Teacher Quality and Value-added Measurement

Dan Goldhaber
University of Washington and The Urban Institute
dgoldhab@u.washington.edu
April 28-29, 2009

Prepared for the TQ Center and REL Midwest Technical Assistance Workshop:

www.crpe.org
We Know Teachers Matter!

• Controlling for family background factors, teacher quality is the single most important schooling factor explaining student achievement
  – Teacher quality can explain more than one grade-level equivalent in test performance (Hanushek, 1992)
  – Impacts of teacher quality can persist for many years (Sanders and Rivers, 1996)
  – Tremendous variation in teacher effectiveness (Bembry et al., 1998; Hanushek, 1992; Sanders and Rivers, 1996)
  – Impact of teacher quality is far larger than any other quantifiable schooling input (Goldhaber, 2002)
Teacher Quality Appears to be Primarily “Unobservable”

Source: Goldhaber et al., 1999
Teacher Quality Appears to be Primarily “Unobservable”

Source: Goldhaber et al., 1999
What Policy Debates Arise From Teacher Quantity Challenge?

• Proper role of state regulation of entry into teaching profession
  – Abel, Fordham, Darling-Hammond, Ballou and Podgursky debates

• Level and structure of teacher salaries
  – Increase teacher salaries, restructure compensation, or do both
Teacher Licensure ("Certification")

- Licensure system designed to screen out low-quality applicants
  - Completion of approved teacher training program
  - Pre- and post-licensure tests
  - Requirements vary considerably by state

- Debate over licensure system
  - Effectiveness of teachers with standard vs. alternative licensure
  - Increasing standard licensure requirements and closing of "loopholes"
    - Misses the point by ignoring the relevant alternatives for many systems
Licensure Theory

- Protects consumers (ultimately students) from poor choices
  - Localities may make poor or purposeful hiring decisions
    - Bad information or nepotism
- Limits choices of localities and may dissuade talented individuals from considering teaching
  - Localities may have better information than states over who should be hired
  - Limits labor mobility from state to state
- Problem of false negatives and positives
Hypothetical Relationship Between Teacher Licensure-Test Performance & Teacher Quality
“…We know that teachers are the most important thing, but teacher quality is not stamped on someone's forehead.”
Comparison of Teacher Effects in Math by Passing Status

![Comparison of Estimated Teacher Effects in Math by Passing Status](chart)

- Difference in mean teacher effect: 0.1832
- Teachers who pass current standard
- Teachers who fail current standard

www.crpe.org
Experience Levels

1st year mean-2nd year mean: 0.059** sd
2nd year mean-3rd year plus mean: 0.026* sd

1st year mean-2nd year mean: 0.050* sd
2nd year mean-3rd year plus mean: 0.039** sd
Degree Levels

Estimated Teacher Effectiveness in Reading by Degree Status

- Density
- Teacher Effect

- Hold Advanced Degree
- Bachelor's Degree only

Difference in means: .005 sd

Estimated Teacher Effectiveness in Math by Degree Status

- Density
- Teacher Effect

- Hold Advanced Degree
- Bachelor's Degree Only

Difference in means: .014 sd
NBPTS Certification Status

Difference in means: 0.19** sd of teacher quality
Arguments for Using VAMs to Assess Teacher Job Performance

• Teachers are the most important *schooling* factor explaining variation in student achievement, but …
  – (Easily quantifiable) teacher characteristics used to determine teachers’ employment eligibility and compensation don’t strongly predict teacher effectiveness
  – Even when there are statistically significant differences, the differences between the best and worst teachers who hold a particular credential swamp the differences between those with and without the credential

• VAMs may draw different people into teaching, thus helping to address the long-term downward trend in the academic skills of the U.S. teacher workforce
Using VAMs for Policy Purposes

• Pay, tenure, and teacher “de-selection” reforms
  – Tennessee and Dallas using individual teacher as unit of analysis
  – Pay-for-performance in Florida, Texas, and Minnesota; TIF grantee districts
  – New York City vs. New York State on student test scores
  – De-selection/selective retention ideas associated with researchers (Gorden et al., 2006; Hanushek, forthcoming)

• Underlying tenure/de-selection is the notion that teacher quality is relatively stable characteristic
But… Significant Potential Problems with Using VAMs

- Logistical issues (timing of tests; # of tested grades/subjects)
- Perverse incentives/unintended consequences (reclassification of students; too-narrow focus on tested items; discourage collaboration)
- Theoretical/practical issues measuring teacher contributions (cross-subject complements)
- Defining the constructed counterfactual (within or between school/district comparisons)
- Measurement issues/stability of teacher performance
  - Signal-to-noise ratio
  - Year-to-year changes in estimated performance
  - Sensitivity of performance ranking to changes in sample, subject, or teaching context
Thoughts on VAMs in Practice

• For policy purposes we probably don’t care about precise estimates of teacher effects
  – We care about where in the effectiveness distribution teachers fall
  – VAM estimates can be wrong, but not so wrong that they radically change the estimated teacher-effectiveness distribution
  – We don’t know much about how or whether VAM errors influence where teachers fall in the distribution

• Are we holding VAMs to a higher standard?
  – Estimates of productivity may be as imprecise and vary as much in the private sector
Focus of this Work

Assess the stability of (value-added) teacher job performance estimates over time, including a focus on pre- and post-tenure

North Carolina Data

- Administrative records for all NC teachers and students for grades 3-8 from 1995-96 to 2005-06
  - Fifth-grade performance for students with full history of test scores & in classes with 10-29 students
- Track teachers for whom we observe for at least two years pre-tenure and one year post-tenure
  - 281 unique teachers in this select sample

Analytic Approach

- \( A_{i,j,t,s,g=5} = \alpha A_{i(history)} + X_{i,t,g=5}y + \tau_{j,t,g=5} + \varepsilon_{i,j,t,s,g=5} \) where \( A_{i(history)} = \begin{bmatrix} A_{i,R,g=4} | A_{i,M,g=4} | A_{i,R,g=3} | A_{i,M,g=3} \end{bmatrix} \)
- Specification is consistent with the unbiased estimates from Kane and Staiger (2008) and the bias-minimizing specification in Rothstein (2008)
• One standard deviation increase in TQ is estimated to increase student achievement by .2 standard deviations (which is approximately 30-40% of the average yearly gain in achievement, so equivalent to about 3 months of learning)
• Variation between teachers explains 52% of overall variance in teacher effects in reading and 63% in math
• Decomposition of teacher effects shows time-varying teacher characteristics explain only a trivial proportion of the variation in the teacher effect estimates
• Average correlation of teacher job performance is 0.32 in reading and 0.54 in math
  – Estimates of stability of job performance are not terribly different from private sector estimates
Components of Estimated Year-By-Year Teacher Effects

Decomposition of Variance in Teacher Value-added Estimates in Math over Time

Variance between teachers: 62.3%
Variance within teachers: 37.7%
Unobserved changes and noise: 97.6%
Observed variation: 2.4%
Transition Matrices on Adjacent-Year Quintile Rankings

**Panel A. Reading Performance**

<table>
<thead>
<tr>
<th>Quantile in Year t</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.82</td>
<td>4.39</td>
<td>3.67</td>
<td>2.71</td>
<td>1.82</td>
<td>3,197</td>
</tr>
<tr>
<td>2</td>
<td>4.20</td>
<td>4.53</td>
<td>4.10</td>
<td>3.90</td>
<td>2.87</td>
<td>3,440</td>
</tr>
<tr>
<td>3</td>
<td>3.37</td>
<td>4.07</td>
<td>4.65</td>
<td>4.35</td>
<td>3.91</td>
<td>3,500</td>
</tr>
<tr>
<td>4</td>
<td>2.72</td>
<td>3.44</td>
<td>4.28</td>
<td>4.91</td>
<td>4.91</td>
<td>3,603</td>
</tr>
<tr>
<td>5</td>
<td>1.72</td>
<td>2.95</td>
<td>3.72</td>
<td>4.98</td>
<td>8.02</td>
<td>3,717</td>
</tr>
<tr>
<td>Total</td>
<td>3,138</td>
<td>3,358</td>
<td>3,570</td>
<td>3,619</td>
<td>3,772</td>
<td>17,457</td>
</tr>
</tbody>
</table>

**Panel B. Math Performance**

<table>
<thead>
<tr>
<th>Quantile in Year t</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7.63</td>
<td>4.73</td>
<td>3.28</td>
<td>1.91</td>
<td>0.77</td>
<td>3,213</td>
</tr>
<tr>
<td>2</td>
<td>4.78</td>
<td>5.22</td>
<td>4.55</td>
<td>3.47</td>
<td>1.67</td>
<td>3,421</td>
</tr>
<tr>
<td>3</td>
<td>3.08</td>
<td>4.34</td>
<td>4.91</td>
<td>4.61</td>
<td>3.11</td>
<td>3,551</td>
</tr>
<tr>
<td>4</td>
<td>1.86</td>
<td>3.34</td>
<td>4.79</td>
<td>5.45</td>
<td>5.21</td>
<td>3,538</td>
</tr>
<tr>
<td>5</td>
<td>0.63</td>
<td>1.60</td>
<td>2.92</td>
<td>5.30</td>
<td>10.84</td>
<td>3,734</td>
</tr>
<tr>
<td>Total</td>
<td>3,112</td>
<td>3,383</td>
<td>3,563</td>
<td>3,640</td>
<td>3,759</td>
<td>17,457</td>
</tr>
</tbody>
</table>
### Pre- and Post-Tenure Job Performance Rankings: Reading

#### Panel A. Using first two years of performance to predict post-tenure performance

<table>
<thead>
<tr>
<th>Pre-tenure Quintile Rank</th>
<th>Bottom Quintile</th>
<th>Second Quintile</th>
<th>Third Quintile</th>
<th>Fourth Quintile</th>
<th>Top Quintile</th>
<th>Total Teachers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bottom Quintile</td>
<td>32%</td>
<td>23%</td>
<td>19%</td>
<td>16%</td>
<td>11%</td>
<td>57</td>
</tr>
<tr>
<td>Second Quintile</td>
<td>27%</td>
<td>14%</td>
<td>27%</td>
<td>18%</td>
<td>14%</td>
<td>56</td>
</tr>
<tr>
<td>Third Quintile</td>
<td>21%</td>
<td>23%</td>
<td>30%</td>
<td>18%</td>
<td>7%</td>
<td>56</td>
</tr>
<tr>
<td>Fourth Quintile</td>
<td>16%</td>
<td>27%</td>
<td>18%</td>
<td>18%</td>
<td>21%</td>
<td>56</td>
</tr>
<tr>
<td>Top Quintile</td>
<td>5%</td>
<td>13%</td>
<td>5%</td>
<td>30%</td>
<td>46%</td>
<td>56</td>
</tr>
<tr>
<td>Total Teachers</td>
<td>57</td>
<td>56</td>
<td>56</td>
<td>56</td>
<td>56</td>
<td>281</td>
</tr>
</tbody>
</table>

#### Panel B. Using first three years of performance to predict post-tenure performance

<table>
<thead>
<tr>
<th>Pre-tenure Quintile Rank</th>
<th>Bottom Quintile</th>
<th>Second Quintile</th>
<th>Third Quintile</th>
<th>Fourth Quintile</th>
<th>Top Quintile</th>
<th>Total Teachers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bottom Quintile</td>
<td>26%</td>
<td>30%</td>
<td>18%</td>
<td>14%</td>
<td>12%</td>
<td>50</td>
</tr>
<tr>
<td>Second Quintile</td>
<td>28%</td>
<td>14%</td>
<td>38%</td>
<td>12%</td>
<td>8%</td>
<td>50</td>
</tr>
<tr>
<td>Third Quintile</td>
<td>26%</td>
<td>24%</td>
<td>16%</td>
<td>22%</td>
<td>12%</td>
<td>50</td>
</tr>
<tr>
<td>Fourth Quintile</td>
<td>12%</td>
<td>18%</td>
<td>22%</td>
<td>24%</td>
<td>24%</td>
<td>50</td>
</tr>
<tr>
<td>Top Quintile</td>
<td>8%</td>
<td>14%</td>
<td>6%</td>
<td>28%</td>
<td>44%</td>
<td>50</td>
</tr>
<tr>
<td>Total Teachers</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>250</td>
</tr>
</tbody>
</table>
### Panel A. Using first two years of performance to predict post-tenure performance

<table>
<thead>
<tr>
<th>Pre-tenure Quintile Rank</th>
<th>Post-tenure Quintile Rank</th>
<th>Total Teachers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bottom Quintile</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>Second Quintile</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>Third Quintile</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>Fourth Quintile</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>Top Quintile</td>
<td>56</td>
</tr>
<tr>
<td>Bottom Quintile</td>
<td>44%</td>
<td></td>
</tr>
<tr>
<td>Second Quintile</td>
<td>25%</td>
<td></td>
</tr>
<tr>
<td>Third Quintile</td>
<td>14%</td>
<td></td>
</tr>
<tr>
<td>Fourth Quintile</td>
<td>14%</td>
<td></td>
</tr>
<tr>
<td>Top Quintile</td>
<td>4%</td>
<td></td>
</tr>
</tbody>
</table>

### Panel B. Using first three years of performance to predict post-tenure performance

<table>
<thead>
<tr>
<th>Pre-tenure Quintile Rank</th>
<th>Post-tenure Quintile Rank</th>
<th>Total Teachers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bottom Quintile</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Second Quintile</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Third Quintile</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Fourth Quintile</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Top Quintile</td>
<td>50</td>
</tr>
<tr>
<td>Bottom Quintile</td>
<td>42%</td>
<td></td>
</tr>
<tr>
<td>Second Quintile</td>
<td>36%</td>
<td></td>
</tr>
<tr>
<td>Third Quintile</td>
<td>16%</td>
<td></td>
</tr>
<tr>
<td>Fourth Quintile</td>
<td>4%</td>
<td></td>
</tr>
<tr>
<td>Top Quintile</td>
<td>2%</td>
<td></td>
</tr>
<tr>
<td>Total Teachers</td>
<td>50</td>
<td></td>
</tr>
</tbody>
</table>
De-selecting Poor Performers in Either Subject
De-selecting Poor Performers in Both Subjects
Tradeoffs

- Multiple years of job performance data certainly improves reliability of estimates
  - More information & ability to use more sophisticated statistical approaches
    - But, no VAM information on first-year teachers & potential dampening of performance incentives
- Comparisons within and between schools
  - May be few good within district comparisons (in small districts) but allows districts to implement policies (sample issue)
  - Within and between school comparisons conflate school and teacher effects but effective teacher in one school might have been ineffective in another (statistical approach issue)
    - Decisions about comparisons have potentially important policy implications for level of policy implementation
      - States could assist by estimating VAMs, but leaving it up to localities to decide how to use the estimates
In the Eye of the Beholder

- Year-to-year job performance estimates are modest (0.3 in reading and 0.5 in math); pre- and post-tenure estimates are somewhat higher (0.4 in reading and 0.6 in math)
  - We can’t know whether these fluctuations represent true changes in job performance
- Inter-temporal estimates are not out of line with those found in other sectors of the economy that use them for policy purposes; and pre-tenure estimates clearly do predict estimated post-tenure performance
- More holistic assessment (complementing VAMs) would be nice, but…
  - Structural impediments to serious evaluation
  - Mistrust of subjective judgments
- How did we get here?
  - Poor evaluation/little use of evaluation today
  - Policymakers hope: VAMs are objective evaluation tool, which allows schools to do what they did not do when left to their own devices
- More research needed on using VAM to identify individual teacher effectiveness
  - Perfect can be the enemy of the good; we cannot learn all of what we need to know outside of actual policy variation
For More Detail…

- [www.crpe.org](http://www.crpe.org)
- [www.caldercenter.org](http://www.caldercenter.org)
1. Are student tests important measures of learning?
2. How should we evaluate teachers in non-tested subjects/grades?
3. What are the ways of mitigating perverse incentives/unintended consequences?
4. What are the right VAM teacher comparisons?
5. How much teacher-student information is enough to make judgments about teachers?