INFORMATION CAPSULE
Research Services

Vol 1609
February 2017

Christie Blazer, Supervisor

USING VALUE-ADDED MODELS TO EVALUATE TEACHERS

At a Glance

The majority of states include value-added models (VAMs) or some other measure of student academic growth as a component of their teacher evaluation systems. However, there is considerable disagreement among researchers about whether states and school districts should use student growth measures to make high-stakes personnel decisions. Many researchers have concluded that VAMs and other student growth models are not appropriate as the primary measure for evaluating individual teachers. Others maintain that VAMs provide important and useful information on the effects that teachers have on their students' achievement. Studies on the consequences of using VAM scores in teacher evaluation systems are still accumulating. However, evidence has begun to emerge that teachers' VAM scores may depend more on the students they teach and the schools where they work than on the effectiveness of their teaching.

This Information Capsule reviews the problems researchers have documented with using VAMs as measures of teacher effectiveness. For example, teachers' VAM scores vary depending on the students in their class, the school where they work, the specific achievement test used, and the statistical model used in the calculations; a teacher's VAM score can vary substantially from year to year; VAM scores are not highly correlated with other measures of teacher effectiveness; and it is difficult to isolate the impact of a single teacher on students' academic growth.

Researchers have urged caution when basing teacher evaluations on VAM scores. Their recommendations for calculating and interpreting VAM scores, summarized in this Information Capsule, include using VAM scores as only one component of a comprehensive teacher evaluation system, using multiple years of data when calculating VAM scores, and calculating VAM scores only in grades and subjects where there are highly reliable and valid assessments that are comparable over time.

There is a growing consensus that teacher evaluations should include an objective measure of teachers' contributions to student learning. Until recently, most states and school districts have failed to design teacher evaluation systems that distinguish between effective and ineffective teachers - some studies have found that as many as 99% of teachers receive evaluation ratings of "satisfactory" (Ryshke, 2013; Goldhaber & Theobald, 2012; Glazerman et al., 2010). As a result, the majority of states include a measure of student academic growth as a component of their teacher evaluation systems and use teachers' evaluation ratings to make decisions regarding compensation, promotion, tenure, and dismissal (Barnum, 2016a; Strauss, 2016; American Educational Research Association, 2015; American Statistical Association, 2014;

Darling-Hammond, 2012; Lomax & Kuenzi, 2012; Goldhaber & Hansen, 2010).

Teacher evaluation systems that incorporate measures of student academic growth have been supported by federal resources disseminated through several programs, including the State Fiscal Stabilization Fund (under the American Recovery and Reinvestment Act), Race to the Top, the Teacher Incentive Fund, and the Investing in Innovation Fund (Anderson et al., 2016).

Use of Student Academic Growth Measures Nationwide

The majority of states partially judge teachers based on student learning gains. The National Council on Teacher Quality (NCTQ) tracks trends in educator evaluations. According to their 2017 report, 40 states incorporate evidence of student learning into teachers' evaluations. Thirty of the 40 states require measures of student academic growth to be a significant factor in teacher evaluations (at least 30% of the overall evaluation framework) and 10 states require some student growth (less than 30% of the overall evaluation framework). Eleven states do not require objective measures of student growth to be included in teacher evaluations (Walsh et al., 2017).

States and school districts choose among a number of statistical approaches to measuring teacher effectiveness based on student test scores. The most widely used models are VAMs and student growth percentiles (SGPs). SGPs use regression analysis to calculate students' growth compared to other students who received similar prior test scores. While some SGP models consider student and school factors, they usually do not control for as many variables as their VAM counterparts (Gitomer et al., 2014; Guarino et al., 2014; Hull, 2013).

Guarino and colleagues (2014) used simulated data and actual data from a large diverse anonymous state to compare teacher rankings using VAMs and SGPs. They found that VAMs and SGPs ranked teachers very similarly under random assignment of students to teachers. However, when students were nonrandomly assigned to teachers, VAMs outperformed SGP models. Walsh and Isenberg (2013) used data from the District of Columbia Public Schools and compared teacher evaluation scores based on VAMs and SGPs. They found that use of SGPs lowered the effectiveness scores of teachers who had larger proportions of English language learners and increased the evaluation scores of teachers of low-achieving students.

Based on the information available on State Department of Education websites, approximately 18% of the 40 states that require measures of student growth as a component of teacher evaluation systems use VAMs and 28% use SGPs. The rest of the states (approximately 55%) do not mandate the use of a specific student growth model in teacher evaluations.

Some teacher evaluation systems also include Student Learning Objectives (SLOs). SLOs are measures of long-term academic growth that demonstrate a teacher's impact on student learning over the course of the school year. SLOs vary considerably in design. They may be based on commercially available tests, developed individually by the teacher, or developed at the state, district, or school level. Based on students' academic growth on the designated assessment from the beginning to the end of the school year, an evaluation score is assigned to the teacher. SLOs are popular for measuring students' academic growth in grade levels and subject areas in which state achievement tests are not administered (Gitomer et al., 2014).

The NCTQ's *Running in Place: How New Teacher Evaluations Fail to Live Up to Promises,* includes a summary of each state's teacher evaluation policy, along with links to websites where the Council obtained information for each state (Walsh et al., 2017). The report can be

accessed at http://www.nctq.org/dmsView/Final_Evaluation_Paper. Based on the Council's report, the table in the Appendix at the conclusion of this report lists each state, whether it includes student growth measures in its teacher evaluations, and if so, the weight it assigns to student growth measures.

In late 2015, there was speculation that states would stop using measures of students' academic growth to evaluate teachers due to the passage of the Every Student Succeeds Act (ESSA), as well as an increase in anti-testing sentiment across the country. The ESSA does not require states to establish teacher evaluation systems based on students' test scores – a key requirement of the U.S. Department of Education's state waiver system in connection with ESSA's predecessor, the No Child Left Behind Act (NCTQ, 2017; Sawchuk, 2016a). But the NCTQ (2017) reported that by January 2017, there had been little change in states' formal teacher evaluation systems. Only four states – Alaska, Mississippi, North Carolina, and Oklahoma – reversed their decision to include student learning in teachers' evaluation ratings.

The push by states to include student learning growth measures in teacher evaluations has led to the filing of lawsuits in both federal and state courts across the country, including states such as Florida, Louisiana, Nevada, New Mexico, New York, Tennessee, and Texas (Association of Texas Professional Educators, 2016; Barnum, 2016a; Collier, 2016; Sawchuk, 2016b; Schwennesen, 2016; Texas American Federation of Teachers, 2016; Education Week, 2015; Nashville Public Radio, 2015; Quintano, 2015; Sawchuk, 2014).

Although the majority of judges have upheld states' rights to use VAMs in teacher evaluations, a New York state court recently sided with a teacher who challenged the student growth component of her teacher evaluation. New York's teacher evaluation system bases 50% of teachers' evaluations on VAM scores. The teacher filed a lawsuit after receiving the lowest possible score on the student growth portion of her evaluation, one year after receiving a high score. In May 2016, the judge ruled that the teacher's VAM score was "arbitrary" and "capricious." According to the judge, the state failed to explain how the teacher's score could vary so widely from one year to the next. After the court's ruling, the state suspended the use of *state* test scores in teacher evaluations for four years. However, there is no moratorium on the use of other test scores, so other measures of student learning – mostly local assessments – will continue to be a critical component of teachers' evaluations (Barnum, 2016b; Kugler, 2016; Sawchuk, 2016c; Strauss, 2016). The New York State Allies for Public Education, a pro opt-out group (cited in Barnum, 2016b), stated that the moratorium will actually lead to increased testing, since new exams will be needed to judge teachers who were previously evaluated by state tests (which will still be given).

Definition of Value-Added Models

Value-added analysis was designed to estimate teachers' contributions to student learning by tracking students' progress on standardized tests from year to year. Each student's performance is compared with his or her own performance in prior year(s). VAMs do not focus on how students test at a single point in time, but rather on how much improvement they make from one year to the next. The approach is designed to estimate teachers' effects on student learning while statistically controlling for other factors that affect achievement, such as prior learning, income level, and parental support (American Educational Research Association, 2015; American Statistical Association, 2014; RAND Corporation, 2012; Song & Felch, 2011).

The VAMs used by states and school districts vary based on the variables they include and the weight assigned to each variable, but all models attempt to isolate the impact a teacher has on

student growth. For example, VAMs may control for students' prior achievement, gender, ethnicity, English language or special education status, past educational experiences, attendance, suspensions, income level, and family environment. School-level variables, such as class size, percentage of students eligible for free or reduced price lunch, and expenditure per student, may also be included in VAMs. The value-added score represents the teacher effect not explained by the other variables in the model (Amrein-Beardsley et al., 2016; Pivovarova et al., 2016; American Educational Research Association, 2015; American Statistical Association, 2014; Hull, 2013; Lomax & Kuenzi, 2012).

As the Reform Support Network (2013) explained, "VAMs aim to predict what student growth can be expected from an average or typical teacher, and then compare actual student achievement with that prediction. A teacher's value-added score is intended to convey how much individual teachers contribute to student learning in a particular subject in a particular year. Teachers who produce more than this typical teacher are thought to have added value. Teachers whose effects on students result in less growth than the typical teacher is expected to yield are considered less effective."

Research on the Accuracy of VAM Results

There is considerable disagreement among educational policymakers about whether to include VAMs in teacher evaluation systems (Amrein-Beardsley et al., 2016; American Educational Research Association, 2015; Ewing, 2011; Goldhaber & Hansen, 2010). According to Powell (2015), studies from credible sources "make it plainly clear that the question of whether VAMs can accurately and reliably help us identify effective teachers is very much an open one."

Many researchers have concluded that VAMs are not appropriate as the primary measure for evaluating individual teachers. They report a weak to nonexistent relationship between teachers' VAM scores and the content or quality of classroom instruction. In fact, studies have found that VAMs often misclassify up to 25% of teachers (Strauss, 2016; Powell, 2015; Pennsylvania State Education Association, 2014; Polikoff & Porter, 2014; Horn & Wilburn, 2013; Hull, 2013; Amrein-Beardsley & Collins, 2012; Darling-Hammond et al., 2012; Lomax & Kuenzi, 2012; American Educational Research Association & National Academy of Education, 2011; Ewing, 2011; Baker et al., 2010; Rothstein, 2008).

On the other hand, proponents of VAMs maintain that they provide important and useful information on the effects teachers have on their students' outcomes. They contend that teacher evaluation systems based solely on classroom observations do little to help teachers improve or to support personnel decision-making. VAM advocates acknowledge that neither method is able to discriminate between effective and ineffective teachers with 100% accuracy, but they argue that VAMs are as reliable as classroom observations, and more objective (Barnum, 2016a; Harris & Herrington, 2015; Kane et al., 2013; Raudenbush & Jean, 2012; Sparks, 2012; Baker et al., 2010; Glazerman et al., 2010; Goldhaber & Hansen, 2010).

The American Educational Research Association (2015) called for "substantial investment in research on VAM and alternative methods and models." In the meantime, researchers continue to advise states and school districts that there are a number of difficulties surrounding these measures when they are used to make high-stakes personnel decisions and that results should be interpreted with caution. These difficulties are summarized later in this report.

More research is needed on the accuracy of student growth measures and the consequences of using them in teacher evaluation systems. Although teacher evaluations based at least in part

on student growth demonstrate a wider range of teacher performance relative to previous evaluation systems that relied solely on teacher observations, evidence of the reliability and validity of student growth measures is limited (McCullough et al., 2015; American Educational Research Association & National Academy of Education, 2011).

Difficulties with Value-Added Models

The American Educational Research Association's (2015) "Statement on the Use of Value-Added Models for the Evaluation of Educators and Educator Preparation Programs" discussed the "scientific and technical" limitations of VAMs. The statement listed specific psychometric problems with VAM, addressed the validity of inferences from VAM (given the challenges of isolating teachers' contributions to student learning), and cautioned against VAM results having "a high-stakes, dispositive weight in evaluations."

The American Statistical Association's (2014) "Statement on Using Value-Added Models for Educational Assessment" recommended, "Estimates from VAMs should always be accompanied by measures of precision and a discussion of the assumptions and possible limitations of the model. These limitations are particularly relevant if VAMs are used for high-stakes purposes." The American Statistical Association also stated, "Ranking teachers by their VAM scores can have unintended consequences that reduce quality."

Researchers have documented a number of problems with using VAMs as measures of teachers' effectiveness. These problems are summarized below.

- **Effects attributed to a teacher may actually be caused by other factors that are not included in the VAM.** The validity of VAM results depends on their ability to isolate the contributions of teachers to student learning from the contributions of other factors. Student achievement and learning gains are affected by a long list of influences both inside and outside of school – the influence of previous teachers, the quality of curriculum materials and other resources, class size, attendance, availability of instructional specialists and tutors, the attitudes of peers, parental support, food and housing security, summer learning loss, student health, and others. Given the large number of potential influences on student learning growth, it is highly unlikely that all of the relevant student, classroom, school, and community characteristics will be included in any value-added model (Pivovarova et al., 2016; American Educational Research Association, 2015; Darling-Hammond, 2015; American Statistical Association, 2014; Pennsylvania State Education Association, 2014; Haertel, 2013; Amrein-Beardsley & Collins, 2012; Ballou et al., 2012; Ewing, 2011; Baker et al., 2010). The American Educational Research Association & National Academy of Education (2011) stated, "Value-added ratings cannot disentangle the many influences on student progress."

  Studies have found that the teacher-explained proportion of variability in standardized test scores ranges from 1% to 14%. In other words, between 86% and 99% of the variance in students' test scores can be explained by factors other than the classroom teacher, such as student, school, family, and other unmeasured influences. This does not mean that teachers have a small effect on students' academic growth, but that *variation* among teachers accounts for only a small part of the variation in test scores (Pivovarova et al., 2016; American Statistical Association, 2014).

- **VAM results are biased when students are not randomly assigned to classroom teachers.** Researchers have concluded that nonrandom assignment of students to teachers increases VAM score bias. Educators acknowledge that assignment of students to classrooms is generally nonrandom, and is frequently the result of deliberate choices by principals or parents to pair specific types of students with specific teachers. In addition, certain students, such as English language learners and students with special educational needs, are rarely randomly assigned to teachers. Researchers have concluded that the value-added models are unable to fully adjust for these nonrandom assignments (Amrein-Beardsley et al., 2016; Pivovarova et al., 2016; American Educational Research Association, 2015; Pennsylvania State Education Association, 2014; Texas Classroom Teachers Association, 2014; Darling-Hammond et al., 2012; Sparks, 2012; Baker et al., 2010; David, 2010; Braun, 2005).

- **Teachers' VAM scores depend on the students in their classrooms.** Despite statistical controls put in place to prevent bias, studies suggest that teachers' VAM scores differ depending on the characteristics of the students assigned to their classrooms. Studies have found that teachers with higher numbers of students who are difficult to teach (such as students with poor attendance, students with high rates of mobility and severe difficulties at home, and those who are English language learners or have special educational needs) receive lower VAM scores than those teaching less disadvantaged students. In addition, teachers of high-achieving and gifted students receive lower VAM scores because their students are already near the top of the test score range and are therefore unable to demonstrate much academic growth (Amrein-Beardsley et al., 2016; Pivovarova et al., 2016; Pennsylvania State Education Association, 2014; Texas Classroom Teachers Association, 2014; Darling-Hammond et al., 2012; Sparks, 2012; American Educational Research Association & National Academy of Education, 2011; Baker et al., 2010; David, 2010; Braun, 2005).

- **VAM scores are impacted by peer effects**. Researchers have concluded that VAMs do not isolate the effects that classroom peers have on students' academic growth. They have identified two types of peer effects that have the greatest potential to impact teachers' VAM scores:

    - The members of the class collectively influence the teacher's pacing of instruction. The average achievement level in the class as a whole affects the amount of content delivered to all of the students over the course of the school year. When classrooms are grouped by achievement level, teachers of low-performing students receive lower VAM scores because they cannot deliver instruction as quickly.

    - A second kind of peer effect occurs when some students in the classroom are disruptive and slow down the pace of instruction. In these cases, teachers receive lower VAM scores because problematic students were assigned to their

classrooms, not because their teaching was ineffective (Haertel, 2013; Raudenbush, 2013; Amrein-Beardsley & Collins, 2012; RAND Corporation, 2012; Ewing, 2011).

- **Teachers' VAM scores vary depending on the school where they work.** Studies have found that teachers' VAM scores are strongly influenced by school-level variables, such as the physical condition of the school, the resources available to the teacher, the learning culture created by the school, the strength of the principal's leadership, and the school's commitment to communication and collaboration. Teachers assigned to more effective schools have been found to have higher VAM scores than teachers assigned to less effective schools (Raudenbush, 2013; Braun, 2005). Haertel (2013) stated that no statistical manipulation can ensure fair comparisons of teachers working in very different types of schools.

- **It is difficult to isolate the impact of a single teacher on students' academic growth.** Experts have noted that teams of teachers, social workers, guidance counselors, media specialists, and others work together. Classroom teachers also build on the efforts of previous teachers. For example, if a student has significant academic growth in grade 5, it might largely be due to the effectiveness of his/her third and fourth grade teachers (FairTest, 2016; Ewing, 2011; Baker et al., 2010).

- **Students' standardized test scores are not accurate measures of teachers' true effectiveness.** Because VAM scores are based on standardized achievement tests, they are subject to certain limitations. For example:

    - Standardized tests are very narrow indicators of student achievement. They cover only a small selection of material from each content area. They also do not measure important attributes, such as creativity, initiative, persistence, curiosity, and self-discipline (American Statistical Association, 2014; Pennsylvania State Education Association, 2014; Harris et al., 2012; Ewing, 2011).

    - Most states' standardized achievement tests measure only grade-level standards. They do not include items needed to measure growth for students who perform well above or well below grade level (American Educational Research Association, 2015; Darling-Hammond, 2015; Baker et al., 2010).

    - Studies have found that test scores can be increased without a corresponding increase in student learning. For example, providing strategies for test-taking has been found to improve students' test performance and narrowing the curriculum to match the test's content has been found to have an even greater effect on students' test scores (Ewing, 2011).

- **VAM scores depend on the specific achievement test used.** A number of studies have documented that teachers' VAM scores differ significantly when different standardized achievement tests are used, even when the tests are within the same

content area (Darling-Hammond, 2015; American Educational Research Association & National Academy of Education, 2011; David, 2010). As part of the Measures of Effective Teaching (MET) Project, teachers' VAM scores were calculated using both state achievement tests and project-administered tests in grades 4-8 in six school districts. The researchers reported that the correlation between VAM scores using the two different tests was weak – 0.38 for math and 0.21 for reading (cited in McCaffrey, 2013).

- **VAM scores vary substantially from year to year.** Many researchers believe that this annual variability reflects VAMs' unreliability, not changes in individual teacher's effectiveness from one year to the next (Amrein-Beardsley et al., 2016; Darling-Hammond, 2015; Pennsylvania State Education Association, 2014; Horn & Wilburn, 2013; American Educational Research Association & National Academy of Education, 2011; David, 2010). According to the Texas Classroom Teachers Association (2014), a teacher classified as "adding value" has a 25% to 50% chance of being classified as "subtracting value" the following year and vice versa. Examples of VAM score instability include:

    - A study of five large urban school districts found that among teachers who were ranked in the top 20% of effectiveness in the first year, fewer than one-third were in that top group the following year, and another third moved down to the bottom 40%. There was similar movement for teachers who received low rankings in the first year – among teachers who were ranked in the bottom 20% of effectiveness in the first year, fewer than a third were in that bottom group the next year, and another third moved up to the top 40% (cited in Baker et al., 2010).

    - Lomax and Kuenzi's (2012) review of the literature reported that when teachers are divided into quintiles based on their VAM scores, the rankings change over time. In general, only about one-quarter to one-third of teachers remain within the same quintile from one year to the next; approximately 10% to 15% of teachers move from the bottom quintile to the top, and an equal number move from the top quintile to the bottom.

- **VAM scores depend on the statistical model used.** Different VAM models account for external factors that impact student learning in very different ways. For example, some models control for preexisting differences in student characteristics, such as ethnicity, gender, English language proficiency, income level, and students' prior achievement. Others also control for school level variables, such as mobility, class size, and percentage of low-income students (Darling-Hammond, 2015; Ballou et al., 2012; Goldhaber & Theobald, 2012; Lomax & Kuenzi, 2012; American Educational Research Association & National Academy of Education, 2011). Data on approximately 11,500 VAM elementary teacher rankings published by the *Los Angeles Times* showed that only 46% of reading teachers - and 60% of mathematics teachers - remained in the same effectiveness category when two different statistical models were used to estimate VAM scores (cited in Pivovarova et al., 2016). Value-added model developers have not

reached a consensus regarding which model most accurately identifies effective and ineffective teachers (Pennsylvania State Education Association, 2014).

- **VAM scores are not highly correlated with other measures of teacher effectiveness.** Research suggests that there are only weak relationships between VAM scores and other measures of teacher effectiveness, such as supervisors' observational assessments, administrators' opinions, and students' survey-based assessments. For example, studies have found low correlations (all below 0.40) between VAM scores and classroom observation scores, even when highly trained and monitored classroom observers are used. Low correlations between VAM scores and other teacher effectiveness measures indicate that teachers who are ranked highly on one measure have a good chance of receiving a low ranking on another measure, and vice versa (Amrein-Beardsley et al., 2016; Pivovarova et al., 2016; Pennsylvania State Education Association, 2014; Polikoff & Porter, 2014; Rothstein & Mathis, 2013; Darling-Hammond, 2012; American Educational Research Association & National Academy of Education, 2011).

- **VAMs are difficult for non-statisticians to understand.** Because VAMs are complex statistical models that use multiple years of data, most teachers do not understand how their performance is measured. Teachers are unlikely to trust or accept their VAM scores when they don't understand how they are calculated (Amrein-Beardsley et al., 2016; Barnum, 2016a; Jennings & Pallas, 2016; Texas Classroom Teachers Association, 2014; Hull, 2013; Reform Support Network, 2013). Amrein-Beardsley and colleagues (2016) also noted that the public has not been adequately educated on how to interpret VAM scores. The American Statistical Association (2014) concluded, "Perceptions of transparency, fairness and credibility will be crucial in determining the degree of success of the system as a whole in achieving its goals of improving the quality of teaching."

- **VAM scores do not lead to instructional improvements.** Researchers have noted that VAM scores do not provide teachers with any information they can use to improve specific aspects of their instructional practice (American Educational Research Association, 2015; Pennsylvania State Education Association, 2014; Texas Classroom Teachers Association, 2014; Amrein-Beardsley & Collins, 2012). Jennings and Pallas' (2016) interviews with 13 New York City teachers found that no teachers reported that their VAM scores would help them improve their instructional practice. Because teachers felt they did not have control over their VAM scores, they said they did not know what they could do differently to improve them. Lomax & Kuenzi (2012) stated, "The teacher effect . . . cannot determine *why* a teacher is effective or ineffective, nor does it provide any information on the specific characteristics of what makes a teacher effective."

- **Teachers have negative perceptions of VAMs.** Studies have found that teachers report a strong distrust of VAM scores. For example:

    - Jennings and Pallas' (2016) interviews with New York City teachers revealed that teachers believed the tests upon which their VAM scores were based lacked

legitimacy as measures of students' and teachers' performance. Furthermore, teachers reported that VAM scores seemed entirely out of their control.

- Another study examined Tennessee teachers who volunteered to be evaluated based on VAMs and to have a substantial share of their compensation tied to their VAM results. After three years, 85% of the teachers said that the VAM evaluation ignored important aspects of their performance that test scores did not measure, and two-thirds thought VAMs did not do a good job of distinguishing effective from ineffective teachers (cited in Darling-Hammond et al., 2012).

- A survey of nearly 3,000 teachers in 48 states conducted by the Network for Public Education (2016) found that 83% of respondents said that the inclusion of students' standardized test scores in teacher evaluations negatively affected classroom instruction.

- **VAMs have unintended consequences on students and teachers.** Research in this area is just beginning to accumulate, leading experts to recognize that the decision to include VAMs in teacher evaluation systems may have unintended consequences, including:

  - The curriculum becomes narrower when teachers spend more time on test preparation. Some teachers also focus on content that is included in the test and exclude other content that may lead to more long-term learning gains (American Statistical Association, 2014; Haertel, 2013; Horn & Wilburn, 2013; Chetty et al., 2011; Baker et al., 2010). The Network for Public Education's (2016) nationwide survey found that 88% of respondents said that more time is spent on test preparation than ever before.

  - High-needs schools become harder to staff because it is difficult for teachers to receive high VAM scores when they work with low-achieving or disadvantaged students (Darling-Hammond, 2015; American Statistical Association, 2014; Baker et al., 2010).

  - The classroom roster gains exaggerated importance. Some students with "growth" potential are seen as more beneficial to teach, while others are less desirable due to criteria that limit growth, such as learning disabilities or limited English proficiency (Network for Public Education, 2016; Haertel, 2013; Rothstein, 2008).

  - Reliance on VAM scores fosters a competitive environment that discourages teacher collaboration (Amrein-Beardsley et al., 2016; American Statistical Association, 2014; Haertel, 2013; Baker et al., 2010). The Network for Public Education's (2016) survey found that 61% of respondents noted that the use of student standardized test scores in teacher evaluations had a negative impact on their relationships with their colleagues, citing reasons such as forced

collaboration and competition. Seventy-two percent of respondents said they were less likely to share instructional strategies with their colleagues.

Recommendations for Calculating, Interpreting, and Using VAM Scores

Researchers have urged caution when including VAM scores as a component of teacher evaluation systems (American Educational Research Association, 2015; Ewing, 2011; Baker et al., 2010). They have made the following recommendations:

- **Use VAM scores as only one component of a comprehensive teacher evaluation system.** A consensus has emerged that teacher evaluation systems should be comprised of multiple indicators of effective teaching practices, as well as a variety of student outcomes. There is broad agreement among statisticians and psychometricians that student test scores alone are not sufficiently reliable and valid indicators of teaching effectiveness to be used as the sole basis for making high-stakes personnel decisions, such as compensation, tenure, promotion, and dismissal (Amrein-Beardsley et al., 2016; American Educational Research Association, 2015; American Statistical Association, 2014; Baker et al., 2010; David, 2010; Glazerman et al., 2010; Rothstein, 2008; Braun, 2005).

- **Limit comparisons to similar groups of teachers.** VAM rankings that mix teachers from different grade levels or those who teach in schools with different demographics place heavy demands on the statistical model's assumptions. To reduce the amount of error in VAM estimates, researchers suggest that comparisons be limited to teachers in a single subject area and grade level within an individual school district. Furthermore, VAM analyses have been found to be more accurate when they are based on the test scores of students who are from similar backgrounds and have comparable prior skill levels (Haertel, 2013; Raudenbush, 2013; Darling-Hammond et al., 2012; American Educational Research Association & National Academy of Education, 2011).

- **Use multiple years of data when calculating VAM scores.** Studies have found that VAM estimates more reliably predict teacher effectiveness when they are based on multiple years of student test scores (American Educational Research Association, 2015; Haertel, 2013; Goldhaber & Hansen, 2010). Amrein-Beardsley and colleagues (2016) reported that accurate VAM estimates are based on at least three years of data. According to the American Statistical Association (2014), "The VAM scores themselves have large standard errors, even when calculated using several years of data. These large standard errors make rankings unstable, even under the best scenarios for modeling. Combining VAMs across multiple years decreases the standard error of VAM scores."

- **Only calculate VAM scores in grades and subjects where there are highly reliable and valid assessments that are comparable over time.** Experts recommend that because the validity of VAM scores is so dependent on the quality of the tests administered to students, VAM scores should only be calculated in grade levels and

subjects where there are valid and reliable assessments. States do not administer achievement tests at all grade levels (K-12) and in all subjects (for example, social studies, health, and art). In order to calculate VAM scores in untested subjects and grades, many districts develop their own alternative assessments. Researchers caution that locally administered tests should only be used to calculate VAM scores when they are accompanied by evidence of high reliability, validity, precision, and fairness (Amrein-Beardsley et al., 2016; American Educational Research Association, 2015).

While researchers have found that locally developed tests are not always well-constructed, they acknowledge that it is preferable to try to develop valid and reliable assessments for untested subjects and grades than to evaluate teachers based on other teachers' VAM scores. Most researchers agree that teacher evaluations lack validity when they are based on student achievement in courses in which the teacher had little or no involvement or impact (Green & Oluwole, 2015; Pennsylvania State Education Association, 2014; Reform Support Network, 2013).

Amrein-Beardsley and colleagues (2016) reported that about 70% of all public school teachers are valued-added ineligible. Haertel (2013) stated, "One of the most troubling aspects of some current reform proposals is the insistence on *universal* application of value-added to all teachers in a district or state. For most teachers, appropriate test data are not available, period. They teach children so young that there are no prior year scores, or they teach untested subjects, or they teach high school courses for which there are no pretest scores that it makes any sense to use."

- **Use VAM results to research groups of teachers, not individual teachers.** Researchers have stated that while educators should not use VAM results to make high-stakes decisions about individual teachers, VAM results are useful for looking at groups of teachers for research purposes; for example, to examine how specific teaching practices influence the learning of large numbers of students or to investigate the effects of teacher training approaches or educational policies. The larger scale of these studies reduces error and their use of a greater number of outcome measures allows more understanding of the effects of specific strategies and interventions (Doherty & Jacobs, 2015; Haertel, 2013; Darling-Hammond et al., 2012; American Education Research Association & National Academy of Education, 2011).

Summary

Value-added analysis was designed to estimate teachers' contributions to student learning by tracking students' progress on standardized tests from year to year, while statistically controlling for other factors that affect achievement, such as prior learning, income level, and parental support. The majority of states include student learning growth measures as a component of their teacher evaluation systems and use teachers' evaluation ratings to make decisions regarding compensation, promotion, tenure, and dismissal. According to the National Council on Teacher Quality, 40 states require the use of student achievement growth measures as a component of teacher evaluations. Eleven states have no formal policy requiring that teacher evaluations take student achievement into account.

Studies on the accuracy of VAM scores, and the consequences of their use in teacher

evaluation systems, are still accumulating. However, evidence has begun to emerge that teachers' VAM scores may depend more on the students they teach and the schools where they work than on the effectiveness of their teaching.

This Information Capsule summarized the specific problems researchers have documented with using VAM scores to evaluate teachers. For example, teachers' VAM scores depend on the students they teach, the school where they work, the achievement test used, and the statistical model used in the calculations; VAM scores vary substantially from year to year; VAM scores are not highly correlated with other measures of teacher effectiveness; and it is difficult to isolate the impact of a single teacher on students' academic growth.

Researchers have urged caution when including VAM scores in teacher evaluation systems and have offered several recommendations, such as using VAM scores as only one component of a comprehensive teacher evaluation system, using multiple years of data when calculating VAM scores, and calculating VAM scores only in grades and subjects where there are highly reliable and valid assessments that are comparable over time.

References

American Educational Research Association. (2015). *AERA Statement on Use of Value-Added Models (VAM) for the Evaluation of Educators and Educator Preparation Programs.* Retrieved from http://journals.sagepub.com/doi/pdf/10.3102/0013189X15618385.

American Educational Research Association & National Academy of Education. (2011). *Getting Teacher Evaluation Right: A Brief for Policymakers.* Retrieved from https://edpolicy.stanford.edu/sites/default/files/publications/getting-teacher-evaluation-right-challenge-policy-makers.pdf.

American Statistical Association. (2014). *ASA Statement on Using Value-Added Models for Educational Assessment.* Retrieved from http://www.amstat.org/asa/files/pdfs/POL-ASAVAM-Statement.pdf.

Amrein-Beardsley, A., & Collins, C. (2012). The SAS Education Value-Added Assessment System (SAS EVAAS) in the Houston Independent School District (HISD): Intended and Unintended Consequences. *Education Policy Analysis Archives, 20*(12). Retrieved from http://epaa.asu.edu/ojs/article/view/1096/982.

Amrein-Beardsley, A., Pivovarova, M., & Geiger, T.J. (2016). Value-Added Models: What the Experts Say. *Phi Delta Kappan, 98*(2), 35-40.

Anderson, L.M., Butler, A., Palmiter, A., & Arcaira, E. (2016). *Study of Emerging Evaluation Systems.* Policy and Program Studies Service, Office of Planning, Evaluation, and Policy Development, U.S. Department of Education, Washington, DC. Retrieved from http://www2.ed.gov/rschstat/eval/teaching/emerging-teacher-evaluation/report.pdf.

Association of Texas Professional Educators. (2016). *ATPE Files T-TESS Legal Challenge, Asserts That New Evaluation Rules Violate State Law.* Retrieved from http://www.teachthevote.org/news/2016/04/25/atpe-files-t-tess-legal-challenge-asserts-that-new-evaluation-rules-violate-state-law/.

Baker, E.L., Barton, P.E., Darling-Hammond, L., Haertel, E., Ladd, H.F., Linn, R.L., et al. (2010). *Problems with the Use of Student Test Scores to Evaluate Teachers.* Economic Policy Institute,

Washington, DC. ERIC Document Reproduction Service No. ED516803.

Ballou, D., Mokher, C.G., & Cavalluzzo, L. (2012). *Using Value-Added Assessment for Personnel Decisions: How Omitted Variables and Model Specification Influence Teachers' Outcomes.* Retrieved from https://aefpweb.org/sites/default/files/webform/AEFP-Using%20VAM%20for%20personnel%20decisions_02-29-12.pdf.

Barnum, M. (2016a). NY Teacher Wins Court Case Against State's Evaluation System, But May Appeal to Set Wider Precedent. *The 74,* May 11, 2016. Retrieved from https://www.the74million.org/article/ny-teacher-wins-court-case-against-statesevaluationsystem-but-she-may-appeal-to-set-wider-precedent.

Barnum, M. (2016b). No, Gov. Cuomo Didn't Kill Test-Based Teacher Evals. In Fact, His 'Moratorium' Could Fuel More Testing. *The 74,* January 7, 2016. Retrieved from https://www.the74million.org/article/no-gov-cuomo-didnt-kill-test-based-teacher-evals-in-fact-his-moratorium-could-fuel-more-testing.

Braun, H.I. (2005). *Using Student Progress to Evaluate Teachers: A Primer on Value-Added Models.* Educational Testing Service, Princeton, NJ. Retrieved from https://www.ets.org/Media/Research/pdf/PICVAM.pdf.

Chetty, R., Friedman, J.N., & Rockoff, J.E. (2011). *The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood.* National Bureau of Economic Research, Working Paper 17699. Retrieved from http://www.nber.org/papers/w17699.pdf.

Collier, K. (2016). Teacher Group Sues Texas Over Plan to Tie Evaluations to Tests. *The Texas Tribune,* April 21, 2016. Retrieved from https://www.texastribune.org/2016/04/20/teacher-group-sues-morath-over-educator-evaluation/.

Darling-Hammond, L. (2015). Can Value Added Add Value to Teacher Evaluation? *Educational Researcher, 44*(2), 132-137.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating Teacher Evaluation. *Phi Delta Kappan, 3*(6), 8-15.

David, J.L. (2010). What Research Says About . . . Using Value-Added Measures to Evaluate Teachers. *Educational Leadership, 67*(8), 81-82.

Doherty, K.M., & Jacobs, S. (2015). *2015 State of the States: Evaluating Teaching, Leading and Learning.* National Council on Teacher Quality, Washington, DC. Retrieved from http://www.nctq.org/dmsView/StateofStates2015.

Education Week. (2015). *Teacher Evaluation Heads to the Courts.* Retrieved from http://www.edweek.org/ew/section/multimedia/teacher-evaluation-heads-to-the-courts.html.

Ewing, J. (2011). Mathematical Intimidation: Driven by Data. *Notices of the American Mathematical Society, 58*(5), 667-673.

FairTest. (2016). *Teacher Evaluation Should Not Rest on Student Test Scores.* Retrieved from http://www.fairtest.org/teacher-evaluation-fact-sheet.

Gitomer, D.H., Crouse, K., & Joyce, J. (2014). *A Review of the DC IMPACT Teacher Evaluation System.* Retrieved from http://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_160397.pdf.

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G.J. (2010). *Evaluating Teachers: The Important Role of Value-Added.* Brown Center on Education Policy at Brookings, Washington, DC. Retrieved from https://www.brookings.edu/research/evaluating-teachers-the-important-role-of-value-added/.

Goldhaber, D., & Hansen, M. (2010). *Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions.* National Center for Analysis of Longitudinal Data in Education Research (CALDER) Working Paper 31, Washington, DC. ERIC Document Reproduction Service No. ED509689.

Goldhaber, D., & Theobald, R. (2012). Do Different Value-Added Models Tell Us the Same Things? *Carnegie Knowledge Network Knowledge Brief,* October 15, 2012. Retrieved from http://www.carnegieknowledgenetwork.org/briefs/value-added/different-growth-models/.

Green, P.C.-III, & Oluwole, J. (2015). Response to 'Applying Value-Added Methods to Teachers in Untested Grades and Subjects.' *Journal of Law and Education, 44*(1), 401-412.

Guarino, C.M., Reckase, M.D., Stacy, B.W., & Wooldridge, J.M. (2014). *A Comparison of Growth Percentile and Value-Added Models of Teacher Performance.* The Education Policy Center at Michigan State University, East Lansing, MI. Retrieved from http://education.msu.edu/epc/publications/documents/WP39AComparisonofGrowthPercentileandValue-Added Model.pdf.

Haertel, E.H. (2013). *Reliability and Validity of Inferences About Teachers Based on Student Test Scores.* Paper presented at the 14th William H. Angoff Memorial Lecture, National Press Club, Washington, DC, March 2013. Retrieved from https://www.ets.org/Media/Research/pdf/PICANG14.pdf.

Harris, D.N., & Herrington, C.D. (2015). Editors' Introduction: The Use of Teacher Value-Added Measures in Schools: New Evidence, Unanswered Questions, and Future Prospects. *Educational Researcher, 44*(2), 71-76.

Harris, P., Harris, J., & Smith, B.M. (2012). Standardized Tests Do Not Effectively Measure Student Achievement. In D. Bryfonski (Ed.), *Standardized Testing.* Detroit, MI: Greenhaven Press.

Horn, J., & Wilburn, D. (2013). *The Mismeasure of Education.* Summarized in Schools Matter, January 2, 2016. Retrieved from http://www.schoolsmatter.info/2016/01/value-added-modeling-goes-to-court.html.

Hull, J. (2013). *Trends in Teacher Evaluation: How States are Measuring Teacher Performance.* Center for Public Education, Alexandria, VA. Retrieved from http://www.centerforpubliceducation.org/Main-Menu/Evaluating-performance/Trends-inTeacher-Evaluation-At-A-Glance/Trends-in-Teacher-Evaluation-Full-Report-PDF.pdf.

Jennings, J.L., & Pallas, A.M. (2016). How Does Value-Added Data Affect Teachers? *Educational Leadership, 73*(8). Retrieved from http://www.ascd.org.

Kane, T.J., McCaffrey, D.F., Miller, T., & Staiger, D.O. (2013). *Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment.* MET Project Research Paper, Bill & Melinda Gates Foundation, Seattle, WA. ERIC Document Reproduction Service No. ED540959.

Kugler, J. (2016). New York Court Decision Throws Out Use of 'VAM' for Teacher Evaluations . . . New York's Higher Court Notes that the VAM System is Bogus. *Substance News,* May 12, 2016. Retrieved from http://www.substancenews.net/articles.php?page=6331.

Lomax, E.D., & Kuenzi, J.J. (2012). *Value-Added Modeling for Teacher Effectiveness.* Congressional Research Service Report for Congress, Washington, DC. Retrieved from https://fas.org/sgp/crs/misc/R41051.pdf.

McCaffrey, D.F. (2013). *Will Teacher Value-Added Scores Change When Accountability Tests Change?* Carnegie Foundation for the Advancement of Teaching, Stanford, CA. Retrieved from http://www.carnegieknowledgenetwork.org/briefs/value-added/accountability-tests/.

McCullough, M., English, B., Angus, M.H., & Gill, B. (2015). *Alternative Student Growth Measures for Teacher Evaluation: Implementation Experiences of Early-Adopting Districts.* National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, Washington, DC. Retrieved from https://ies.ed.gov/ncee/edlabs/regions/midatlantic/pdf/REL_2015093.pdf.

Nashville Public Radio. (2015). *Tennessee Teachers File Federal Lawsuit Against Value-Added Evaluations.* Retrieved from http://www.schoolsmatter.info/2015/02/tennessee-teachers-file-federal-lawsuit.html.

Network for Public Education. (2016). *Teachers Talk Back: Educators on the Impact of Teacher Evaluation.* Retrieved from http://networkforpubliceducation.org/wp-content/uploads/2016/04/NPETeacherEvalReport.pdf.

Pennsylvania State Education Association. (2014). *Using Value-Added or Growth Measures for Teacher Evaluation: What are the Issues?* Retrieved from https://www.psea.org.

Pivovarova, M., Amrein-Beardsley, A., & Broatch, J. (2016). Value-Added Models (VAMs): Caveat Emptor. *Statistics and Public Policy, 3*(1), 1-9.

Polikoff, M.S., & Porter, A.C. (2014). Instructional Alignment as a Measure of Teaching Quality. *Educational Evaluation and Policy Analysis, 36*(4), 399-416.

Powell, D. (2015). What Does the Research Say About Value-Added Models and Teacher Evaluation? *Education Week,* March 11, 2015. Retrieved from http://blogs.edweek.org/edweek/K-12_Contrarian/2015/03/vam_research.html.

Quintano, C. (2015). Judge Suspends Penalties Linked to State's Teacher Evaluation System. *Santa Fe New Mexican,* December 2, 2015. Retrieved from http://www.santafenewmexican.com

RAND Corporation. (2012). *Value-Added Modeling 101: Using Student Test Scores to Help Measure Teaching Effectiveness.* Retrieved from http://www.rand.org/education/projects/measuring-teacher-effectiveness/value-added-modeling.html.

Raudenbush, S.W. (2013). *What Do We Know About Using Value-Added to Compare Teachers Who Work in Different Schools?* Carnegie Foundation for the Advancement of Teaching, Stanford, CA. Retrieved from http://www.carnegieknowledgenetwork.org/wp-content/uploads/2013/08/CKN_Raudenbush-Comparing-Teachers_FINAL_08-19-13.pdf.

Raudenbush, S.W., & Jean, M. (2012). *How Should Educators Interpret Value-Added Scores?* Carnegie Foundation for the Advancement of Teaching, Stanford, CA. Retrieved from http://www.carnegieknowledgenetwork.org/briefs/value-added/interpreting-value-added/.

Reform Support Network. (2013). *Measuring Teaching Matters: What Different Ways of Looking at Student Results Tell Us About Teacher Effectiveness.* U.S. Department of Education, Washington, DC. Retrieved from https://www2.ed.gov/about/inits/ed/implementation-support-unit/tech-assist/measuring-teaching-matters.pdf.

Rothstein, J. (2008). *Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement.* National Bureau of Economic Research, Working Paper 14442. Retrieved from http://www.nber.org/papers/w14442.pdf.

Rothstein, J., & Mathis, W.J. (2013). *Review of Two Culminating Reports from the MET Project.* National Education Policy Center, Boulder, CO. Retrieved from http://nepc.colorado.edu/thinktank/review-MET-final-2013.

Ryshke, R. (2013). *Value-Added (VAM) Teacher Evaluations: Will They Be a Game-Changer?* The Westminster Schools, Atlanta, GA. Retrieved from https://rryshke.blog/2013/03/03/value-added-vam-teacher-evaluations-will-they-be-a-game-changer/.

Sawchuk, S. (2014). Florida Judge: Teacher-Evaluation System Unfair, But Legal. *Education Week,* May 7, 2014. Retrieved from http://blogs.edweek.org/edweek/teacherbeat/2014/05/florida_judge_teacher-evaluati.html.

Sawchuk, S. (2016a). ESSA Loosens Reins on Teacher Evaluations, Qualifications. *Education Week, 35*(15), 14-15.

Sawchuk, S. (2016b). Teacher-Evaluation Lawsuit in Tennessee Nixed by Federal Judge. *Education Week*, February 22, 2016. Retrieved from http://blogs.edweek.org/edweek/teacherbeat/2016/02/teacher-evaluation_lawsuit_in_.html.

Sawchuk, S. (2016c). Court Sides With Teacher in N.Y. Evaluation-Rating Lawsuit. *Education Week,* May 10, 2016. Retrieved from http://blogs.edweek.org/edweek/teacherbeat/2016/05/court_sides_with_teacher_in_ny.html?r=490806560.

Schwennesen, T. (2016). State Teachers Association Files Lawsuit to Block New Teacher Evaluation System. *Texas Public Radio,* April 20, 2016. Retrieved from http://tpr.org/post/state-teachers-association-files-lawsuit-block-new-teacher-evaluation-system#stream/0.

Song, J., & Felch, J. (2011). Times Updates and Expands Value-Added Ratings for Los Angeles Elementary School Teachers. *Los Angeles Times,* May 7, 2011.

Sparks, S.D. (2012). Caution Urged in Using 'Value Added' Evaluations: Scholars Say Districts Must Be More Careful. *Education Week, 32*(10), 6.

Strauss, V. (2016). Judge Calls Evaluation of N.Y. Teacher 'Arbitrary' and 'Capricious' in Case Against New U.S. Secretary of Education. *The Washington Post,* May 10, 2016.

Texas American Federation of Teachers. (2016). *Lawsuits Multiply Against Value-Added Teacher Evaluation in Texas.* Retrieved from http://www.texasaft.org/lawsuits-multiply-against-value-added-teacher-evaluation-in-texas/.

Texas Classroom Teachers Association. (2014). *Issues with Test-Based, Value-Added Models of Teacher Assessment.* Retrieved from https://tcta.org/sites/tcta.org/files/vam_final_final.pdf.

Walsh, E., & Isenberg, E. (2013). *How Does a Value-Added Model Compare to the Colorado Growth Model?* Mathematica Working Paper 22. Retrieved from https://www.mathematica-mpr.com.

Walsh, K., Joseph, N., Lakis, K., & Lubell, S. (2017). *Running in Place: How New Teacher Evaluations Fail to Live Up to Promises.* National Council on Teacher Quality, Washington, DC. Retrieved from http://www.nctq.org/dmsView/Final_Evaluation_Paper.

Appendix: Teacher Evaluation Policies on Student Growth, by State

State	Student Growth Component Required	Weight of Student Growth Component
Alabama	No	N/A
Alaska	No	N/A
Arizona	Yes	33-50%
Arkansas	Yes	Not Specified
California	No	N/A
Colorado	Yes	50%
Connecticut	Yes	45%; Evaluation system begins 2017-2018.
Delaware	Yes	20%
District of Columbia	Yes	Not Specified
Florida	Yes	One-third
Georgia	Yes	30%
Hawaii	Yes	50%
Idaho	Yes	33%
Illinois	Yes	30%
Indiana	Yes	Not Specified
Iowa	No	N/A
Kansas	Yes	Not Specified
Kentucky	Yes	Not Specified
Louisiana	Yes	50%
Maine	Yes	Not Specified
Maryland	Yes	50%
Massachusetts	Yes	Not Specified
Michigan	Yes	40%
Minnesota	Yes	35%
Mississippi	No	N/A
Missouri	Yes	Not Specified
Montana	No	N/A
Nebraska	No	N/A
Nevada	Yes	40%
New Hampshire	No	N/A
New Jersey	Yes	45%
New Mexico	Yes	25-50%
New York	Yes	50%; New and revised evaluation system to begin 2019-2020.
North Carolina	No	N/A
North Dakota	Yes	Not Specified
Ohio	Yes	35-50%
Oklahoma	No	N/A
Oregon	Yes	Not Specified
Pennsylvania	Yes	50%
Rhode Island	Yes	30%
South Carolina	Yes	20%
South Dakota	Yes	Not Specified
Tennessee	Yes	50%
Texas	Yes	20%; Evaluation system begins 2017-2018
Utah	Yes	20%
Vermont	No	N/A
Virginia	Yes	Not Specified
Washington	Yes	Not Specified
West Virginia	Yes	20%
Wisconsin	Yes	50%
Wyoming	Yes	Not Specified; Evaluation system begins 2019-2020

Source: Walsh et al., 2017 (National Council on Teacher Quality).