Inferences About Teachers Based on Student Test Scores

Edward Haertel
Stanford University

Washington, DC
March 22, 2013
Policy Context

- Sense of urgency about reform
  - Persistent achievement gaps
  - International rankings
  - 21st century skills

- Evidence of powerful teacher effects

- Weak, ineffective teacher evaluation

- Finding and removing ineffective teachers as a potent path for education policy?
The Allure of VAM

- **Value-Added Models (VAM)** use student test score gains to provide direct evidence of teacher effectiveness
  - Data linking student test scores to teachers are much improved
  - Statistical models seem to promise reliable and valid results
  - Extraordinary benefits are projected
Overview

- How big are teacher effects?
- A close look at test score gains
- How do VAMs work?
- Interpretive Argument
  - What do scores mean?
  - How reliable are they?
  - What do they leave out?
  - How should they be used?
- Sound teacher evaluation
Overview

- How big are teacher effects?
- A close look at test score gains
- How do VAMs work?
- Interpretive Argument
  - What do scores mean?
  - How reliable are they?
  - What do they leave out?
  - How should they be used?
- Sound teacher evaluation
How Big are Teacher Effects?

Influences on Student Test Scores
- Teacher
- Other School Factors
- Out-of-School Factors
- Unexplained Variation

2013 William H. Angoff Memorial Lecture
How Big are Teacher Effects?

Out of reach?

Influences on Student Test Scores

- Teacher
- Other School Factors
- Out of reach?

BIG, as potential policy variable

2013 William H. Angoff Memorial Lecture
How Big are Teacher Effects?

Influences on Student Test Scores
- Teacher
- Other Factors

Random noise plus systematic bias

SMALL, as “signal” relative to “noise”
Problems With the Logic

- Cannot reliably identify top-quintile teachers
- Effects fade over time
- “Top-quintile” teachers are in short supply
- Proposal to replace “worst” with average teachers is similarly flawed
Overview

- How big are teacher effects?
- A close look at test score gains
- How do VAMs work?
- Interpretive Argument
  - What do scores mean?
  - How reliable are they?
  - What do they leave out?
  - How should they be used?
- Sound teacher evaluation
A Close Look at Test Score Gains

2013 William H. Angoff Memorial Lecture
Consequences of Nonlinear Scale

- A nonlinear scale means teachers are rewarded or penalized, depending on where their students start out
  - Especially problematical for teachers of students above or below grade level, or with special needs

Measured Growth = 6 points

Measured Growth = 7 1/2 points
Overview

- How big are teacher effects?
- A close look at test score gains
- How do VAMs work?
- Interpretive Argument
  - What do scores mean?
  - How reliable are they?
  - What do they leave out?
  - How should they be used?
- Sound teacher evaluation

2013 William H. Angoff Memorial Lecture
## Comparison to a Familiar Testing Situation

<table>
<thead>
<tr>
<th></th>
<th>Typical Test</th>
<th>Simplified Teacher VAM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Examinees</strong></td>
<td>Students</td>
<td>Teachers</td>
</tr>
<tr>
<td><strong>Items</strong></td>
<td>Test questions</td>
<td>Students</td>
</tr>
<tr>
<td><strong>Test</strong></td>
<td>Items in a test form</td>
<td>Students in a classroom</td>
</tr>
<tr>
<td><strong>Administration</strong></td>
<td>Student answers items</td>
<td>Teacher teaches students</td>
</tr>
<tr>
<td><strong>Item Scoring</strong></td>
<td>Item responses scored according to key</td>
<td>Student learning “scored” by giving each student a standardized test</td>
</tr>
<tr>
<td><strong>Test Score</strong></td>
<td>Sum of Item scores</td>
<td>Average of student test scores</td>
</tr>
</tbody>
</table>

*This simplified version cannot work, because teachers get “tests” (classes) of varying difficulties (prior knowledge, educational challenge)*

2013 William H. Angoff Memorial Lecture
Accounting for Student Differences

- Begin with average score for all students
- Make Adjustments
  - Remove whatever teacher is not responsible for
  - Leave whatever teacher is responsible for
- Assume that everything left over (i.e., not explained or accounted for) is:
  - “Effect” of particular teacher and/or
  - Random (or nonsystematic) variation

In other words, we assume that once adjustments are made, assignments of students to teachers may be regarded as random.

2013 William H. Angoff Memorial Lecture
What to Adjust For?

- Prior-year test scores
- Test scores from two or more years earlier
- Absences, suspensions, grade retentions
- “English learner” or “special education” status
- Title I eligibility
- Student mobility
- Summer school attendance
- Gender
- …
Los Angeles Value-Added Model

“Stripped of the Greek symbols and statistical jargon, [the] Los Angeles Value-Added Model … , in essence, claims that once we take into account five pieces of information about a student, the student’s assignment to any teacher in any grade and year can be regarded as occurring at random.”

(Briggs & Domangue, 2011, p. 4)
Background Variables in the LA Value Added Model

- Prior-year test scores
- Gender
- English language proficiency
- Eligibility for Title I
- Whether student entered LAUSD schools after kindergarten
Besides Student Adjustments…

- Available instructional materials, resources
- Classroom aides
- Other teachers
- Student peers
- School safety, climate, policies
- Out-of-school influences during the year
- …
Estimating Unobserved Scores

- How well can VAM models adjust for student and school differences?

There is essentially no “mixing” of students from most affluent versus least affluent schools. While students do change teachers and schools, they rarely make large moves up or down across social strata.

“Given the reality of school segregation on the basis of various demographic characteristics of students, including family socioeconomic background, ethnicity, linguistic background, and prior achievement … in practice, some students [may] have no access to certain schools.”

Reardon & Raudenbush (2009, p. 494)
Peer Effects Due To...

- Group work
- Peer culture
- Collective influence on pacing
- Classroom “chemistry”
Overview

- How big are teacher effects?
- A close look at test score gains
- How do VAMs work?
- Interpretive Argument
  - What do scores mean?
  - How reliable are they?
  - What do they leave out?
  - How should they be used?
- Sound teacher evaluation
Interpretive Argument

- Scoring
  - Observed score meaning
    (bias? other evidence of effectiveness?)

- Generalization
  - From observed score to universe score
    (stability/consistency? reliability?)

- Extrapolation
  - From universe score to target score
    (target qualities fully captured?)

- Implication
  - From target score to specific interpretations or decisions
    (usefulness for specific purposes? unintended effects?)
Interpretive Argument

1. Scoring
   This class, this year, this test
   *Bias?*

2. Generalization
   Other classes, other years
   *Random Error?*

3. Extrapolation
   Other tests, non-test outcomes
   *Other Measures?*

4. Implication
   Validity of possible uses
   *Consequences?*

2013 William H. Angoff Memorial Lecture
Interpretive Argument

1. Scoring
   - This class, this year, this test
     Bias?

2. Generalization
   - Other classes, other years
     Random Error?

3. Extrapolation
   - Other tests, non-test outcomes
     Other Measures?

4. Implication
   - Validity of possible uses
     Consequences?
Scoring

- Do teachers’ VAM scores accurately reflect their actual effectiveness, this year with these students, in teaching the content covered on this test?
  - Is bias acceptably small?
  - Do scores reflect quality of teaching versus characteristics of students and schools?
Bias vs. Noise

Systematic Error (Bias)  
*(Scoring concern)*
- + more likely for some teachers, - for others
- Cannot be reduced by averaging over more observations
- Several plausible examples

Random Error (Noise)  
*(Generalization concern)*
- +/- equally likely for any teacher
- Tends to average out over time or across classes
- Example: random variations across classes

2013 William H. Angoff Memorial Lecture
Falsification Test

- Logically, future teachers cannot influence past achievement
- Thus, if a model predicts significant effects of current-year teachers on prior-year test scores, then it is flawed or based on flawed assumptions
Falsification Test Findings

- Using each of three different VAM specifications, Rothstein (2010) found large “effects” of students’ fifth grade teachers on their fourth grade test score gains.
Falsification Test Findings

- Briggs & Domingue (2011) applied Rothstein’s test to LAUSD teacher data analyzed by Richard Buddin for the LA Times
  - For Reading, ‘effects’ from next year’s teachers were about the same as from this year’s teachers
  - For Math, ‘effects’ from next year’s teachers were about 2/3 to 3/4 as large as from this year’s teachers
Possible Sources of Bias

- Massively nonrandom assignment of students to teachers
- Teachers with special qualifications work with students with particular needs
- Nonrandom assignment of teachers to schools
- Peer effects not adequately accounted for
- Tests insensitive to growth of very low-performing or very high-performing students
- Differential summer learning loss
Summer Learning Loss

- Low-income families:
  - Summer learning loss
  - Spring-to-spring gain understates school year gain

- High-income families:
  - Summer learning gain in reading
  - Spring-to-spring gain overstates school year gain

Measured Spring-to-Spring test score gain = Spring-to-Fall (summer) loss or gain + Fall-to-Spring (school year) gain
Scoring Conclusion

- VAM scores do capture meaningful differences in teaching effectiveness, but …
- There is compelling evidence of systematic bias in teacher VAM estimates
  - Falsification test findings
  - Documented patterns of summer learning loss
- VAM scores reward or penalize teachers not only for how well they teach, but also for whom they teach and where they teach
Interpretive Argument

1. Scoring
   This class, this year, this test
   Bias?

2. Generalization
   Other classes, other years
   Random Error?

3. Extrapolation
   Other tests, non-test outcomes
   Other Measures?

4. Implication
   Validity of possible uses
   Consequences?
Reliability

- Describes score stability or consistency
- The *correlation* between two measurements indicates reliability
  - reliability (correlation) coefficient goes from zero (no linear relation) to one (perfect linear relation)

\[ r = \begin{align*}
.00 & \quad r = .50 \\
.80 & \quad r = .90 \\
1.00 & 
\end{align*} \]
Year-to-Year Changes in Ranks

Next-Year Distribution of One Year’s Bottom-Quintile Elementary Teachers, in Five Florida Counties

Next-Year Distribution of One Year’s Top-Quintile Elementary Teachers, in Five Florida Counties

Based on McCaffrey, Lockwood, Sass, & Mihaly, 2009, Table 4 (p. 591)

2013 William H. Angoff Memorial Lecture
Effectiveness Varies Year-to-Year

- Does each teacher have an “effectiveness” that is constant over time?

“Approximately one-third to one-half of the variation in teacher effects is simply due to sampling error or noise in student achievement. Of the remaining variance, between one-third and two-thirds is attributable to variation in effectiveness within teachers over time.”

McCaffrey, Sass, Lockwood, & Mihaly, 2009, p. 599
## Reliabilities from MET Project

<table>
<thead>
<tr>
<th>Test</th>
<th>Same Year, Different Course Sections</th>
<th>Different Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>State Math Test</td>
<td>0.381</td>
<td>0.404</td>
</tr>
<tr>
<td>State English Language Arts Test</td>
<td>0.180</td>
<td>0.195</td>
</tr>
<tr>
<td>Balanced Assessment in Mathematics</td>
<td>0.228</td>
<td></td>
</tr>
<tr>
<td>Stanford 9 Open-Ended Reading</td>
<td>0.348</td>
<td></td>
</tr>
</tbody>
</table>

Findings from *Measures of Effective Teaching (MET) Project*, Bill & Melinda Gates Foundation, 2010, Tables 6, 7, and 8
Averaging Over 2-3 Years Helps

<table>
<thead>
<tr>
<th>Reliability of Single-Year Data</th>
<th>Reliability of 2-Year Average</th>
<th>Reliability of 3-Year Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>.20</td>
<td>.33</td>
<td>.43</td>
</tr>
<tr>
<td>.30</td>
<td>.46</td>
<td>.56</td>
</tr>
<tr>
<td>.40</td>
<td>.57</td>
<td>.67</td>
</tr>
<tr>
<td>.50</td>
<td>.67</td>
<td>.75</td>
</tr>
</tbody>
</table>

\[ r = .40 \quad r = .57 \quad r = .67 \]
From a Teacher in Houston, TX

“I do what I do every year. I teach the way I teach every year. [My] first year got me pats on the back; [my] second year got me kicked in the backside. And for year three, my scores were off the charts. I got a huge bonus, and now I am in the top quartile of all the English teachers. What did I do differently? I have no clue.”

*Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein (2012, p. 11.)*
Interpretive Argument

1. Scoring
This class, this year, this test
Bias?

2. Generalization
Other classes, other years
Random Error?

3. Extrapolation
Other tests, nontest outcomes
Other Measures?

4. Implication
Validity of possible uses
Consequences?

2013 William H. Angoff Memorial Lecture
Extrapolation

- How do teacher VAM scores correlate with other indicators of teaching quality?
- How much do rankings change if a different student achievement test is used?
- Does achievement test content capture valued learning outcomes?
- How do VAM scores relate to valued non-test (non-cognitive) outcomes?
VAM Scores vs. Other Evidence

Teacher Description Based on Classroom Observations

“She reasons incorrectly about unit rates. She concludes that an answer of 0.28 minutes must actually be 0.28 seconds …. She tells students that integers include fractions. She reads a problem out of the text as $3/8 + 2/7$ but then writes … and solves it as $3.8 + 2.7$. She calls the commutative property the community property. She says proportion when she means ratio. She talks about denominators being equivalent when she means the fractions are equivalent.”

Hill, Kapitula, & Umland (2010, p. 820)

And she ranks in the second-to-top quartile on value-added.
## MET Project Correlations with Classroom Observations

<table>
<thead>
<tr>
<th>Subject Area</th>
<th>Classroom Observation System</th>
<th>Correlation of Overall Quality Rating with Prior-Year VAM Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematics</td>
<td>CLASS</td>
<td>0.18</td>
</tr>
<tr>
<td>Mathematics</td>
<td>FFT</td>
<td>0.13</td>
</tr>
<tr>
<td>Mathematics</td>
<td>UTOP</td>
<td>0.27</td>
</tr>
<tr>
<td>Mathematics</td>
<td>MQI</td>
<td>0.09</td>
</tr>
<tr>
<td>English Language Arts</td>
<td>CLASS</td>
<td>0.08</td>
</tr>
<tr>
<td>English Language Arts</td>
<td>FFT</td>
<td>0.07</td>
</tr>
<tr>
<td>English Language Arts</td>
<td>PLATO</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Bill & Melinda Gates Foundation (2012, pp. 46, 53)
# MET Project Correlations with Classroom Climate

<table>
<thead>
<tr>
<th>Test</th>
<th>Same Year, Same Section</th>
<th>Same Year, Different Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>State Math Test</td>
<td>.21</td>
<td>.22</td>
</tr>
<tr>
<td>State English Language Arts Test</td>
<td>.10</td>
<td>.07</td>
</tr>
<tr>
<td>Balanced Assessment in Mathematics</td>
<td>.11</td>
<td>.11</td>
</tr>
<tr>
<td>Stanford 9 Open-Ended Reading</td>
<td>.14</td>
<td>.06</td>
</tr>
</tbody>
</table>
Changes if a Different Achievement Test Is Used


“These correlations are uniformly low, … the two achievement outcomes lead to distinctly different estimates of teacher effects.”

Lockwood, et al. (2007, p. 54)
Changes if a Different Achievement Test Is Used

Papay (2007) compared VAM scores using three different reading tests, with similar results.

“Correlations between teacher value-added estimates derived from three separate reading tests … range from 0.15 to 0.58 …. if a school district were to reward teachers for their performance, it would identify a quite different set of teachers … depending simply on the specific reading assessment used.”

Papay (2011, p. 187)
MET Project Findings

Correlation between VAM scores based on two different math tests (same students, same year) = .38

Correlation between VAM scores based on two different English language arts tests (same students, same year) = .22

State Tests vs. Content Standards

- Too much memorization
- Too little complex reasoning
- Overall alignment indices low

Polikoff, Porter, & Smithson (2011)

Teaching to these tests will not foster desired range of student learning outcomes
Noncognitive Outcomes

- Work Attitudes and Behavior
  - Punctuality
  - Self-Discipline
  - Listening
  - Taking responsibility
- Motivation, Goal Setting, Planning
- Problem Solving
- Teamwork, Positive Social Behavior
- Social and Emotional Skills
- Self-monitoring/Self-regulation

Examples from Levin (2012)
80% - 85% of teachers agree that

“The POINT experiment ignores important aspects of my performance that are not measured by test scores.”

Springer, et al. (2010, p. 38)
Interpretive Argument

1. Scoring
   This class, this year, this test *Bias?*

2. Generalization
   Other classes, other years *Random Error?*

3. Extrapolation
   Other tests, nontest outcomes *Other Measures?*

4. Implication
   Validity of possible uses *Consequences?*

2013 William H. Angoff Memorial Lecture
How Should VAM Scores Be Used?

- **Inappropriate Uses**
  - Teacher VAM scores should **not** be included as a substantial factor with a fixed weight in consequential teacher personnel decisions.
  - Teacher VAM scores should **not** be included as a substantial factor in evaluations of principals.
  - Individual teachers’ VAM scores should **not** be made public.
  - VAM scores should **not** be used to compare teachers from very different sorts of schools, or working with very different student populations.
Likely Consequences of Misuse

- Increased pressure to teach to the test
- Reduced cooperation among teachers within a school
- Teacher resentment of students who struggle with academic content
- Manipulation of assignments of students to teachers
  - Includes “push-out” of low achievers
How Should VAM Scores Be Used?

• **Appropriate Uses:** For large-scale research studies
  ◦ Studying alternative teacher training programs, educational policies, curricula, etc.
    • Researchers should understand VAM well
How Should VAM Scores Be Used?

- **Appropriate** Uses: For teacher evaluation, *if*:
  - Scores based on sound, appropriate student tests
  - Comparisons limited to homogeneous teacher groups
  - *No fixed weight*—Flexibility to interpret VAM scores *in context* for each individual case
  - Users well trained to interpret scores
  - Clear and accurate information about uncertainty (e.g., “margin of error”)
Overview

- How big are teacher effects?
- A close look at test score gains
- How do VAMs work?
- Interpretive Argument
  - What do scores mean?
  - How reliable are they?
  - What do they leave out?
  - How should they be used?
- Sound teacher evaluation

2013 William H. Angoff Memorial Lecture
Sound Teacher Evaluation

- Attends to actual teaching practice
- Is research-based
- Provides constructive feedback to guide improvement
Cautions re Classroom Observation

- Requires observing multiple lessons
- Requires observer training
- May have poor reliability
- Susceptible to some of the same biases as VAM
VAM as “Trigger”?  

- Might work as stopgap, but limited…  
  - VAM scores available for only a minority of teachers  
  - Weak teachers with high VAM scores will be missed  
  - Makes teacher evaluation a remediation issue
Sound Evaluation for All Teachers

All teachers have a right to expect sound professional evaluation and opportunities for continuous improvement, at all career stages.

The work of teaching children is far too important to settle for anything less.
Thank you