

Robert Gordon
November 2008
National Academy of Sciences Panel

Comments on Use of Value-Added Methodology to Improve Student Achievement

I appreciate the opportunity to comment on these papers. I cannot compete with all of you in my knowledge of statistics or economics, but as an outsider from the world of law and policy, I can perhaps add a usefully different perspective.

Several of the papers for this session seem to me to proceed as follows. They note, at some length, the problems with value-added measurement. Scores are subject to significant measurement error. Tests cover different subjects in different years and are not always ideally scaled. Results are infected by selection among students based on relevant characteristics for which researchers cannot control. Scores bounce around from year to year. Teachers can game tests. And so on.

The papers point out all of these things, with far more nuance than I can capture. But then the papers say something simple and clear. They say something to this effect: *therefore*, value-added measurement may be used for low-stakes purposes, but not high-stakes purposes.” Sometimes this is stated as a self-evident truth. Sometimes it is adorned with a few policy arguments. But the conclusion about the proper use of value-added data seems invested with the authority of science.

My own view is that authors may be right to believe that high-stakes use of value-added data will ultimately fail to take root in our schools. But they may also be wrong. The only way to know for sure is to adjust as best we can for the many legitimate concerns, and then to try out VAM in carefully designed settings, with the involvement of teachers every step of the way. This is already happening in hundreds of schools around the country, without the cataclysmic effects that critics expect. In my view, between the facts in these papers and the value-judgment that value-added data have no role in high-stakes decisions, there is a logical gap. The unwarranted leap across that gap would interfere with the advance of learning, for both science and students.

What these papers miss, it seems to me, is that high-stakes decisions must be made, and they must be made now, whether researchers would prefer further study or not. The kids are in school, now. They are learning, or perhaps not learning, now. And their teachers must be hired and rehired. They must be paid. So either we will use value-added data in these decisions or we will use something else. But, and I apologize if I sound like one of those existentialists you may have read late nights in college, we cannot escape the responsibility to decide.

Let me try to use an analogy to clarify this point. My apologies on two fronts: Not only is my analogy highly flawed—it is also from sports.

Imagine that tomorrow you become the coach of an NFL team with a mediocre record. You have a decent starting quarterback, but there are also some decent free agents elsewhere who might move to your team. You want to know how your guy stacks up. Would you be interested to know how many touchdowns and interceptions your starter had thrown compared to the competition? I am not statistics expert, but I can think of a lot of objections to these numbers. Your right guard allows three sacks a game, and two of your guy’s interceptions were passes batted down. Your wide receivers have butterfingers. The team’s schedule was way tougher than average this year. And the numbers don’t capture what a great motivator your quarterback is, how he brings out the best in other players. Globally, the use of touchdowns and interceptions as a metric will tend to encourage quarterbacks to

throw more passes in the red zone, even when a running play is safer. And, by the way, these are just inert numbers on the page, not tools for improving performance. Knowing that you are supposed to throw more touchdowns does not help you do so.

These are all legitimate concerns well worth discussing. They are the reason we have sports radio. They are also reasons no one in his or her right mind would use touchdowns and interceptions *alone* as a basis for decisions. That said, I think it's unlikely that you would willfully blind yourself to these figures. You might find some other better (though less transparent) way to measure success, like the "passer rating." But you probably would not decide these figures are altogether irrelevant.

That is because, as a coach, you would need to put a starting quarterback on the field next Sunday. And you would be expected to do everything in your power to win.

I know there are myriad problems with this analogy. The touchdown count is a lot simpler than a value-added score. And identifying the winner in football is a lot easier than identifying a good school. So I want to change up the analogy a little bit.

Instead of becoming a football coach, you become the superintendent of a typical big-city school district. Like so many urban districts, you've got a racial achievement gap of three or four grade levels by high school, and you've got a dropout rate of 40 or 50 percent.

Your task is to do better. What do you do? I'm going to briefly rehearse some the things you know:

The most important players on your team are the teachers. In spite of all our important efforts to improve practice for everyone, teachers differ enormously. Professor Wilms recounts the evidence that the variance among teachers is greater than the variance among schools.

Existing policies often do not distinguish effective and ineffective teachers. In many systems, the great majority of teachers are granted tenure as a matter of course, with little serious evaluation or review. (Teachers themselves complain about their managers' failure to give meaningful feedback.) And teacher compensation depends on only two things. One is education in education, even though the research is clear that getting such a master's degree does not improve performance. (Goldhaber 2007; Harris 2008) The other is experience, with the greatest compensation very late in careers. (Clotfelter 2008) That's true even though the research generally suggests that the productivity returns to experience sharply diminish within three to five years.

What are the other facts on the ground? We are neither attracting nor retaining the most talented Americans in the teaching profession. Even with encouraging improvements in their test scores, America's most able young people generally are not going into teaching. (Dillon 2007) And many of America's most talented teachers leave for other professions. (Boyd 2008)

So this is an argument—not an airtight argument, but what lawyers like me call *prima facie* argument—for trying something different. What we might try instead—together with efforts to improve practice and development for *all* teachers—is increasing salaries for teachers who excel, and making sure we grant tenure only to teachers who improve student learning in the classroom. We would do these things primarily not to improve teacher effort, an idea about which I am skeptical, but to attract and keep better more effective teachers in the classroom. Common sense, and some research, suggest that able people with many options prefer to work in professions that honor excellence and promise a higher standard of living. (Duffett et al., 2008)

So now we come to the hard question. If you want to try out this approach, how are you going to measure success and failure? The controversial proposal on the table is to use value-added measures *in conjunction with others*, including principal evaluations and, if you can afford it, peer review or other external review. I know that it is sometimes suggested that my paper with Tom Kane and Doug Staiger from 2006 called for exclusive use of value-added data. It didn't, and I wouldn't. The question is whether a school system could use value-added measures in conjunction with others.

There are three possible positions on this question:

1. Yes, use value-added data in conjunction with other information.
2. No, just use the other information (evaluations, peer review, etc.)
3. No, don't do any selection at all. Keep the single salary schedule. Keep tenuring teachers as a matter of course

Not to belabor the point, but there is no fourth option: "Study the matter further."

In fact, very few people take the third position either. Rote and unreasoned grants of tenure are hard to defend. Plus there is wide support for giving extra pay to groups of teachers who succeed, or to individual teachers who receive National Board Certification. The support for these kinds of differentiation suggests wide discomfort with the uniform treatment of teachers that prevails today. The only real question is whether value-added data can come into the mix.

Here I think there are two arguments against bringing in these data. One is that these measures do not enhance the accuracy of our identification of effective teachers. The other is that they do, but the costs of using the data outweigh the benefits. Let me take these in turn.

An outstanding reason not to use value-added measures would be that they are useless pseudoscience, akin to phrenology or astrology. Those of you who use value-added measurement to make judgments about programs or teachers seem to me to testify, through your work, to the difference between measuring a teacher's effectiveness in the classroom and, say, gazing at his palm.

But more importantly, value-added measures do predict teachers' results in the classroom. As I understand it, one of the biggest concerns about the measures is that they are picking up sorting among students, rather than differences in impacts among teachers. This is the concern highlighted by Jesse Rothstein's work on "predicting the past." And of course it is a real problem. But if this student sorting were *all* that value-added measures represented, then we would find that value-added effects disappear when students are randomly assigned to classes. But this is not what we find. Research from the Tennessee STAR experiments suggests that "teacher effects are real and are of a magnitude that is consistent with that estimated by previous studies." (Nye, Konstantopoulos, and Hedges 2004). And recent research in Los Angeles similarly finds that, at the level of individual teachers, teachers with higher value-added scores based on their prior work achieved higher average gains with randomly assigned groups of students than teachers with lower value-added scores. (Kane & Staiger 2008). But if value-added gains predict actual gains under experimental conditions, then value-added measures are different from skull measures.

Surely others can and should try to repeat these findings. I strongly suspect they will be able to do so, because these findings are corroborated elsewhere. A few researchers have compared value-added measures with principal evaluations. Papers by Brian Jacob and Lars Lefgren (2005), and by Doug Harris and Tim Sass (2007), find correlations between value-added measures and principal evaluations.

Jacob and Lefgren find, for example, that the top category of value-added scores are in math are correctly predicted by principals 69% of the time in math and 52% of the time in reading. If principals were choosing randomly, they say we would expect correct predictions 26% and 14% of the time, respectively. Harris and Sass note that while the correlation of value-added scores to principal evaluations is modest, it is stronger than the correlation for the measures we use to make pay determinations, like experience and advanced degrees.

So value-added measures seem to tell us something useful. Now we come to harder questions, about whether their use yields enough benefits to justify the costs. I think we need a lot more research here—but this research can only succeed if it includes some learning by doing. Let me mention four questions.

First, does good student performance on tests lead to better life achievement down the road? In other words, do tests measure “higher-order thinking” and all that other good stuff? As I understand it, research does regularly correlate test scores, employment, and earnings. (Hanushek 2006; Rose 2006; Jencks & Phillips 1999) But these studies may fail to disentangle achievement from underlying ability that teachers can’t change. I imagine there is interesting research still to be done here. Query once more whether the qualities measured by value-added research are more or less likely to lead to better life outcomes than the qualities captured in master’s degrees, which are already the basis for billions in payouts by school systems today.

Second, will another high-stakes use of testing lead to a whole bunch of bad consequences, like more teaching to the test and less cooperation in the classroom? Even if value-added measurement is supposed to be only one component in a mix, will it crowd out alternatives, with harmful consequences? These are very real concerns too. They can be mitigated by smart policy—using multiple measures so the test doesn’t count too much, for example, and providing rewards to whole schools, not only individual teachers. And investing more in the creation of better tests is also an imperative. As I mentioned earlier, actual practice suggests that high-stakes uses of tests can work out just fine. For example, the Teacher Advancement Program, which uses value-added measures as a basis for some pay and advancement decisions, now operates in more than 200 schools. As a matter of policy, TAP now goes into schools only when teachers support their entry, which they often do. TAP does not represent the only way to use value-added measures, though. The only way to know whether other approaches will cause implementation problems is to implement them.

Third, as John Easton asks, is value-added measurement too esoteric to be useful to teachers in the real world? Can it be a tool for improvement and not merely for judgment? As a lay reader of value-added literature, I agree this stuff is complicated. But I also believe it can be simplified. For all the controversy surrounding them, the value-added reports in New York City today reflect good progress toward simplification. But there is more simplification to do, and the people in this room could be enormous assets.

It is also surely true that we should use these metrics to identify good practice and try to spread it around, not only to evaluate teachers. We badly need to figure out better ways to help teachers improve their practice. But formative and summative evaluation are not mutually exclusive, nor need the tools for one and the other be the same. To return to my old analogy, football teams need quarterback coaches to help passers with their throwing motions, but they also need head coaches and general managers to decide who will be on the team and who will play each week. We do not necessarily insist that the person making the trades know how to throw a pass.

Finally, is the error rate just too darn high? Researchers often point to intertemporal instability and say

that it is self-evidently too great to allow high-stakes uses of the data. But a few points in response. First, the instability declines dramatically when we use multiple years of data, which is better practice anyway. Second, the instability is smaller at the top or the bottom, which are the most relevant parts of the distribution for decisions about added pay and retention. I don't know anybody who is proposing to make fine-grained distinctions between the 40th and 50th or 50th and 60th percentiles.

It is worth unpacking why we might say that a high error rate should disqualify the use of data altogether. One argument is that use of such a flawed measure extracts a high demoralization cost, by making teachers feel that they are subject to an arbitrary fate. Again, that concern needs to be addressed by minimizing the error rate through improved data systems, by using value-added measures in conjunction with others, and by making the data as transparent as possible. The experience of TAP and other places, like Denver, suggests the concern may be overblown. It can be addressed only in practice, not in theory.

Another concern, rarely articulated, seems to be that “getting it wrong” is too great an injustice to teachers. And make no mistake, it *is* an injustice to deny tenure to a teacher who is getting the job done. To a lesser extent, it is an injustice to deny a pay bonus to a teacher who performs well. But it is also an injustice for a child to keep an ineffective teacher in the classroom, or to lose an effective teacher to another profession because we choose not to pay her enough. These are tradeoffs, and we need to keep both sides in mind.

I am a lawyer, so I will close with a legal analogy. In the law, we make rules based on our judgments about the relative importance of different interests. We really don't like to put innocent people in jail, so we tip the scales against wrongful convictions. That's why we apply the beyond a reasonable doubt standard. It's why criminal courts usually refuse to admit polygraph tests. And it's one reason why our courts regularly acquit people who have committed crimes. It's not a bug, it's a feature.

When the stakes do not involve the stigma and loss of liberty from a criminal conviction, we act differently. In civil cases, we just want to get the right answer. So we apply a preponderance of the evidence standard, which is 50% plus 1, and we are more likely to admit polygraphs. We just want to get it right.

Tenure and pay are high-stakes decisions for teachers at risk of losing their jobs or part of their pay. If we focus on the risk of erroneously taking away a job or a pay raise, then we likely will not want to use evidence as flawed as value-added measures. On the other hand, tenure and pay are also high stakes decisions for children who need an education to fulfill their potential in life. They also can be important decisions for the teachers who would benefit from new earnings opportunities and more effective colleagues. If we worry mostly about these groups, then we should at least *try* using the data to increase the likelihood that students will have effective teachers.