

The Value of Value-Added Methods in Identifying Effective Teachers

Barnett Berry and Ed Fuller

Introduction

Many policymakers and school reformers are calling for the use of student standardized test scores as a primary—and in some cases, sole—means to identify and/or reward good teachers and root out bad ones. Often, they propose to rate teacher performance by using a value-added methodology (VAM) that measures how individual teachers influence learning for each child. VAM draws on new statistical techniques that use multiple years of student achievement data to estimate the effects of schools or teachers. Students are tracked as individuals over time, not as cohorts, and serve as their own controls. By tracking individual students' academic growth over several years and different subjects, researchers can estimate the contributions that teachers make to that growth. A number of studies using these methods have surfaced, with stunning results. For example, William Sanders, a pioneering VAM statistician, found in one of his earlier studies that students assigned to the most effective teachers for 3 years in a row performed 50 percentile points higher than comparable students assigned to the least effective teachers for 3 years in a row (Sanders & Rivers, 1996).

One of the most important applications of VAM to date is emerging in Louisiana, where the state's Board of Regents, under the leadership of Jeanne Burns (Associate Commissioner for Teacher Education Initiatives) and LSU researcher, George Noell, has been testing the use of a new Value-Added Teacher Preparation Program Assessment Model that has the "capacity to examine the growth of achievement of children and link growth in student learning to teacher preparation programs." The state has been developing a comprehensive teacher quality (TQ) data system so that they can track their teacher education graduates and determine how long they stay in teaching and where they go if they leave. Then, using a specially designed VAM, researchers can determine the effectiveness of the state's teacher preparation (both traditional and alternative) programs in helping increase student achievement.

Using his VAM model, Noell found in his 2003-04 analyses that students of some teachers from some universities systematically demonstrated greater academic achievement, as measured by state standardized tests in English/language arts, than other universities.¹ His preliminary analyses also identified one teacher preparation program whose new graduates taught students whose growth in learning in mathematics surpassed the growth of learning in mathematics of children taught by experienced teachers. Additional analyses in 2004-05 replicated the initial findings. Based upon results of this study, the state decided to explore further the use of a Value-Added Teacher Preparation Program Assessment Model in Louisiana and funded a study to examine the technical qualities and adequacy of the model when using data across all of the state's 68 school districts. We believe strongly with the idea that VAM models help focus attention on TQ and teacher education where it belongs: on increasing student learning.

VAM offers, in the eyes of some reformers and policymakers, a simple methodology that can identify reliably and accurately the effect of individual teachers and teacher preparation

programs on student achievement. Although we strongly believe that VAM holds great promise in multiple areas and states should move forward in this area, we also strongly believe that using VAM is far from a “simple” process and is often fraught with technical problems. Therefore, high-stakes decisions about teachers or preparation programs should not be made in isolation from other robust sources of data.

Problems With VAM in Assessing Teacher Effects

1. Standardized tests—especially those relying on a multiple-choice format—are not perfect measures of student achievement. Standardized tests can capture, although not with perfect precision, whether or not students have mastered the “basics” or have memorized facts or applied formulas in routine ways. These tests, however, typically do not do a very good job of determining whether or not students have developed higher order thinking skills or advanced reasoning. Although Sanders and other researchers have found that “good instruction” is a far more powerful predictor of student achievement than income and race, the tests that are used to assess learning must be of high quality and must measure meaningful outcomes reliably.
2. These tests all have what is called *random error*, which often limit them in measuring the performance of both students and the teachers who teach them. A number of today’s standardized tests, used in states and school districts across the nation, have significant statistical sampling problems—to the point that they should not be employed for many of the high-stakes decisions for which they are now being used (Kane & Staiger, 2001).
3. Some researchers, such as Sanders, argue that student background characteristics should not be taken into account when assessing the effect of teachers on student performance. However, when Dale Ballou (2002), an economist noted for his support of market principles in improving TQ, examined the often-heralded system of value-added assessment in Tennessee, he noted that value-added assessments may not adequately control for factors like poverty and limited English proficiency that may affect a student’s rate of progress as well as his or her absolute performance. Another review (Kupermintz, 2003) found similar problems and also pointed to other factors beyond the control of a teacher (like student attendance and high student mobility) that can greatly affect whether a value-added accountability system will accurately gauge an individual teacher’s direct impact on the learning gains of a large group of students. Indeed, a recent RAND report concluded, “Models that fail to account for differences in student population across schools can yield biased estimates of teacher effects. This is the case even for complex multivariate models that jointly model student outcomes” (McCaffrey, Lockwood, Koretz, & Hamilton, 2004, p. xvii). In other words, models that omit controls for student background characteristics can provide results that do not accurately identify the true effects of teachers on student achievement.

Problems with VAM in Assessing Teacher Preparation Program Effects

1. First, as stated previously, any use of VAM is only as good as the standardized tests upon which VAM relies. Unfortunately, too many state-mandated tests assess only lower level thinking skills and rely too often on multiple-choice formats. This is not necessarily the fault of the state: Creating more useful tests like the National Assessment of Educational Progress (NAEP) is usually far too expensive for state education agencies to develop and score.
2. There is the problem of missing data. Not only is there a large amount of missing data with respect to individual student test scores, but there is also the problem of missing data on teachers. Not all states have an effective data system that can track teachers into schools and classrooms. Moreover, if a preparation program sends graduates to private rather than public schools, these teachers most likely will show up as “missing” in any state data system.
3. Only a small percentage of teachers produced by a teacher preparation program could be assessed using VAM. Typically, only elementary teachers and secondary mathematics and reading/language arts are assigned to teach students for which there would be an adequate amount of test data to conduct a VAM. Moreover, only teachers assigned to teach in grade levels for which there were scores from the previous grade level could be included in any VAM analysis. As we have found from our work in Texas, this results in a relatively small percentage of teachers included in a VAM analysis. Making any judgments about a teacher preparation program based on a small subset of the program’s teachers seems rather unfair. The results of such an analysis, however, could be used in a formative manner to improve practice.
4. Finally, any VAM used to assess the effectiveness of a teacher preparation program must also take into account the types of schools in which graduates are employed. Typically, graduates from a teacher preparation program take jobs that are fairly close to the geographic location of the program. If the schools surrounding one preparation program differ systematically from schools surrounding another preparation program, and if those differences are not controlled for in the VAM, then the results of the VAM may be inaccurate. In fact, the VAM may conclude the difference in teacher effectiveness between two programs is due to differences in the preparation programs, when, in fact, the difference in teacher effectiveness is more related to the difference in capacity of the two sets of schools. Thus, a comprehensive VAM must collect data on the capacity of schools to effectively support new teachers and to provide the necessary instructional resources.

Conclusions About VAM

While the use of VAM becomes more prevalent, there are serious problems with the implementation of such efforts. If the fundamental technical issues underlying the VAM process are not fully explored and addressed, “VAM is likely to misjudge the effectiveness of teachers

and schools and could produce incorrect generalizations about their characteristics, thus hampering systematic efforts to improve achievement” (McCaffrey et al., 2004, p. iii).

Despite the potential pitfalls of using VAM, we concur with RAND and others that VAM holds great promise, and researchers and policymakers should work together carefully to implement and research the use of VAM to accurately identify teacher and school effects. We do not, however, condone the sole use of VAM to make high-stakes decisions about teachers, schools, or preparation programs. Indeed, the RAND Corporation, after a 2-year investigation of VAM, pointed to a number of other technical problemsⁱⁱ and concluded that while the model is more sound than many other methods currently being used for test-based accountability, “the research base is currently insufficient to support the use of VAM for high-stakes decisions” (McCaffrey et al., 2004, p. 8). Dale Ballou also has been critical of relying solely on value-added standardized achievement scores for teacher accountability purposes.

Thoughtful analysts and reformers like Ted Hershberg (2005) call for value-added assessments to be part of individual accountability only if it is coupled with other measures of teaching performance. As Hershberg said eloquently, as long as teachers are treated “fairly as individuals,” then “we must be willing to innovate, take risks, and not let perfect be the enemy of the good.”

References

- Ballou, D. (2002, Summer). Sizing up test scores. *Education Next*, 2(2). Available at <http://www.educationnext.org/20022/10.html>
- Hershberg, T. (2005, December). Value-added assessment and systemic reform: A response to the challenge of human capital development. *Kappan*, 87, 276-283.
- Kane, T. J., & Staiger, D. O. (2001). *Volatility in school test scores: Implications for test-based accountability systems*. Paper presented at a Brookings Institution Conference, Washington, DC.
- Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis*, 25, 287-298.
- McCaffrey, D., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2004). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND.
- Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement*. University of Tennessee, Value-Added Research and Assessment Center.

ⁱ Dr. Noell also found, not surprisingly, that students of new teachers overall did not demonstrate as much growth in English/language arts as students of experienced teachers.

ⁱⁱ RAND, after a 2-year investigation of VAM, funded by Carnegie, pointed out that the way VAM controls for “context” may not be sufficient for precisely measuring teacher effects; in other words, the models used may not properly distinguish the effects of teachers from other effects of the school in which the teacher works. The researchers expressed concern about incomplete data, which can arise in two areas: data for individual students over time and information on the linking of students to teachers. The researchers noted that several other problems, including the infrequent testing (once a year), the limited number of topics tested, and the way test scores are scaled, may be biasing the estimated teacher effects.