Introduction

Since the original writing of this document to today (Friday April 3, 2015) there are three updates I will share at the end of this introduction.

This collection of research, articles, and links is designed to provide some insight into how New York State's "Student growth Formula" works. A more accurate statement would be how it doesn't work. As you may or may not be aware of this formula is used as a major component in ranking teachers and students in NYS. The formula is based on VAM, or Value – Added Measures.

NYS went as far as posting a video explaining how it works:

https://www.engageny.org/resource/animated-video-student-growth-on-state-tests-2012-13

It is a slickly produced video that states that it doesn't matter who takes the test – high, low, or in between - they get a growth score that is a percentile of all the kids who are "like" them. To be honest it looks good, but when examined carefully, at its core, it is dependent on the idea of VAM. The first four parts of this document clearly show why the tests are not worth the paper they are written on.

The first document, by the ASA (American Statistical Association) shows without question that the VAM used to determine growth scores is clearly neither reliable nor valid, and is not accepted as model of evaluation by the ASA What NYSED is doing is complete junk science. There are some who think that think we need the scores of good students to help teachers. Those scores are completely invalid as well due to the simple fact that VAM is not reliable. If the formula truly worked as it is claimed to it wouldn't matter which of your students do and don't take the test as long as you have been educating all of your students. The bottom line, however, is that VAM simply does not work. Lastly, common sense itself dictates against the validity of a VAM as it pertains to both student growth and teacher effectiveness. How the state really believes it can find a student so much like each and every one of one teacher's to match so perfectly with students of others that they can be compared side by side is incredulous. Another point of contention is that a child who achieved one score on one test at one time being compared to another child, from a different community and social system, who achieves one score on one test at one time does not even have the appearance of being a legitimate comparison. There are simply too many variables to account for.

The second, third, and fourth sections are a mix of articles and research. The second is an article detailing a law suit currently happening in New York State that seriously calls into question the lack of VAM reliability and has evidence to show it. The law suit calls into question the unreliability of VAM as it pertains to even "good" students. The third and fourth sections give more accounts of VAM shortcomings and lack of validity and reliability. The final section was included to show that not only teachers with low VAM scores are upset about the current reform agenda being pushed by Governor Cuomo. It is an open letter from seven former NYS Teachers of the Year written to Governor Cuomo that clearly expresses their professional misgivings in regard to his "reforms." These are seven of the last ten recipients of this prestigious award, which in itself is a statistic that should not be ignored.

The updates are as follows:

1) *The legislators passed Governor Cuomo's budget. In doing so it essentially accepted making the tests 100% of a teachers' evaluations no matter what the other evaluative parts score. That is because the way it is written says that if a teacher has an ineffective score on the NYS Assessment (remember it is based on an unreliable, statistically unaccepted, and flawed formula) that the teacher will be labeled as ineffective…in essence the test score supersedes all other potions of evaluations.*
(http://dianeravitch.net/2015/04/01/here-is-the-new-york-state-teacher-evaluation-bill/)

2) As of yesterday 2 NYS REGENTS have spoken very publicly about their dismay and the ttsts themselves

   a) *A strong and very accurate statement was released by NYS Regent Dr. Kathleen Cashin:*

> *"As a Regent of the State of New York, I cannot endorse the use of the current state tests for* ***teacher/principal evaluation since that was not the purpose for which they were developed****. It is axiomatic in the field of testing that tests should be used only for the purpose for which they were designed. They were designed to measure student performance, not teacher effectiveness.* ***The American Statistical Association, the National Academy of Education, and the American Educational Research Association have cautioned that student tests should not be used to evaluate individual teachers****. Nor should these tests be used for student growth measures until there is clear evidence that they are valid and reliable****. The Board of Regents should commission an independent evaluation of these tests to verify their reliability and validity before they are used for high-stakes purposes for students, teachers, principals, and schools****. How can we criticize people for opting out when the tests have not been verified? We need to cease and desist in the use of these tests until such time as we can be confident of their reliability and validity****. If tests do meet those criteria, the tests must be released to teachers and to the public after they are given, in the spirit of transparency and accountability."*
>
> http://dianeravitch.net/2015/04/02/regent-cashin-of-new-york-speaks-out-against-high-stakes-testing/

   b) Her statement is supported by Regent Betty Rosa http://dianeravitch.net/2015/04/02/regent-cashin-of-new-york-speaks-out-against-high-stakes-testing/ Furthermore, in December 2013 (last school year) Regent Rosa herself stated that, she thought "the Common Core program is based on incomplete, manipulated data." She went on to say, "They are using false information to create a crisis, to take the state test and turn it on its head to make sure the suburbs experience what the urban centers experience: failure," http://www.timesunion.com/local/article/Common-Core-divides-state-s-Regents-board-5067470.php

c) As early as 2011 Regent Roger Tilles stated "our present state tests, and the way they are scaled, are not designed to measure growth from year to year. We are years away from actually having in place valid state tests designed to measure growth."……"**At some point a pushback is necessary. For me the time is now."** http://www.washingtonpost.com/blogs/answer-sheet/post/ny-regent-dont-link-teacher-evaluation-to-test-scores/2011/05/17/AFsJFr5G_blog.html

*More recently in September of 2014 Mr. Tillis addressed the fact in an OpEed piece in Newsday that "s*ome students being set up to fail." http://www.newsday.com/opinion/oped/some-students-being-set-up-to-fail-roger-tilles-1.9414375

3) *The law suit against NYS brought by a teacher who was highly effective and then labeled ineffective is essentially stalled….by NYS.*

The state has allowed this teacher to use the "local" score twice and throw out her NYS Growth score; in effect admitting their growth formula is flawed..

Even with these updates I hope you read the document as originally written. The bottom line, however, is simply this:

As stated in my introduction, and more eloquently so by Dr. Cashin, the formula does not work.

This document is merely proof.

VAM: Value-Added Measure

1) American Statistical Association (April, 2014)
    a) Using Value-Added Models for Educational Assessment
    b) http://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf?1420416000028

2) Washington Post Article (October, 2014)
    a) High-achieving teacher sues state over evaluation labeling her "ineffective"
    b) http://www.washingtonpost.com/blogs/answer-sheet/wp/2014/10/31/high-achieving-teacher-sues-state-over-evaluation-labeling-her-ineffective/

3) What's wrong with using data to grade teachers (December, 2014)
    a) by Mercer hall and Gina Shipley
    b) http://america.aljazeera.com/opinions/2014/12/education-data-teachers.html

4) Dangerously Irrelevant/Value-Added Measures (VAM) (April, 2014)
    a) Scott McLeod, J.D., Ph.D., is widely recognized as one of the nation's leading experts on K-12 school technology leadership issues
    b) http://dangerouslyirrelevant.org/resources/value-added-measures

5) Open letter to Governor Cuomo from seven former New York teachers of the year (February, 2015)
    a) article written by seven New York State teachers of the year: Ashli Dreher, 2014 New York State Teacher of the Year; Katie Ferguson, 2012 New York State Teacher of the Year; Jeff Peneston, 2011 New York State Teacher of the Year; Rich Ognibene, 2008 New York State Teacher of the Year; Marguerite Izzo, 2007 New York State Teacher of the Year; Steve Bongiovi, 2006 New York State Teacher of the Year; and Liz Day, 2005 New York State Teacher of the Year. The letter has been published in theAlbany Times Union and, with permission from the authors, reprinted in Washington Post
    b) http://www.washingtonpost.com/blogs/answer-sheet/wp/2015/02/09/you-have-made-us-the-enemy-this-is-personal-7-n-y-teachers-of-the-year-blast-cuomo/

American Statistical Association (April, 2014)

American Statistical Association Promoting the Practice and Profession of Statistics ASA Statement on Using Value-Added Models for Educational Assessment April 8, 2014 Executive Summary Many states and school districts have adopted Value-Added Models (VAMs) as part of educational accountability systems. The goal of these models, which are also referred to as Value-Added Assessment (VAA) Models, is to estimate effects of individual teachers or schools on student achievement while accounting for differences in student background. VAMs are increasingly promoted or mandated as a component in high-stakes decisions such as determining compensation, evaluating and ranking teachers, hiring or dismissing teachers, awarding tenure, and closing schools. The American Statistical Association (ASA) makes the following recommendations regarding the use of VAMs:

• The ASA endorses wise use of data, statistical models, and designed experiments for improving the quality of education.

• VAMs are complex statistical models, and high-level statistical expertise is needed to develop the models and interpret their results.

• Estimates from VAMs should always be accompanied by measures of precision and a discussion of the assumptions and possible limitations of the model. These limitations are particularly relevant if VAMs are used for high-stakes purposes.

o VAMs are generally based on standardized test scores, and do not directly measure potential teacher contributions toward other student outcomes.

o VAMs typically measure correlation, not causation: Effects – positive or negative – attributed to a teacher may actually be caused by other factors that are not captured in the model.

o Under some conditions, VAM scores and rankings can change substantially when a different model or test is used, and a thorough analysis should be undertaken to evaluate the sensitivity of estimates to different models.

• VAMs should be viewed within the context of quality improvement, which distinguishes aspects of quality that can be attributed to the system from those that can be attributed to individual teachers, teacher preparation programs, or schools. Most VAM studies find that teachers account for about 1% to 14% of the variability in test scores, and that the majority of opportunities for quality improvement are found in the system-level conditions. Ranking teachers by their VAM scores can have unintended consequences that reduce quality.

As the largest organization in the United States representing statisticians and related professionals, the American Statistical Association (ASA) is making this statement to provide guidance, given current knowledge and experience, as to what can and cannot reasonably be expected from the use of VAMs. This statement focuses on the use of VAMs for assessing teachers' performance but the issues discussed here also apply to their use for school or principal accountability. The statement is not intended to be prescriptive. Rather, it is intended to enhance general understanding of the strengths and limitations of the results generated by VAMs and thereby encourage the informed use of these results.

Value-Added Models and Their Interpretation

In recent years test-based accountability for schools and educators has become a prominent feature of the education landscape. In particular, the use of sophisticated statistical methods to create performance measures from student achievement data, often through VAMs, has become more prevalent. VAMs attempt to predict the "value" a teacher would add to student achievement growth, as measured by standardized test scores, if each teacher taught comparable students under the same conditions. VAM results are often regarded as more objective or authoritative than other types of information because they are based on student outcomes, use quantitative complex models, and rely on standardized test scores and common procedures for all teachers or schools.

This statement by the American Statistical Association provides guidance as to what can and cannot be reasonably expected, given current knowledge and experience, from use of VAMs. It is intended to enhance general understanding of the strengths and limitations of the results generated by VAMs and thereby encourage the informed use of these results. It is not meant to be prescriptive or advocate any particular VAM specification or promote or condemn specific uses of VAM.

Value-added models typically use a form of regression model predicting student scores or growth on standardized tests from background variables (including prior test scores), with terms in the model for the teachers who have taught the student. The model coefficients for the teachers are used to calculate their VAM scores. In related models known as "growth models" a regression model is fit to predict students' current test scores from previous test scores. A percentile is calculated for each student from the model, relating his or her growth to the growth of other students with similar previous test scores. The median or average of the percentiles of a teacher's students is then used to calculate the teacher's VAM score. The statistical issues underlying the use of these various types of models are similar, and in this statement, the term "VAM" is used to describe both traditional value-added models and growth models. In both types of models, if a teacher's students have high achievement growth relative to other students

with similar prior achievement, then the teacher will have a high VAM score. Some VAMs also include other background variables for the students.

There are a number of key questions states and districts should address regarding the use of any type of VAM. VAMs are being used for the evaluation of individual teachers on the basis of claims that they can measure those teachers' effects on student achievement growth. These questions are concerned with how well VAMs measure these effects and how the results should be interpreted.

• The measure of student achievement is typically a score on a standardized test, and VAMs are only as good as the data fed into them. Ideally, tests should fully measure student achievement with respect to the curriculum objectives and content standards adopted by the state, in both breadth and depth. In practice, no test meets this stringent standard, and it needs to be recognized that, at best, most VAMs predict only performance on the test and not necessarily long-range learning outcomes. Other student outcomes are predicted only to the extent that they are correlated with test scores. A teacher's efforts to encourage students' creativity or help colleagues improve their instruction, for example, are not explicitly recognized in VAMs.

• VAM scores are calculated from classroom-level heterogeneity that is not explained by the background variables in the regression model. Those classroom-level differences may be due in part to other factors that are not included in the model (for example, class size, teaching "high-need" students, or having students who receive extracurricular tutoring). The validity of the VAM scores as a measure of teacher contributions depends on how well the particular regression model adopted adjusts for other factors that might systematically affect, or bias, a teacher's VAM score. The form of the model may lead to biased VAM scores for some teachers. For example, "gifted" students or those with disabilities may exhibit smaller gains in test scores if the model does not accurately account for their status.

• VAM scores are calculated using a statistical model, and all estimates have standard errors. VAM scores should always be reported with associated measures of their precision, as well as discussion of possible sources of biases.

VAMs are complicated statistical models, and they require high levels of statistical expertise. Sound statistical practices need to be used when developing and interpreting them, especially when they are part of a high-stakes accountability system. These practices include evaluating model assumptions, checking how well the model fits the data, investigating sensitivity of estimates to aspects of the model, reporting measures of estimated precision such as confidence intervals or standard errors, and assessing the usefulness of the models for answering the desired questions about teacher effectiveness and how to improve the educational system.

Quality Improvement and Value-Added Models

Statistical science has a rich and continuing history of successful contributions to quality improvement undertakings. While the methods and approaches vary, consensus exists that:

1. The quality improvement process should be monitored and informed using relevant quantitative information;

2. Almost all systems of measurement contain random variation;

3. Attaching too much importance to a single item of quantitative information is counterproductive—in fact, it can be detrimental to the goal of improving quality. In particular, making changes in response to aspects of quantitative information that are actually random variation can increase the overall variability of the system.

When used appropriately, VAMs may provide quantitative information that is relevant for improving education processes. For example, the models can provide information on important sources of variability, and they can allow teachers and schools to see how their students have performed on the assessment instruments relative to students with similar prior test scores. Teachers and schools can then explore targeted new teaching techniques or professional development activities, while building on their strengths.

Using VAM scores to improve education requires that they provide meaningful information about a teacher's ability to promote student learning. For instance, VAM scores should predict how teachers' students will progress in later grades and how their future students will fare under their tutelage. Various studies have demonstrated positive correlations between teachers' VAM scores and their students' future academic performance and other long term outcomes. In a limited number of studies, teachers have been randomly assigned to classes within schools, thus reducing systematic effects that might arise because of assignment of students to teachers.

These studies indicate that the VAM score of a teacher in the year before randomization is positively correlated with the test score gains of the teacher's students in the year after randomization, but the correlations are generally less than 0.5. Also, studies have shown that teachers' VAM scores in one year predict their scores in later years. These studies, however, have taken place in districts in which VAMs are used for low-stakes purposes. The models fit under these circumstances do not necessarily predict the relationship between VAM scores and student test score gains that would result if VAMs were implemented for high-stakes purposes such as awarding tenure, making salary decisions, or dismissing teachers.

The quality of education is not one event but a system of many interacting components. The impact of high-stakes uses of VAMs on the education system depends not only on the statistical properties of the VAM results but on their deployment in the system, especially with regard to

how various types of evidence contribute to an overall evaluation and to consequences for teachers.

It is unknown how full implementation of an accountability system incorporating test-based indicators, such as those derived from VAMs, will affect the actions and dispositions of teachers, principals and other educators. Perceptions of transparency, fairness and credibility will be crucial in determining the degree of success of the system as a whole in achieving its goals of improving the quality of teaching. Given the unpredictability of such complex interacting forces, it is difficult to anticipate how the education system as a whole will be affected and how the educator labor market will respond. We know from experience with other quality improvement undertakings that changes in evaluation strategy have unintended consequences. A decision to use VAMs for teacher evaluations might change the way the tests are viewed and lead to changes in the school environment. For example, more classroom time might be spent on test preparation and on specific content from the test at the exclusion of content that may lead to better long-term learning gains or motivation for students. Certain schools may be hard to staff if there is a perception that it is harder for teachers to achieve good VAM scores when working in them. Overreliance on VAM scores may foster a competitive environment, discouraging collaboration and efforts to improve the educational system as a whole.

Research on VAMs has been fairly consistent that aspects of educational effectiveness that are measurable and within teacher control represent a small part of the total variation in student test scores or growth; most estimates in the literature attribute between 1% and 14% of the total variability to teachers. This is not saying that teachers have little effect on students, but that variation among teachers accounts for a small part of the variation in scores. The majority of the variation in test scores is attributable to factors outside of the teacher's control such as student and family background, poverty, curriculum, and unmeasured influences.

The VAM scores themselves have large standard errors, even when calculated using several years of data. These large standard errors make rankings unstable, even under the best scenarios for modeling. Combining VAMs across multiple years decreases the standard error of VAM scores. Multiple years of data, however, do not help problems caused when a model systematically undervalues teachers who work in specific contexts or with specific types of students, since that systematic undervaluation would be present in every year of data.

A VAM score may provide teachers and administrators with information on their students' performance and identify areas where improvement is needed, but it does not provide information on how to improve the teaching. The models, however, may be used to evaluate effects of policies or teacher training programs by comparing the average VAM scores of teachers from different programs. In these uses, the VAM scores partially adjust for the

differing backgrounds of the students, and averaging the results over different teachers improves the stability of the estimates.

Statistical science has an important role to play in raising the quality of education, through developing and refining statistical models for use in education, providing guidance on designing experiments and interpreting statistical results, and applying quality and process improvement expertise to help guide judgments in the presence of uncertainty. The ASA promotes sound use of statistical methodology for improving education.

Washington Post Article (October, 2014): High-achieving teacher sues state over evaluation labeling her "ineffective"

By Valerie Strauss October 31, 2014

Sheri G. Lederman has been teaching for 17 years as a fourth-grade teacher in New York's Great Neck Public School district. Her students consistently outperform state averages on math and English standardized tests, and Thomas Dolan, the superintendent of Great Neck schools, signed an affidavit saying "her record is flawless" and that "she is highly regarded as an educator."

Yet somehow, when Lederman received her 2013-14 evaluation, which is based in part on student standardized test scores, she was rated as "ineffective." Now she has sued state officials over the method they used to make this determination in an action that could affect New York's controversial teacher evaluation system.

How is it that a teacher known for excellence could be rated "ineffective"?

The convoluted statistical model that the state uses to evaluate how much a teacher "contributed" to students' test scores awarded her only one out of 20 possible points. These ratings affect a teacher's reputation and at some point are supposed to be used to determine a teacher's pay and even job status.

The evaluation method, known as value-added modeling, or VAM, purports to be able to predict through a complicated computer model how students with similar characteristics are supposed to perform on the exams — and how much growth they are supposed to show over time — and then rate teachers on how much their students compare to the theoretical students. New York is just one of the many states where VAM is one of the chief components used to evaluate teachers.

If it sounds as if it doesn't make a lot of sense, that's because it doesn't. Testing experts have for years now been warning school reformers that efforts to evaluate teachers using VAM are not reliable or valid. But reformers, including Education Secretary Arne Duncan, have embraced the method as a "data-driven" evaluation solution championed by some

economists. Earlier this year, the American Statistical Association issued a report slamming the use of VAM for teacher evaluation, saying in part:

*VAMs are generally based on standardized test scores and do not directly measure potential teacher contributions toward other student outcomes.

*VAMs typically measure correlation, not causation: Effects – positive or negative – attributed to a teacher may actually be caused by other factors that are not captured in the model.

Lederman filed her lawsuit against New York State Education Commissioner John King Jr., Assistant Commissioner Candace H. Shyer and the Office of State Assessment of the New York State Education Department, challenging the rationality of the VAM model being used to evaluate her and, by extension, other teachers. The suit alleges that the New York State Growth Measures "actually punishes excellence in education through a statistical black box which no rational educator or fact finder could see as fair, accurate or reliable."

The lawsuit shows that Lederman's students traditionally perform much higher on math and English Language Arts standardized tests than average fourth-grade classes in the state. In 2012-13, 68.75 percent of her students met or exceeded state standards in both English and math. She was labeled "effective" that year. In 2013-14, her students' test results were very similar but she was rated "ineffective." The lawsuit says:

This simply makes no sense, both as a matter of statistics and as a matter of rating teachers based upon slight changes in student performance from year to year.

Superintendent Dolan supported Lederman, saying in an affidavit:

As superintendent of the GNPS, I have personally known Dr. Lederman for approximately 4 years. I have had the opportunity to meet with her personally. I have also reviewed her record of teaching, particularly the performance of her students on New York State

assessment tests. I can personally attest that she is highly regarded as an educator by the administration of GNPS. Her classroom observations have consistently identified her as an exceptional educator. She is widely regarded in the GNPS as someone who brings out the best in her students. She has taught for seventeen (17) years in the GNPS and her record is flawless.

Sharon Fougner, the principal at Elizabeth M. Baker Elementary School, where Lederman teaches, signed an affidavit saying that she believes the awarding of 1 out of 20 possible points to Lederman under VAM is"arbitrary and capricious" and agreed with Dolan that Lederman is an excellent teacher.

Still, the state of New York says she is "ineffective," and offers a teacher no way to appeal the result.

The lawsuit will be worth watching because it is taking on the entire notion of VAM. If VAM were to fall in New York, more legal challenges would be likely in other states.

(Correction: Earlier version said in one place that superintendent called Lederman an administrator. He didn't. He called her an educator.)

What's wrong with using data to grade teachers (December, 2014)

"Ill-conceived ratings systems can wreak havoc on educators' careers"

by Mercer hall and Gina Shipley

New York educators are pushing back forcefully against the state's controversial teacher evaluation system. This spring, the Teachers Association of the cities of Rochester and Syracuse filed a lawsuit against the state, arguing that the ratings metrics unfairly penalize teachers of disadvantaged students. Now Sheri G. Lederman (PDF), a lifelong teacher from Long Island, is challenging her "ineffective" rating as arbitrary and capricious, based on an ill-conceived and misapplied statistical model of teaching quality.

These suits converge on the issue of whether teachers should be judged on the basis of student test scores, and New York state is poised to set a nationwide precedent on the use of value-added testing data in teacher evaluations. While most parents and administrators would agree that educational accountability is essential, thorny questions persist about how the art of teaching should be appraised in a data-driven culture.

Value-added evaluation systems have been celebrated by U.S. Secretary of Education Arne Duncan and his 2010Race to the Top initiative for their potential to distinguish between highly effective and ineffective teachers. Value-added models (VAMs) draw from students' prior test scores and their backgrounds, such as race and socioeconomic status, to forecast how well they ought to score on a current year's standardized exam. If math or English students fail to reach these benchmarks, then their teachers are deemed ineffective. This expectation game, however, springs from Byzantine formulas that have been denounced by the American Statistical Association as wrongly measuring "correlation, not causation."

For example, because tests are given only in certain subjects to certain age groups, 70 percent of educators in Florida last year received VAM rankings based on students or subjects they didn't even teach. New York's system determines whether a teacher is highly effective, effective, developing or ineffective, using a triad of measures: 20 percent based on value-added modeling of students' state test scores, 20 percent on district level assessments and 60 percent on an array of other measures, such as classroom observations. Lederman's value-added classification dropped two rungs in just one year despite having student test scores that were consistently more than double the state average for meeting standards.

One explanation for this change is that New York statisticians rejigger the VAM formula each year, effectively moving the goalposts without informing teachers. Furthermore, researchers at the University Of Colorado at Boulder found that tweaks to the formula for reading outcomes would alter the effectiveness ratings for more than 50 percent of Los Angeles public school teachers. In New York an ineffective rating cannot be appealed, which explains why the impetus behind Lederman's suit is not monetary or political; rather, she seeks to have her score clarified and recalculated.

Erroneous evaluations can have real-world ramifications. The public release of teachers' ratings can damage their professional reputations and set up future employment challenges, including denial of tenure or dismissal. Lederman has job security after a 17-year career, but green teachers, who are increasingly the norm in nationwide classrooms, face serious risks. They have no existing file of other evaluations and, without seniority, are often slotted into classrooms with underperforming students of different learning needs.

The danger of VAMs can be seen in the verdict of a May 2014 study published by the American Educational Research Association, which found no consistent correlation between teachers with high-scoring students and teachers who excelled in other metrics of effective schooling. Across six sample states, the report concluded, "The tests used for calculating VAM are not particularly able to detect differences in the content or quality of classroom instruction."

Teacher evaluations should provide educators with actionable guidance, not simply grade them on their effectiveness in prepping kids for tests.

VAM-style evaluations might work well for internal diagnostics in painting broad-brush district comparisons or in pinpointing areas for teacher training. Yet the shoddiness of specific VAM forecasts raises serious doubts about their use in determining an individual teacher's worth. A 2010 report (PDF) commissioned by the U.S. Department of Education (DOE) found that the error rate for value-added scores can be as high as 35 percent when using only one year of data. A system that could rate 1 in 3 teachers incorrectly is one that essentially plays pin the tail on the donkey with their livelihoods.

One major concern made apparent in the Lederman case is that the metrics of VAMs are opaque. The exact inputs remain unknown, since children enter school with a host of differences in ability, attendance and background. Carol Burris, the principal of South Side High School in Rockville Centre, New York, and a vigorous opponent of high-stakes testing, explained in an interview with The Washington Post that the recipe of the annual value-added

growth measure is unclear. This is a problem: A big goal of evaluations should be staff development, but if administrators do not know the metrics, they cannot mentor their teachers.

Jennifer Wallace Jacoby, an assistant professor of education and psychology at Mount Holyoke College who researches socioeconomically diverse groups of children, suggested to us that in many ways the VAM is a measure of convenience — looking to capture those data points that are easy to compartmentalize (such as standardized test scores) and ignoring those aspects of the classroom dynamic that are messier and more difficult to quantify.

Rather than shy away from learning complexity, states could take into account data streams that encompass the full range of a child's schooling. Vivienne Ming, a visiting scholar at the University of California at Berkeley's Redwood Center for Theoretical Neuroscience and a co-founder of the educational research firm Soccos, explained in an interview that naturalistic data-collection methods begin with scientific inquiry and an assumption that all data points in a student's learning environment, not just standardized assessments, are relevant.

In other words, the goal of teacher evaluations should be to provide educators with feedback and actionable guidance rather than to simply grade them on their effectiveness in prepping kids for tests. Soccos, for instance, develops tech tools to assess behavioral data such as motivation, creativity, social intelligence and metacognitive ability, using algorithms based, in part, on University of Pennsylvania professor Angela Duckworth's measures of student grit and self-regulation and Stanford University professor Carol Dweck's measures of mindset. This behavioral data can be used to predict or improve quality of life outcomes for students beyond Race to the Top's narrow focus on college readiness, including a student's social, emotional and physical health.

In an ideal world, all public servants, not just teachers, would be a part of this endeavor. But as Burris emphasizes, quality-of-life outcomes are not the sole responsibility of schools. "We are an important piece of the puzzle, but we are not the whole puzzle. The courts, social services and politicians are equally responsible for the social well-being of children," she said.

Education is about relationships, not statistics. Reducing Lederman's outstanding classroom observations and her award-winning doctoral dissertation to a mathematical equation devalues the influence she has had in her years of service.

States must build better evaluation models that include a more thoughtful use of data about the true drivers of contemporary learning. If Lederman is victorious, New York might just become the forge for these new kinds of tools.

Mercer Hall is a teacher and co-founder of the American Society for Innovation Design in Education. He is co-editor of theASIDEblog and his work is regularly featured in EdSurge, Edutopia, EdTechMagazine and other forums.

Gina Sipley is a lifelong teacher who has been nationally recognized as a teacher of the future for her commitments to technology, sustainability and social justice. She writes about educational technology for EdSurge and Mic.

Dangerously Irrelevant/Value-Added Measures (VAM) (April, 2014)

~ Scott McLeod, J.D., Ph.D., is widely recognized as one of the nation's leading experts on K-12 school technology leadership issues.

While it seems to make intuitive sense to evaluate teachers based on students' standardized test scores (aka using 'value-added measures,' or VAM), in practice it doesn't seem to work very well. At this time, researchers do not support the incorporation of student test scores into teacher evaluations except in carefully-designed, low-stakes pilot experiments.

Extreme rating volatility

Rating instability in value-added models is very high, resulting in extreme year-to-year and even multi-year volatility:

- United States Department of Education, *Error rates in measuring teacher and school performance using student test score gains*: Value-added estimates for teacher-level analyses are subject to a considerable degree of random error when based on the amount of data that are typically used in practice for estimation. If three years of data are used for estimation, more than 1 in 4 teachers who are truly average in performance will be erroneously identified for special treatment.
- Di Carlo, *The war on error*: A recent analysis of VAM scores in New York City shows that the average error margin is plus or minus 30 percentile points. That puts the "true score" (which we can't know) of a 50th percentile teacher at somewhere between the 20th and 80th percentile – an incredible 60 point spread.
- Economic Policy Institute, *Problems with the use of student test scores to evaluate teachers*: VAM estimates have proven to be unstable across statistical models, years, and classes that teachers teach. One study found that across five large urban districts, among teachers who were ranked in the top 20% of effectiveness in the first year, fewer than a third were in that top group the next year, and another third moved all the way down to the bottom 40%. Another found that teachers' effectiveness ratings in one year could only predict from 4% to 16% of the variation in such ratings in the following year.
- Baker, *You've been VAM-ified*: Even in the more consistently estimated models, half or more of [New York City] teachers move into or out of the good or bad categories from year to year, between the two years that show the highest correlation in recent years. And this finding still ignores whether other factors may be at play in keeping teachers in certain categories. For example, whether teachers stay labeled as 'good' because they continue to work with better students or in better environments.
- Di Carlo, *Reign of error*: When you're looking at the single-year teacher estimates (in this case, for 2009-10), the average spread is a pretty striking 46 percentile points in math and 62 in ELA [English-Language Arts]. Furthermore, even with five years of data, the intervals are still quite large – about 30 points in math and 48 in ELA.

- National Education Policy Center, *Due diligence and the evaluation of teachers*: It is likely that there are a significant number of false positives (teachers rated as effective who are really average), and false negatives (teachers rated as ineffective who are really average) in the *L.A. Times'* [value-added] rating system. Only 46% of reading teachers – and only 60% of math teachers – retain the same effectiveness rating [when the model is altered to better account for students' past performance, peer influence, and other school factors].
- ETS, *Using student progress to evaluate teachers*: If making causal attributions is the goal, then no statistical model, however complex, and no method of analysis, however sophisticated, can fully compensate for the lack of randomization [in schools]. Other identified problems with VAM include inappropriate attribution, missing data, inappropriate assumptions underlying VAM models, and difficulty in obtaining precise estimates of teacher effects, all of which lead to bias in the data.
- Baker, *AIR pollution in New York State?*: The measures are neither conceptually nor statistically accurate. They suffer significant bias … And inaccurate measures can't be fair.

Educational research and policy institutions do not support the use of VAM for teacher evaluation

Because the ratings are so unstable, those organizations that actually look at the peer-reviewed research – and not just ideologically-driven policy advocacy papers – have come out strongly against the use of student test scores for teacher evaluation. These include many of our most respected assessment experts (such as James Popham, Gerald Bracey, and Robert Linn) and educational policy and research institutions such as the National Research Council, the American Educational Research Association, the National Academy of Education, and RAND:

- American Statistical Association, *ASA statement on using value-added models for educational assessment*: Most VAM studies find that teachers account for about 1% to 14% of the variability in student test scores. VAM scores have large standard errors, even when calculated using several years of data. These large standard errors make rankings unstable, even under the best modeling scenarios. Multiple years of data do not help problems caused when models systematically undervalue teachers who work in specific contexts or with specific types of students.
- ETS, *Reliability and validity of inferences about teachers based on student test scores*: Teacher VAM scores should emphatically not be included as a substantial factor with a fixed weight in consequential teacher personnel decisions. Scores may be systematically biased for some teachers and against others.
- National Research Council, *Value-added methods to assess teachers not ready for use in high-stakes decisions*: Too little research has been done on these methods' validity to base high-stakes decisions about teachers on them. VAM estimates of teacher effectiveness should not be used to make operational decisions because such estimates are far too unstable to be considered fair or reliable.
- American Educational Research Association & National Academy of Education, *Getting teacher evaluation right*: Value-added models of teacher effectiveness are highly unstable. Teachers' value-added ratings are significantly affected by differences in the students who are assigned to them, even when models try to control for prior achievement and student demographics. Value-added ratings cannot disentangle the many influences on student progress. Other [teacher evaluation] tools have been found to be more stable. [Using VAM for] high-

stakes, individual-level decisions, as well as comparisons across highly dissimilar schools or student populations, should be avoided.

- RAND, *Evaluating value-added models for teacher accountability*: The research base is currently insufficient for us to recommend the use of VAM for high-stakes decisions. In particular, the likely biases from the factors we discussed … are unknown, and there are no existing methods to account for either the bias or the uncertainty that the possibility of bias presents for estimates. Furthermore, the variability due to sampling error of individual teacher-effect estimates depends on a number of factors — including class sizes and the number of years of test-score data available for each teacher — and is likely to be relatively large. Similarly, rankings of teachers should be avoided because of lack of stability of estimated rankings.
- Annenberg Institute for School Reform, Brown University, *Can teachers be evaluated by their students' test scores?*: In the abstract, value-added assessment of teacher effectiveness has great potential to improve instruction and, ultimately, student achievement. The notion that a statistical model might be able to isolate each teacher's unique contribution to their students' educational outcomes – and by extension, their life chances – is a powerful one. However, the promise that value-added systems can provide such a precise, meaningful, and comprehensive picture is not supported by the data. Annual value-added estimates are highly variable from year to year, and, in practice, many teachers cannot be statistically distinguished from the majority of their peers. Persistently exceptional or failing teachers – say, those in the top or bottom 5 percent – may be successfully identified through value-added scores, but it seems unlikely that school leaders would not already be aware of these teachers' persistent successes or failures.
- Popham, *Teacher evaluation pitfalls*: Despite the current clamor to evaluate teachers' effectiveness on the basis of their students' test scores, no evidence currently exists to show that the tests intended for use in such evaluations are up to the job. Put simply, there is no proof – none at all – that these tests can accurately distinguish between welltaught and badly taught students.
- National Education Policy Center, *Review of two culminating reports from the MET project*: Randomization was significantly compromised, and participating teachers were not representative of teachers as a whole. [Regarding] how best to combine value-added scores, classroom observations, and student surveys in teacher evaluations, the data do not support the MET project's premise that all three primarily reflect a single general teaching factor, nor do the data support the project's conclusion that the three should be given roughly equal weight. . . . Evaluating teachers requires judgments . . . that are not much informed by the MET's masses of data. While the MET project has brought unprecedented vigor to teacher evaluation research, its results . . . offer little guidance about how to design real-world teacher evaluation systems.
- Bracey, *What's the value of growth measures?*: [VAM] cannot permit causal inferences about individual teachers. At best, it is a first step toward identifying teachers who might need additional professional development or low performing schools in need of technical assistance.

Predictable, harmful results from the use of VAM

Despite researchers' and statisticians' strong recommendations against doing so, some states have forged ahead with 'value-added' teacher evaluation systems anyway. Ignoring numerous warnings to the contrary has resulted in predictable, harmful outcomes. For example:

- Washington Post, *A 'value-added' travesty for an award-winning teacher*: Teacher of Year rated unsatisfactory under VAM system.
- Orlando Sentinel, *Teacher evaluation process: unsatisfactory*: One of nation's top high schools, with highest FCAT scores in high-achieving Seminole County, rated as 'needs improvement' under state's teacher evaluation system.
- Amrein-Beardsley, et al., *Value-added model research for educational policy*: Policymakers throughout the country are increasingly embedding score-based (VAM) approaches within educational evaluation and accountability systems. On the other hand, social science researchers are increasingly questioning the methodological, technical, and inferential attributes of these same VAM approaches. . . . Policymakers have come to accept VAM as an objective, reliable, and valid measure of teacher quality. At the same time, [they ignore] the technical and methodological issues.

Legal concerns

Until the measures are more stable, policymakers should note that the legality of VAM is very much in question:

- Baker, Oluwole, & Green, *The legal consequences of mandating high-stakes decisions based on low quality information*: Student growth percentile measures being adopted by states for use in teacher evaluation are, on their face, invalid for this particular purpose. . . . [and] are likely to open the floodgates to new litigation over teacher due process rights. This is likely despite the fact that much of the policy impetus behind these new evaluation systems is the reduction of legal hassles involved in terminating ineffective teachers.
- Pullin, *Legal issues in the use of student test scores and value-added models to determine educational quality*: If VAM is used for high-stakes consequences like salary differentiation, termination, or damage to professional reputation, the potential for successful legal challenge is high. Given the scientific issues associated with VAM methodologies, it is possible that the use of VAM to make a high-stakes decision about an educator would not even survive a rational basis review under Equal Protection analysis.
- NPR, *Teachers union files federal lawsuit challenging Florida teacher evaluations*: Current teacher evaluation system violates equal protection and due process rights of teachers.

Using VAM as one of 'multiple measures'

Some VAM advocates (such as StudentsFirst and the Gates Foundation) have proposed using student test scores as just one of 'multiple measures' to evaluate teachers, along with student surveys, administrator observations, professional portfolios, and other factors. Unfortunately, the instability of the test score component still means that a significant percentage of teachers' evaluations is highly volatile. Do we ask doctors, lawyers, and other professionals to adopt systems in which a large percentage of their evaluation is based on a component that has been shown repeatedly by researchers to be statistically invalid, operationally unreliable, and disproportionately impactful? It's like asking them to eat an ice cream sundae with two scoops of ice cream and one scoop of horse droppings. Even though it's only one part of many, we're still asking them to eat manure…

Another issue worth noting is that even if teacher effects could be teased out, decades of peer-reviewed research show that teachers only account for about 10% of overall student achievement (give or take a few percentage points). Another 10% or so is attributable to other school factors such as leadership, resources, and peer influences. The remaining 80% of overall

student achievement is attributable to non-school factors such as individual, family, and neighborhood characteristics. A few exceptional 'beating the odds' schools aside, these ratios have remained fairly stable (i.e., within a few percentage points) since they were first noted by the famous Coleman Report of the 1960s. Given the overwhelming percentage of student learning outcomes that is attributable to non-teacher factors, it is neither ethical nor legally-defensible to base teacher evaluations on factors outside of their control.

Using VAM as a screening measure

Right now, the best way to use VAM appears to be as a screening mechanism, much like in medicine. Screening procedures used by doctors often have high error rates so they simply are used to identify patients who warrant further investigation. As Douglas Harris, endowed chair at Tulane University and author of *Value-Added Measures in Education*, explains:

[In medicine,] those who are positive on the screening test are given another "gold standard" test that is more expensive but almost perfectly accurate. They do not average the screening test together with the gold standard test to create a combined index. Instead, the two pieces are considered in sequence.

Ineffective teachers could be identified the same way.

Value-added measures could become the educational equivalent of screening tests. They are generally inexpensive and somewhat inaccurate. As in medicine, a value-added score, combined with some additional information, should lead us to engage in additional classroom observations to identify truly low-performing teachers and to provide feedback to help those teachers improve. If all else fails, within a reasonable amount of time, after continued observation, administrators could counsel the teacher out or pursue a formal dismissal procedure.

[read more at *Creating a valid process for using teacher value-added measures*]

Conclusion

Legislation or policies that advocate for the inclusion of student test scores as part of teacher evaluation will have to somehow overcome the significant limitations outlined above in order to be both ethically and legally defensible. In particular, the rating volatility that results in large percentages of teachers bouncing from year to year between excellent, average, and unsatisfactory categories must be drastically reduced. Standardized test scores that purport to be fair, objective, valid, and reliable for student learning purposes appear to be much less so when it comes to evaluating teachers' contributions to that learning. The fact that these technical, methodological, statistical, and implementation challenges still loom large after nearly two decades of work underscores the difficulty of the task. At this point, 'value-added' teacher evaluation is an idea that makes sense in theory but remains unworkable in practice. As such, no state should be incorporating student test scores into teacher evaluations in anything other than carefully-designed, low stakes pilot experiments.

Other resources that may be helpful

- Petrilli, *All or nothing on teacher accountability (teacher improvement v. teacher accountability)*
- Amrein-Beardsley, *Why VAM is a sham (top 10 reasons VAM doesn't work)*
- New York Times, *Confessions of a 'bad' teacher*
- Baker, *On misrepresenting (Gates) MET to advance state policy agendas*

- Amrein-Beardsley, *Methodological concerns about [Tennessee's] education value-added assessment system*
- Baker, *The toxic trifecta, bad measurement, and evolving teacher evaluation policies*
- Baker, *Gates still doesn't get it! Trapped in a world of circular reasoning and flawed frameworks*

Open letter to Governor Cuomo from seven former New York teachers of the year (February,2015)

‘You have made us the enemy. This is personal.’ — 7 N.Y. Teachers of the Year blast Cuomo

By Valerie Strauss February 9 ~ The Answer Sheet published in The Washington Post

“Seven New York State Teachers of the Year have written an open letter to Gov. Andrew Cuomo, blasting his new proposed education reforms that, among other things, link half of a teacher’s evaluation to student standardized test scores. Rich Ognibene, 2008 New York State Teacher of the Year, said in an e-mail that he wrote the first draft and six others contributed to the effort because “we were deeply hurt by the governor’s proposed education reforms.”

Writing and signing the letter along with Ognibene are Ashli Dreher, 2014 New York State Teacher of the Year; Katie Ferguson, 2012 New York State Teacher of the Year; Jeff Peneston, 2011 New York State Teacher of the Year; Marguerite Izzo, 2007 New York State Teacher of the Year; Steve Bongiovi, 2006 New York State Teacher of the Year; and Liz Day, 2005 New York State Teacher of the Year. The letter has been published in theAlbany Times Union and, with permission from the authors, I am publishing it here.”

Dear Governor Cuomo:

We are teachers. We have given our hearts and souls to this noble profession. We have pursued intellectual rigor. We have fed students who were hungry. We have celebrated at student weddings and wept at student funerals. Education is our life. For this, you have made us the enemy. This is personal.

Under your leadership, schools have endured the Gap Elimination Adjustment and the tax cap, which have caused layoffs and draconian budget cuts across the state. Classes are larger and support services are fewer, particularly for our neediest students.

We have also endured a difficult rollout of the Common Core Standards. A reasonable implementation would have started the new standards in kindergarten and advanced those standards one grade at a time. Instead, the new standards were rushed into all grades at once, without any time to see if they were developmentally appropriate or useful.

Then our students were given new tests—of questionable validity—before they had a chance to develop the skills necessary to be successful. These flawed tests reinforced the false narrative that all public schools—and therefore all teachers—are in drastic need of reform. In our many years of teaching, we've never found that denigrating others is a useful strategy for improvement.

Now you are doubling down on test scores as a proxy for teacher effectiveness. The state has focused on test scores for years and this approach has proven to be fraught with peril. Testing scandals erupted. Teachers who questioned the validity of tests were given gag orders. Parents in wealthier districts hired test-prep tutors, which exacerbated the achievement gap between rich and poor.

Beyond those concerns, if the state places this much emphasis on test scores who will want to teach our neediest students? Will you assume that the teachers in wealthier districts are highly effective and the teachers in poorer districts are ineffective, simply based on test scores?

Most of us have failed an exam or two along life's path. From those results, can we conclude that our teachers were ineffective? We understand the value of collecting data, but it must be interpreted wisely. Using test scores as 50 percent of a teacher's evaluation does not meet this criterion.

Your other proposals are also unlikely to succeed. Merit pay, charter schools and increased scrutiny of teachers won't work because they fundamentally misdiagnose the problem. It's not that teachers or schools are horrible. Rather, the problem is that students with an achievement gap also have an income gap, a health-care gap, a housing gap, a family gap and a safety gap,

just to name a few. If we truly want to improve educational outcomes, these are the real issues that must be addressed.

Much is right in public education today. We invite you to visit our classrooms and see for yourself. Most teachers, administrators and school board members are doing quality work. Our students and alumni have accomplished great things. Let's stop the narrative of systemic failure.

Instead, let's talk about ways to help the kids who are struggling. Let's talk about addressing the concentration of poverty in our cities. Let's talk about creating a culture of family so that our weakest students feel emotionally connected to their schools. Let's talk about fostering collaboration between teachers, administrators and elected officials; it is by working together, not competing for test scores, that we will advance our cause.

None of these suggestions are easily measured with a No. 2 pencil, but they would work. On behalf of teachers across the state we say, these are our kids, we love them, and this is personal.

Ashli Dreher 2014 New York State Teacher of the Year

Katie Ferguson 2012 New York State Teacher of the Year

Jeff Peneston 2011 New York State Teacher of the Year

Rich Ognibene 2008 New York State Teacher of the Year

Marguerite Izzo 2007 New York State Teacher of the Year

Steve Bongiovi 2006 New York State Teacher of the Year

Liz Day 2005 New York State Teacher of the Year