Problems with the Use of Student Test Scores to Evaluate Teachers

Edward Haertel
School of Education
Stanford University

California Educational Research Association
Anaheim, California
December 1, 2011
Economists, statisticians, psychometricians, and policy experts all worked together to write this EPI Briefing Paper, released in August 2010.

Thanks to my co-authors for contributing to my own education on this important issue.
Framing the Problem

- Teacher quality is central to student success
- There is broad consensus that teacher support and evaluation need improvement
  - Teachers need better support and targeted assistance to identify and remediate deficiencies
  - Principals are challenged by the sheer number of teachers they must monitor and evaluate
  - Contracts and labor laws make teacher dismissal difficult
Framing the Problem

• Looking directly at student outcomes to judge teachers has intuitive appeal
  ◦ Test scores are already used to evaluate students and schools, why not teachers?
  ◦ Numbers appear objective and impartial
  ◦ Complex statistical models lend an aura of scientific rigor
  ◦ Value Added Models (VAMs) are actively promoted as scientific tools that can distinguish good teachers from bad
VAM Logic

If prior achievement is “held constant” by building prior-year test scores into a statistical model, then student score gains should reflect teacher effectiveness.

- The difference between last year’s score and this year’s score represents the “value added” by this year’s teacher.
Two Simplified Assumptions

- Teaching matters, and some teachers teach better than others

Simplified to

- There is a stable construct we may refer to as a teacher’s “effectiveness”
  - that can be estimated from students’ test scores
  - that can predict future performance
Two Simplified Assumptions

- Student achievement is a central goal of schooling
- Valid tests can measure achievement

Simplified to

- Achievement is a one-dimensional continuum
- Brief, inexpensive achievement tests locate students on that continuum
It’s not that simple

- Student growth is not:
  - One-dimensional
  - Steady
  - Linear
  - Influenced by the teacher alone
  - Well measured using brief, inexpensive tests
  - Independent from growth of classmates
Sorting Out Teacher Effects

- Start-of-year student achievement varies due to
  - Home background and community context
  - Individual interests and aptitudes
  - Peer culture
  - Prior teachers and schooling
  - Differential summer loss
Sorting Out Teacher Effects

- End-of-year student achievement varies due to
  - Start-of-year differences
  - Continuing effects of out-of-school factors, peers, and individual aptitudes and interests
  - Instructional effectiveness
Sorting Out Teacher Effects

- **Instructional effectiveness** reflects:
  - District and state policies
  - School policies and climate
  - Available instructional materials and resources
  - Student attendance
  - **The teacher**
Logic of the Statistical Model

- What is a “Teacher Effect”?
  - Student growth (change in test score) attributable to the teacher
  - I.e., *caused* by the teacher
Logic of the Statistical Model

Teacher Effect on One Student = Student’s Observed Score − Student’s Predicted Score

“Predicted Score” is *Counterfactual* – an estimate of what would have been observed with a hypothetical average teacher, *all else being equal*

These (student-level) “Teacher Effects” are averaged up to the classroom level to obtain an overall score for the teacher.
Value-Added Models rely on formidable statistical assumptions, unlikely to hold in the real world.
Some Statistical Assumptions

- Manipulability
- No interference between units
- Interval scale metric
- Strongly Ignorable Treatment Assignment
- Various additional assumptions re
  - functional form of model
  - rate of decay of teacher effects over time
  - other matters
Manipulability

- It is meaningful to conceive of any student being assigned to any teacher in the comparison set ...
  - without changing any of that student’s pre-enrollment characteristics.
  - Otherwise, some potential outcomes are undefined, which undermines the logical and statistical basis of the intended causal inference.
No Interference Between Units

- “Units” here are students.
- “No Interference” means a student’s end-of-year test score is not affected by which other students were assigned to the same classroom
  - Closely related to the “Stable Unit Treatment Value Assumption (SUTVA)"
Interval Scale Metric

- Effects for different teachers occur in different regions of the test score scale
  - Fair comparison requires assuming that “a point is a point is a point,” all along the scale
- Untenable due to:
  - Floor and ceiling effects on the test
  - Failure to test below- (or above-) grade-level content
Strongly Ignorable Treatment Assignment

- We must assume that once variables in the model are accounted for, assignment of students to teachers is independent of potential outcomes
  - In other words, a student with a particular set of background characteristics who is assigned to teacher X is on average no different from all the other students with that same set of background characteristics (with regard to potential end-of-year test score outcomes)
In or out?

- District leadership
- School norms, academic press
- Quality of school instructional staff
- Early childhood history; medical history
- Quality of schooling in prior years
- Parent involvement
- Assignment of pupils (to schools, to classes)
- Peer culture
- Students’ school attendance histories
- …
Controlling for prior-year score is not sufficient

- **First problem**—Measurement Error: prior-year achievement is imperfectly measured

- **Second problem**—Omitted variables: models with additional variables predict different prior-year true scores as a function of
  - additional test scores
  - demographic / out-of-school factors
Controlling for prior-year score is not sufficient

Third problem—Different trajectories: students with identical prior-year true scores have different expected growth depending on:

- individual aptitudes
- out-of-school supports for learning
- prior instructional histories
- variation in summer learning loss

Two students’ knowing the same amount of last year’s content is not the same as their being equally well prepared to make sense of this year’s instruction.
A small digression: Student Growth Percentiles

• Construction
  ◦ Each student’s SGP score is the percentile rank of that student’s current-year score within the distribution for students with the same prior-year score
Student Growth Percentiles

• Interpretation
  ◦ How much this student has grown relative to others who began at the “same” (prior-year) starting point

• Advantages
  ◦ Invariant under monotone transformations of score scale
  ◦ Directs attention to distribution of outcomes, versus point estimate
Is anything really new here?

Thanks to Andrew Ho and Katherine Furgol for this graphic
Examining the Evidence

- Stability of “effectiveness” estimates
  - *That first “simplified assumption”*

- Problems with the tests
  - *That second “simplified assumption”*

- Strongly Ignorable Treatment Assignment

- Professional consensus
Examining the Evidence

- Stability of “effectiveness” estimates
  - *That first “simplified assumption”*

- Problems with the tests
  - *That second “simplified assumption”*

- Strongly Ignorable Treatment Assignment

- Professional consensus
Stability of “Effectiveness” Estimates

- Newton, Darling-Hammond, Haertel, & Thomas (2010) compared high school math and ELA teachers’ VAM scores across:
  - Statistical models
  - Courses taught
  - Years

Findings from Newton, et al.

Sample* for Math and ELA VAM Analyses

<table>
<thead>
<tr>
<th>Academic Year</th>
<th>2005-06</th>
<th>2006-07</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math teachers</td>
<td>57</td>
<td>46</td>
</tr>
<tr>
<td>ELA teachers</td>
<td>51</td>
<td>63</td>
</tr>
<tr>
<td>Students</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 9</td>
<td>646</td>
<td>881</td>
</tr>
<tr>
<td>Grade 10</td>
<td>714</td>
<td>693</td>
</tr>
<tr>
<td>Grade 11</td>
<td>511</td>
<td>789</td>
</tr>
</tbody>
</table>

*Sample included all teachers who taught multiple courses. Ns in table are for teachers x courses. There were 13 math teachers for 2005-06 and 10 for 2006-07. There were 16 ELA teachers for 2005-06 and 15 for 2006-07.
% of Teachers Whose Effectiveness Ratings Change …

<table>
<thead>
<tr>
<th></th>
<th>By at least 1 decile</th>
<th>By at least 2 deciles</th>
<th>By at least 3 deciles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Across models*</td>
<td>56-80%</td>
<td>12-33%</td>
<td>0-14%</td>
</tr>
<tr>
<td>Across courses*</td>
<td>85-100%</td>
<td>54-92%</td>
<td>39-54%</td>
</tr>
<tr>
<td>Across years*</td>
<td>74-93%</td>
<td>45-63%</td>
<td>19-41%</td>
</tr>
</tbody>
</table>

*Depending on the model
One Extreme Case: An English language arts teacher

- Comprehensive high school
- Not a beginning teacher
- White
- Teaching English I
- Estimates control for:
  - Prior achievement
  - Demographics
  - School fixed effect
Teacher “effectiveness” bounces around from one year to the next

- Value-added estimates are extremely noisy.
- Consider classification of teachers into 5 categories (A-F) in two consecutive years.

<table>
<thead>
<tr>
<th>Grade in first year:</th>
<th>Grade in second year:</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>F</td>
<td>A</td>
</tr>
</tbody>
</table>

Average across 5 Florida districts. Grades A-F correspond to quintiles 1-5. Source: Sass (2008). Thanks to Jesse Rothstein for the original version of this slide.
Many teachers indicated as effective or ineffective in one year are not for others

- 27% of “A” teachers one year get D or F next year. 45% get C or lower.
- 30% of “F” teachers one year get A or B next year. 51% get C or better.

Average across 5 Florida districts. Grades A-F correspond to quintiles 1-5. Source: Sass (2008). Thanks to Jesse Rothstein for the original version of this slide.
Examining the Evidence

- Stability of “effectiveness” estimates
  - That first “simplified assumption”

- Problems with the tests
  - That second “simplified assumption”

- Strongly Ignorable Treatment Assignment

- Professional consensus
7th Grade History / Social Studies

WH7.8.5.

Detail advances made in literature, the arts, science, mathematics, cartography, engineering, and the understanding of human anatomy and astronomy (e.g., by Dante Alighieri, Leonardo da Vinci, Michelangelo di Buonarroti Simoni, Johann Gutenberg, William Shakespeare).
Item Testing WH7.8.5

- Artist, architect, mathematician
- Studied anatomy to draw more realistic human figures
- Painted a mural depicting the last meeting of Jesus and his disciples
- Painted the portrait known as "Mona Lisa"

The information in the chart above best describes which of these individuals of the Renaissance?

A  Raphael
B  Michelangelo
C  da Vinci
D  Botticelli
US11.11.2.

Discuss the significant domestic policy speeches of Truman, Eisenhower, Kennedy, Johnson, Nixon, Carter, Reagan, Bush, and Clinton (e.g., education, civil rights, economic policy, environmental policy).
This administration, today, here and now, declares unconditional war on poverty in America, and I urge this Congress and all Americans to join with me in that effort.

—President Lyndon B. Johnson
State of the Union Address
January 8, 1964

The program President Lyndon B. Johnson created to wage his unconditional war on poverty was the

A  Alliance for Progress.
B  Fair Deal.
C  Great Society.
D  New Deal.
9th Grade English-Language Arts

9RC2.8

**Expository Critique:** Evaluate the credibility of an author’s argument or defense of a claim by critiquing the relationship between generalizations and evidence, the comprehensiveness of evidence, and the way in which the author’s intent affects the structure and tone of the text (e.g., in professional journals, editorials, political speeches, primary source material).
Which of the following statements from the passage supports the author’s conclusion that carrier pigeons sometimes had a dangerous job?

A  In 1815 an English banker named Nathan Rothschild made his fortune by relying on messages sent to him by carrier pigeons.

B  Since they could easily be released from airplanes or ships, every branch of the armed services used the birds.

C  On his last mission, though wounded, he carried a message that saved the lives of 194 American soldiers.

D  Many people find carrier pigeons ugly because of their big wattle, a knobby buildup of skin on the beak.
25.1

Students use properties of numbers to construct simple, valid arguments (direct and indirect) for, or formulate counterexamples to, claimed assertions.
Item Testing 25.1

16. John’s solution to an equation is shown below.

Given: \( x^2 + 5x + 6 = 0 \)

Step 1: \((x + 2)(x + 3) = 0\)

Step 2: \(x + 2 = 0\) or \(x + 3 = 0\)

Step 3: \(x = -2\) or \(x = -3\)

Which property of real numbers did John use for Step 2?

A. multiplication property of equality

B. zero product property of multiplication

C. commutative property of multiplication

D. distributive property of multiplication over addition
Students know at each link in a food web some energy is stored in newly made structures but much energy is dissipated into the environment as heat. This dissipation may be represented in an energy pyramid.
57. Which of these organisms would *most* likely be found at the top of an energy pyramid?

A. clams
B. sardines
C. sharks
D. kelp
Problems With Tests Will Persist

- PARCC and SBAC assessments aligned to the CCSS should be better than most existing state assessments, but not good enough to solve these problems
  - Content standards are not all to blame
  - Testing limitations arise due to
    1. costs of some alternative item formats;
    2. inevitable differences between teaching to the test and teaching to the standards;
    3. technical challenges in measuring some key skills
Examining the Evidence

- Stability of “effectiveness” estimates
  - *That first “simplified assumption”*
- Problems with the tests
  - *That second “simplified assumption”*
- **Strongly Ignorable Treatment Assignment**
- Professional consensus
Student Assignments Affected By

- Student ability grouping (tracking)
- Teachers’ particular specialties
- Children’s particular requirements
- Parents’ requests
- Principals' judgments
- Need to separate children who do not get along
Teacher Assignments Affected By

- Differential salaries / working conditions
- Seniority / experience
- Match to school’s culture and practices
- Residential preferences
- Teachers’ particular specialties
- Children’s particular requirements
Does Non-Random Assignment Matter? A 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

- Rothstein (2010) examined three VAM specifications using a large data set and found “large ‘effects’ of fifth grade teachers on fourth grade test score gains.”
  - In addition to North Carolina, similar results have been found in Texas and Florida, as well as in San Diego and New York City.
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
Examining the Evidence

- Stability of “effectiveness” estimates
  - That first “simplified assumption”

- Problems with the tests
  - That second “simplified assumption”

- Strongly Ignorable Treatment Assignment

- Professional consensus
Professional Consensus

- We do not think that their analyses are estimating causal quantities, except under extreme and unrealistic assumptions.
  – Donald Rubin
Professional Consensus

• The research base is currently insufficient to support the use of VAM for high-stakes decisions about individual teachers or schools. – Researchers from RAND Corp.
Professional Consensus

• VAM estimates of teacher effectiveness that are based on data for a single class of students should not be used to make operational decisions because such estimates are far too unstable to be considered fair or reliable.

– 2009 Letter Report from the Board on Testing and Assessment, National Research Council
Unintended Effects

- Narrowing of curriculum and instruction
  - What doesn’t get tested doesn’t get taught
- Instructional focus on students expected to make the largest or most rapid gains
  - Student winners and losers will depend on details of the model used
- Erosion of teacher collegial support and cooperation
Valid and Invalid Uses

- **VALID**
  - Low-stakes
  - Aggregate-level interpretations
  - Background factors as similar as possible across groups compared

- **INVALID**
  - High-stakes, individual-level decisions, comparisons across highly dissimilar schools or student populations
Unintended Effects

“The most pernicious effect of these [test-based accountability] systems is to cause teachers to resent the children who don’t score well.”

—Anonymous teacher, in a workshop many years ago
Thank you

This PowerPoint will soon be available at http://www.stanford.edu/~haertel, under “Selected Presentations”