three, the global wave of decentralization that led to increased demand for competition and accountability as government services were auctioned off to the most efficient bidder. These three threads combined to promote the view of education as a commodity, a good that is designed and tested by efficiency-minded managers to benefit all student-consumers. With a more aggressive analysis of the American context in the last 30 years, Gerson (2012) argued that this reform agenda is a neoliberal model of accountability and

Address correspondence to Daniel Gleason, Illinois Mathematics and Science Academy, 1500 Sullivan Rd., Aurora, IL 60506, USA. E-mail: gleason@imsa.edu

choice, and that it has been motivated by two broad claims: one, that our schools are failing and putting the nation at risk of economic decline; and two, that this failure is caused by poor and erratic teaching, and such failure can be corrected by standards, testing, and reward-and-punishment schemes.

In 2001, No Child Left Behind (NCLB) became law, instituting a vast system of student measurement and demanding that schools raise test scores or suffer penalties. With NCLB as the strongest lever of reform to date, schools were under intense pressure to improve test scores however they could, and to measure the progress of specific student groups in order to stay ahead of federal improvement mandates. As Ravitch (2010) noted, schools scrambled for "growth models," or systems that would track students' progress over time. A value-added model developed by statistician William Sanders, who had previously used statistical modeling to inform the agricultural and manufacturing sectors, was available; districts could use his approach to determine how much their students had grown. Crucially, however, Sanders's model did not just track student growth over time, but rather tracked that growth and calculated the extent to which individual teachers contributed to student score gains (Ravitch, 2010, p. 179).

The value-added method appealed to many supervisors, as well as policy makers of all political stripes. With an emphasis on using numbers provided by the district and analyzed by a computer, the method is thus much cleaner and more efficient than time-consuming classroom visits or analyses of student portfolios. Further, the mathematical method can prompt many appealing inferences. Several studies (Gordon, Kane, & Staiger, 2006; Hanushek & Rivkin, 2004) used statistics to calculate the cumulative effect of having top-scoring teachers in multiple years; the studies asserted that the achievement gap between races and income groups could be closed in four to five years. Such heady calculus appealed across the political spectrum, as Ravitch (2010) noted: "Liberals liked the prospect of closing the achievement gap, and conservatives liked the possibility that it could be accomplished with little or no attention to poverty, housing, unemployment, health needs, or other social and economic problems" (p. 182). Liberals and conservatives alike were seduced by the notion that a statistical focus on teacher quality could advance educational outcomes, and quickly.

In many states today, value-added modeling (VAM) is a key aspect—and in some cases the centerpiece—of teacher evaluation. In 2009, President Obama's Race to the Top initiative made VAM a central component of state applications for large federal education grants; states were encouraged to "differentiate effectiveness using multiple rating categories that take into account data on student growth . . . as a significant factor" (U.S. Department of Education, 2009, p. 9). Since 2010, VAM has counted for 50% of a teacher's rating in Louisiana; since 2012, VAM has counted for 20–25% of the rating in New York and 35% in Tennessee; since 2013, in Ohio, VAM has counted for at least 50%; and in Florida, VAM will count for at least 50% of the overall rating starting in 2014.

As policy makers build VAM ratings into more and more teacher evaluation systems across the country, the debate rages over the use of these rankings in high-stakes decisions.

The Case for Value-Added Modeling

Reformers base their arguments on three main principles: (a) teaching quality is central to educational improvement, (b) VAM gives weight and objectivity to otherwise flimsy teacher evaluations, and (c) the test-based modeling aligns well with higher-order student skills and even desirable outcomes later in life.

The reformers begin with the premise that teaching quality is of the utmost importance for student growth. Indeed, in introducing their "Measures of Effective Teaching" (MET) study, the Bill and Melinda Gates Foundation (2010), a leader among test-based reformers, declared: "For four decades, educational researchers have confirmed what many parents know: Children's academic progress depends heavily on the talent and skills of the teacher leading their classroom. Although parents may fret over their choice of school, research suggests that their child's teacher assignment in that school matters a lot more" (p. 3). The point, they argued, is both intuitive and well-researched: teacher quality is crucial to student learning.

Accordingly, reformers seek the best and most objective methods to assess teacher quality, so that districts are able to more effectively winnow the least effective teachers and provide more students with high-quality experiences. The current system, they argue, tends to promote the status quo because its methods are not rigorous. In their report on their MET study, the Gates Foundation noted that teacher evaluation is generally a "perfunctory exercise" (2010, p. 3), guided by a checklist and almost destined to produce near-universal "satisfactory" ratings. Against this backdrop of gauzy sameness, reformers project an image of a bold new system that will bring greater certainty and clarity to teacher evaluation. The Gates Foundation wrote about "objective information" and "new ways to diagnose . . . strengths and weaknesses" as key elements in VAM systems (2010, p. 3).

The numbers, they argued, help cut through the noise of personality and sampling error, allowing principals to see just how effective teachers are at their main task. Principals may be moved by fond feelings for certain teachers, and those feelings may skew qualitative ratings; however, fond feelings will not affect student scores. Principals may also be wowed by a particularly impressive lesson that they have seen, despite the fact that it may not be representative; the numbers are not susceptible to this bias and reflect broader swaths of teaching effectiveness. Further, a lesson may make sense to a principal, but not to the child; the numbers will indicate if the students have absorbed the idea or skill. Overall, reformers have faith in the numbers to identify the real ability of teachers over time; numbers do not play favorites, and they reflect larger patterns of classroom experience than several supervisory visits do.

Reformers point to several studies that highlight the real-world value of VAM evaluation systems. One, the MET study, found a correlation between student score gains on the basic standardized tests and student score gains on a higher-order test of critical thinking. This finding suggests that VAM illustrates not only the teacher's ability to help students grasp narrower, more focused math and English skills, but also the teacher's ability to teach deeper skills and concepts. Thus, the reformers argue that VAM identifies rigorous, thoughtful teaching, not just basic-skills teaching.

Another study, by Harvard University researchers (Chetty, Friedman, & Rockoff, 2011), found that teachers with high VAM ratings are more linked than lower-rated teachers to better adult outcomes for their students, noting, "We find that students assigned to higher VA teachers are more . . . likely to attend college, earn higher salaries, live in better neighborhoods, and save more for retirement." According to this evidence, the VAM approach to teaching is validated not only by a variety of tests in school, but also by the varying life outcomes for students of high- and low-rated teachers. Together, these studies seem to confirm the notion that the VAM numbers provide a more objective, rigorous, and clear view of the value of teachers to their students.

The Case Against Value-Added Modeling

Not so fast, say the critics of statistical accountability measures. These critics, largely union and educational groups, argue that VAM systems are impractical, statistically unreliable, logistically untested, undermined by serious practicality concerns, and dangerous in their pedagogical effects. Assigning score gains is not as easy as it may sound. Many teachers engage with the students throughout the day, some covering similar material. Should the math score gains be attributed to the math teacher or the physics teacher who covered similar ground? Should an ineffective English teacher benefit from his students' gains in verbal scoring when the students have befitted from a strong history teacher and writing center? Even more problematic for the system are teachers not connected at all to the subjects tested. Rothstein (2012) wondered, "What about teachers who don't teach math or reading and so who don't have standardized value-added scores?" These teachers (e.g., of art and gym) seem to exist outside the system, suggesting that VAM numbers can only be used for some teachers. To the critics, such an uneven system hardly seems fair and equitable.

Further, critics argue that even when there is a clear classroom teacher to assign score changes to, such assignment is fraught with causal unreliability. They note that students are the products of home and neighborhood environments, not just classroom experiences; these factors, they argue, exist outside the teacher's control yet have a significant effect on test scores. As Darling-Hammond, Amrein-Beardsley, Haertel, and Rothstein (2011) noted, many factors—class size, instructional time, access to resources, home environment, peer culture, and student health, among others—influence the students' test scores, perhaps even more than the quality of instruction: "the teacher's effort and skill, while important, constitute a relatively small part of this complex equation" (p. 1).

Even if causation could be assured, critics argue the numbers are not statistically meaningful. The wide swings in reported quality for a given teacher sow doubt about the VAM process itself, not the teacher. Darling-Hammond found that one-third of teachers see their ratings change by 30% or more from year to year, nearly half encounter a similar amount of variation from class to class, and—most troubling—14% of teachers could see their ratings change by 30% or more based only on which mathematical model is used to calculate their rating (2010, p. 2). Ratings shift wildly over time: of teachers rated in the top or bottom quintile, only 20–30% of that group will be in the same quintile the next year; the majority of the group will shift outside that band. The researchers also noted that, despite the system's promise to account for various populations equally well, teachers

with special needs students and English language learners tend to score lower than teachers without such students. Another critique takes issue with the method of ranking teachers. Because teachers are plotted on a scale (1–99) according to what percentage of other teachers they outperform, the rankings are inherently relative and may misrepresent quality: the lowest quintile may actually have good gains (just lower than other teachers), and the highest quintile could have little gains (again, but higher than other teachers). Corcoran (2010) noted, "A district with uniformly declining test scores will still have 'high' and 'low' value-added teachers; a district's logical aspiration to have exclusively 'high value-added' teachers is a technical impossibility" (p. 9). Critics have also challenged the connection between VAM results and higher-order thinking. Rothstein (2011), a Berkeley economist, analyzed the MET data and found the correlations to be very weak: More than 30% of teachers in the bottom quartile of the state English Language Arts test were in the top half for the alternative, critical thinking-focused assessment. Rothstein argued that these results are "only slightly better than coin tosses" (2011, p. 5). Opponents of VAM would thus argue that the VAM numbers have seduced policy makers with the promise of objectivity, but actually manifest a fatal unreliability for both teachers and leaders.

The numbers may not present a clear image of the teacher's ability, and no one knows exactly how those numbers will combine with other evaluative measures to make a coherent package, argue the critics. Here is a key distinction between the positions: Where the reformers assume that VAM numbers will bring objectivity to the whole evaluative process (the Gates Foundation's MET analysis claims that "benchmarking against student achievement gains is the best way to know when the evaluation system is getting closer to the truth" [p. 5]), critics urge caution and wonder if those numbers might actually leach objectivity out of the observations. For example, given that school principals will know a teacher's latest value-added ratings when observing a class, Rothstein (2012) wondered whether such knowledge may skew the principals' judgment: "Will they tend to give high ratings to teachers with high value-added scores in order not to call attention to possible flaws in their observational skills, will they tend to offset value-added conclusions in order to save favored teachers who have low value-added, or will they tend to sink unfavored teachers with high value-added?" The fact that these questions are hypothetical and still unanswered is a key point of criticism. The critics would argue that the effect of VAM numbers on the total evaluative package needs to be understood before the statistical systems start rolling out and affecting teachers.

Finally, critics take issue not just with what is dangerously unknown about VAM accountability, but also what is dangerously predictable about it. The high-stakes value-added system encourages teachers and districts to narrow the curriculum in search of higher scores. The VAM system is already linked to rewards and sanctions for teachers (e.g., tenure, bonuses, and dismissals); schools may face similar incentives in the forms of extra funding, restructuring, and even closure. Given these pressures, teachers and administrators are likely to serve their own best interests by focusing on test preparation. A report by the Economic Policy Institute (EPI) (Baker et al., 2010) noted that test preparation leads to changes both *between* the subjects (as time spent on history, science, and art is diverted to math and ELA) and *within* the subjects (as math and ELA instructional materials grow closer to the lower-order, multiple-choice format of the tests), diluting the richness of the educational

world. In addition, the value-added system may lead all too predictably to decreased collaboration, collegiality, and morale among colleagues. It is not hard to imagine that when test scores are used to help determine teacher retention, teachers may want their students to test particularly well and may feel less inclined to share their new ideas and innovative materials with each other, if doing so undercuts their own job security. The EPI report warned that high-stakes VAM systems may distort teacher and district motivations: "Their interest becomes self-interest, not the interests of students" (Baker et al., 2010, p. 18).

The Weight of Analogies: Baseball Players and Doctors

So who is to be believed? Proponents of VAM ratings argue that the numbers are meaningful and relevant, while critics argue that the numbers are simply too unstable and come with too much instructional baggage to be useful. On the surface, this impasse indicates the need for caution because using a controversial system that inspires such rabid disagreement invites real trouble. On a basic level, the strong criticism of the approach should at least encourage a wait-and-see attitude; given the controversies, the theorists and statisticians need more time to hash out the systems, improving reliability while minimizing ancillary problems such as narrowed curricula.

However, the problem is not just that the method spurs disagreement. Further analysis indicates problems with the method that are greater than mere controversy. In fact, analogies with baseball players and doctors suggest that even more weight should be given to the critics of VAM. In different ways, the analogies with these professions help to reveal the deep flaws in using test-score–based statistical modeling to help determine teacher retention and bonuses.

Baseball Players: Unstable Averages and Blurry Glimpses

Proponents of VAM argue that the numbers remain useful despite noise in the system; many other professions, they note, use metrics that are not perfect but are still helpful. In particular, likening a teacher's average to a baseball player's batting average is a popular analogy for reformers. Despite some year-to-year change, the argument goes, both teachers' VAM ratings and batting averages help leaders understand how these professionals are doing, year to year, in their jobs. Chetty and Friedman (2012) argued, "The manager of a baseball team pays attention to a player's batting average even though it too is an imperfect statistic that bounces around over time." Harvard economist Thomas Kane, a leading VAM proponent, makes the very same point: "A teacher's average may vary year to year, but so do the batting averages of professional baseball players. In each case, the measure provides a glimpse (albeit imperfect) of future performance" (Kane & Darling-Hammond, 2012).

Just how imperfect is this glimpse? Are batting averages and VAM scores still useful at predicting future performance, despite the variability in the system? Before addressing questions, consider the analogy: Is batting average a fair comparison to a teacher's VAM score? In many ways it is not. Certainly, hitting a 95-mph fastball is an incredibly daunting task, but a player's batting average nonetheless is both a simpler and more consistent measure of a hitter's ability than VAM is of a teacher. While standing in the batter's box, a hitter really has one primary focus: the ball that is arriving at high speed from 60 feet away. Yes, there may

be choices to bunt, sacrifice fly, or single, but the principal factor to engage with is the ball coming across the plate. A teacher, on the other hand, may have a unitary focus (e.g., teach these eighth-graders about money systems), but the inputs are vastly more varied: 20 or 30 different minds, 20 or 30 willful bodies, different language backgrounds, varying states of care and hunger, teaching resources of different quality, different technology systems, etc. The factors involved in teaching success multiply in a way that the factors involved in batting success do not; classrooms are simply more chaotic and fragmented than the batter's box. Further, while the active and healthy baseball player is measured hundreds of times a year (every time he steps up to the plate), the teacher is measured—or rather, his or her 80 students are measured—just once per year, on one day, no matter the conditions. Certainly, the number of students taking the test does provide some statistical power to these numbers, but as we've noted before, confounding factors within the students themselves also skew the results. Unlike baseball players, teachers do not have the luxury of being assessed hundreds of times a year on a focused task over which they have near-total control.

However, even if we disregard the big gaps in the VAM/batting average analogy, the numbers are still not that impressive. That is, neither of the ratings systems remains very stable year to year, which suggests (to answer our major question above) that the "glimpse" provided by the numbers is imperfect. Batting averages have a relatively middling correlation year to year: Analysis by famed statistician Bill James finds the year-to-year correlation (in statistical terms, *r*, or the coefficient of correlation) as .56. (Other statisticians find a lower number, usually between .3 and .5.) This number is not very large: While it is a positive correlation, it is just above .5, and thus counts as a poor correlation; generally, only correlations over .8 are considered strong. To get a better understanding of that number, statisticians use the coefficient of determination (*r* squared), which indicates just how related the two sets of numbers are: How much of the variance in the second set is accounted for by the variance of the first set? A correlation of .56, it turns out, is quite low here: The coefficient of determination for batting averages is .31, or 31% (.56 × .56), which means that only 31% of a given year's batting average is related to the prior year's average. How well a player hits this year accounts for less than one-third of hitting success next year; more than two-thirds of the subsequent average will be determined by factors beyond performance this year. Obviously enough, batting averages are very imprecise (and dangerously imperfect) tools for predicting future performance.

The VAM numbers are even less stable and predictive. Given all the disputes within the VAM world, the correlations between ratings from year to year are more varied. In fact, this wide variance in correlation numbers is enough to give any statistician or policy maker pause. In addition, however, many of these correlations are even lower than those for batting averages. Goldhaber and Hansen (2010) found a year-to-year correlation of .55 for elementary school math ratings, and .32 for elementary reading ratings. (They also found a slight increase in correlation, to .59 for math and .38 for reading, when three years' worth of data was used as the baseline against which the next year was measured. Interestingly, they found that while the correlation for reading scores continued to rise with baselines longer than three years, the correlations for math scores declined with baselines longer than three years.) A large study of test score data from five Florida counties, 2000–2005 (McCaffrey, Sass, Lockwood, & Mihaly, 2009), found similar correlations, between .2 and .5 for elementary

Table 1. Correlation and Determination in Accounts of Value-Added Modeling

Study	*Correlation coefficient*	*Coefficient of determination*
Goldhaber and Hansen (2010)	.32–.59	.10–.35
McCaffery et al. (2009)	.2–.6	.04–.36
Gary Rubinstein (2012)	.35	.12
Value-Added Research Center (2010)	.2–.62	.04–.38

school teachers and between .3 and .6 for middle school teachers. Working with three years of New York City score data released by *The New York Times*, Rubinstein (2012) found an overall correlation of .35. Using similar data but for two school years, researchers from the University of Wisconsin Madison (Value-Added Research Center, 2010) who worked with New York's Department of Education found correlations between .45 and .62 for grades 4–8 math scores, and correlations between .2 and .33 for grades 4–8 English scores. (These correlation coefficients, as well as the associated coefficients of determination, are collected in Table 1 for your convenience.) Again, to put these correlation numbers in perspective, only correlations of .8 and above count as strong; a correlation of .8 means a coefficient of determination of .64, meaning that 64% of one year's rating has been explained by the previous year's rating. As correlations drop, the coefficient of determination drops much more rapidly: A correlation of .5 means that only 25% of the value has been accounted for by the previous rating: .4, 16%; .3, 9%. A correlation of .3 thus means that a teacher's rating this year only explains about 10% of the next year's rating.

Needless to say, correlations of even .5, a relatively high finding for these data sets, indicate that teacher ratings are only minimally interdependent, and thus very unstable. The "glimpse" that correlations under .5, or even .7, provide is blurry, indeed. If a baseball player's future performance at bat is only marginally linked (31%) to this year's batting average, a teacher's VAM rating is even less certain: This year's rating may indicate, on the high end, a strong amount of overlap and influence (38%), but on the low end, an amount of overlap and influence (4%) that is effectively zero. To summarize Thomas Kane's analogy justifying the variance in VAM ratings through related variance in batting averages: Batting averages only marginally explain future batting averages, and many teacher ratings are even more unstable and disconnected from year to year.

Doctors: Pay-for-Performance and Legal Liability

Doctors provide another rich analogy for our examination of value-added methods. The case of doctors helps to reveal just how difficult (and possibly unfair) it is to rate teachers on the basis of their students' outcomes. While the batting average analogy indicates the great variance within year-to-year performance, the analogy with doctors indicates the myriad and complex factors hidden within apparently clear numbers.

Consider the following thought experiment: What would happen if we tried to determine the effectiveness of doctors by examining their patients' health over time? Problems

in assigning responsibility (and, accordingly, praise or blame) would quickly arise. Does the heart surgeon get credit for the patient's longevity, or the primary physician who has been checking in with the patient and clearing up smaller problems before they gain momentum? And do both doctors deserve poor ratings if the patient, for reasons of his or her own stubbornness or confusion or poverty, fails to take his medicine as instructed? Can ratings systems account for all of these variables, and can we trust these systems to help determine which doctors get raises and which are fired?

Various pay-for-performance (P4P) schemes have been developed by healthcare providers and insurers, but these systems have been criticized for misjudging quality and even undermining care. In general, P4P measures reward doctors and hospitals for delivering the most effective care and for obtaining good patient outcomes. But many problems have been noted, and many of these concerns mirror the problems within VAM. For example, many doctors note that patients may undermine their own treatment. One physician, struggling to get his patient into his office for diabetes check-ins, asked, "She just can't afford to take that much time off from work. Does that make me a worse doctor?" (Chen, 2010). Indeed, a study published in the *Journal of American Medical Association* (Hong et al., 2010) found that the demographic profile of a doctor's patients can seriously influence that doctor's performance rating. The study found that older, insured patients were likely to boost a doctor's rating by visiting often and submitting to a host of tests and procedures; on the other hand, minority, non-English speaking, and uninsured patients were likely to bring a doctor's performance rating down. In addition to unreliability, critics have noted that such performance incentive structures may lead to dangerous and unintended consequences. For example, an outcomes-based payment system may lead doctors to avoid higher-risk patients, who may lower effectiveness ratings (Rosenthal & Frank, 2006). Just as high-stakes tests can encourage curricular narrowing, incentives based on limited and objective physician performance (e.g., the implementation of diagnostic tests) may lead doctors to overemphasize narrow, non-holistic responses. Such a limited focus could undercut strong holistic care, especially for the most vulnerable patients. As the American College of Physicians Ethics warned, "Pay-for-performance initiatives that provide incentives for good performance on a few specific elements of a single disease or condition may lead to neglect of other, potentially more important elements of care for that condition or a comorbid condition" (Snyder & Neubauer, 2007).

The analogy between teacher ratings and doctor ratings is pretty strong in terms of the flaws of the systems. But there is a key difference that undercuts the analogy and reveals that the VAM system is even more dangerous. At issue are the consequences: While doctor ratings have been used primarily for bonus pay or informational purposes, teacher ratings often have a much greater effect on professional livelihood. Most P4P systems are linked to incentives that doctors can earn for meeting or beating benchmarks. On the other hand, teachers have been fired when their VAM ratings mar their evaluations. For this reason alone, the VAM rating system is a more dangerous and ruthless system than the P4P systems rolled out by various healthcare groups.

In addition to the shared flaws in the statistical rating systems for doctors and teachers, doctors reveal the unreliability of teacher ratings in another way. In the legal system,

medical malpractice exists, with huge payouts for victims of negligence and errors by doctors, but "educational malpractice" is almost entirely a fiction. In other words, though teachers may be fired for coming up short in their VAM ratings, the courts have held that teachers are largely immune from legal liability for poor educational outcomes. (This statement does not imply that teachers have not harmed their students or been convicted of crimes; rather, "poor educational outcomes" refers to failures of learning and academic improvement, not criminal behavior that occurs in the classroom.) In these ways the analogy with doctors is the (revealing) photo negative for teachers: statistical ratings mean little for doctors, but they may be legally liable for their mistakes; on the other hand, statistical ratings can cost teachers their jobs, but they have almost never been held liable, in legal terms, for poor performance.

The legal rationale is important. Significantly for the critics of VAM, courts have held that responsibility for the student's poor education cannot be laid at the feet of a single teacher or school system; rather, the failure is likely caused by many other factors. Unlike a surgery in which a passive patient is operated on by a scalpel-wielding doctor, education demands the cooperation and mutual involvement of teacher, student, family, and community. As noted in the research study by Darling-Hammond et al. (2011), nonpedagogical factors (e.g., poverty, nutrition, access to resources, peer culture) intrude upon and affect learning. Indeed, when a plaintiff in California sued his school district for negligence in allowing him to graduate from high school with a fifth-grade reading level (despite the state's requirement of an eighth-grade level for graduation), the Court declared as much:

> *The "injury" claimed here is plaintiff's inability to read and write. Substantial professional authority attests that the achievement of literacy in the schools, or its failure, are influenced by a host of factors which affect the pupil subjectively, from outside the formal teaching process, and beyond the control of its ministers. They may be physical, neurological, emotional, cultural, environmental; they may be present but not perceived, recognized but not identified.* (Peter W. v. San Francisco School District, 1976)

In other words, the school district cannot be held responsible for Peter W.'s illiteracy because too many other influences play a role in educational success. Three years later, the New York Appeals Court noted in *Donohue v. Copiague School District* that many "collateral factors" make causation very difficult, if not impossible, to prove in cases of poor educational outcomes. Admittedly, courts often have many reasons for denying tort claims: In *Peter W.*, the court worried about promoting a flood of litigation; other courts have worried about interfering with the executive branch. However, the multiple factors argument is a highly relevant piece of the puzzle. Unlike negligence charges in medicine, which lay the blame (often) on one doctor's egregious error in judgment, schools and teachers cannot bear the burden alone for educational weakness—the equation for academic success is simply too complicated.

The reluctance of courts to justify notions of "educational malpractice" speaks, at least in part, to the complex interrelation of students, schools, families, peers, and cultures. While the ratings systems for teacher and doctors may be quite similar, especially

in their limitations, the judicial branch suggests that assigning responsibility for teaching outcomes is very complex, if not impossible, and thus blame cannot be laid at the door of individual teachers (or even districts). If baseball players and batting statistics reveal the dangerous instability in VAM ratings, the analogy with medical malpractice reveals the unreliability of those ratings, and the sheer complexity in linking learning (or the failure to learn) to individual teachers. The analogies poke gaping holes in the argument for value-added methods, and suggest that policy makers need to tread carefully in using statistical modeling in their districts.

Proposed Policy: District-Wide Snapshots, Not Spurious Glimpses of Teachers

Obviously enough, the data presented above reveal that caution is in order. So one must consider: In what ways are VAM ratings useful to teachers and districts? And how might one capture the benefits of statistical modeling without suffering its flaws?

I suggest that VAM numbers are useful to districts, not individual teachers. As noted above, these value-added rankings are quite unstable and misleading about individuals. Yet these statistics are very useful to broader informational purposes: How well is the district serving all of its students? As Darling-Hammond (2010) acknowledged, despite her larger problems with NCLB, its disaggregation of student statistics has helped reveal troubling aspects of our schools: "By flagging differences in student performance by race and class, [NCLB] shines a spotlight on long-standing inequalities and has triggered attention to the needs of students neglected in many schools" (p. 67). Certainly, such a spotlight is very useful, if sometimes painful, for a district, and continued attention through broad VAM numbers can help schools track student growth. Further, collecting VAM numbers can help districts and policy makers understand the effectiveness of groups of teachers and help reveal the benefits of certification, professional development programs, and teacher experience for student learning. Keeping these numbers school- or district-wide would help districts measure student progress and understand factors involved in teaching quality without penalizing or stigmatizing teachers for a potentially unreliable test result. The breadth of these sample sizes would also help to ensure reliable results for both student and teacher groups, given the instability and unreliability of any given numbers for individuals.

I also propose that these district VAM ratings be used only for informational purposes. NCLB has helped to reveal the folly of punishing schools for their failures to improve: The schools often descend into chaos, and strong teachers and school leaders are embittered in the process. Forcing a school to restructure often compounds the problems, making failure even more likely in the future. Further, high-stakes punishments such as school closures encourage teaching to the test, gaming the system by hiding low-performing students, and even outright cheating, as some teachers break the rules to avoid losing their jobs. In addition, district-wide incentives, such as shared bonus pay, for meeting certain benchmarks are vulnerable to the same problems as teachers strive to earn extra money. Removing school closures and incentive pay from district-level VAM would allow districts and policy makers to focus on more productive ways of improving schools—methods that do not promote curricular narrowing and trickery.

These two recommendations—recording value-added growth at the district level, and using findings for informational purposes only—would benefit students, teachers, and districts. First, the recommendations would help districts minimize all of the problems listed just above—namely, chaotic and damaging school closings, curricular narrowing, and cheating—while allowing districts to gather important information on students and teacher groups. Second, removing the specter of high-stakes accountability pressures would help teachers move away from internecine competition and back toward collegial cooperation; the teachers would no longer feel that sharing a good idea might lead others to benefit at their own expense. Third, and closely linked to the second reason, teacher morale would improve as teachers and districts were no longer punished for complex test score issues, and as teachers were no longer competing with each other for the best results.

Critics of this proposal may claim that it is too generous and lenient; it will not be tough enough on bad teachers. To that critique, one must counter that in terms of the numbers, it is better to be lenient with bad teachers than strict with good ones. That is, the dangerous consequences of strictness are more grave than the consequences of leniency. With a lenient approach, one will find some poor teachers retained when their low value-added ratings do not count against them; with a strict approach, one will find some capable teachers fired on the basis of low score gains, especially if those statistical findings count (as they do in several states) for 50% or more of an evaluation. But leniency with unreliable numbers and rankings does not mean leniency overall. One must note here that if a bad teacher cannot be fired without using VAM numbers to support the dismissal, the case must be quite weak, indeed.

Conclusion and Next Steps

While proponents of VAM systems argue that the models bring sorely needed objectivity and rigor to otherwise subjective evaluation, the critics reject such notions of objective reliability and also warn against dangerous unintended consequences for students and teachers alike. The weight of evidence supports the critics' position, and suggests that policy makers should consider using VAM models only for informational purposes, and only at the district level. Such a policy will reduce the negative consequences associated with unreliable statistical modeling and high-stakes accountability while also boosting cooperation and morale.

With this caution in mind, one must seek a more definitive assessment of value-added modeling. The next few years offer a wonderful natural experiment that can shed brighter light on the effects of VAM systems. With Race to the Top, many states have implemented value-added components to their teacher evaluation models. Many states have not. Clearly, states with VAM programs can be compared against states without these models, using a wide variety of data points as indicators of success: standardized test scores, including their change over time; measures of critical thinking; measures of curricular breadth and depth; student retention, graduation, and college attendance rates; and teacher retention and job satisfaction data. Will schools in value-added states outperform, through this broad range of indicators, the schools in states without value-added systems? Detailed analyses using large data sets will contribute to the debate about VAM systems, and policy makers should look forward to these studies. Though political considerations are never far from educational policy, one hopes that these studies will ultimately play a decisive role in the fate of value-added modeling.

References

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

Bill and Melinda Gates Foundation. (2010). *Learning about teaching: Initial findings from the Measures of Effective Teaching project*. Seattle, WA: Author.

Chen, P. W. (2010, September 30). Paying doctors for patient performance. *The New York Times*. Retrieved from: http://www.nytimes.com/2010/10/01/health/01chen.html?_r=0.

Chetty, R., & Friedman, J. N. (2012, January 17). The value of test scores. *The New York Times*. Retrieved from: http://www.nytimes.com/roomfordebate/2012/01/16/can-a-few-years-data-reveal-bad-teachers/the-value-of-data-in-teacher-evaluations.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood (NBER Working Paper No. 17699). Cambridge, MA: National Bureau of Economic Research.

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 (Education Policy for Action Series). Retrieved from Brown University Annenberg Institute for School Reform Web site: http://annenberginstitute.org/pdf/valueaddedreport.pdf.

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

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E. H., & Rothstein, J. (2011, September). Getting teacher evaluation right: A background paper for policy makers (Research Briefing: American Educational Research Association & National Academy of Education). Retrieved from http://tx.aft.org/files/gettingteacherevaluationright.pdf.

Gerson, J. (2012). The neoliberal agenda and the response of the teachers unions. In W. H. Watkins (Ed.), *The assault on public education* (pp. 97–124). New York: Teachers College Press.

Goldhaber, D., & Hansen, M. (2010). Is it just a bad class? Assessing the stability of measured teacher performance (Working Paper No. 2010-3). Retrieved from Center for Education Data & Research Web site: http://cedr.us/papers/working/CEDR%20WP%202010-3_Bad%20Class%20Stability%20%288-23-10%29.pdf.

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. Retrievedfrom www.brookings.edu/views/papers/200604hamilton_1.pdf.

Hanushek, E. A., & Rivkin, S. G. (2004). How to improve the supply of high-quality teachers. In D. Ravitch (Ed.), *Brookings papers on education policy* (pp. 7–44). Washington, DC: Brookings Institution Press.

Hong, C. S., Atlas, S. J., Chang, Y., Subramanian, S. V., Ashburner, J. M., Barry, M. J., & Grant, R. W. (2010). Relationship between patient panel characteristics and primary care physical clinical performance rankings. *Journal of the American Medical Association, 304*(10), 1107–1113. doi:10.1001/jama.2010.1287.

Kane, T., & Darling-Hammond, L. (2012, June 24). Should student test scores be used to evaluate teachers? *The Wall Street Journal*. Retrieved from http://online.wsj.com/news/articles/SB10001424052702304723304577366023832205042.

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.

Peter W. v. San Francisco Unified Sch. Dist., 131 Cal. Rptr. 854 (Cal. Ct. App. 1976).

Ravitch, D. (2010). *The death and life of the great American school system: How testing and choice are undermining education*. New York: Basic Books.

Rosenthal, M. B., & Frank, R. G. (2006). What is the empirical basis for paying for quality in health care? *Medical Care Research and Review, 63*(2), 135–157. doi: 10.1177/1077558705285291.

Rothstein, J. (2011). Review of *Learning about Teaching: Initial findings from the Measures of Effective Teaching project*. Boulder, CO: National Education Policy Center. Retrieved from: http://nepc.colorado.edu/thinktank/review-learning-about-teaching.

Rothstein, R. (2012). Teacher accountability and the Chicago teachers strike. *The Economic Policy Institute Blog.* Retrieved from http://www.epi.org/blog/teacher-accountability-chicago-teachers/.

Rubinstein, G. (2012, February 26). Analyzing released NYC value-added data part 1 [blog post]. Retrieved from http://garyrubinstein.teachforus.org/2012/02/26/analyzing-released-nyc-value-added-data-part-1/.

Sahlberg, P. (2011). *Finnish lessons: What can the world learn from educational change in Finland?* New York, NY: Teachers College Press.

Snyder, L., & Neubauer, R. L. (2007). Pay-for-performance principles that promote patient-centered care: An ethics manifesto. *Annals of Internal Medicine, 147*(11), 792–794.

U.S. Department of Education. (2009, November). Race to the top program: Executive summary. Washington, D.C.: U.S. DOE.

Value-Added Research Center. (2010). NYC teacher data initiative: Technical report on the NYC value-added model. Madison, WI: University of Wisconsin–Madison, Wisconsin Center for Education Research. Retrieved from http://schools.nyc.gov/NR/rdonlyres/A62750A4-B5F5-43C7-B9A3-F2B55CDF8949/87046/TDINYCTechnicalReportFinal072010.pdf.