Carnegie Foundation for the Advancement of Teaching

Will Teacher Value-Added Scores Change when Accountability Tests Change?

DANIEL F. MCCAFFREY
EDUCATIONAL TESTING SERVICE

CARNEGIE KNOWLEDGE NETWORK
What We Know Series:
Value-Added Methods and Applications

ATTL ADVANCING TEACHING - IMPROVING LEARNING

HIGHLIGHTS

- There is only a moderate, and often weak, correlation between value-added calculations for the same teacher based on different tests.
- The content, timing, and structure of tests all contribute to differences in value-added calculations based on different tests.
- The stakes attached to a test affect the correlations between value-added estimates based on different tests.
- Conclusions drawn from one test might not serve as an accurate picture of a teacher's true effectiveness.
- More studies are needed to better assess the potential that "teaching to the test" has for distorting value-added estimates.
- Composite measures may mitigate some of the distortions caused by using different tests, and they may be better predictors of student learning gains.
- States should expect large changes in value-added calculations in 2014-15 when they switch to tests aligned with the Common Core State Standards.

INTRODUCTION

Value-added evaluations use student test scores to assess teacher effectiveness. How we judge student achievement can depend on which test we use to measure it.[1] Thus it is reasonable to ask whether a teacher's value-added score depends on which test is used to calculate it. Would it change if we used a different test? Specifically, might a teacher admonished for poor performance or recognized for good performance have been treated differently if a different test had been used? It's an important question, particularly because most states will soon be adopting new tests aligned with the Common Core State Standards.[2,3]

In this article we discuss what is known about how sensitive value-added scores are to the choice of test and what more needs to be known. We also discuss issues about the choice of test that might not be resolved through empirical investigation, as well as the implications of these findings for states and school districts.

WHAT IS KNOWN ABOUT TEACHER VALUE-ADDED ESTIMATES FROM DIFFERENT TESTS?

In this section, we discuss the research studies that compared value-added calculated with one test to value-added calculated for the same teachers using a different test. These studies found the correspondence between the two sets of value-added estimates to be moderate at best. (In the next section, we discuss possible reasons for these differences.)

The Measures of Effective Teaching (MET) Project calculated teacher value-added scores using state accountability tests and, separately, using project-administered tests in grades four through eight in six school districts.[4] The study used the correlation coefficient to describe the level of agreement between the two measures for each teacher. A value of 1 represents perfect correspondence, and a value of 0

means no correspondence. The MET Project found that the correlation between value-added using the two different tests administered to the same class of students was 0.38 for math and 0.21 for reading.[5] Associations in this range typically are considered weak: teachers with value-added in the top quartile on the state reading test would have about a 40 percent chance that their value-added on the alternative test would be below the 50th percentile.

Researchers also have taken advantage of the multiple tests administered by some states and school districts to investigate how much value-added changes when it is calculated with different tests. The studies used data from Hillsborough County, Florida; Houston, Texas; and a large urban district in the Northeast. In Hillsborough County, students completed two tests administered by the state: the Sunshine State Standards Test, a criterion-referenced test that assessed student mastery of Florida standards and served as the primary test for school accountability, and a norm-referenced test used to compare Florida students with those in other states.[6] In Houston, students completed the state accountability test and a standardized norm-referenced test administered by the district.[7] In the Northeast urban district, students completed the state test and two tests administered by the district: a standardized norm-referenced test in reading and math and a separate reading test.[8] Each of the studies presented the correlation between teachers' value-added based on the state accountability test and the alternative test. The studies found that correlations ranged from .20 to .59.[9] The highest correlation coefficients were for math and reading teachers in Houston – .59 for math and .50 for reading – where value-added was calculated by pooling up to eight years of data for a teacher. The smallest values, of around .20, were for reading teachers in the Northeastern city. That case compared value-added on the state test, which was administered in the spring, to value-added on a test administered by the district the following fall. It used only one year of data from both tests.

Taken together, this evidence suggests that a teacher who taught the same curriculum to the same students, and who is rated at a given level based on value-added calculated from one test has a strong likelihood of earning a different level based on value-added calculated from a different test.

Why is value-added sensitive to the test?

No standardized achievement test covers all the content and skills that a student might learn in a year. Consequently, value-added from one test might not fully reflect a teacher's effectiveness at promoting learning. It will not include content or skills not covered by the test. However, a teacher's effectiveness at teaching the content on one test may be very similar to her effectiveness at teaching the content on other tests. That is, a "good teacher" is a "good teacher" regardless of the content, skills or the test. Alternatively, some teachers may be more effective at teaching some content and skills and less effective at others. A teacher may be good at teaching mathematical computations but less good at teaching problem-solving. If value-added scores from different tests lead to different conclusions about a teacher, then we may worry that value-added from any single test provides an incomplete picture of a teacher's effectiveness, and that using it to make decisions about teachers may be inefficient or, for some teachers, unfair.

From the results of the studies mentioned above, we might at first conclude that value-added on one test is a poor measure of a teacher's effectiveness at teaching the content and skill measured by other tests. However, we consider six possible reasons for the weak correspondence between value-added calculated with two different tests: 1) the timing of the tests; 2) statistical imprecision; 3) test content; 4) the cognitive demands of the tests; 5) test format; and 6) the consequences of the test for students,

teachers, or schools. These other possible reasons have different implications for what value-added from one test might tell us about a teacher. We will discuss each in turn:

1. Test timing. We have seen that the lowest correlation between value-added calculated with two different tests was for tests administered at two different times in the school year, one in the fall and the other in the spring. Thus, value-added scores are sensitive to the timing of the tests, and any comparisons of value-added among teachers or for the same teachers across years should use tests given at the same time of the school year. Should states use fall-to-fall or spring-to-spring testing? There is no research on whether testing in either period leads to value-added scores that better reflect teachers' true effectiveness; however, spring tests do allow for calculating value-added closer to when the students were in their teachers' classes.
2. Statistical imprecision. All the studies except the Houston study calculated value-added for teachers using a single year of data and two different tests administered to the same group of students. Each measure is imprecise because it is calculated with a small number of students responding to a particular test form[10] on a particular day. The imprecision in each test contributes to disagreements in the value-added that is based on them. However, imprecision does not mean that value-added on one test is a poor measure of a teacher's effectiveness at teaching other content. If we remove the statistical noise that creates the imprecision, the agreement should be stronger. The Houston study used multiple years of data to calculate a teacher's value-added on each test. Each year the state test used a different test form, and each year the teacher had different students. Thus, by combining data across multiple years, the value-added in the Houston study reduced the imprecision. As a result, the correlation between tests in that study was higher than it was in other studies. The MET Project made adjustments to the correlation between tests to estimate what the correlation would be between the average of a teacher's value-added calculated for several years using the state test and the average of her value-added calculated for several years using alternative tests. The estimates were .54 for math and .37 for reading--again higher than the correlation for value-added from a single year. These correlations are still not strong. Value-added scores from one test might not provide the full picture of a teacher.
3. Test content. As discussed above, disagreement between value-added calculated with different tests might be due to differences in what is tested: A teacher may be differently effective at promoting achievement depending on the content measured. Two studies directly compared valued-added based on different content. In both of the studies, researchers calculated teacher value-added on the problem-solving subtest scores from a math assessment and, separately, on the procedures subtest scores on the same assessment, using the same students, tested under the same conditions, on the same day. In both studies, the agreement between the value-added using the two different subtests was modest, suggesting that a teacher who effectively promotes growth on problem-solving might not be equally effective at promoting growth on procedures.[11] These studies suggest that content does matter and that teachers who are effective at teaching the content of one test might not be equally effective with other content. This has implications for what is tested and the efficiency of decisions made using value-added.
4. Cognitive demands. Assessments vary in the cognitive demands made by the material on the test. A teacher may be differently effective at promoting growth in different types of skills, and this could result in differences in value-added from different tests. The research is unclear about how these differences in a test's cognitive demands contribute to differences in value-added. The alternative tests chosen by the MET Project had different cognitive demands than did the state tests, and agreement between the value-added from these two tests was moderate at

best. However, the tests also had some differences in content, so it is unclear how much the difference in cognitive demand contributed to the low correlation in value-added.[12]

5. Formats. Tests may have different formats. For instance, some tests may use multiple choice items, and others may use constructed response items. The different formats used by tests might contribute to difference in value-added. Some teachers may spend more time promoting the skills students need to succeed with constructed responses; other teachers may spend more time promoting the skills students need for multiple choice tests. The supplemental tests used in the MET Project were open-ended tests with constructed responses and no multiple-choice items, but the state tests contained mostly multiple choice items. Such differences could have weakened the agreement between the value-added calculated from different tests. Variation in value-added due to the test format does provide meaningful information about teachers. Tests used for value-added should include multiple formats, and any comparisons of teachers' value-added should be restricted to tests using similar formats.
6. Consequences. Tests can differ according to the consequences for students, teachers, and schools that are attached to their outcomes. Tests with consequences are called "high-stakes" tests; tests with limited or no consequences are called "low-stakes" tests. The Sunshine State Standards test in Florida and the state tests in the other studies were used to hold schools and districts accountable for performance; they faced penalties if their students scored poorly. The stakes of the other tests used in the studies were not as high. Teachers report focusing on test preparation for high-stakes tests, and there is concern that this focus can inflate scores.[13,14]Teachers who focus narrowly on the high-stakes state test may not do as well promoting growth on the lower-stakes test. The current research provides some evidence that value-added on high-stakes tests may be distorted somewhat by a narrow focus on the tested material. One study calculated a teacher's value-added to her students' test scores at the end of the year and scores produced at the end of the next several school years. The study showed that value-added fades over time as students progress through school. For example, a fourth grade teacher's value-added based on her students' fourth grade test scores is larger than her value-added based on those students' fifth or sixth or higher grade scores. The study also calculated value-added for the current and future years using both a high-stakes and a low-stakes test. It found that teachers' contributions to learning appeared to fade more quickly on a high-stakes test than on a low-stakes test. [15] If some teachers narrowly focus on features specific to a test, such as its format for math problems or its limited set of vocabulary, then we might not expect their students' achievement growth to carry over to future tests.[16]

What are the consequences of sensitivity to the test?

The research suggests that conclusions about a teacher's effectiveness drawn from one test might be specific to that test, and might not provide an accurate picture of that teacher's effectiveness overall. For instance, the study in Hillsborough found that only 43 percent of teachers who ranked in the top 20 percent of teachers according to value-added from the norm-referenced test also ranked in the top 20 percent according to value-added estimates from the Sunshine State Standards test.[17] Similarly, the data from the Northeastern city found that if the district had been using a pay-for-performance system,[18] changing tests would have changed the bonuses for nearly 50 percent of teachers, for an average salary difference of $2,000.[19] Some of the variation is due to the imprecision in calculating value-added, but even without these errors, the correspondence among conclusions would be less than perfect. The Houston study, which substantially reduced imprecision by pooling multiple years of data, still found

that only about 46 percent of reading teachers classified in the top 20 percent on the state test were also among the top performers on the district test.[20]

Although conclusions about individual teachers can be highly sensitive to the test used for calculating value-added, value-added on state tests can be used to identify groups of teachers who, on average, have students who show different growth on alternative tests. The MET Project found that, on average, if two teachers' value-added calculated using the state test in year one of the study differed by 10 points, then, in year two, their students' achievement on the low-stakes alternative test would differ by about 7 points.[21] If we use the state test to identify teachers who have low value-added, the selected teachers' students would also have lower growth on the alternative tests than would students of other teachers. The differences between the two groups of teachers would be about 70 percent as large on the alternative test as on the state test. Some of the individual teachers identified as low-performing might actually have students with substantial growth on the alternative test, but as a group, their students would tend to have lower growth on both tests. Thus, the sensitivity of value-added to a test does not mean it cannot be used to support decisions about teachers that could potentially help student outcomes.

WHAT MORE NEEDS TO BE KNOWN ON THIS ISSUE?

Multiple sources can contribute to differences in value-added on different tests. Determining the contributions of each would be valuable for designing effective evaluation systems.

Differences in value-added from high- and low-stakes tests might be due to some teachers focusing more than others on superficial aspects of the tests and practices that improve student test scores but not student achievement. These practices result in what is sometimes referred to as "score inflation."[22] An egregious example is of teachers changing student test scores in the highly publicized cheating scandals in Atlanta and New York.[23] Practices that lead to test-score inflation can also be less overt. For instance, teachers may have students practice test-taking skills.[24] But differences in value-added from high- and low-stakes tests might not be due to score inflation. They may also be due to differences in the content on the tests. A teacher's effectiveness at helping students learn the state standards might be only weakly related to his effectiveness at teaching other material, especially if he focuses on the state standards and his instructional materials are related to those standards.

The two scenarios have different implications for the utility of value-added for improving student outcomes and the consequences of switching to the rigorous Common Core State Standards and their associated tests. If some teachers have high value-added on the high-stakes test because of score inflation, then value-added might not be useful for identifying effective teachers. And using it in evaluations may have negative consequences for students. If teachers are focused on the content of the state standards and if value-added reflects their effectiveness at teaching that material, then setting rigorous standards should have positive effects on student learning. The research discussed above offers limited evidence in support of both scenarios. States and districts would benefit from knowing how much each scenario contributes to value-added on high stakes tests.

As value-added estimates become more common in consequential teacher evaluations, the motivation for teachers to take steps that might inflate their value-added will increase. There are many examples in education, and other fields, in which the use of performance indicators leads to their distortion.[25] So,

understanding this risk, and designing systems to prevent it, will be important for maintaining the integrity of teacher evaluation. One particular concern is peer competition: the extent to which teachers feel they need to take steps to inflate their scores because they think their colleagues are.

If differences in value-added with different tests are due to differences in the skills measured by the tests, the choice of skills to be measured is a critical one. We have some evidence that teacher effectiveness differs according to the two math areas of problem-solving and procedures, but we need to further understand the contributions of measurement error to those findings and to extend them to other subjects and skills. We also need to better understand which skills lead to better long-term outcomes and to determine how best to measure those skills so that value-added can provide the most meaningful measure of a teachers' effectiveness.

It would be helpful to have more systematic evaluations of how value-added changes when states introduce new tests. In recent years, some states have made significant changes to their tests, and these might hint at what states can expect when they change to the Common Core tests. However, there is no research that documents how value-added changed with the change in tests. New studies might explore the correlation in value-added with the new and old tests, using these states as examples: What types of teachers have value-added that is different with the new test than with the old test? The studies should also consider the differences in the tests themselves and explore how these contribute to differences in value-added.

WHAT CAN'T BE RESOLVED BY EMPIRICAL EVIDENCE ON THIS ISSUE?

We cannot accurately test students on all the skills they need to be ready for college and careers and to lead happy and productive lives. Even the best test will measure only a limited set of skills, and value-added that is based on that test will evaluate teachers only on those skills. When considering skills to be tested, ideally we would select those that are required for students to achieve the long-term outcomes that society values. Empirical analyses might someday help us determine how skills relate to long-term outcomes, but they cannot tell us what those outcomes should be.

TO WHAT EXTENT, AND UNDER WHAT CIRCUMSTANCES, DOES THIS ISSUE IMPACT THE DECISIONS AND ACTIONS THAT DISTRICTS AND STATES CAN MAKE ON TEACHER EVALUATION?

The research presented above raises two concerns for states: 1) Is the value-added for teachers from one test providing an adequate measure of the teacher's contributions to learning, and how can the data be used to provide the most accurate measure of teacher effectiveness? 2) How should states prepare for the transition to new tests aligned with the Common Core State Standards and the impact this will have on value-added?

Implications for using value-added from one test

As discussed above, value-added from one test might not agree with that from another because 1) value-added depends on factors that are specific to each test – test format, for instance – but unrelated to teacher effectiveness, or 2) value-added does not measure aspects of teacher effectiveness related to

content not covered by the test. States should choose tests that ensure that student scores are determined by content knowledge and not by extraneous factors. They also should select tests with a very broad range of content so that meaningful aspects of teacher effectiveness are not missed. The tests that two consortia are developing to measure the Common Core State Standards are designed to meet both these goals.

However, states and districts might still worry that even with the new tests, value-added might not reflect all aspects of a teacher's contributions to student outcomes. States might combine value-added with other measures to reduce this risk. The MET Project found that composite measures of teacher effectiveness that combined, with roughly equal weights, value-added calculated with state tests, classroom observations, and student responses to surveys were somewhat better predictors of a teacher's future value-added calculated from an alternative test than was value-added calculated from the state test alone.[26]

Implications for the transition to the new test

The transition to the new Common Core tests will introduce many conditions that could lead to changes in teachers' value-added. The new tests will measure content aligned with the new standards, not current standards. They will also use a different test format. Administered on computer, they will have fewer multiple choice items and more performance tasks and constructed response items. The new tests also are to be more cognitively demanding than the current tests and could be administered at different times than some tests are now. Consequently, student scores on these new tests will likely differ from what scores would have been on the current state tests, and teachers' value-added is likely to be different, as well. Already, evidence shows that there is often a precipitous change in student achievement when districts change from one test to another.[27] If a similar drop follows the switch to new Common Core tests, it could further contribute to instability in value-added estimates.

With these conditions in mind, states should prepare for larger year-to-year changes in value-added in the 2014-15 school year when they switch to tests aligned with the Common Core. They should also be prepared for greater year-to-year variability in value-added for a few years after the change when they will be using both old and new tests.[28] Large year-to-year variability makes value-added hard to interpret. A teacher doing well one year may appear to be doing poorly the next. It can also weaken the credibility of value-added among teachers.[29]

There is no formal research on what states should do to ease this transition. The following paragraphs provide insights from an assistant superintendent from a large urban school district that uses value-added for performance-based pay and from analysts who supply value-added to states to address questions states might have about the upcoming changes.[30]

One question states have is how they should modify their value-added models when they switch tests. Most of the analysts contacted for this brief said that when states changed tests, they simply applied the same methods they did in other years, except the prior year scores were from the old test and the current year scores were from the new test.[31] All of the analysts noted the importance of making sure that scores on the old tests are predictive of students' performance on the new tests before calculating value-added using data from both tests. If students' scores on the old tests are weak predictors of their performance on the new tests, it could lead to error in value-added. One analyst recommended that to improve the predictive power of the prior tests, states could use prior achievement scores in math,

reading, and other available subjects in calculating value-added for either math or reading. He also advised that states account for measurement error in the prior scores when creating value-added.

Many value-added methods evaluate teachers' contributions to student learning by comparing their students' growth in achievement with the average growth of all students in the district or state during the school year. As a result, value-added cannot be used to determine if the average teacher is improving across time. To avoid this problem, some value-added methods including those used by Tennessee and Ohio, compare student achievement growth to the average growth of students from a prior school year called a base year. The base year remains constant over time so that across years the average value-added will increase with any improvements in teaching. The analysts recommended that the base year be changed when the states switch tests, noting that scores from the new test cannot be compared with base-year scores from the old test.[32] The analysts further suggested waiting a few years before establishing a new base year because, in their experiences, large year-to-year changes in the student score distributions can occur in the first few years of a new testing program.

The assistant superintendent cited above noted that when the state made a significant change to its annual test, teachers whose classes had high prior achievement saw their value-added go up relative to other teachers, and teachers whose classes had low prior achievement saw their value-added go down. The effects were detrimental to the credibility of value-added in the district overall. They fed existing suspicions that value-added was too low for some teachers because of the students they taught.[33] District leaders were not prepared to address the problems created by the new test.

Because changing the test is very likely to lead to changes in value-added for some teachers, states may want to prepare: they may want to thoroughly investigate how value-added changes from the years before the Common Core test to the first years after it. If teachers with certain types of students have systematically higher or lower value-added on the new test, the state might want to take steps to reduce negative repercussions. The state might not release value-added for some teachers for a few years, instead waiting for multiple prior years of data on the new test to better adjust for differences among classes. The state might follow the recommendations of analysts and use tests from multiple subjects and control for measurement error in their value-added calculations. The state might use a weighted average of value-added calculated using the old and the new tests to smooth out the transition in tests. The state might follow the MET Project and use a composite estimate with less weight on value-added, or if the effects of the new test are concentrated on the value-added for a subset of teachers, the state might give these teachers' value-added less weight or allow districts greater flexibility in how they use value-added for performance evaluations.

ENDNOTES

[1] The report by Koretz and colleagues provide several examples of differences in scores for students on high- and low-stakes tests. See: Koretz, Daniel, Robert L. Linn, Stephen B. Dunbar, and Lorrie A. Shepard. The Effect of High-stakes Testing on Achievement: Preliminary Findings about Generalization Across Tests. US Department of Education, Office of Educational Research and Improvement, Educational Resources Information Center, 1991. Retrieved March 4, 2013 from http://www.colorado.edu/UCB/AcademicAffairs/education/faculty/lorrieshepard/testing.html

[2] Two consortia are developing the tests aligned with the Common Core State Standards. The Partnership for Assessment in Readiness in College and Careers includes 21 states and the District of Columbia

(http://www.achieve.org/parcc-states) and the SmarterBalanced Consortium includes 23 states (http://www.smarterbalanced.org/about/member-states). Two states, North Dakota and Pennsylvania are participating in both efforts.

[3] Issues about how the change to Common Standards and the associated tests will affect school and teacher accountability are already gathering considerable attention. Examples include blogs in the *Hechinger Report*, http://eyeoned.org/content/its-the-curriculum-stupid_394/#more-394 and a response to it in a blog at *Education Week Teacher,* http://blogs.edweek.org/teachers/teaching_now/2013/02/will_the_common_core_skew_value-added_scores.html?cmp=ENL-TU-VIEWS2.

[4] See: Bill and Melinda Gates Foundation, *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project*, 2010. Retrieved March 4, 2013 from http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf

[5] These correlation coefficients were not adjusted for measurement error in value-added. As discussed in the section, *Why Is Value-added Sensitive to the Test,* the MET report also presents correlation coefficients that are corrected for the measurement error and the corrected values are larger.

[6] Tim R. Sass, "The Stability of Value-Added Measures of Teacher Quality and Implications for Teacher Compensation Policy," (The National Center for Analysis of Longitudinal Data in Education Research, Brief 4, 2008). Retrieved March 4, 2013 from http://www.caldercenter.org/about/Tim-Sass.cfm

Sean P. Corcoran, Jennifer L. Jennings, and Andrew A. Beveridge, "Teacher Effectiveness on High- and Low-Stakes Tests," (In paper originally delivered to the SREE Conference, Washington, DC, 2011). Retrieved March 4, 2013 from https://files.nyu.edu/sc129/public/research.htm

Papay, John P, "Different Tests, Different Answers The Stability of Teacher Value-Added Estimates Across Outcome Measures," *American Educational Research Journal* 48 (1) (2011): 163-193. Retrieved March 4, 2013 from http://aer.sagepub.com/content/early/2010/03/31/0002831210362589.full.pdf

[7] The district administered the Stanford Achievement Test Series, Tenth Edition. For additional details, see: Corcoran, et al, 2008, ibid.

[8] The district administered the Stanford Achievement Test of math and reading and the Scholastic Reading Inventory. For details, see: Papay, 2011, ibid.

[9] The correlation between value-added calculated using the SSS and the NRT for math teachers in Hillsborough County was 0.48. For teachers in Houston the study considered multiple model specification and the correlation between value-added calculated using the state test and the district administered norm-referenced test ranged from .45 to .57 with a value of .50 for the baseline specification for reading teachers and the correlation ranged from .58 to .62 with a value of .59 for the baseline specification for math teachers. The study of the Northeastern city calculated value-added for reading teachers using three tests and used multiple model specification for the calculations. The correlation between value-added calculated using the district administered norm-referenced test and either of the other tests (the state test or the reading test) ranged from .16 to .28 across different model specifications. The correlation between value-added calculated using the reading test and the states test ranged from .44 to .51 depending on the model specification.

[10] Tests often have different test forms. For example, the test booklet used in one year has different specific items than the booklets used in other years. Each year's test is a different test form. Test forms are constructed using common design guidelines to measure the same content and they use very similar item types and the same format each year. They are equated so that scores from one form can be used interchangeably with scores from other forms. However, because the forms use different items, a student's score on one form will differ from the score she or he would receive if tested on a different form. This variability in scores due to the test form is known as test measurement error. It creates instability in value-added that contributes to statistical error. We distinguish between sensitivity to test form and sensitivity to tests because test forms are designed to measure the same content and be exchangeable. The variability in due to test forms is included in the statistical errors. Different tests are designed to measure related but somewhat different content and statistical errors do not account for the contributions of different tests to variation in value-added.

[11] In the paper, "Different Tests, Different Answers: The Stability of Teacher Value-Added Estimates Across Outcome Measures," John Papay finds the correlation ranges from .52 to .59 depending on the specification of the

model. He uses the Spearman correlation of ranks but this is likely to have limited impact on the estimate of the correlation. J.R. Lockwood and colleagues find that the correlation ranges from .01 to .46 depending on the value-added model specification.
See: Lockwood, J. R., Daniel F. McCaffrey, Laura S. Hamilton, Brian Stecher, Vi-Nhuan Le, and José Felipe Martinez, "The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics Achievement Measures," *Journal of Educational Measurement* 44(1) (2007): 47-67.
http://www.copymail.wceruw.org/news/events/VAM%20Conference%20Final%20Papers/StudentTeacherInteractions_LockwoodMcCaffrey.pdf
Neither study adjusts the reported correlations for the test measurement error due to using a small number of items to measure the subtest content. However, Lockwood and colleagues report high reliability for the each subscale and correlation between individual student's scores on the two subscales that is much higher than the correlation in the two sets of value-added measures.

[12] The MET study conducted a formal evaluation of the alignment of the content covered by its supplemental test and the content covered by the fourth and eighth grade state tests administered in the two participating districts. There was overlap in the content but the emphasis of specific areas differed between the state and supplemental tests. Hence, the MET results cannot clearly distinguish between the effects of difference in content and difference in cognitive demands of the state and supplemental tests on value-added.

[13] Daniel M. Koretz, "Alignment, High Stakes, and the Inflation of Test Scores," (National Center for Research on Evaluation, Standards, and Student Testing (CRESST) Center for the Study of Evaluation (CSE), Report 655, 2005). Retrieved March 4, 2013 from http://cse.ucla.edu/products/reports/r655.pdf

[14] The MET Project surveyed students about test preparation in their classrooms. In first year of the study, Elementary school students were asked to evaluate the statements "We spend a lot of time practicing for the state test" and "Getting ready for the state test takes a lot of time in our class" on a 1 to 5 scale from "No, Never" to "Always". The average scores for classrooms for these items were 4.3 and 3.8, indicating that elementary students perceived they spent a considerable amount of time practicing for the test. For secondary classes the means were 3.6 and 3.4 on the 1 to 5 scale from "Totally Untrue" to "Totally True" suggesting that secondary students perceived less time spent on test preparation. Secondary students were also asked evaluate the statement "I have learned a lot this year about the state test," which had an average score of 3.7 across classes.

[15] The study by Corcoran et al. (2008) finds that about 40% of a fourth grade teacher's value-added as measured on high stakes tests carried over to his/her students fifth grade scores but about 60% of the teacher's value-added as measured by low stakes tests carried over to fifth grade scores.

[16] Koretz, Daniel M, "Limitations in the use of achievement tests as measures of educators' productivity." *Journal of Human Resources* 37(4) (2002): 752-777.
http://standardizedtests.procon.org/view.resource.php?resourceID=004348

[17]Sass, Tim R, "The Stability of Value-Added Measures of Teacher Quality and Implications for Teacher Compensation Policy," (The National Center for Analysis of Longitudinal Data in Education Research, Brief 4, 2008).

[18] The study assumed a system like ASPIRE used in Houston, Texas.

[19] Papay, 2011, ibid.

[20] Corcoran, et al, 2008, ibid.

[21] Kane, Thomas J., Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment," (Bill & Melinda Gates Foundation, MET Project Research Paper, 2013). Retrieved March 4, 2013 from http://www.metproject.org/downloads/MET_Validating_Using_Random_Assignment_Research_Paper.pdf.
Students were randomly assigned to teachers' classrooms within schools in year two of the MET study, so we can conclude that being assigned a teacher with higher value-added on the state test has a positive effect on student achievement on both the state test and other tests.

[22] Koretz, 2002, ibid.

[23] An article by Lois Beckett gives a brief summary of these and other high profile cheating scandals. See: Beckett, Lois, "America's most outrageous teacher scandals," *Propublica April 2013*.
http://www.propublica.org/article/americas-most-outrageous-teacher-cheating-scandals

[24] See Koretz, 2002, ibid, for examples of other forms of score inflation.
[25] The pressure for people to corrupt performance indicators is so common that back in 1976 Donald T. Campbell summarized it in what is now known as Campell's law: "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor." See: Campbell, Donald T, "Assessing the impact of planned social change," *Evaluation and Program Planning* 2(1) (1979): 67-90. Retrieved May 24, 2013 from https://www.globalhivmeinfo.org/CapacityBuilding/Occasional%20Papers/08%20Assessing%20the%20Impact%20of%20Planned%20Social%20Change.pdf. Sheila Bird and her colleagues noted that when journey times were use to evaluate the performance of ambulance services in the Great Britain, the services made perverse changes to the start time to improve their performance on the indicators. See:
Bird, Sheila M., David Cox, Vern T. Farewell, Harvey Goldstein, Tim Holt, and C. Peter, "Performance indicators: good, bad, and ugly," *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 168(1) (2005): 1-27.
[26] Kata Mihaly, Daniel F. McCaffrey, Douglas O. Staiger, and J.R. Lockwood, "A Composite Estimator of Effective Teaching," (MET Project Research Paper, 2013). Retrieved March 4, 2013 from http://www.metproject.org/downloads/MET_Composite_Estimator_of_Effective_Teaching_Research_Paper.pdf
[27]Koretz, 2002, ibid.
[28]Some states and districts are using growth measures, such as median growth percentiles, as an alternative to value-added. Those measures also will rely on new and old tests together when states transition tests and will be susceptible to some of the same issues as value-added scores.
[29] In their paper, "Evaluating Teacher Evaluations," Linda Darling-Hammond, et al provide examples of teachers' responses to year-to-year variability in their value-added scores and the negative consequences of these unstable scores. See: Linda Darling-Hammond, Audrey Amrein-Beardsley, Edward Haertel, and Jesse Rothstein, "Evaluating Teacher Evaluations", *Phi Delta Kappan* 93(6) (2012): 8-15.
[30] We received informal comments on this issue from Damian Betebenner, at the National Center for the Improvement of Educational Assessment, Inc., Robert Meyers at the Value-Added Research Center at the University of Wisconsin, Mary Peters at Battelle for Kids, Carla Stevens at the Houston Independent School District, and John White at SAS.
[31] These analysts use residual growth models to calculate value-added. In a residual growth model, prior year test scores are used to predict students' current year scores and the difference between the predicted score and the actual score determines a student's growth relative to other students. Value-added typically uses variables in addition to prior achievement when predicting current scores. A teacher's value-added equals the average of her students' residual growth. For additional details on alternative formulations of growth models, see: Castellano, K. E., Andrew D. Ho, "A Practitioner's Guide to Growth Models," (Council of Chief State School Officers, 2013). Retrieved May 23, 2013 from
http://scholar.harvard.edu/files/andrewho/files/a_pracitioners_guide_to_growth_models.pdf
[32] The analysts noted that there are methods for linking the scores on the old and new tests and retaining the base year. However, they advised against using this approach because it can yield unstable results.
[33] The educator explained that the old state test had a "ceiling effect" for high-achieving students; many of them got all the questions right. Teachers in the district questioned the accuracy of value-added from a test with a ceiling effect. The new test was more challenging and did not have a ceiling effect. Some people in the district interpreted the changes following the introduction of the new test as confirmation of bias in the value-added that was calculated with the old test and evidence that it should not be used in teacher evaluations.

AUTHORS

Daniel F. McCaffrey is a Principal Research Scientist at Educational Testing Service. Previously, he was a Senior Statistician and PNC Chair in Policy Analysis at the RAND Corporation. He is a fellow of the American Statistical Association and is nationally recognized for his work on value-added modeling for estimating teacher performance. McCaffrey oversaw RAND's efforts as part of the Gates Foundation's Measures of Effective Teaching study to develop and validate sophisticated metrics to assess and improve teacher performance. He is currently working on additional studies comparing value-added measures to other measures of teaching, including classroom observations. He recently completed work on a four year project funded by the Institute of Education Sciences (IES) that developed alternative value-added models of teachers' effectiveness. McCaffrey is also the principal investigator of a National Institute on Drug Abuse–funded study, and recently led RAND's efforts as a major partner in the National Center on Performance Incentives, which conducted random control experiments to test the effects of using value-added to reward teachers with bonuses. He led an evaluation of the Pennsylvania Value-Added Assessment Pilot Program (PVAAS) and was the lead statistician on other randomized field trials of school-based interventions; including evaluations of the Cognitive Tutor geometry curriculum, the Project ALERT Plus middle and high school drug prevention program, and the teen dating violence prevention curriculum, Break the Cycle. McCaffrey received his Ph.D. in statistics from North Carolina State University.

ABOUT THE CARNEGIE KNOWLEDGE NETWORK

The Carnegie Foundation for the Advancement of Teaching has launched the Carnegie Knowledge Network, a resource that will provide impartial, authoritative, relevant, digestible, and current syntheses of the technical literature on value-added for K-12 teacher evaluation system designers. The Carnegie Knowledge Network integrates both technical knowledge and operational knowledge of teacher evaluation systems. The Foundation has brought together a distinguished group of researchers to form the *Carnegie Panel on Assessing Teaching to Improve Learning* to identify what is and is not known on the critical technical issues involved in measuring teaching effectiveness. Daniel Goldhaber, Douglas Harris, Susanna Loeb, Daniel McCaffrey, and Stephen Raudenbush have been selected to join the Carnegie Panel based on their demonstrated technical expertise in this area, their thoughtful stance toward the use of value-added methodologies, and their impartiality toward particular modeling strategies. The Carnegie Panel engaged a User Panel composed of K-12 field leaders directly involved in developing and implementing teacher evaluation systems, to assure relevance to their needs and accessibility for their use. This is the first set of knowledge briefs in a series of Carnegie Knowledge Network releases. Learn more at **carnegieknowledgenetwork.org**.

CITATION

McCaffrey, Daniel F. Carnegie Knowledge Network, "Will Teacher Value-Added Scores Change when Accountability Tests Change?" Last modified June 2013. URL = <http://www.carnegieknowledgenetwork.org/briefs/value-added/accountability-tests/>

Carnegie Foundation
for the Advancement of Teaching

Carnegie Foundation for the Advancement of Teaching
51 Vista Lane
Stanford, California 94305
650-566-5100

Carnegie Foundation for the Advancement of Teaching seeks to vitalize more productive research and development in education. We bring scholars, practitioners, innovators, designers, and developers together to solve practical problems of schooling that diminish our nation's ability to educate all students well. We are committed to developing networks of ideas, expertise, and action aimed at improving teaching and learning and strengthening the institutions in which this occurs. Our core belief is that much more can be accomplished together than even the best of us can accomplish alone.

www.carnegiefoundation.org

We invite you to explore our website, where you will find resources relevant to our programs and publications as well as current information about our Board of Directors, funders, and staff.

Knowledge Brief 8
June 2013
carnegieknowledgenetwork.org

Funded through a cooperative agreement with the Institute for Education Science. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.