

RESEARCH BRIEFS

Brief #8

Fall 2011

Value-Added Measurement

Types of Models Used to Measure Changes in Test Scores

Several types of test-based evaluation models currently are used for education decision-making. Each type of model is designed to answer a different set of policy-relevant questions. Only the last is a value-added model.¹

1. Status models give a **snapshot** of student performance at a point in time, which is often compared with an established target. For example, the mean test score for a subgroup of students or a school can be compared with the state's annual target to determine if the school has met the state goal. A status model (as with the WKCE) is useful if one wants to answer such questions as "What percentage of students in the state is performing at the proficient level this year?" "Has school X met the state proficiency target this year?"

2. Cohort-to-cohort change models can be used to measure the change in test results for a teacher, school, or state by comparing status at two points in time—but not for the same students. For example, the percentage proficient for this year's fourth graders in reading on the WKCE can be compared with that of last year's fourth graders. A cohort -to-cohort change model answers the question, "Are students at a certain grade level doing better this year in comparison to the students who were in the same grade last year?"

3. Growth models measure student achievement by tracking the test scores of the same students from one year to the next to determine the extent of their progress. Gain scores can be computed to compare the performance of the current year's fourth graders with that of the same group of students last year, when they were in third grade. This type of model is preferable if one wants to know "how much, on average, did students' performance change between grade X and grade Y?" There might also be a statewide growth target that subgroups or school systems must meet. Accountability systems built on growth models give teachers and schools credit if their students show improvement, regardless of whether they were high-performing or low-performing to begin with.

RESEARCH BRIEFS

However, growth models usually do not control for student or school background factors, and therefore they do not attempt to address [explain] which factors are responsible for student growth.

4. Value-Added Measurement (VAM) refers to different statistical procedures (or models) that are applied to the scores of students on standardized achievement tests in order to estimate the relative contributions of specific teachers, schools, or programs to student test performance. VAM is intended to answer questions such as the following: how did teacher A compare with teacher B in terms of how much the test scores of students improved over some period of time, or how much of the change in students' performance can be attributed to attending one school compared with another?

Proponents claim that VAM is fair and objective because it does not penalize high poverty, low-achieving schools by comparing them with their more affluent and higher-achieving neighbors. Instead, VAM allows us to honor low achieving schools that show significant improvement, even though many students fail to reach some arbitrary standard, such as Proficient or Advanced on state tests.

All this sounds very good; however, we need to recognize that VAM is based exclusively on standardized achievement tests--which are predominantly multiple choice with a few short answer questions thrown into the mix. These types of tests provide us with a snapshot about how students are doing in a few key areas. However, they do not measure problem solving skills, divergent thinking, communication skills, the ability to give performances or create projects, citizenship, character, or the capability to work with and get along with others. These are the skills and knowledge that the public wants students to have when they leave school.

Proceed With Caution: Using VAM to Evaluate or Compensate Educators

Several respected organizations have questioned the use of VAM to measure teacher effectiveness. For example, researchers at the Rand Corporation recently noted that “. . . the estimates from VAM modeling of achievement will often be too imprecise to support some of the desired inferences. . .” They also state that “. . . the research base is currently insufficient to support the use of VAM for high stakes-decisions about individual teachers or schools” (Problems with the Use of Student Test Scores to Evaluate Teachers, p. 3).

RESEARCH BRIEFS

A similar comment comes from the Board on Testing and Assessment of the National Research Council of the National Academy of Sciences: "...VAM estimates of teacher effectiveness should not be used to make operational decisions because such estimates are far too unstable to be considered fair or reliable" (Problems with the Use of Student Test Scores to Evaluate Teachers, p. 8).

National Reports on Value-Added Measurement

In 2008 and 2010 there were two major reports written by experts in assessment, statistics, measurement, teacher evaluation, and economics:

1. In 2008, a workshop was held on Value Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability. The report of this group ("Getting Value out of Value-Added") was approved by the Governing Board of the National Research Council.²
2. In 2010, a report entitled "Problems with the Use of Student Test Scores to Evaluate Teachers" was released.³

Listed below are observations made by the members of these panels. Their observations fall into four broad categories: (1) Problems with Value-Added Measurement (Models); (2) Concerns about Validity and Reliability; (3) Concerns about Unintended Consequences of Value-Added Testing; (4) Concerns about What the Tests Actually Measure. Added to this list of four observations are two additional issues that are important to any discussion of value added: (5) Attribution and (6) Ceiling Effect.

RESEARCH BRIEFS

1. Problems with Value-Added Measurement (Models)

- “To nobody’s surprise, there is not one dominant VAM. Each major class of models has shortcomings, there is no consensus on the best approaches, and little work has been done on synthesizing the best aspects of each approach. There are questions about the accuracy and stability of value added estimates of schools, teachers, or program effects. More needs to be learned about how these properties differ, using different value-added techniques and under different conditions” (Getting Value out of Value-Added, p. 25).
- “For a variety of reasons, analyses of VAM results have led researchers to doubt whether the methodology can accurately identify more and less effective teachers. VAM estimates have proven to be unstable across statistical models, years, and classes that teachers teach” (Problems with the Use of Student Test Scores to Evaluate Teachers, p. 2).
- “. . . even when methods are used to adjust statistically for student demographic factors and school differences, teachers have been found to receive lower ‘effectiveness’ scores when they teach new English learners, special education students, and low-income students than when they teach more affluent and educationally advantaged students” (Problems with the Use of Student Test Scores to Evaluate Teachers, p. 3).
- “. . . The claim that [VAM] can “level the playing field” and provide reliable, valid, and fair comparisons of individual teachers is overstated. Even when student demographic characteristics are taken into account, the value-added measures are too unstable (i.e., vary widely) across time, across the classes that teachers teach, and across tests that are used to evaluate instruction, to be used for the high-stakes purposes of evaluating teachers (Problems with the Use of Student Test Scores to Evaluate Teachers, p. 9).
- “Several studies show that VAM results are correlated with the socioeconomic characteristics of the students. This means that some of the biases that VAM was intended to correct may still be operating” (Problems with the Use of Student Test Scores to Evaluate Teachers, p. 10).

RESEARCH BRIEFS

2. Concerns about Validity and Reliability

- "... there is broad agreement among statisticians, psychometricians, and economists that student test scores alone are not sufficiently reliable and valid indicators of teacher effectiveness to be used in high stakes personnel decisions, even when the most sophisticated statistical applications of value added modeling are employed" (Problems with the Use of Student Test Scores to Evaluate Teachers, p.2).
- "The larger the number of students in a tested group, the smaller will be the average error because positive errors will tend to cancel out negative errors. But the sampling error associated with small classes of, say, 20-30 students could well be too large to generate reliable results. Most teachers, particularly those teaching elementary or middle school students, do not teach enough students in any year for average test scores to be highly reliable"(Getting Value out of Value-Added, p. 12).
- "Small sample sizes are particularly a problem when trying to estimate teacher effects. . . Because longitudinal student data are needed, missing data can further shrink the sample size. For a classroom of 25 students, the effective sample size may dwindle down to 10 because of missing data and student mobility" (Getting Value out of Value-Added, p. 33).

The typical classroom has too few students to produce reliable results. Because the number of students is so important in VAM, the standard error of measurement always should be reported (along with the 95 percent and 99 percent confidence intervals). This is necessary because all test scores are estimates that are subject to error (similar to a public opinion poll that reports the margin of error).

- "... because of the broad agreement by technical experts that student test scores alone are not a sufficiently reliable or valid indicator of teacher effectiveness, any school district that bases a teacher's dismissal on her students' test scores is likely to face the prospect of drawn-out and expensive arbitration and/or litigation in which experts will be called to testify, making the district unlikely to prevail. The problem that advocates had hoped to solve will remain, and could perhaps be exacerbated" (Getting Value out of Value-Added, p. 20).

RESEARCH BRIEFS

- “All value-added models produce estimates of school or teacher effects that vary from year to year. This raises the question of the degree to which this instability reflects real variation in performance from year to year, rather than error in the estimates. McCaffrey discussed research findings demonstrating that only about 30 to 35 percent of teachers ranked in either the top or bottom quintile in one year remain there in the next year. If estimates were completely random, 20 percent would remain in the same quintile from one year to the next. If the definition of a weak teacher is one in the bottom quintile, then this suggests that a significant proportion of teachers identified as weak in a single year would be falsely identified” (Getting Value out of Value-Added, p. 45).
- “Instability will tend to erode confidence in value-added results on the part of educators because most researchers and education practitioners will expect that true school, teacher, or even program performance will change only gradually over time rather than display large swings from year to year. Moreover, if estimates are unstable, they will not be as credible for motivating or justifying changes in future behavior or programs” (Getting Value out of Value-Added, p. 46).
- “Research on the precision of value-added estimates consistently finds large sampling errors. As McCaffrey reported, based on his prior research, standard errors are often so large that about two-thirds of estimated teacher effects are not statistically significantly different from the average” (Getting Value out of Value-Added, p. 45).

3. Concerns about Unintended Consequences of Value-Added Testing

- Because “. . . a value-added system compares the performance of teachers relative to one another, it could reduce teacher cooperation within schools, depending on how the incentives are structured. On the other hand, if school-level value-added is rewarded, it can create a “free rider” problem whereby some shirkers benefit from the good work of their colleagues, without putting forth more effort themselves” (Getting Value out of Value-Added, p. 15).
- “Using test scores to evaluate teachers unfairly disadvantages teachers of the neediest students. Because of the inability of value-added methods to fully account

RESEARCH BRIEFS

- for the differences in student characteristics and in school supports, as well as the effects of summer learning loss, teachers who teach students with the greatest educational needs will appear less effective than they are” (Problems with the Use of Student Test Scores to Evaluate Teachers, p. 15).
- “Narrowing of the curriculum to increase time on what is tested is another negative consequence of high-stakes uses of value-added measures for evaluating teachers” (Problems with the Use of Student Test Scores to Evaluate Teachers, p. 16).
- “Pressure to raise student test scores, to the exclusion of other important goals, can demoralize good teachers and, in some cases, provoke them to leave the profession entirely” (Problems with the Use of Student Test Scores to Evaluate Teachers, p. 19).

4. Concerns about What the Tests Actually Measure

- “To date, most value-added research in education has been conducted by specialists in education statistics, as well as by economists who work in the area of education policy analysis. At the workshop, Dale Ballou, an economist, pointed out that ‘the question of what achievement tests measure and how they measure it is probably the [issue] most neglected by economists. . . . If tests do not cover enough of what teachers actually teach (a common complaint), the most sophisticated statistical analysis in the world still will not yield good estimates of value-added unless it is appropriate to attach zero weight to learning that is not covered by the test” (Getting Value out of Value-Added, p 27).
- “A test covers only a small sample of knowledge and skills from the much larger subject domain that it is intended to represent (e.g., fourth grade reading, eighth grade mathematics), and the test questions are typically limited to a few formats (e.g., multiple choice or short answer). The measured domains themselves represent only a subset of the important goals of education; a state may test mathematics, reading, and science but not other domains that are taught, such as social studies, music, and computer skills. Furthermore, large-scale tests generally do not measure other important qualities that schools seek to foster in students but are more difficult to measure, such as intellectual curiosity, motivation, persistence in tackling

RESEARCH BRIEFS

difficult tasks, or the ability to collaborate well with others. For these reasons, value-added estimates are based on a set of test scores that reflect a narrower set of educational goals than most parents and educators have for their student” (Getting Value out of Value-Added, p. 29).

- “. . . if a state’s science standards emphasize scientific inquiry as an important goal, but the state test primarily assesses recall of science facts, then the test results are not an appropriate basis for using value-added models to estimate the effectiveness of science teachers with respect to the most valued educational goals. A science teacher who focuses instruction on memorization of facts may achieve a high value-added. . . whereas one who emphasizes scientific inquiry may obtain a low value added” (Getting Value out of Value-Added, p. 30).

5. Attribution

Attributing (or crediting) the gains or losses in students’ test scores to a specific teacher, program, or a group of teachers in a building, is not an easy task. This is due to the fact that numerous variables affect gains in student learning, not just the teacher(s) to whom students currently are assigned. In fact, what happens outside the school has a stronger overall effect on student achievement than anything that happens within the school.

Among the more important factors affecting students’ achievement are the following: a student’s previous teachers, home and family environment (including parents’ education and expectations), one’s past and present friends, personal motivation, specialists within the school, and the expectations of teachers and friends. The complexity of factors that influence student success or failure makes it risky to attribute gains in student learning to a single teacher or group of teachers.

Attribution is especially problematic when tests such as Wisconsin Knowledge & Concepts Examinations (WKCE) are used to measure student gains. The WKCE is administered only once each year in November. Results are returned to districts approximately four months later, in February or March. In order to use these, it becomes necessary to use the test scores from two different years, when nearly all students will have been taught by different teachers. This makes attribution even more of a challenge.

RESEARCH BRIEFS

6. Ceiling Effect

In any discussion of value-added or growth measures of student achievement, the issue of ceiling effect often is overlooked. Many standardized achievement tests (including the WKCE) have been designed to measure the knowledge and skills of the typical or “average” student. They never were intended to measure the upper limits of what students know or are able to do. As a result, it is difficult for high achieving students to show improved test scores even when their knowledge and skills have grown significantly.

As Kodel and Betts point out, this characteristic has significant consequences if these tests are used to evaluate schools, programs, or individual teachers:

“We refer to the tendency for gains in a student’s test score to be smaller if the student’s initial score is toward the top end of the distribution, simply because the student has little room for improvement given the difficulty level of the test, as a ‘ceiling effect.’ Ceiling effects will be most pronounced in minimum-competency or proficiency-based tests which are being used increasingly across the United States” (p. 2370).⁴

They further state: “. . . consider a testing instrument where the top 20 percent of the student population is at or near the maximum possible score. Teachers and schools charged with raising these students’ test scores will have little opportunity to add value” (p. 2371).

Recommendations for the Use of VAM:

The measurement experts cited in this paper express numerous concerns about VAM, suggesting the following:

- VAM may be appropriate for low stakes purposes, such as measuring how well a particular program or school is doing. However, it should not be used for teacher evaluation or identification of teacher effectiveness. VAM has significant error when small samples of students are tested. In addition, different value-added models produce different estimates of teacher effects.

RESEARCH BRIEFS

- Whenever value added is calculated and reported, the standard error of measurement should be reported (along with the 95 and 99% confidence intervals). This is necessary because all test scores are estimates that are subject to error. The confidence intervals should be reported for both pre-tests and post-test results.
- Value added models that collect demographic information about the students (e.g., family income, parents' levels of education, etc) should be used so that the effects of schooling can be separated from the effects of family or other outside sources. Teachers and administrators should not be credited/blamed for gains or losses over which they have no direct influence.
- Whenever VAM is used, unintended consequences should be expected. Users of this methodology should determine if value added measurement is resulting in unintended effects, and if so, whether these effects are negative or positive.
- If VAM is used, at least three years of data should be collected and used. This recommendation is made because research has shown value added results tend to be unstable (e.g., teachers/schools judged as adding high value in one year can be identified as adding low value in subsequent years).

Written by: Russ Allen PhD and Jeffrey Leverich PhD

Endnotes

¹ "Getting Value out of Value-Added," Members who participated in the workshop were drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. Available online: http://www.nap.edu/catalog.php?record_id=12820.

² The authors of this report included four former presidents of the American Educational Research Association, two former presidents of the National Council on Measurement in Education, the current and two former chairs of the Board of Testing and Assessment of the National Research Council of the National Academy of Sciences, the president-elect of the Association for Public Policy Analysis and Management, the former director of the Educational Testing Service's Policy Information Center, a former associate director of the National Assessment of Educational Progress, a former assistant U.S. secretary of education, a member of the National Assessment Governing Board; and the vice president, a former president, and three other members of the National Academy of Education. The report is available online: http://www.nap.edu/catalog.php?record_id=12820.

³ The report is available online: <http://www.epi.org/publications/entry/bp278>.

⁴ Corey Koedel and Julian Betts. Test Score Ceiling Effects and Value-Added Measures of School Quality. Available online: <http://www.amstat.org/sections/srms/proceedings/y2008/Files/301495.pdf>.