Value Added?

UCLA | IDEA

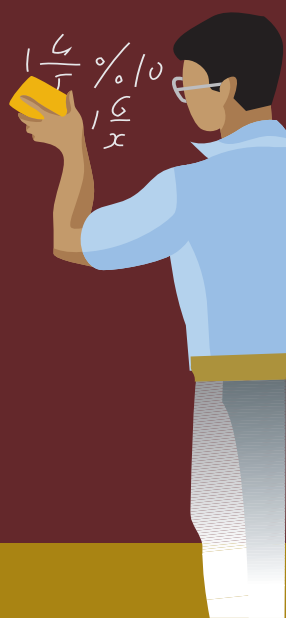
What makes an effective teacher?

A: How might we systematically assess teacher "effectiveness"? Is it accurate or fair to evaluate teachers based on student test scores? These questions lie at the heart of recent debates surrounding teacher quality and evaluation. Most centrally, these debates have focused on the appropriate use of "value added measures" (hereafter VAM) in judging teacher effectiveness and making decisions about teacher compensation, promotion, and dismissal.

VAM uses changes in student test scores to determine how much "value" an individual teacher has "added" to student growth during the school year. Some policymakers, school districts, and educational advocates have applauded VAM as a straightforward measure of teacher effectiveness: the better a teacher, the better students will perform on standardized tests. However, many prominent researchers and educators have expressed concern and urged caution.

Below, we present a series of questions and answers (as well as resources for more in-depth analysis) aimed at disentangling this complex issue. In particular, we are concerned that a narrow use of "value added" as the single measure of teacher effectiveness will have a detrimental effect on student learning, teacher retention, and educational equity. In other words, without more careful implementation and use, VAM could exacerbate the very problems they are alleged to help address.

Q: What are "Value Added Measures"?



A: VAM is a new statistical tool for quantifying teacher effectiveness on the basis of student gains on standardized tests. VAM compares students' test scores at the beginning of the year with their results on a comparable test at the end of the year, thus *isolating* the "value added" by a particular teacher.[2] In theory, a teacher's "value added" is the unique contribution she makes to her students' academic progress.[3]

VAM marks an improvement over methods that evaluate teacher effectiveness based on average ("raw achievement") scores. For example, comparing two teachers' average student test scores to one another does not take into account where each group of students began. Teacher A may have a higher class average than Teacher B, but Teacher B's students may have began the year with much lower scores. Thus, Teacher B's students may have actually made greater gains.

Proponents of VAM (including some equity-minded educational advocates) therefore point to its improved accuracy and fairness. Unlike previous approaches, VAM attempts to account for 1) where each group of students began and 2) the influence of external factors on student growth (greater family resources, instruction in previous grades, out of school support, etc.).[4] The teacher ends up with a score that is supposed to reflect her *individual* impact on student achievement.[5]

This is the potential appeal of VAM: Evaluate teachers on the basis of how much academic growth their students experience over the course of the school year.[6] Use these evaluations to identify and reward "effective" teachers, and dismiss or target those who are deemed "ineffective" for professional development. VAM has also gained popularity for its relative statistical sophistication.[7]

But this is not just an academic exercise. Policy makers and educational leaders are increasingly talking about using VAM to make high-stakes decisions — decisions that will shape the quality of education students receive. To gain a clearer sense of these measures, including the potential unintended consequences of evaluating teachers based on student test scores, we offer a closer look at the methodology and practical implementation of VAM.

Q: Does VAM provide a reliable and valid measure of teacher effectivenes?

A: Many researchers and statisticians argue that VAM does not provide a sufficiently reliable and valid measure of teacher effectiveness, particularly when used to make high-stakes personnel decisions.[8] Methodological problems with VAM include:

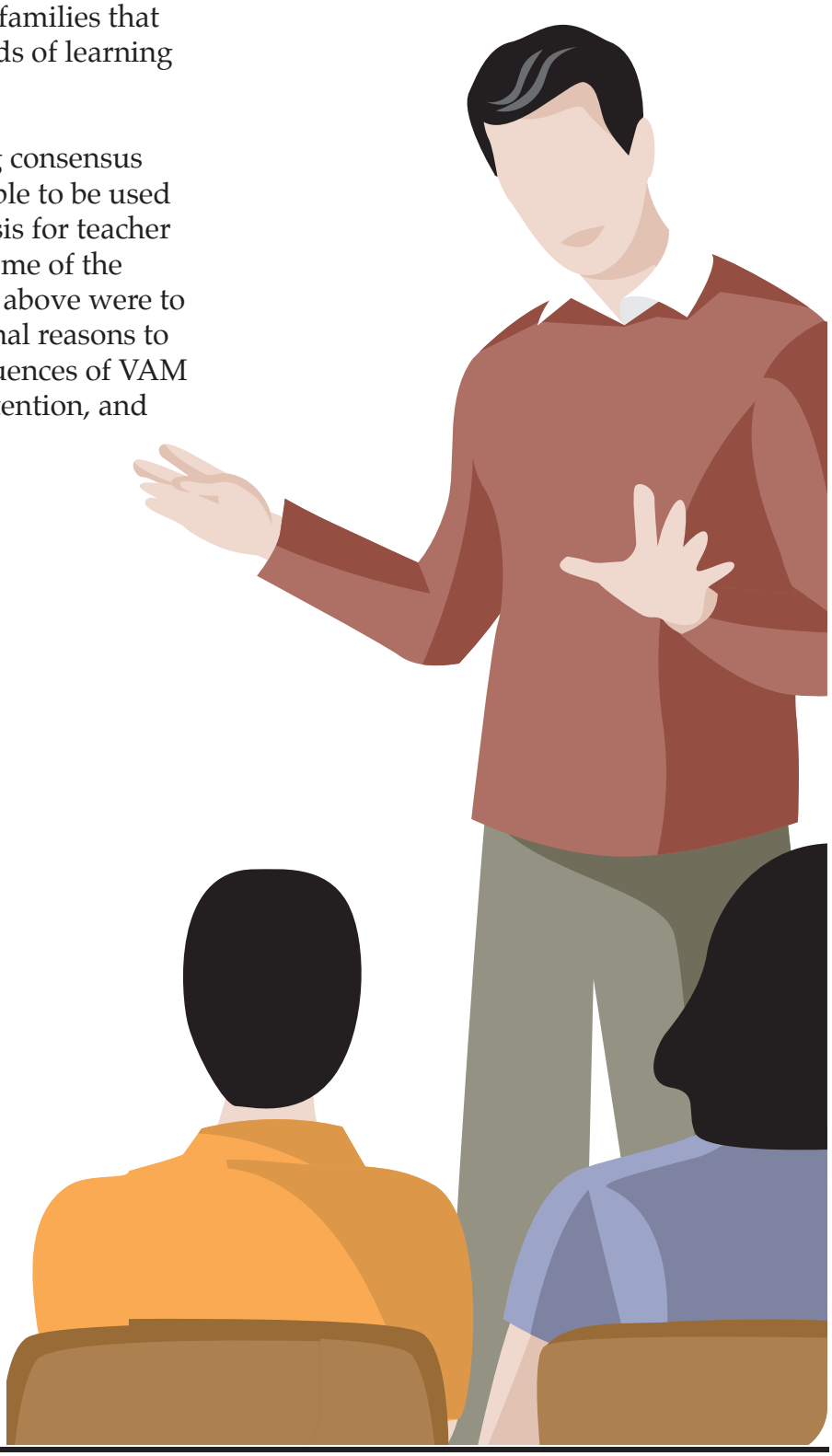
- **The non-random sorting of teachers and students:** VAM assumes that what teachers do in the classroom has a *causal effect* on student test scores. Increasing scores are a result of greater teacher effectiveness. Decreasing scores are a result of teacher ineffectiveness. However, such causal interpretation requires random sorting.[9] VAM is most credible when students are randomly sorted into classes, and teachers are randomly assigned to those classes. Without random sorting, it is impossible to know whether rising or falling test scores can actually be attributed to the individual teacher.[10] Importantly, non-random sorting is often a deliberate practice (on the part of schools and parents) used to ensure that students are assigned to the classroom most likely to meet their learning needs.

- **The instability of teachers' scores:** VAM is relatively unstable over time. In one study, a large percentage of the teachers who were identified as "most effective" one year were then identified as "least effective" the next year.[11] This is partially because the impact of a teacher simply cannot be separated from other influences (both inside and outside the school).[12] If test scores were an accurate measure of teacher effectiveness, one would expect much greater stability in teachers' scores from year to year.[13]

- **The difficulty of isolating teacher effects:** Fundamentally, the impact of teachers cannot (and perhaps *should not*) be separated from external influences on student growth. There are many reasons why students score well on standardized tests. Certainly one reason is that their teacher effectively taught the material. But students also score well because they have access to learning opportunities outside their classroom. Even within the same classroom, students may not be getting the same educational experiences and supports:

- Students are exposed to more adults than just the teacher at school, including other teachers, classroom aides, tutors, etc.[14]
- Students attend after-school, summer, and weekend educational programs.
- Students go home to families that provide different kinds of learning opportunities.[15]

There is, consequently, growing consensus that VAM is simply too unreliable to be used widely or to form the single basis for teacher evaluation. However, even if some of the methodological issues outlined above were to be addressed, there are additional reasons to be concerned about the consequences of VAM for student learning, teacher retention, and educational equity.



Q: What are the potential unintended consequences of VAM for educational equity?

A: Evaluating teachers based solely on student test scores prioritizes test preparation at the expense of more enriching and challenging curriculum. VAM assumes that gains in student test scores are synonymous with meaningful forms of learning. However, the tests used to determine teacher effectiveness often focus on "testable skills" rather than deep and broad conceptual understanding.

For example, whereas mathematical knowledge may be easier to assess on short answer or multiple choice tests, subjects such as history, civics, English literature, writing, and critical thinking require distinct forms of assessment.[16] Evaluating teachers based on student test scores creates incentives to diminish instruction in these areas.[17] A focus on "testable skills" also narrows the curriculum *within* the subjects most emphasized by recent policies: math and reading. In the domain of literacy, high-stakes tests often accompany scripted curriculum that emphasize fluency and speed over reading comprehension.

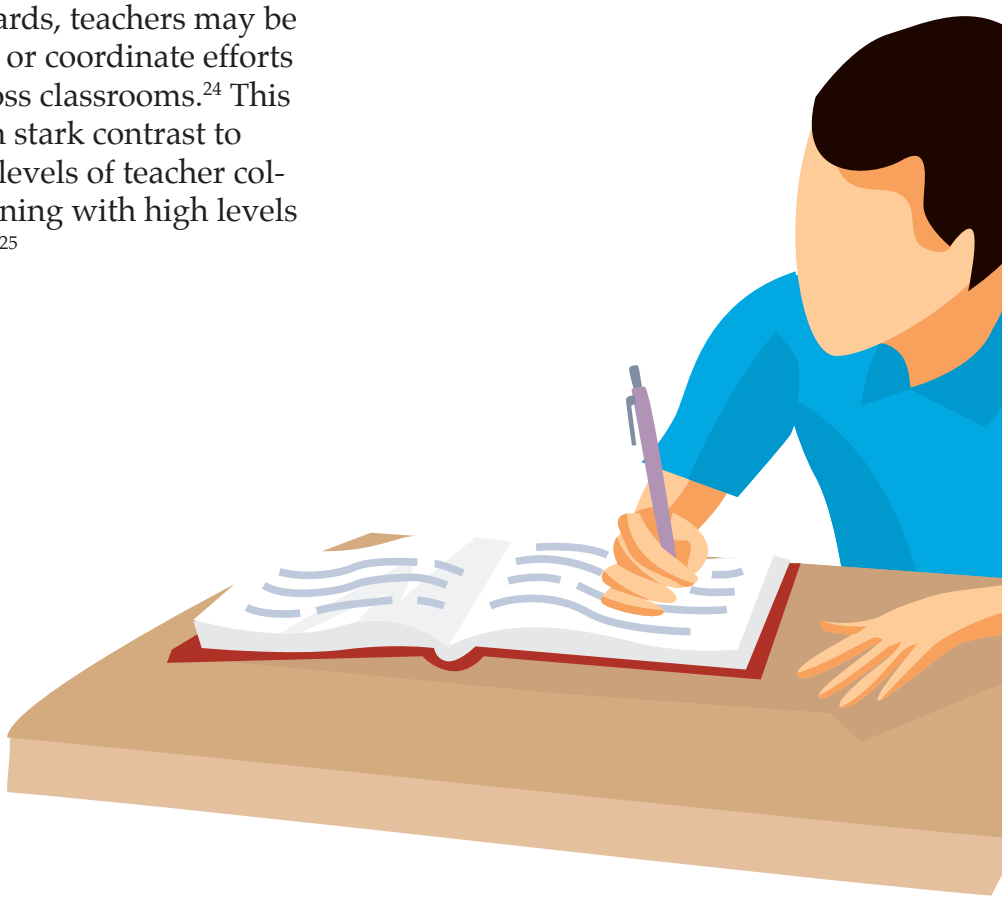
What, then, does VAM *value?* While not dismissing that information from tests can sometimes be useful, we are concerned that VAM directs curriculum and instruction towards lower-level skills. This is reflected in the widespread (and often encouraged) practice of "teaching to the test." In addition to drilling students on test-type questions, teachers who gain familiarity with the test may focus on 'likely-to-be-tested' topics and organize learning in the format of common test questions.[18] Ultimately, the skills being tested offer a very limited representation of the kinds of thinking, knowledge, and practices we aim to cultivate in classrooms.[19]

Using student test scores as the single indicator of teacher effectiveness may exacerbate educational inequity. Under NCLB, schools enrolling large numbers of low-income students and students of color often have focused on a narrow set of "testable skills" to avoid sactions. As educational researcher Mike Rose writes, "You can prep kids for a standardized test, get a bump in scores, yet not be providing a very good education. The end result is the replication of a troubling pattern in American schooling: poor kids get an education of skills and routine, a lower-tier education, while students in more affluent districts get a robust course of study."[20] Equity oriented, high-quality teaching and learning must be defined as more than doing well on a narrow set of measures.

Further, linking teacher evaluation with test scores provides a disincentive for working with the most vulnerable populations of students. According to the Economic Policy Institute, "teachers have been found to receive lower 'effectiveness' scores when working with English language learners, special education students and low-income students than when they teach more affluent and educationally advantaged students."[21] Thus, teachers may be further discouraged from working in the most high-need schools. Within schools and classrooms, students with greater or special educational needs may be perceived as 'pulling down' teachers' VAM scores.[22] High-stakes accountability has already led some schools to pressure their most struggling students to transfer or drop out.[23]

Finally, a narrow use of VAM may have a detrimental effect on teacher collaboration and morale. As stated, VAM aims to isolate the contributions of *individual* teachers on student outcomes. If increasing test scores are linked to monetary rewards, teachers may be less likely to collaborate or coordinate efforts to support students across classrooms.[24] This potential trend stands in stark contrast to research that links high levels of teacher collaboration and peer learning with high levels of student achievement.[25]

Ultimately, basing professional evaluation on VAM is likely to result in the demoralization and attrition of teachers possibly and students**.** Teachers will face what is legitimately perceived as arbitrary and unfair forms of evaluation, without adequate attention to the conditions within which they work.[26] Rather than creating opportunities for teachers to hone their craft, VAM demands an even greater emphasis on raising student test scores. Thus, the narrowing of curriculum and instruction leads to the deskilling and devaluing of teachers.[27] This shift will hinder teachers' ability to create intellectually rich contexts where all students have an opportunity to learn – the kind of education many joined the teaching force to help cultivate, and the kind of education students deserve.



Conclusion

For all these reasons, we believe equity-minded educational advocates ought to challenge the use of VAM as the single measure of teacher effectiveness, particularly in the context of high-stakes personnel decisions. As reflected in the *Los Angeles Times*' (2010) recent publication of teachers' scores, singling out individual teachers as "effective" or "ineffective" based on unreliable information is not a fair or useful strategy for improving teacher quality. Information is a good thing as long as we know exactly what that information is telling us, and how we can use it to better the educational experiences of all students.

VAM might be useful as *one* piece of a much larger plan for improving teacher quality and student learning. For example, rather than focusing on individual teachers, VAM could be a useful tool for school- or district-level assessment. [28] Focusing on school-level change and formative evaluation would help circumvent some of the threats to collaboration and equity mentioned above.

For VAM to help individual teachers reflect on and improve their practice, it must be part of a more comprehensive approach to evaluation. This approach ought to include:

- **Well-analyzed test data:** Overall value added scores do not tell us where to focus improvement efforts. Instead, we need to provide teachers with specific data about how particular groups of students perform on particular tasks. If a teacher can see that all third graders made mistakes in long division or that English Learners had particular difficulty with word problems, he can take action to provide targeted assistance in these areas. Schools and districts can also take action to provide specific supports.

- **Classroom observations:** Provide teachers with quality feedback about their classroom practice. This includes offering specific suggestions about what to improve on and how to improve on it. This should take place in a low-stakes environment where teachers receive professional support to continue developing their practice.

- **Professional development:** Create high-quality professional development experiences where teachers can build their repertoire of skills, particularly in those areas that test observation data have identified as needing improvement. This includes creating opportunities for teacher collaboration and peer learning.

- **Comprehensive assessment of students:** Using student portfolios and other formative assessments would address concerns that a narrow focus on standardized outcome measures can lead to "teaching to the test" or a narrowing of the curriculum as mentioned above.

Endnotes

1 According to McCaffrey, et. al., (2004) the teacher's contribution to student outcomes is defined as the difference between a student's achievement in the teacher's class and his/her predicted achievement with a teacher of "average" effectiveness. Also, see Daniel Willingham's short video for a succinct explanation of VAM and Merit Pay: http://www.youtube.com/watch?v=uONqxysWEk8

2 Corocan, 2010, p. 4.

3 As Corocan explains, "If we assume that many of the external factors influencing a student's fourth grade achievement are the same as those influencing her third grade achievement, then the change in the student's score will cancel out these effects and reveal only the impact of changes since the third grade test, with the year of fourth grade instruction being the most obvious" (2010, p. 4).

4 According to Braun (2005, p. 7), "that number, expressed in scale score points, may take on both positive and negative values. It describes how different that teacher's performance is from the performance of the typical teacher, with respect to the average growth realized by the students in their classes."

5 Braun, 2005, p. 2

6 According to the Economic Policy Institute (EPI), "Value added approaches are a clear improvement over *status* test-score comparison (that simply compare the average student scores of one teacher to the average student scores of another); over *change* measures (That simply compare the average student scores of a teacher in one year to her average student scores in the previous year); and over *growth* measures (that simply compare the average student scores of a teacher in one year to the same students' scores when they were in an earlier grade the previous year)…Although value added approaches improve over these other methods, the claim that they can 'level the playing field' and provide reliable, valid, and fair comparisons of individual teachers is overstated" (2010, p. 9).

7 Economic Policy Institute (EPI), (2010), p. 2.

8 Random sorting is similar to experiments that designate a "control" group and a "variable" group, with the goal of identifying the unique effects of a particular variable (in this case, the individual teacher).

9 Braun (2005). As economist Jesse Rothstein (2009) argues, in order for Value Added Measures to be of use, "they must reflect teachers' causal effects on the student outcomes of interest, not preexisting differences among students for which the teacher cannot be given credit or blame."

10 Berry (2010) and Sass, (2008).

11 Amrein-Beardsley (2008) and McCaffrey, et. al., 2004 (RAND). This is also due to the problem of missing data.

12 As educational researcher Wayne Au (2011) writes, "The year-to-year instability that Sass [2008] highlights shows that test scores have very little to do with the effectiveness of a single teacher and have more to do with the change of students from year to year (unless, of course, one believes that one-third of the highest ranked teachers in the first year of the study simply decided to teach poorly in the second)."

13 Sometimes termed the "spill-over effect," this is an especially important factor to consider in middle and high school, where students' learning and growth in distinct subjects and classrooms may be (and, ought to be) mutually influential. For example, learning how to develop an argument in the context of history or social studies may positively influence a students' development in English. Or, practice with problem solving in one content area might fruitfully support students' learning in another.

14 EPI, 2010, p. 9.

15 As Sean Corocan of the Anneburg Institute for School Reform argues, "it makes little educational sense to force such skills to conform to such a structure purely for value added assessment" (2010, p. 14).

16 EPI, 2010, p. 16.

17 EPI, 2010, p. 17; Corocan, 2010; McCaffrey, et. al., 2004 (RAND) For example, teachers who do try to teach the full curriculum (or who might be focused on preparing their students for the type of work they will encounter in future grades) may find their students not gaining as much as others, whose teachers resort to some form of teaching to the test (Braun, 2005, 16).

18 An increasingly narrow focus on testing may also contribute to student disengagement and teacher demoralization. As one teacher states, "Children have not stopped doing what children do but teachers don't have time to deal with it. They don't have time to talk to their class, and help the children figure out how to resolve things without violence. Teachable moments to help the schools and children function are gone" (EPI, 2010, 19).

19 Rose (In press, *Dissent*).

20 EPI (2010), p. 3. "Other human service sectors, public and private, have also experimented with rewarding professional employees by simple measures of performance, with comparably unfortunate results. In both the United States and Great Britain, governments have attempted to rank cardiac surgeons by their patients' survival rates, only to find that they had created incentives for surgeons to turn away the sickest patients" (p. 7).

21 EPI (2010), p. 16.

22 Hinchey, (2010), 1.

23 EPI (2010), p. 18.

24 Metlife Foundation (2009); Jackson, C.K. & Bruegmann, E. (2009) As Barnett Barry of the Center for Teaching Quality reports, "Over 90 percent of the nation's teachers report that their colleagues contribute to their teaching effectiveness. New teachers, in particular, were more likely to strongly agree that their success in the classroom hinged on the effectiveness of others" (2010, p. 5).

25 As educational researcher Mary Kennedy writes, "We measure and track their value added test scores but we do not measure their teaching loads, planning time, student absences, proportion of difficult-to-teach or resistant students, frequency of outside interruptions, access to textbooks or equipment of good quality, or whether their instructional materials arrived before the school year began" (2010, p. 596).

26 Describing the experience of one veteran teacher, Rose writes, "The school's test scores were not adequate last year, so the principal, under immense pressure, mandated a "scripted" curriculum, that is, a regimented curriculum focused on basic math and literacy skills followed by all teachers. The principal also directed the teachers not to change or augment this curriculum. So Priscilla cannot draw upon her cabinets full of materials collected over the years to enliven, extend, or individualize instruction. (Though like any experienced teacher, she figures out ways to use what she can when she can.) The teachers have also been directed by the principal to increase the time spent on the literacy and math curriculum and trim back science and social studies. Art and music have been cut entirely. "There is no joy here," she told me, "only admonition." (Rose, in press, 2011)

27 Garcia (2010) http://educationadvocacy.wordpress.com/2010/09/09/all-eyes-in-education-on-los-angeles-monica-garcia/

28 This touches on one of the central criticisms of VAM: "When teachers receive data based on once-a-year standardized tests, they rarely are informed of why they are or are not effective in teaching their students. They simply have raw scores, absent any deeper analytics that can help their improve their classroom teaching practices" (Berry, 2010, p. 4).

Institute for Democracy, Education, and Access

UCLA IDEA is a research institute seeking to understand and challenge pervasive racial and social class inequalities in education. In addition to conducting independent research and policy analysis, IDEA supports educators, public officials, advocates, community activists, and young people as they design, conduct, and use research to make high-quality public schools and successful college participation routine occurrences in all communities. IDEA also studies how research combines with strategic communications and public engagement to promote widespread participation in civic life. www.ucla-idea.org

For further information, contact UCLA IDEA
phone: (310) 206-8725; fax: (310) 206-8770;
email: idea@ucla.edu
www.ucla-idea.org

UCLA | IDEA