



NATIONAL COMPREHENSIVE CENTER
FOR **TEACHER QUALITY**

Measuring Teacher Effectiveness

Laura Goe, Ph.D.
Research Scientist, ETS

April 28, 2009

Copyright © 2008 National Comprehensive Center for Teacher Quality. All rights reserved.



Measuring teacher effectiveness

- Measuring teacher effectiveness is impacted by
 - What is valued
 - Our technological advances and limitations
 - The data, evidence, and information we have or can acquire
 - The resources (staff, money, time, policy levers) available to us
 - The cooperation of the teachers themselves

What is valued

- An examination of the research on teacher effectiveness suggests that it is usually defined in terms of teachers' contributions to students' learning *as measured by test scores*
- This unfortunately means that other ways teachers contribute to the growth, well-being, and character of students or to the culture and stability of the school are often not measured at all, and may be given too little consideration

Measuring teacher effectiveness

- If we adopt a definition of teacher effectiveness that goes *beyond* teachers' contribution to student achievement growth, we find a dearth of valid, reliable measures
- Whatever we value (which may vary at the SEA and LEA level) will drive the development of measures
- Testing the validity and reliability of new measures is challenging!

Other measures of teacher effectiveness

- Many types of evidence*—including classroom observations, portfolios, administrator recommendations, analysis of teachers' assignments, analysis of students' work, documentation of teachers' positive contributions to the school, documentation of teacher leadership and mentoring—can be used in addition to student test scores

*For descriptions and discussions of instruments for measuring various aspects of teacher performance, see Goe, Bell & Little (2008)

Commonly used measures

- **Value-added:** of little value in helping teachers improve their practice, since value added scores *tell us nothing about what goes on in teachers' classrooms*
- **Teacher observations:** great for formative evaluation, but are more expensive to conduct (personnel time, training, calibrating) and *are only as good as the instruments and the observers*

Using value-added models

- Bill Sanders developed TVAAS in the '80s and has sold the technology to many states
- He contends that it allows districts to distinguish among good and bad teachers
- Teachers are ranked within a district
 - High-ranking teachers' students did much better on the state test than their previous test scores would have predicted
 - Low-ranking teachers' students did much worse

Drawbacks to using VAMs to evaluate teacher effectiveness

- VAMs are really measuring *classroom* effects, not *teacher* effects
- VAMs 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!*

Data requirements for VAMs

- To use VAMs, you must have the following
 - Unique identifiers for each *teacher*
 - Unique identifiers for each *student*
 - A link between the student and each of his or her teachers in the data system
 - Accurate, complete data going as far back as possible, linking students with teachers
 - Student achievement scores for several years (to be used to predict next year's score)

VAMs don't measure all teachers

- Many teachers cannot be accurately evaluated for “effectiveness” using VAMs
 - Teachers in subject areas that are not regularly tested (science, social studies, arts)
 - Teachers in grade levels (lower elementary) where no prior test scores are available
 - Teachers with fewer students
 - Special education teachers
- And VAMs don't work well for subjects where tests are not vertically aligned (science)

Questions about validity of VAMs for measuring teacher effectiveness

- Many researchers have raised questions about the models used and whether they are accurately measuring teachers' contributions to student achievement growth (see Goe, Bell & Little for issues and references to research)

Goe, Bell, & Little (June, 2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.
<http://www.tqsource.org/publications/teacherEffectiveness.php>

Teacher observations

- There are a number of good teacher observation instruments (Charlotte Danielson's, CLASS, Praxis III)
- But even if you have great instruments, the usefulness of the scores for *high-stakes evaluation* depends on
 - How well trained the observers are
 - How many times the teacher is observed
 - Whether calibration and retraining occurs

Validity of observations

- **Validity:** The extent to which evidence and theory support an interpretation of scores *for the use of the measure*, i.e., a higher standard for evidence should be used when high stakes decisions are associated with scores
- Do higher observation scores mean more effective teaching? The research base shows mixed results

Instrumentation is key to validity

➤ Potential problems with observation instruments

- The instrument doesn't measure what is *important and valued*, given the purpose of the observation
 - What is being measured does not match with what *teachers* believe to be important
 - What is being measured does not match with what *observers/raters/evaluators* believe to be important

Training is important to validity

- Raters are poorly trained in the use of the observation instrument
 - In order for scores to be valid for any purpose, and particularly for teacher compensation, *raters must all be scoring in the same way*
 - In other words, a teacher should get virtually the same score, no matter which rater does the observation
 - Proper training ensures that each rater looks for the *same evidence and scores it in the same way*

Calibration is important to validity

Raters have “drifted,” meaning the raters have forgotten how to use the instrument properly

- Including extraneous evidence
- Neglecting to include important evidence
- Allowing personal bias to creep in

➤ Result: the rater may be too “hard” or too “easy”

➤ Rater calibration means evaluating raters’ scores and retraining when appropriate

Measuring teacher effectiveness: System design suggestions

1. Involve teachers and stakeholders in developing the evaluation system
 - Increases teacher/stakeholder buy-in
 - Increases *validity* of the system
2. Use *multiple* indicators, not just an observation score or a value-added rank
 - There are many other important things you can measure economically

Suggestions (continued)

3. Differentiate among teachers by grade level and experience
 - Standards may be the same, but what constitutes effectiveness may be different
4. Use appropriate weights to give more importance to components of the system that are most important, i.e., some aspects of teacher effectiveness may be more important to the SEA or LEA than others

Suggestions (continued)

5. Measure what's most important to you, your administrators, your teachers, and other education stakeholders
 - The system will *drive improvement*
 - Teachers will strive to improve in areas that they know will count
 - Ensure that what they are striving for is truly important in your definition of successful teaching

Suggestions (continued)

6. Give teachers opportunities to improve in areas where they score poorly
 - Provide assistance in determining problem areas and planning strategies to address them
 - Provide resources (time, dollars, people) to help them improve
 - Mentors
 - Targeted professional development
 - Subsidized college courses

Suggestions (continued)

7. For observations:

- Set aside funds to support training and calibrating of observers
- Devise a system that involves multiple observations over the year, not just one
- Use at least two different evaluators for each teacher and average the scores

Suggestions (continued)

8. For value-added:

- Examine vertical alignment
- Look for “ceiling effects”
- Consider what to do about evaluating teacher effectiveness for the 50% or more of teachers who cannot be accurately evaluated with VAMs

9. Triangulate data:

- Do value-added scores correlate with observation scores and other measures?

Summary

- There are no shortcuts to measuring teacher effectiveness
- Design a system that *drives* effective instruction, not just measures it
- With only one indicator, there is a greater potential for misidentifying effective teachers
- Include other outcomes in your teacher evaluation system, not just value-added scores or observation scores



Laura Goe, Ph.D.

P: 609-734-1076 >

E-Mail: lgoe@ets.org