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

Daniel F. McCaffrey

Educational Testing Service

Carnegie Knowledge Network Webinar

August 14, 2013

Background

- Value-added (VA) models use student test scores to estimate teachers' contributions to student achievement
- How we judge student achievement can depend on which test we use to measure it
- Will VA also be sensitive to the test?
- Many states have plans for a substantial change to their tests over the next few years

Guiding Questions

- What do we know about VA estimates from different tests?
- What more needs to be known on this issue?
- What can't be resolved by empirical evidence on this issue?
- How, and under what circumstances, does this issue impact the decisions and actions that districts can make on teacher evaluation?

What do we know about VA estimates from different tests?

Two Approaches

- Compare VA from different tests
- Study what happened when states changed tests in the past

The MET Project

- Compared VA estimates on the state math or reading test with estimates on study administered tests
- Project tests were
  - Balance Assessment of Mathematics or SAT-9 Open-Ended Reading test
  - More cognitively demanding and open-ended
  - Used only by the project without consequences for teachers or students
- Association between VA estimates from different tests was weak
  - Correlation of .38 for math and .21 for reading

Other Comparisons

- Three studies compared VAM estimates from two different tests administered by the school district or state
  - Houston, TX, Hillsborough County, FL and large urban district in the Northeastern US
  - Compared VA on state accountability test to VA on alternative state test (FL) or district administered test
  - Correlation between VA from tests was generally low to modest (0.15 to 0.59)

Difference in VA Across Tests Could Lead to Different Conclusions about Teachers

- In the Houston and Hillsborough studies, less than half of the teachers ranked in the top 20% of teachers on one test received the same ranking on the other test
- Had the Northeastern urban district been using pay for performance, changing the test used to calculate VA would change the bonuses of 50% of teachers

Differences Are Not Just Statistical Errors

- VA has statistical errors due to test measurement error, the sample of students, and other chance factors
- Statistical error suppresses correlation
- MET project adjusted correlations for the statistical error, the resulting values were 0.54 for math and 0.37 for reading
- Houston study which combined multiple years of VA on each test for each teacher which reduces statistical errors had high correlation of 0.50 for reading and 0.59 for math

Multiple Possible Sources of Differences in VA

- Tested content
- Other features of the test: timing of the test, item format, and cognitive demand
- Consequences associated with test outcomes

Content

- Teachers might not be equally effective at promoting growth on all content
  - Teachers might focus on only some of the possible content such as content on the state standards
- Content evaluations of tests suggest overlap but difference in tests used in the research studies we reviewed
- Studies have found that teacher VA differs on different content from the same test administered to the same students
  - Estimated VA from the “procedures” and “problem-solving” subtests of a standardized math test
  - Weak correlation between VA from the two subtests
  - Replicated in two different studies

Other Features of the Test

- Test can differ on timing of administration (fall-to-fall vs. spring-to-spring), cognitive demand of the items, and item format
- Teachers' effectiveness may vary with these factors
- Urban district study had differences in timing and found this was an significant contributor to difference in VA
- Tests administered by MET project were chosen to be more cognitively demanding and they used open-ended items rather than multiple choice items which dominates state test
- Student scores can be sensitive to even small changes in item format

Consequences

- The outcomes of tests can have consequences for schools, teachers, and students
- These consequences may influence student effort and teacher attention to the specifics of the test content and structure
- Student outcomes and VA may vary with consequences
- All the comparisons in the reviewed studies involved the state accountability test which had significant consequences and another test that had fewer potential consequences
- Literature finds that high-stakes can distort test results and the distortions are not equal across all educators
- In other contexts, there is evidence that students have low motivation on tests with no consequences

What Happened in the Past When States Changed Their Tests?

- In the past, state tests have changed in some districts or states that were estimating teacher VA
- The districts did not report the impact of the changes on VA
- We contacted vendors who calculate VA and one district using VA for performance pay to learn about their experiences
  - Vendors did not report any particular issues that arose
  - Vendors did not change their procedures when tests changed
  - One exception was a vendor who started controlling for both prior math and reading scores after a state changed its test and now does this routinely
  - District contact reported that VA went up for teachers of high achieving students when the test changed and some teachers took this as evidence that VA was invalid

What more needs to be known on this issue?

Two Questions

- What does this mean for using VA to assess teachers?
- What does this mean for VA when states change to tests of the Common Core State Standards (CCSS)?

Need to Know Contribution of Possible Sources of Differences in VA

- Tests used in the research studies differed on multiple features
- How much each contributed cannot be determined
- Knowing how much each contributed is important because sources have different implication for answering the two questions

Potential Sources of Differences in VA Have Implications for Validity of Teachers Evaluations

Content

- **Teacher's effectiveness on limited sample of content may not accurately reflect effectiveness on other content**
- **Selection of tested content is an important decision**

Other Features of the Test

- **Neither spring nor fall testing is more valid for conclusions about teachers but it is a source of error in conclusions about teachers**
- **Tests with low cognitive demands and restrictive item format may limit what know about a teacher's performance**

Potential Sources of Differences in VA Have Implications for Assessing Teachers

Consequences

- **Literature finds high-stakes can distort test results ("score-inflation") and the distortions are not equal across all educators**
  - **Distorted scores cannot be used to make valid decisions about a teacher**
  - **Distortions could bias comparisons among teachers**
- **Increasing use of value-added in consequential teacher evaluations increases the motivation for teachers to take steps to inflates scores**
- **Repeated use of similar items may enable teachers to teach the item types rather than broader content understanding**

Many Sources Could Make VA Differ on New Tests

- Tests of the CCSS are likely to differ from current tests on content, item format, and cognitive demands
- New and current tests will have similar consequences

How, and under what circumstances, does this issue impact the decisions and actions that districts can make on teacher evaluation?

Possible Actions to Reduce Threats to Validity

- Careful selection of test content and other features to align tests with valued outcomes
- Design tests and testing program to discourage score inflation and detect it, if it occurs
  - Use diverse item formats
  - Build in audit testing
- Combining VA on the state test with other measures of teaching
  - MET project found composites with roughly equal weight on VA on the state test, classroom observations, and student surveys improved prediction of VA on the alternative test

Ways to Prepare for Changing to the New Test

Districts and states might

- **Test stability of VA before releasing results**
- **Identify any teachers where large changes in VA occur**
- **Find ways to soften the consequences of any systematic changes in VA that could undermine its credibility**
- **Used standard procedures applied to both old and new tests in their VA calculations**
  - **This is the approach vendors reported using**

Changes to VA Are Not Guaranteed

- **Empirical studies did not test VA following a change in state tests**
- **Concerns about changes to VA are based on deductions and involve some amount of extrapolation**
- **Some states moved to tests of the Common Score standard in the 2012-13 school year, we should encourage those states to study the stability of VA and share their results**