Evaluating Teacher Effectiveness: An Overview

Laura Goe, Ph.D.

Ohio Task Force

Columbus, Ohio ♦ January 26, 2011
Laura Goe, Ph.D.

- Former teacher in rural & urban schools
  - Special education (7th & 8th grade, Tunica, MS)
  - Language arts (7th grade, Memphis, TN)
- Graduate of UC Berkeley’s Policy, Organizations, Measurement & Evaluation doctoral program
- Principal Investigator for the National Comprehensive Center for Teacher Quality
- Research Scientist in the Performance Research Group at ETS
The goal of teacher evaluation

The **ultimate** goal of all teacher evaluation should be... TO IMPROVE TEACHING AND LEARNING
Race to the Top definition of effective & highly effective teacher

**Effective teacher**: students achieve acceptable rates (e.g., at least one grade level in an academic year) of student growth (as defined in this notice). States, LEAs, or schools must include multiple measures, provided that teacher effectiveness is evaluated, in significant part, by student growth (as defined in this notice). Supplemental measures may include, for example, multiple observation-based assessments of teacher performance. (pg 7)

**Highly effective teacher** students achieve high rates (e.g., one and one-half grade levels in an academic year) of student growth (as defined in this notice).
“When all you have is a hammer, everything looks like a nail.”
Value-added research shows that teachers vary greatly in their contributions to student achievement (Rivkin, Hanushek, & Kain, 2005)

The Widget Effect report (Weisberg et al., 2009) “…examines our pervasive and longstanding failure to recognize and respond to variations in the effectiveness of our teachers.” (from Executive Summary)
Despite Lack of Consensus, Respondents Agree Student Outcome Can Be Improved

- Many Washington Insiders focus on teachers; more than two-thirds suggest increasing teacher effectiveness (70 percent) or improving teacher training (66 percent) will improve K-12 student outcome.
- A number of respondents (62 percent) indicate that reducing the number of students per class would translate to better results for K-12 students in the United States.

<table>
<thead>
<tr>
<th>Ways to Improve K-12 Education Outcomes*</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increase teacher effectiveness</td>
<td>70%</td>
</tr>
<tr>
<td>Improve teacher training</td>
<td>66%</td>
</tr>
<tr>
<td>Reduce class size</td>
<td>62%</td>
</tr>
<tr>
<td>Implement college and career ready standards</td>
<td>37%</td>
</tr>
<tr>
<td>Improve student assessment tools</td>
<td>30%</td>
</tr>
<tr>
<td>Increase spending per student</td>
<td>21%</td>
</tr>
<tr>
<td>Other</td>
<td>21%</td>
</tr>
<tr>
<td>None of the above</td>
<td>2%</td>
</tr>
</tbody>
</table>

*In your opinion, which of the following will significantly improve student outcome as it pertains to K-12 education in the United States. Please select all that apply.
In general, Washington Insiders believe that improved resources and more comprehensive evaluations are the best ways to improve teacher effectiveness. More than half of respondents identified “improve curriculum and resources” (68 percent), “more comprehensive evaluation process” (62 percent), or “enhance evaluation process based on in-class observations” (57 percent) as ways to improve teacher effectiveness.

Only 34 percent of respondents believe that increasing certification or training requirements is likely to improve teacher effectiveness.

* Ways to Improve Teacher Effectiveness*

<table>
<thead>
<tr>
<th>Approach</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improve curriculum and resources</td>
<td>68%</td>
</tr>
<tr>
<td>More comprehensive evaluation process (i.e. multiple measures)</td>
<td>62%</td>
</tr>
<tr>
<td>Enhance evaluation process based on in-class observations</td>
<td>57%</td>
</tr>
<tr>
<td>Base compensation on teacher effectiveness</td>
<td>51%</td>
</tr>
<tr>
<td>Improve access to data on student outcome and progress</td>
<td>50%</td>
</tr>
<tr>
<td>Increase compensation</td>
<td>49%</td>
</tr>
<tr>
<td>Enhance evaluation process based on measures of student improvement</td>
<td>47%</td>
</tr>
<tr>
<td>Increase certification/retraining requirements</td>
<td>34%</td>
</tr>
<tr>
<td>Other</td>
<td>6%</td>
</tr>
<tr>
<td>None of the above</td>
<td>0%</td>
</tr>
</tbody>
</table>

* In your opinion, which of the following is likely to improve teacher effectiveness? Please select all that apply.
Multiple measures of *teacher effectiveness*

- **Evidence of *growth in student learning and competency***
  - Standardized tests, pre/post tests in untested subjects
  - Student performance (art, music, etc.)
  - Curriculum-based tests given in a standardized manner
  - Classroom-based tests such as DIBELS

- **Evidence of *instructional quality***
  - Classroom observations
  - Lesson plans, assignments, and student work

- **Evidence of *professional responsibility***
  - Administrator/supervisor reports
  - Surveys of students and/or parents
  - An “evidence binder” created & presented by the teacher
Multiple measures of student learning

- Evidence of growth in student learning and competency
  - Standardized assessments (state/district tests)

- Evidence collected by teachers and scored by groups of educators
  - The 4 Ps: portfolios, projects, products, and performances
  - Essays, written responses to complex questions

- Evidence collected and scored in classrooms
  - Classroom-based assessments such as DRA, DIBELS, curriculum-based tests, unit tests
FOR A FAIR SELECTION EVERYBODY HAS TO TAKE THE SAME EXAM: PLEASE CLIMB THAT TREE
From “Race to the Top” and reiterated in the August 5, 2010 Federal Register (Vol. 75, No. 150) “Secretary’s Priorities for Discretionary Grant Programs”

- Teachers should be evaluated using state standardized tests where possible
- For non-tested subjects, other measures (including pre- and post-tests) can be used but must be “rigorous and comparable across classrooms” and must be “between two points in time”
- Multiple measures should be used, such as multiple classroom evaluations
Student growth means the change in student achievement (as defined in this notice) for an individual student

- between two points in time

A state may also include other measures that are

- rigorous
- comparable across classrooms (pg 11)
Potential Definition

Rigorous

• Rigorous measures may mean that high expectations for student progress towards college- and career-readiness are exhibited
  ▪ Measure designed to measure students’ mastery of grade-level standards for that subject.
Between two points in time

- May mean assessments that occur as close as possible to the beginning and end of a course, so that the maximum growth towards subject/grade standards can be shown.

Examples:
- Pre and post test given in a particular course
- Student portfolio demonstrating mastery of standards-based knowledge and skills (measured over time)
Frequently used growth models

- **Value-added models (requires prediction)**
  - There are many versions of value-added models (VAMs), but results from the different models are quite similar.
  - Most states and districts that use VAMs use the Sanders’ model, also called TVAAS.
  - Prior test scores (3+ years in the Sanders’ model) are used to predict the next test score for a student.

- **Colorado Growth model (no prediction needed)**
  - Focuses on “growth to proficiency”
  - Measures students against “academic peers”
Sample student report: Colorado Growth Model

Slide courtesy of Damian Betebenner at www.nciea.org
Why growth models are better than status models (1)

In terms of value-added, Teachers A and B are performing equally.

Teacher A: “Success” on Ach. Levels
Teacher B: “Failure” on Ach. Levels

Slide courtesy of Doug Harris, Ph.D, University of Wisconsin-Madison
Why growth models are better than status models (2)

A teacher with low-proficiency students can still be high value-added (and vice versa)

Achievement

Proficient

High Ach. Level, Low Value-Added

Low Ach. Level, High Value-Added

Start of School Year

End of Year

Slide courtesy of Doug Harris, Ph.D, University of Wisconsin-Madison
Teacher effects

Grades
3 4 5 6

Student Test Scores

Student-teacher links

VAM

Teacher Effects

A B C D E

100 80 50 30 10
Classroom effects

Grades
3 4 5 6

Student Test Scores

Student-teacher links

VAM

Classroom Effects

A  B  C  D  E

100 80 50 30 10
What Value-Added Models Cannot Tell You

- Value-added models are really measuring classroom effects, not teacher effects
- Value-added models can’t tell you why a particular teacher’s students are scoring higher than expected
  - Maybe the teacher is focusing instruction narrowly on test content
  - Or maybe the teacher is offering a rich, engaging curriculum that fosters deep student learning.
- *How* the teacher is achieving results matters!
Cautions about using value-added for teacher evaluation

- Braun et al. (2010) provides some useful definitions and a good review of research; notes that most researchers are not comfortable with using VAMs as the sole measures of teacher effectiveness.

- Schochet & Chiang (2010) “Type I and II error rates for comparing a teacher’s performance to the average are likely to be about 25 percent with three years of data and 35 percent with one year of data.”
Considerations in using value-added for teacher evaluation

- Koedel & Betts (2009) suggest using multiple years of data for teacher evaluation to mitigate sorting bias; novice teachers cannot be evaluated under this system.

- McCaffrey et al. (2009) “...there are significant gains in the stability [of teachers’ value-added scores] obtained by using two-year average performance measures rather than single-year estimates”
In one state, 69% of teachers (Prince et al., 2006) could not be accurately assessed with VAMs

- Teachers in subject areas that are not tested with annual standardized tests
- Teachers in grade levels (lower elementary) where no prior test scores are available
- Questions about the validity of measuring special education teachers and ELL teachers with VAMs

May be higher in many states—75-80%
Specify the elements of the evaluation system such as

- How often teachers are to be evaluated
- Who will collect and evaluate evidence on teachers’ performance
- What training will be required in order to conduct observations and evaluate other types of evidence
- Whether student learning results (such as standardized tests) will be used as a component of teachers’ scores

(continued, next slide)
Models (continued)

• What percentage of a teachers’ total scores will be based on student achievement vs. other measures

• State teaching standards to be used to guide measurement of performance

• Several levels for teacher performance (such as “highly effective,” “effective,” or “ineffective”)

• Consequences for failure to meet acceptable performance levels, such as referral to a Peer Assistance and Review (PAR) program

• Which categories of measures or which specific measures are to be used, i.e., observations (category) or Charlotte Danielson’s Framework (specific)
Questions to ask about models

- Do they include measures that are “rigorous and comparable across classrooms”? 
- Do they include measures that show learning growth “between two points in time”? 
- Do they include measures that are aligned with and focused on grade level and subject standards? 
- Do they allow teachers from all subjects to be evaluated with evidence of student learning growth? 
- Will using this model help improve teaching and learning?
References


http://www.nap.edu/catalog.php?record_id=12820


http://www.mitpressjournals.org/doi/abs/10.1162/edfp.2009.4.4.572


Race to the Top Application

http://www2.ed.gov/programs/racetothetop/resources.html
References (continued)


Questions?
Laura Goe, Ph.D.
P: 609-734-1076
E-Mail: lgoe@ets.org

National Comprehensive Center for Teacher Quality
1100 17th Street NW, Suite 500
Washington, DC 20036-4632
877-322-8700 > www.tqsource.org