Kappa Delta Pi Record, 48: 156–161, 2012
Copyright © Kappa Delta Pi
ISSN: 0022-8958 print/2163-1611 online
DOI 10.1080/00228958.2012.733929

Measuring Teacher Quality with Value-added Modeling

by Michael Marder

***Michael Marder** is Professor of Physics at the Center for Nonlinear Dynamics at The University of Texas at Austin. Since 1998 he has been co-director of UTeach Natural Sciences, which prepares secondary mathematics and science teachers at UT Austin and has since expanded to more than 30 other universities.*

Value-added modeling carries the promise of measuring teacher quality automatically and objectively, and improving school systems at minimal cost. Yet value-added modeling cannot be carried out without value judgments; and if there are technical errors, they will have human cost.

Computer programs, quickly and quietly reaching their judgments, may soon decide your value, and you should know what they are doing. *Value-added modeling* is a name for the most sophisticated computerized ways of taking into account all known numbers that describe your students, and deciding how much you added to their learning.

Pressure to implement automated methods arises because:

1. "Research tells us that the influence of teachers is the single-most important [in-school] factor in determining student achievement" (Suh and Fore 2002).
2. Teachers' backgrounds do not tell whether they are good or not; nor their degrees, nor even whether they are certified (Gordon, Kane, and Staiger 2006).
3. Conventional evaluation systems give nearly all teachers the same rating (Weisberg et al. 2009).
4. There are computer programs that can find the best and worst teachers by analyzing state tests students already are taking (Wright, Horn, and Sanders 1997).

If you held these beliefs—and many of your governors, state education commissioners, and federal legislators do—wouldn't you act?

Action is coming. Some was triggered by Race to the Top, to which 40 states and the District of Columbia applied in 2009. To win grant funding, states needed to put in place systems to measure student growth over time and use these data as part of a system of teacher and principal evaluation. An analysis sponsored by the Council of Chief State School Officers (CSSO and Learning Point Associates 2010) found that in 2010, twenty-one states already were using tests to measure student growth or were planning to do so. More action will be triggered by the final phase of No Child Left Behind (2002). In 2014, all public schools will be required to succeed in educating children—overall and in disaggregated subgroups—so that at least 95 percent pass state tests in mathematics and reading. This bar is so high that well over half of all public schools will be labeled *Unacceptable* and start down the path to reorganization. The U.S. Secretary of Education has been granting waivers from No Child Left Behind provided states put forward plans that include measurements of teacher performance.

In short, big wheels are in motion. If you are a public school teacher, it is very likely that within the next two years part of your annual evaluation will involve calculations using student growth on standardized tests. The same holds if you are a public school principal. If you are an educator at a public university, for the moment you are exempt. But forces in the media that successfully created the approving climate for test-based accountability of public school teachers are calling for its application to colleges and universities (Brooks 2012).

The Idea of Value-added Modeling

Using computers to evaluate teachers based on student test scores is more difficult than it seems. Value-added modeling is a genuinely serious attempt to grapple with the difficulties.

A first idea one might have is to grade teachers based simply upon the test scores of students in their classrooms. For example, a mathematics teacher might get a grade of A, B, C, D, or F depending on whether 90 percent, 80 percent, 70 percent, or 60 percent of students passed a state mathematics exam—or on whether the average score of students in the class was 90 percent, 80 percent, etc. Either way would be badly unfair. Suppose a teacher gets a class where at the beginning of the year all students already can pass the exam they will take at the end of the year. All the teacher has to do is to ensure the students do not forget what they already know. This is much less demanding than teaching a class of students to whom much of the material is new.

The next idea is to grade teachers based on how much student test scores increase from one year to the next. The bigger the student gains, the better the teacher. This idea is better than the previous one, but it raises many new difficulties. Suddenly the teacher whose students all know the material upon arrival moves from having a huge advantage to having a huge disadvantage. If the whole class has perfect scores at the beginning of the year, the most the teacher can hope is that the students will not move down; and if they do, she will appear to have caused them to go backward (Pallas 2012). There are many other problems as well. What precise test should be used to determine how much students know at the beginning of the year? It is never the exact same test students take at the end of the year, so what do score changes mean?

Addressing such questions leads to value-added modeling. The word *modeling* is important. It means that test scores undergo a large amount of mathematical processing before any conclusions are drawn. That is the strength of the various methods for value-added modeling, but also a weakness. The calculations are so complex that only a handful of specialists know how to carry them out.

The idea of value-added modeling is to take a great deal of available information about a classroom and to create expectations for how well the students should do by the end of the year. Students who all come into the class with nearly perfect scores on last year's mathematics exam should be expected to get nearly perfect scores again, but cannot be expected to gain; that would be impossible. Students who on average were just below failing last year may be expected to rise above failing this year. The expectations are partly a reflection of goals and values for education, partly influenced by what is practically possible, and partly dictated by mathematical feasibility. The end result of a value-added model is a very specific number for each teacher: an expectation, a target, describing the scores his or her students should obtain.

To design a computer program that creates a custom expectation for the classroom of each teacher is not an easy task. The program has to take many things into account. But in an age where Google™ somehow scans through 4.67 million web pages on value-added modeling in 0.2 seconds and does a great job of finding the best one, surely the nation's top researchers, having worked on this problem for decades, have come up with an awfully good solution. They have, but I do not believe it is yet good enough. To explain, I will need to go into some of the details of how value-added models are actually constructed.

What Were You Expecting?

The essence of value-added models lies in the precise way they calculate expected scores for the students of each teacher. The mathematical ideas on which they are based are complicated and appear inaccessible to anyone with less training than upper-division university statisticians. A fairly small community of scholars, made up of both advocates and skeptics, has been responsible for developing the calculations (McCaffrey et al. 2003; National Research Council 2010). All of these experts agree that the results should be used with caution. However, because numbers in official printouts are so specific and appear so authoritative, it will prove problematic in practice to prevent them from dominating decisions about promotion and dismissal.

Some concerns that have been raised previously about value-added modeling include the possible influence of missing information such as student mobility, large variations in results from year to year, the need for many years of data to obtain reliable results, and the absence of suitable pretests in some subject areas. Here I describe a very particular worry I have had for some time, but for which I only recently was able to obtain any evidence. Deem this an invitation to grapple with the sorts of decisions that lurk behind the mathematics.

To begin, consider the following question: "What are the data you can find on a student that most accurately predict how much the student's score will change over the next year?" No researcher should answer, "The identity of the student's teacher." Other things come first. To describe them quantitatively requires a system of units. A conventional way to describe gains on tests is in units of standard deviations—that is, in terms of the typical amount scores vary on an exam for a given year and grade level from one student to another. For many exams I have inspected, the standard deviation is in the range of 15–20 percent, where 100 percent is a perfect score. Because standard deviations have a complicated technical sound, there is another convention widely used to discuss value-added methods, which is to refer to a change of a quarter of a standard deviation as "one year of learning." "One year of learning" is really just a code for the student gaining a quarter of a standard deviation on an exam, which means in a typical case getting 5 more points out of 100.

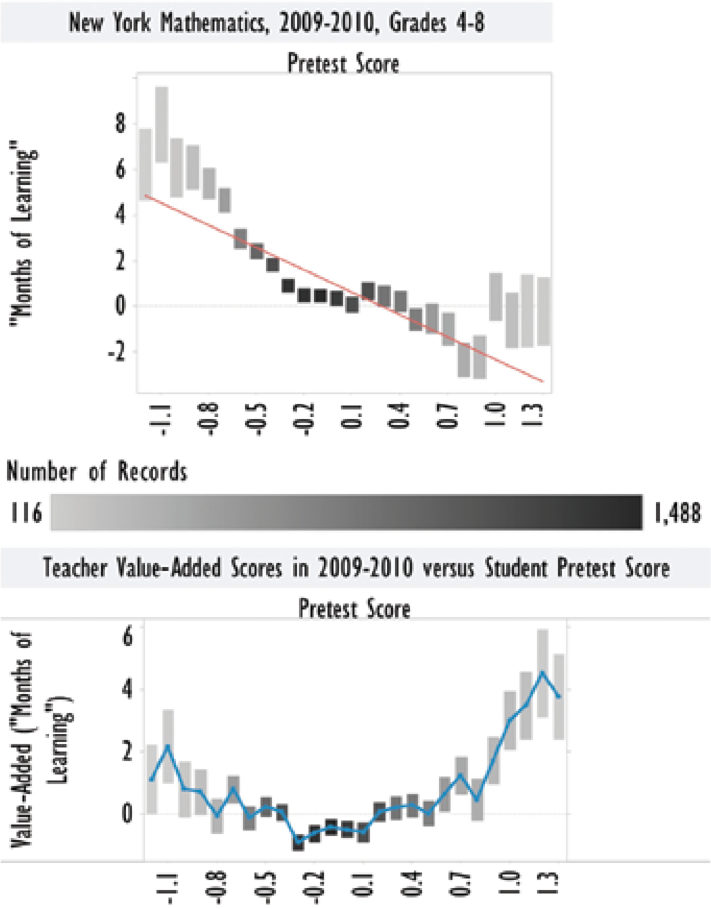
With units of "years of learning" in hand, now to business. The most important predictor of how much a student's test scores will *change* is the student's *score the previous year*. This effect is huge. For example, looking at Texas 5th-graders who obtained the very low raw score of 30–40 percent on the Texas high-stakes mathematics exam, the average score increase the next year is 10 percent, or "two years of learning." This large gain has taken place repeatedly over the last seven years (Marder 2012a). Gains this large are not very surprising. Some students who score 30–40 percent really know little mathematics or struggle to parse the test questions, and it takes a miracle-worker of a teacher for their scores to rise. But other students get low scores because, on the test date, they were terribly ill, could not focus, or were angry or indifferent and decided to fill in bubbles on the answer sheet at random. These students can easily do much better the next year, and it turns out many of them do.

Now switch over to Texas 5th-graders whose mathematics raw score is 70–80 percent. With equal consistency, their score the next year has dropped around 5 percent, or "one year of learning." Maybe the transition to middle school is rough, and many students have some trouble adjusting. In any event, the *difference* between the average raw score gain of 5th-graders scoring 30–40 percent and 5th-graders scoring 70–80 percent is quite large: "three years of learning." Take three solid B students out of a classroom and replace them with three students who failed badly the year before, and the odds of a teacher demonstrating large score gains shoot up.

There are other large effects floating around. One of them is grade level. In certain grades, students have to pass exams in order to advance or to graduate. The effect is strongest in Texas at 11th grade when graduation is at stake. The starting scores of most 11th-graders reflect a gain of 10 percent compared to 10th-graders: "two years of learning"! Mathematics teachers, for example, can face classes with students mixed in from different grade levels because students do not go through high school in lockstep with one another.

Finally, one arrives at the school factors that are most commonly discussed in public: poverty, race, and differences between best and worst teachers (Marder 2012b). The statistical effects of these three factors are all typically "one year of learning" in size. In particular, the difference between teachers whose value-added scores put them in the top and lowest quartiles, is typically 0.2 standard deviations (Hanushek and Rivkin 2010), or around "10 months of learning." This difference between student score gains due to the highest- and lowest-ranked teachers is small compared with the difference in student score gains from the lowest- and highest-scoring students, and about the same as changes in student scores due to other causes I mentioned. In order to conclude from such data that replacing bad teachers with good can transform education, one has to take

Figure 1. Analysis of Scores.



Top, *Average score changes of New York City students as a function of their scores the previous year. The data undershoot, overshoot, and undershoot a best-fit line. The NYC Department of Education rescaled the test scores before reporting them so 0 is an average score. Error bars shown here are two standard errors high.* ***Bottom,*** *Average value-added attributed to New York City teachers as a function of their students' pretest scores. The undershoots and overshoots of the top graph correspond closely to the values here, although a great deal of complicated mathematical processing intervenes. These graphs are my plotting of the public data.*

an additional step and find that gains due to good teachers do add on to one another year after year (Hoxby, Murarka, and Kang 2009).

The premise of value-added modeling is that once student and school characteristics have properly been taken into account, what remains are two things: (1) a mixture of completely random occurrences out of anyone's control (such as a student falling sick); and (2) the skill and influence of the teacher. The influence of the teacher must be teased out of a collection of other influences of equal or greater importance. Make a technical mistake in accounting for them, then attribute that mistake to the teacher, and the results are wrong.

Wrong? Impossible! Right? Yet every value-added model I have seen written down (not all of them are published, so I cannot speak about them all) begins with a bit of technical hastiness. Researchers do know that changes in students' scores depend on the students' prior year scores. The way they account for the change is with a term like this:

$$S_i - S_{i-1} = A(100 - S_{i-1}) + \ldots$$

The change in a student's score from last year *(i–1)* to this year *(i)* is expected to be larger and larger the farther below 100 the student scored, and some constant *A* tells how the expectations change. There are lots of other terms in the equations, often including classroom characteristics such as race and poverty, sometimes not, but no others dealing with this particular point. The equation describes learning gains with a straight line with slope *–A* and intercept *100A*.

A straight line. . . . My research home is in the Center for Nonlinear Dynamics. It's an odd name for a research area, because the name describes what the research is not—it avoids use of straight lines—rather than what it is. But there is a reason, because examining relationships that cannot be described by straight lines turns out to lead to enough research that it has kept a community of physicists busy for 50 years. Aha! Could this be another case? Let's have a look.

So, back to a way value-added models could be wrong. In my experience, test score changes depend on prior year scores in a particular nonlinear way. For low scores, the curve slopes downward steeply, and then it flattens out (Figure 1, *top*). It looks like a child's slide, which slopes down at first to get kids going and then evens out at the end so they do not ram into the ground. Why score changes have a form like this, I do not really know; maybe it is because the rapid improvement of previously ill or disaffected children is so much more likely for very low-starting scores. But one does not have to know. That is the way they are.

So try this. *Draw a shape like a child's slide* with a large slope on the left and then flat on the right. Next, take a ruler and *draw a straight line* passing through the middle that hugs as close to the curve as it can. The two cannot match. The closest fitting straight line must undershoot at the beginning, overshoot in the middle, and undershoot at the end. The *difference* between the two has a particular U-shaped nonlinear form. The worry is that when value-added models compensate for the scores your students had last year, a fraction of the skill attributed to the teacher was in fact an error due to the researcher.

Out of Control in New York City

In practice, what fraction? I had a chance to check when newspapers obtained the value-added scores of all New York City teachers and posted them publicly, together with discussions of teachers by name (NY1 News 2012). Maybe in the course of all the processing steps—transforming of raw scores into scaled scores, dozens of extra terms in the equations—the U-shaped error somehow disappears?

But no, it is there. New York City student score gains versus prior score have exactly the same shape as so many other test-score results, with a steep drop for low scores followed by a flatter plateau (Figure 1, *top*). The data at first overshoot the closest fitting straight line, then undershoot, then overshoot again. Plotting the value-added scores New York City attributed to its teachers against average scores of students the year before, the results look like a U (Figure 1, *bottom*). The size of the error at its largest is about half the difference commonly attributed to highest- versus lowest-quartile teachers.

Thus around half of the student gains and losses being attributed to the skill and shortcomings of teachers can be due to the technical mistake of trying to find a single straight line that describes a curve.

Use with Caution

It is tempting to automate a process that previously has been the province of human judgment. But judgment is always present: if not each detailed decision, then in the rules of automation.

Automating a decision does not make it right. Computers are consistent, but not necessarily correct. The technical problem described here can easily be corrected, but it is just one example of the limitations to computer measures of teacher value. The rules put into the computers may or may not correspond to what we want to achieve. Some objectives of school, such as whether one child learns to speak confidently in public, or another child gains hope and stays in school after thinking of dropping out, are not plausibly measured well by multiple-choice tests of mathematics and reading. What we value in schools cannot completely be decided by technicians drawing curves.

Expert advice on value-added modeling always says that it should at most be used as a component of evaluation, in combination with other factors. Indeed. It provides information. It can flag real problems. But it has a limited view. And like the humans that created it, it is fallible.

References

Brooks, D. 2012. Testing the teachers. *The New York Times*, April 19. Available at: *http://nyti.ms/OfM2Dh.*

Council of Chief State School Officers and Learning Point Associates. 2010. *Measurement of student growth.* Washington, DC, and Naperville, IL: CSSO and Learning Point. Available at: *http://bit.ly/MmPLyV.*

Gordon, R., T. J. Kane, and D. O. Staiger. 2006. *Identifying effective teachers using performance on the job.* Washington, DC: Brookings Institution. Available at: *http://bit.ly/P5Yx1O.*

Hanushek, E. A., and S. G. Rivkin. 2010. Generalizations about using value-added measures of teacher quality. *American Economic Review* 100(2): 267–71. Available at: *http://bit.ly/LpWFUs.*

Hoxby, C. M., S. Murarka, and J. Kang. 2009. *How New York City's charter schools affect achievement,* August 2009 report, second report in series. Cambridge, MA: New York City Charter Schools Evaluation Project. Available at: *http://bit.ly/Kcurxa.*

Marder, M. 2012a. Unpublished analysis of Texas TAKS mathematics scores from 2003 to 2010.

Marder, M. 2012b. Failure of U.S. public secondary schools in mathematics. *AASA Journal of Scholarship and Practice* 9(1): 8–25. Available at: *http://bit.ly/MarderAASA.*

McCaffrey D. F., J. R. Lockwood, D. M. Koretz, and L. S. Hamilton. 2003. *Evaluating value-added models for teacher accountability.* Santa Monica, CA: RAND Corporation. Available at: *http://bit.ly/LizdZj.*

National Research Council. 2010. *Getting value out of value-added: Report of a workshop,* ed. H. Braun, N. Chudowsky, and J. Koenig. Washington, DC: The National Academies Press. Available at: *http://bit.ly/MtBttI*.

No Child Left Behind Act of 2001. 2002. Public Law 107–110. Washington, DC: U.S. Congress. Available at: *www2.ed.gov/policy/elsec/leg/esea02/107-110.pdf.*

NY1 News. 2012. 2007–2010 NYC teacher performance data. NY1 News [Web], February 28. Available at: *http://bit.ly/z9PgIX*.

Pallas, A. 2012. The worst eighth-grade math teacher in New York City. *The Hechinger Report* [blog], May 15. Available at: *http://bit.ly/NyoIUj.*

Suh, T., and R. Fore. 2002. *The National Council on Teacher Quality: Expanding the teacher quality discussion.* Washington, DC: ERIC Clearinghouse on Teaching and Teacher Education. ERIC Digest ED 477 730. Available at: *http://tinyurl.com/NCTQEQa.*

Weisberg, D., S. Sexton, J. Mulhern, and D. Keeling. 2009. *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*, 2nd ed. Brooklyn, NY: The New Teacher Project. Available at: *http://widgeteffect.org/overview.*

Wright, S. P., S. P. Horn, and W. L. Sanders. 1997. Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education* 11(1): 57–67. Available at: *http://www.sas.com/govedu/edu/teacher_eval.pdf.*