Stability of the Louisiana Value Added Model (VAM)

Wayne Free, Louisiana Association of Educators
Wayne.free@lae.org

In a report entitled:

The Status of the Development of the Value Added Assessment Model as specified in Act 54
A report to the Senate Education Committee and the House Education Committee
Of the Louisiana Legislature
February 25, 2011

Dr. George Noell and Dr. Beth Gleason, Strategic Research and Analysis, Louisiana Department of Education (DOE) write: “The results show moderate stability across years. Teachers who fell in the bottom 20% in 2007-2008 were likely to fall in the bottom 20% of results again (mathematics: 45.3%; ELA: 39.8). They were unlikely to move to the top of the distribution one year later. Teachers who were in the top 20% in 2008-2009 were most likely to fall in that range in 2009-2010 (mathematics: 61.6%; ELA: 55.7%). They were unlikely to move to the bottom of the distribution one year later.”

The description of “results show moderate stability” is questionable. The data published by Noell and Gleason indicate that, “Teachers who fell in the bottom 20% in 2007-2008 were likely to fall in the bottom 20% of results again (mathematics: 45.3%; ELA: 39.8)”

Perhaps I am weak in mathematical analysis but when I read data that is not consistent with the conclusions, I question the reasons. Over 50% (54.7%) of the teachers that scored in the bottom 20% of the VAM mathematics score in 2008-2009 scored above the bottom 20% in 2009-2010. In fact, of the teachers who scored below 10% over 70% raised their score the 2nd year without changing teaching practice. So much for “moderate stability” (ELA scores were even wider). The same can be said of the teachers who scored in the top 20% and will be explained in the graphs to come later.

On page 15, first paragraph, of the report Noell and Gleason wrote, “However, the level of correlation across these two consecutive years suggests using caution in reaching conclusions from any single year’s data.”

In an editorial published Tuesday (9/25/2012) the Lake Charles American Press wrote the following, “In 2009-2010, 715 teachers in the trial received a highly effective rating. In 2010-2011, only 277 of those teachers maintained the highly effective rating. And last year, that number fell to 149.” Essentially, this data provided by the DOE indicates a drop from the first year’s top performance was 62% and the drop over 3 years was 79% for the same teachers. Again, a drop of that magnitude would not suggest moderate stability to me.

A report by the Economics Policy Institute (EPI) reinforces the instability of VAM scores. In the announcement of the study results EPI writes, “Student test scores are not reliable indicators of teacher effectiveness, even with the addition of value-added modeling (VAM), a new report by leading testing experts finds. Though VAM methods have allowed for more sophisticated comparisons of teachers than
were possible in the past, they are still inaccurate, so test scores should not dominate the information used by school officials in making high-stakes decisions about the evaluation, discipline and compensation of teachers.

The co-authors make clear that the accuracy and reliability of analyses of student test scores, even in their most sophisticated form, is highly problematic for high-stakes decisions regarding teachers. Consequently, policymakers and all stakeholders in education should rethink this new emphasis on the centrality of test scores for holding teachers accountable.

Analyses of VAM results show that they are often unstable across time, classes and tests; thus, test scores, even with the addition of VAM, are not accurate indicators of teacher effectiveness. Student test scores, even with VAM, cannot fully account for the wide range of factors that influence student learning, particularly the backgrounds of students, school supports and the effects of summer learning loss. As a result, teachers who teach students with the greatest educational needs appear to be less effective than they are. Furthermore, VAM does not take into account nonrandom sorting of teachers to students across schools and students to teachers within schools.

There are further negative consequences of using test scores to evaluate teacher performance. Teachers who are rewarded on the basis of their students’ test scores have an incentive to “teach to the test,” which narrows the curriculum not just between subject areas, but also within subject areas.

Furthermore, creating a system in which teachers are, in effect, competing with each other can reduce the incentive to collaborate within schools—and studies have shown that better schools are marked by teaching staffs that work together. Finally, judging teachers based on test scores that do not genuinely assess students’ progress can demoralize teachers, encouraging them to leave the teaching field. Evaluating teachers accurately is a critical piece of the effort to improve America’s schools, and VAM methods are appealing in that they seem to offer an objective and simplified way of comparing one teacher with another.

However, as EPI’s report makes clear, “There is simply no shortcut to the identification and removal of ineffective teachers.” The authors conclude that, “Although standardized test scores of students are one piece of information that school leaders may use to make judgments about teacher effectiveness, test scores should be only a small part of an overall comprehensive evaluation.”

A recent report released by Marcus A. Winters, a senior fellow at the Manhattan Institute and an assistant professor at the University of Colorado, Colorado Springs, takes a different view. Winters argues in “Transforming Tenure: Using Value-Added Modeling to Identify Ineffective Teachers,” that VAM scores can be used to deny or assign tenure to classroom teachers. However, he points out that his research as well as those of others finds a great deal of instability and error in a single year of VAM scores and he recommends the use of multiple years of scores before any decision is made concerning firing or continuing employment.

In his report Winters writes, “Though VAM is a powerful technique, it is undoubtedly an imperfect measure of a teacher’s effectiveness. VAM is limited partly because it considers student performance only as measured by standardized tests, which are themselves imperfect measures of student achievement and account for only part of what school systems ask teachers to do. But even as a measure of the teacher’s contribution to student test scores, VAM has potentially serious limitations.
Critics of VAM analysis rightly point out that, as a statistical tool, VAM must contend with measurement error—the inevitable fact that measurements of the same thing, taken at different times, will vary, and some of this variation will be essentially random. VAM-based measures of teacher performance can be quite imprecise. When VAM is used to inform tenure decisions, it is likely that some average and even above-average teachers could be removed from the classroom because of a low VAM score caused by random variation in measurement over the years, rather than their own failures. The influence of measurement error can be mitigated by statistical adjustments and by incorporating multiple years of student performance when evaluating any particular teacher. But measurement error cannot be eliminated."

The point is this... VAM as it was proposed by the Governor, adopted by the legislature, and is being administered by BESE, and the DOE is unstable and does not fairly represent teacher/student performance. There are a number of reasons for this and dozens of other research studies/ papers pointing them out.

In this discussion I will focus on a lack of stability which is inherent in Value Added Models and the confidence interval/ margin of error involved in value-added and other teacher performance measures and what that means as you focus on RIF Polices, Salary Schedules, etc.

The following data which was included in the report by Noell and Gleason provides the information. The graphs that break the data down to make it understandable illustrates that stability is lacking.

Graph 1 shows that over 70% of the teachers who score in the bottom 10% on the first year’s evaluation score higher than the bottom 10% the second year. (19% increase to the 10-20% level and 46% increase to the 20-80% level).

Graph 5 shows that over 50% if the teachers who scored in the top 10% dropped below that level the second year. (16% moved down to the 80-90% level and 35% moved down to the 20-80% level)

The rest of the graphs break down the data for the lower 10-20%, 20-80% and 80-90%.

Each of these graphs identifies a range of scores which might be appropriate as an estimation of a teacher’s measure of performance. That range is often called a “confidence interval or margin of error” and based on the data provided I would estimate the confidence interval as plus or minus 20 - 25% and perhaps more. That means a teacher estimated to be at the 50% range of all other teachers in that field might actually score anywhere from 25% to 75%. We should not be surprised at this since the various studies from around the country confirm that VAM scores are inconsistent from year to year in all cases studied and that validity errors are a major issue.

The report provided to the legislature does not include validity data for Social Studies and Science. I originally suspected it was because the data would indicate that the validity of those scores was even weaker than mathematics and ELA and that the confidence interval would be greater. I requested those scores from the DOE. (See the attached email from Renee Greer... The data does not exist).

All of this would be academic if it weren’t for the politics of the issue and the rush to judgment about teacher performance using a system that is inappropriate for “high stakes” decisions.

If you are interested in further information about VAM research or other issues surrounding teacher effectiveness email me at wayne.free@lae.org
Graph # 1

Teachers who ranked in bottom 1 to 10% in Year 1 who:

- Remained in bottom 1-10% in Year 2: 27%
- Moved up to bottom 11-20% in Year 2: 4%
- Moved up to middle 21-80% in Year 2: 4%
- Moved up to top 81-90% in Year 2: 19%
- Moved up to top 91-99% in Year 2: 46%

Graph # 2

Teachers who ranked in bottom 11-20% in Year 1 who:

- Moved down to bottom 1-10% in Year 2: 2%
- Remained in bottom 11-20% in Year 2: 15%
- Moved up to middle 21-80% in Year 2: 62%
- Moved up to top 81-90% in Year 2: 5%
- Moved up to top 91-99% in Year 2: 2%
Graph # 3

Teachers who ranked in middle 21-80% in Year 1 who:

- Moved down to bottom 1-10% in Year 2: 10%
- Moved down to bottom 11-20% in Year 2: 7%
- Remained in middle 21-80% in Year 2: 64%
- Moved up to top 81-90% in Year 2: 9%
- Moved up to top 91-99% in Year 2: 10%

Graph 4

Teachers who ranked in top 81-90% in Year 1 who:

- Moved down to bottom 1-10% in Year 2: 5%
- Moved down to bottom 11-20% in Year 2: 3%
- Moved down to middle 21-80% in Year 2: 16%
- Remained in top 81-90% in Year 2: 54%
- Moved up to top 91-99% in Year 2: 22%
Wayne,

Answers are below, based on the mathematics data (exact numbers vary slightly between content areas.

1. If a teacher scores in the lowest 10% of the VAM score the first year and does nothing different the next year what is the likelihood they will fall in the lowest 10% the second year and remain "ineffective".

26.8%

2. If a teacher scores in the lowest 10 - 20% range of the VAM score the first year and does nothing different the next year what is the likelihood they will fall in the lowest 10% the second year and become "ineffective".

14.8%

3. If a teacher scores in the lowest 20 - 30% range of the VAM score the first year and does
nothing different the next year what is the likelihood they will fall in the lowest 10% the second year and become "ineffective".

I don't have those data and I don't have the to run that since I am back at LSU.

4. If a teacher scores in the highest 10% of the VAM score the first year and does nothing different the next year what is the likelihood they will fall in the highest 10% the second year and remain "highly effective".

45.8%

5. If a teacher scores in the highest 10 - 20% of the VAM score the first year and does nothing different the next year what is the likelihood they will fall in the highest 10% the second year and become "highly effective".

22.1%

6. If a teacher scores in the highest 20- 30% of the VAM score the first year and does nothing different the next year what is the likelihood they will fall in the highest 10% the second year and become "highly effective".

I don't have those data and I don't have the to run that since I am back at LSU.

7. I guess what I'm actually asking is what is the stability range across years based on a 10% differential each year and not the top to bottom analysis given in the report

Numbers are above.

Hope that helps.

George

_______________________
George Noell, PhD, BCBA
Professor
Department of Psychology
Louisiana State University
Baton Rouge, LA 70803
office: (225) 578-4119
FAX: (225) 578-4125
Wayne,

Your request for information for the Value-Added report to the legislature has been forwarded to me. The data related to Science and Social Studies Value-Added Data does not exist. However, I believe you have the Department’s report to the Legislature, which is also posted on our website.

Thank you.

Rene’ Greer
Director of Public Affairs
Louisiana Department of Education
1201 North Third Street
Baton Rouge, Louisiana 70802
225-342-3600
www.louisianaschools.net

From: Wayne.Free@lae.org [mailto:Wayne.Free@lae.org]
Sent: Wednesday, May 09, 2012 10:09 AM
To: Kim Nesmith
Subject: RE: Bulletin 130

Kim,
Thanks, I know what you mean things are busy here too... By the way, I have been looking at George and Beth’s report to the legislative education committees on the VAM validity and I think I need to look at the Science and Social Studies data to get a better idea of what they are trying to convey. It would be great if I could get it in the same format as provided for the mathematics and ELA data.

Thanks again,

Wayne