



Teacher Evaluation in Transition: Using Evaluation to Improve Teacher Effectiveness

Laura Goe, Ph.D.

Delaware Department of Education

Dover, DE ◆ 6/16-17/2010

Copyright © 2009 National Comprehensive Center for Teacher Quality. All rights reserved.



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 Learning & Teaching Research Center at ETS

National Comprehensive Center for Teacher Quality (the TQ Center)

A federally-funded partnership whose mission is to help states carry out the teacher quality mandates of ESEA

- Vanderbilt University
 - Students with special needs, at-risk students
- Learning Point Associates
 - Technical assistance, research, fiscal agent
- ETS

The goal of teacher evaluation

The ultimate goal of all teacher evaluation should be...

**TO IMPROVE
TEACHING AND
LEARNING**

How do we measure teacher effectiveness?

➤ "It's a hard nut to crack. The things that are easy to measure don't matter that much, and the things that matter aren't easy to measure."

- Adam Gamoran, interim Dean at the University of Wisconsin School of Education, talking to Teacher Magazine in December 2008



How we measure teacher effectiveness is impacted by various factors

- What is valued
- Our technological advances and limitations
- The data, evidence, and information we have or can acquire
- Availability of rubrics with demonstrated validity
- Understanding of what it takes to do evaluation rigorously
- The resources (staff, money, time, policy levers) available to us
- The cooperation of the teachers themselves
- Our motivation for measuring effectiveness

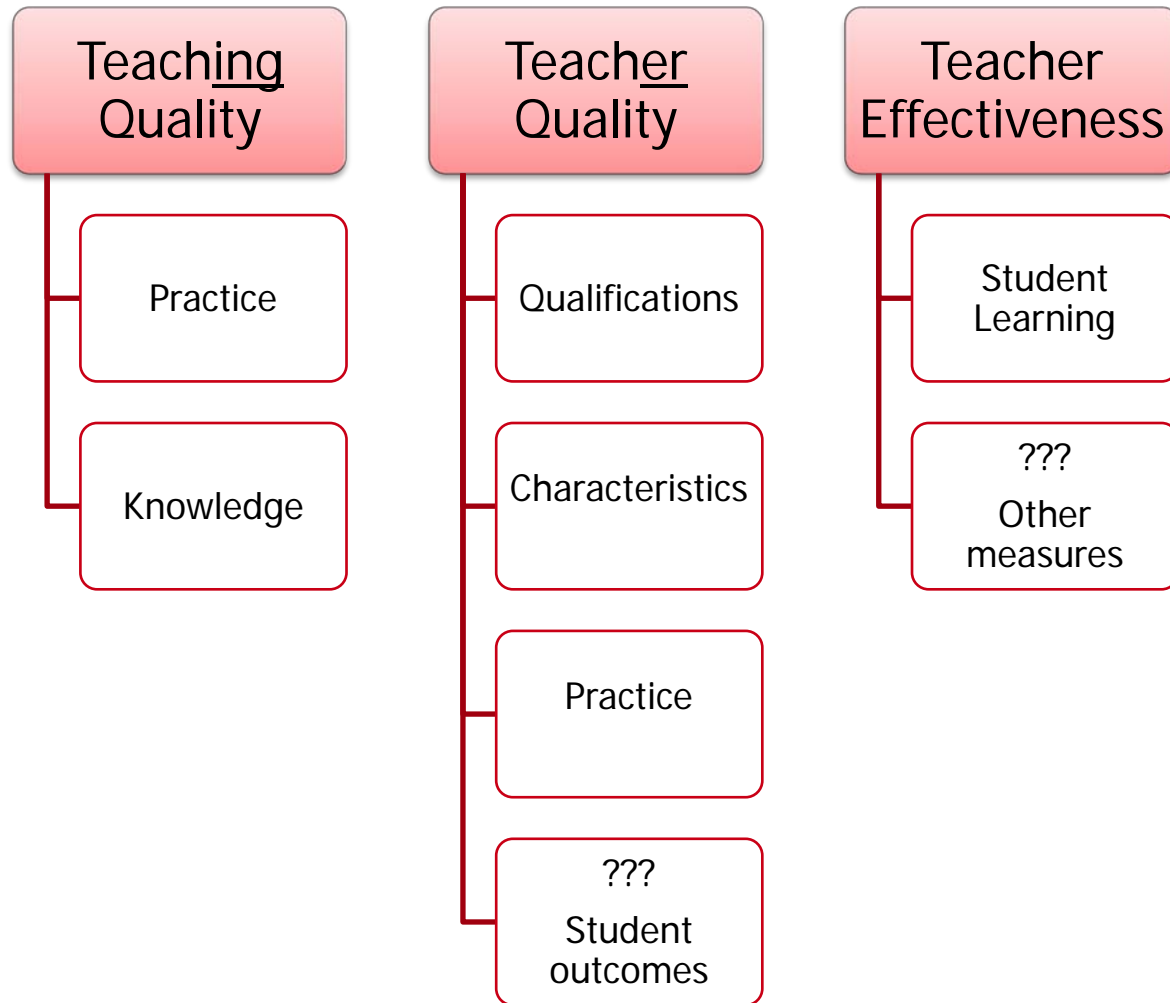
Keys to Measuring teacher Effectiveness

- Measure what is *required* (i.e., federal/state legislation and incentives)
- Measure what is *valued* (i.e., all the things we expect teachers to do)
- Develop and make available to teachers and evaluators the standards by which teachers will be evaluated
- Familiarize teachers with tools and processes of the evaluation
- Measure performance against the standards

The quandary

- Tension between federal (and sometimes state) pressures and teacher (and sometimes union) beliefs about what should constitute measures and evidence
- Teachers want a greater focus on the job they are doing (instructional quality)
 - Assumes shared responsibility for student learning
- Federal pressures are towards a greater focus on outcomes (student learning)
 - Assumes that teachers are primarily responsible for student learning

Terminology for this presentation



What the evidence says about teacher quality (Goe, 2007)

- **Experience** matters, but only for the first five years or so as teachers learn on the job; After that, experience adds little in terms of student achievement
- Teachers' **subject matter knowledge** (as evidenced by course-taking) appears to contribute significantly to math achievement, particularly at the secondary level, but research has not convincingly demonstrated that it matters in other subjects
- Subject matter **certification** contributes significantly to math achievement, but is not significantly and consistently related to student achievement in other subjects

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).

Emphasis on student achievement

- Teacher effectiveness is often discussed by researchers and politicians solely in terms of teachers' contributions to students' learning as measured by test scores
- This unfortunately means that other ways teachers contribute to student learning and well-being or to the culture and stability of the school are often not measured at all or may be given little consideration

Goe, Bell, & Little (2008) definition of teacher effectiveness

1. Have high expectations for all students and help students learn, as measured by value-added or alternative measures.
2. Contribute to positive academic, attitudinal, and social outcomes for students, such as regular attendance, on-time promotion to the next grade, on-time graduation, self-efficacy, and cooperative behavior.
3. Use diverse resources to plan and structure engaging learning opportunities; monitor student progress formatively, adapting instruction as needed; and evaluate learning using multiple sources of evidence.
4. Contribute to the development of classrooms and schools that value diversity and civic-mindedness.
5. Collaborate with other teachers, administrators, parents, and education professionals to ensure student success, particularly the success of students with special needs and those at high risk for failure.

Logic Model 1: Improving student achievement by focusing on accountability and rewards for teachers

Evaluate teacher performance using multiple measures



Use results for performance awards or referral to support services



Teaching and learning improves
because of awards or services

Logic Model 2: Improving student achievement by focusing on improving instructional quality

Evaluate teacher performance using multiple measures



Using results, help all teachers identify areas where they can improve instruction



Provide teachers with time, resources, and support to improve instruction



Teacher instructional performance improves with access to time, resources, and support



As instruction improves, student learning improves—reflected in standardized tests

Questions about Part 1?



Validity in measurement

- Validity: The extent to which evidence and theory support an interpretation of scores for a particular use of the measure
 - We want to know: Is our measure telling us how teachers perform in the areas we care about? What are appropriate uses of these results?
 - Instruments (such as tests) do not “have” validity
 - Validity lies in how well the instrument measures the domain we care about and how the results are used

Measures of teacher effectiveness

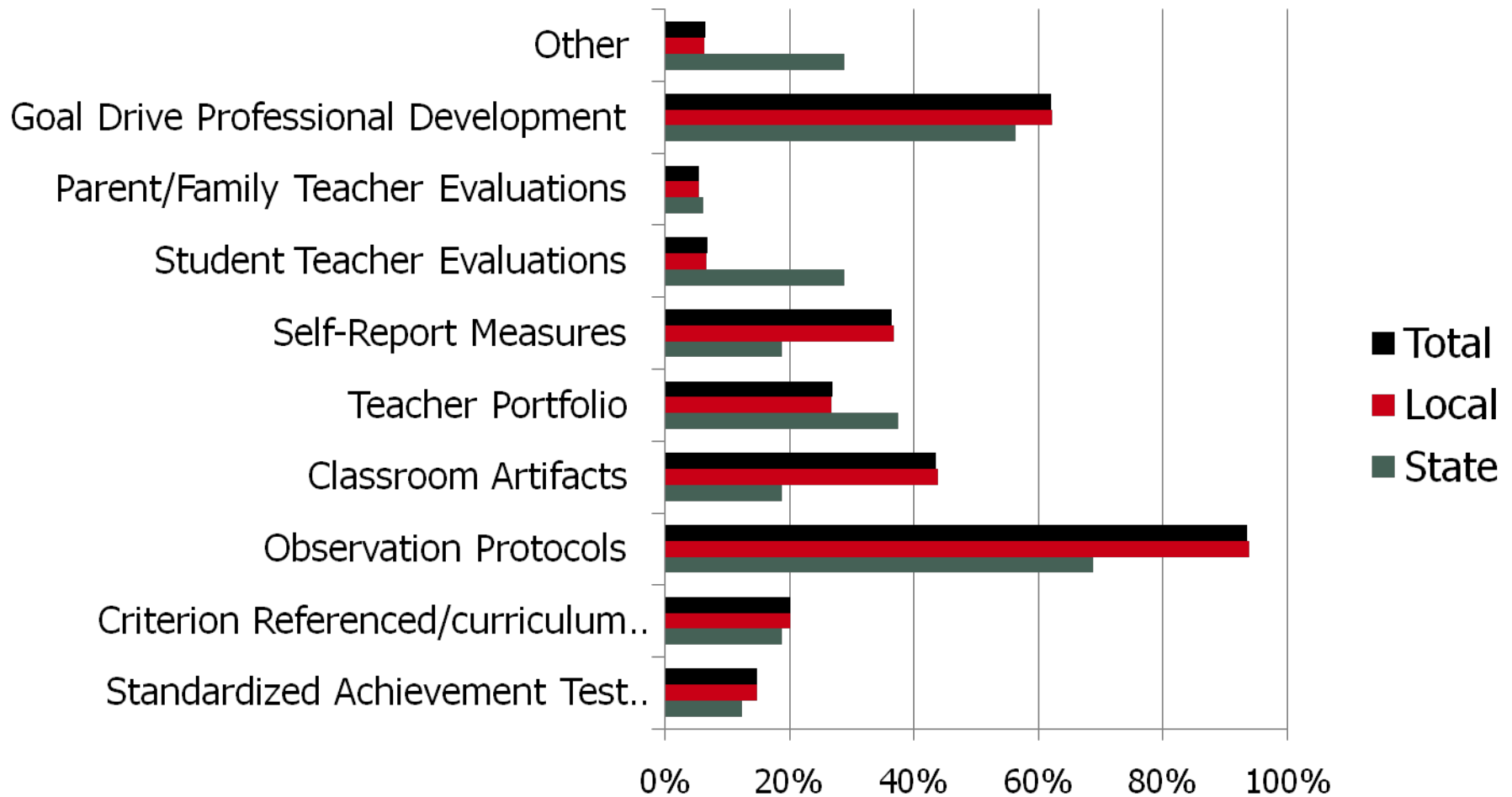
- **Evidence of growth in student learning and competency**
 - Standardized tests, pre/post tests in untested subjects
 - Student performance (art, music, etc.)
 - Curriculum-based tests given in a standardized manner
 - Classroom-based tests such as DIBELS
- **Evidence of instructional quality**
 - Classroom observations
 - Lesson plans, assignments, and student work
- **Other evidence (varies, based on local values)**
 - Administrator/supervisor reports
 - Surveys of students and/or parents
 - An “evidence binder” created & presented by the teacher

Interpretation of evidence

- Is the instrument being used for the same purposes for which it was designed?
- Does the instrument capture what it is intended to, or is it biased by factors unrelated to teaching?
- Do the interpretations being drawn from the scores go beyond what the instrument is actually able to measure?

Measurement Instruments

(from Holdheide et al., 2010)



Contrasting measures of teacher effectiveness

➤ Value-Added Models

- Becoming increasingly common
- Considered “more objective”
- Of little value in helping teachers improve their practice because value-added scores tell us nothing about what goes on in teachers’ classrooms

➤ Teacher Observations

- Great for formative evaluation but expensive to conduct (personnel time, training, calibrating)
- Only as good as the instruments and the observers
- Considered “less objective”

Example: University of Virginia's CLASS observation tool

	Emotional Support	Classroom Organization	Instructional Support
Pre-K and K-3	<p>Positive Climate</p> <p>Negative Climate</p>	<p>Behavior Management</p> <p>Productivity</p>	<p>Concept Development</p> <p>Quality of Feedback</p> <p>Language Modeling</p>
Upper Elementary/ Secondary	<p>Teacher Sensitivity</p> <p>Regard for Student (Adolescent) Perspectives</p>	<p>Instructional Learning Formats</p>	<p>Content Understanding</p> <p>Analysis and Problem Solving</p> <p>Quality of Feedback</p>

Example: Charlotte Danielson's Framework for Teaching

Domain 1: Planning and Preparation

includes comprehensive understanding of the content to be taught, knowledge of the students' backgrounds, and designing instruction and assessment.

Domain 3: Instruction is concerned with the teacher's skill in engaging students in learning the content, and includes the wide range of instructional strategies that enable students to learn.

Domain 2: The Classroom

Environment addresses the teacher's skill in establishing an environment conducive to learning, including both the physical and interpersonal aspects of the environment.

Domain 4: Professional

Responsibilities addresses a teacher's additional professional responsibilities, including self-assessment and reflection, communication with parents, participating in ongoing professional development, and contributing to the school and district environment.

Other Measures

- Many types of evidence*—including portfolios, administrator recommendations, analysis of teachers' assignments, analysis of students' work, documentation of teachers' positive contributions to the school, student and parent reports, and 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, and Little (2008).

Evidence binders

- Teachers collect and organize evidence that demonstrates their proficiency and/or indicates progress in
 - Classroom practice
 - Professional/out-of-class activities
 - Student learning linked to teacher practice
 - Teacher assignment + student work
 - Teacher assignment + set of student work showing growth
 - Pre- and post-test scores showing student progress
 - DIBELS and other classroom-based tests

Measures that improve teaching and learning

- Value-added models (such as Sanders' model) are too far away from the classroom to provide *actionable* evidence to improve learning
- Standardized test scores (broken down by domains) can tell you a lot about students' knowledge and skills *but little about teaching*
- *However, student learning improves because of what teachers do in the classroom*
- Thus, it is necessary to evaluate *teaching practice* in order to improve student learning

Growth-oriented evaluation

- Many evaluation systems currently in use ignore growth opportunities for teachers who are “doing fine”
- For teachers who are struggling, the “help” may be seen as punitive rather than as creating opportunities for teachers to improve practice
- *But some measures of teacher performance are far more useful than others in helping teachers to improve their practice, which will in turn improve student learning*

Measures that help teachers grow

- Measures that motivate teachers to examine their own practice against specific standards
- Measures that allow teachers to participate in or co-construct the evaluation (such as “evidence binders”)
- Measures that give teachers opportunities to discuss the results with evaluators, administrators, colleagues, teacher learning communities, mentors, coaches, etc.
- Measures that are directly and explicitly aligned with teaching standards
- Measures that are aligned with professional development offerings
- Measures which include protocols and processes that teachers can examine and comprehend

Growth opportunities for *all* teachers

The point is *not* that accountability systems lack value. They serve an important purpose. But alone they touch too few teachers. We need evaluation systems that promote the development of *all* teachers, not just those having difficulty. We need teacher evaluations that help and encourage the tenured teacher to perform to maximum capabilities. In addition, we need evaluations that help the outstanding teacher—the virtuoso performer—to (a) use his or her strengths to maximum efficiency *and* (b) share these strengths with other teachers.

Duke, DL; Stiggins, RJ. (1986.) Teacher Evaluation: Five Keys to Growth. West Haven, CT: National Education Association. ERIC # ED275069 (full text, pg 15)

How can schools use evaluation results to improve teacher effectiveness?

Poor

- Evaluator (administrator, peer, district evaluator) gets test scores, does evaluation → puts results in a file cabinet and forgets them

- Discusses results with the teacher → then puts them in a file cabinet and forgets them

- Discusses results with the teacher → they collaboratively decide on a development plan

- Discusses → development plan → provides the teacher with the necessary support, resources, and time to carry out the plan

Best

- Discusses → development plan → support, resources, and time → accountability for progress (for *both* teachers & administrators)

AFT's Innovation Grant work

- Working with New York and Rhode Island to develop Comprehensive Teacher Evaluation Systems that include measures of student learning growth
- District teams (including union representatives, superintendents, and administrators) work with experts to learn about measuring teacher performance
 - Rhode Island requires that 51% of teacher evaluation be based on student achievement growth; RIDE decides measures
 - New York is just developing statewide teacher standards; recently decided that 40% of teacher evaluation must be based on student learning growth, including 20% standardized test scores
 - Incorporating student learning growth is biggest challenge

Practical issues/challenges

- Problem: Administrator doesn't have time to evaluate and follow-up with teachers
 - Solution: Administrator doesn't need to do it all; mentors, partners, teams members, master teachers, consulting teachers, etc. can provide support
- Problem: Lack of resources to support teacher growth
 - Solution: Restructure school day or week to create opportunities for teachers to work together (observe/ be observed, lesson study, RTI, etc.)
 - Solution: Use professional development dollars to support professional development *aligned with evaluation results*
 - Solution: Use classrooms of master teachers as "labs" for teachers to observe effective practices

Challenges in implementation

- Context: politics and policies
- Teacher buy-in
- Valid measures & instruments
- Resources to support implementation
- Time pressure (i.e., must have system in place by a certain date)
- Research base
- Models of good CTES

Questions about Part 2?



Race to the Top definition of student achievement

Student achievement means—

- (a) For tested grades and subjects: (1) a student's score on the State's assessments under the ESEA; and, as appropriate, (2) other measures of student learning, such as those described in paragraph (b) of this definition, provided they are rigorous and comparable across classrooms.
- (b) For non-tested grades and subjects: alternative measures of student learning and performance such as student scores on pre-tests and end-of-course tests; student performance on English language proficiency assessments; and other measures of student achievement that are rigorous and comparable across classrooms.

Race to the Top definition of student growth

- **Student growth** means the change in student achievement (as defined in this notice) for an individual student between two or more points in time. A State may also include other measures that are rigorous and comparable across classrooms. (pg 11)

Using standardized tests to measure teacher effectiveness

- Standardized tests were designed to measure student learning—and they are valid for that purpose, but they were not designed or validated for evaluating teachers' contributions to that learning

Growth models vs. status models

- A “growth model” looks at students’ progress from grade to grade
 - Aggregated student scores are use for school accountability
- A “status model” looks at students’ current proficiency
 - The current proficiency of this year’s 5th graders it compared to last year’s 5th graders
- Both are used for *school-level* accountability

Two Ways VAMs May Be Used for Accountability

➤ *School Focus (state and district level)*

- Identification for “needs improvement”
- Rewards or sanctions
- Could be used for NCLB purposes

➤ *Teacher Focus (district level)*

- One measure in a teacher evaluation system
- Pay for Performance
- Promotion/dismissal

One type of measurement: student achievement gains

- Value-added models (VAMs) are the most sophisticated way to measure student growth
 - VAMs are a version of growth models
 - There are many versions, but results from the different models are generally similar
 - Prior test scores (at least 3 years in the Sanders' model) are used to predict the **next** test score for a student
 - If average student performance in a classroom is better than predicted, they can be said to have an effective teacher

Using value-added models

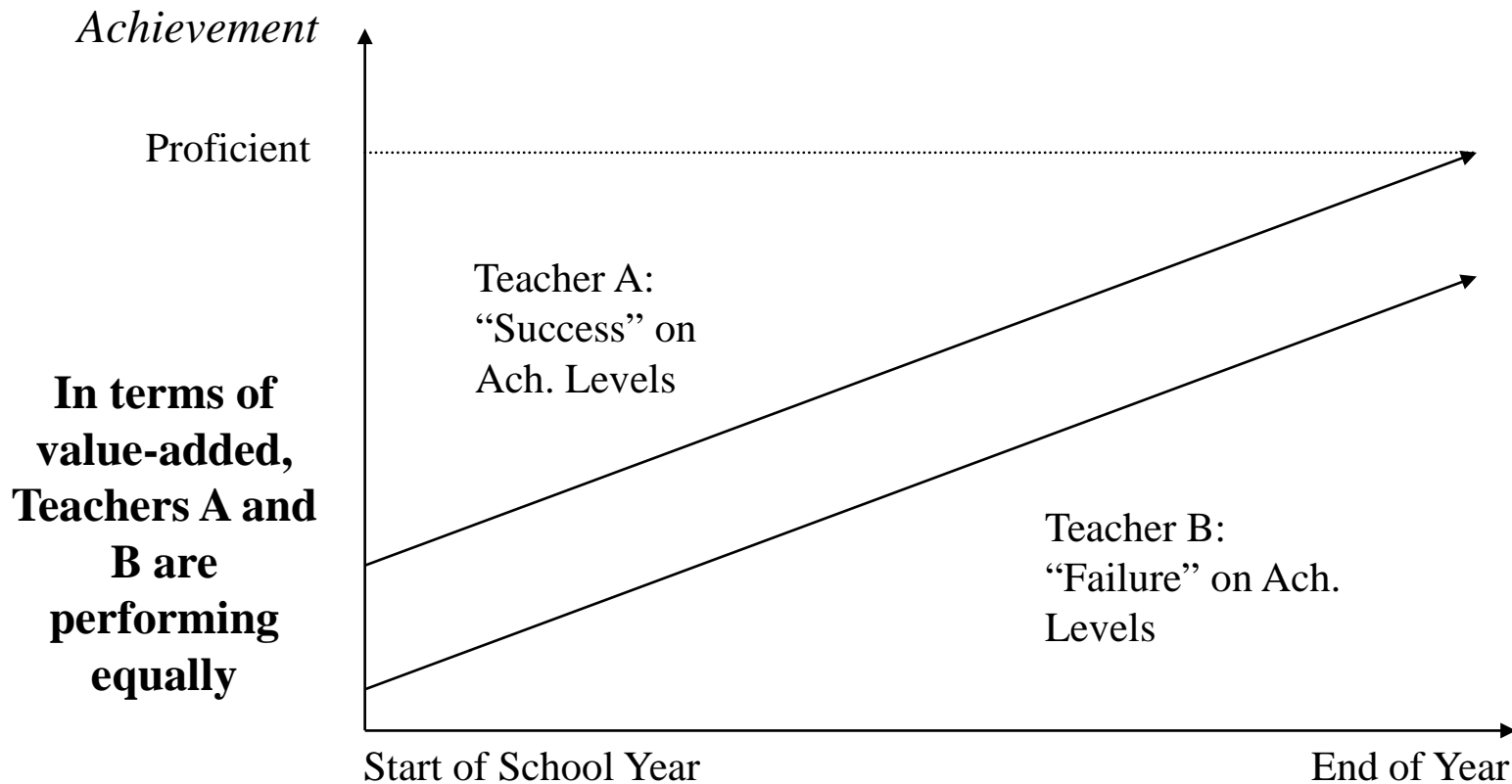
- Bill Sanders, who was an agricultural researcher, developed TVAAS and has sold the technology to a number of states
- He contends that it allows districts within states to distinguish among good and bad teachers
- Teachers are ranked within a district
 - High-ranking teachers' students did much better on the state test than students' previous test scores would have predicted
 - Low-ranking teachers' students did worse



Data requirements for VAMs

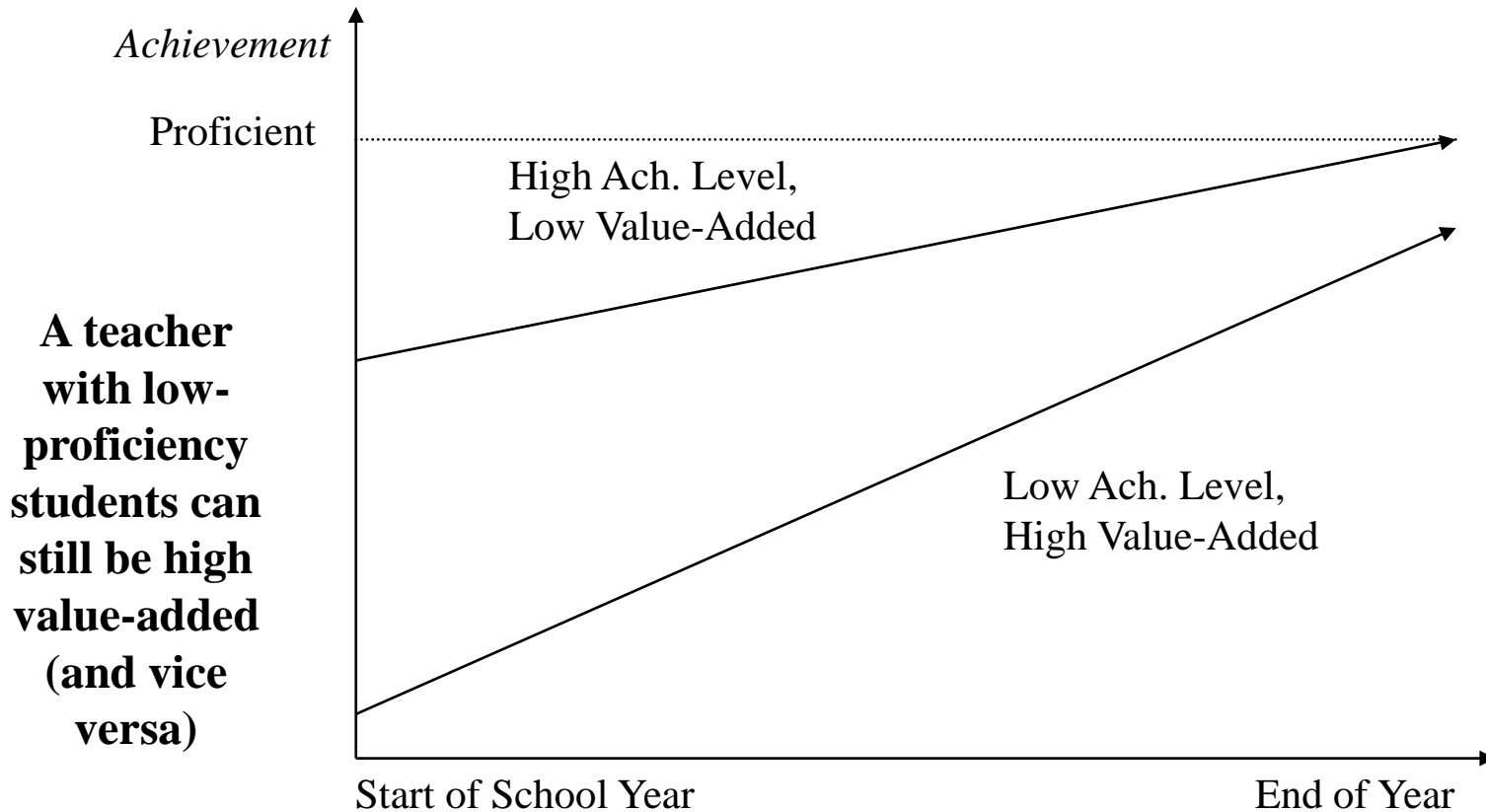
- To use VAMs, you must have the following
 - Unique identifiers for each *teacher*
 - Unique identifiers for each *student*
 - A link between the student and each of his or her teachers in the data system
 - Accurate, complete data going as far back as possible, for both students and teachers
 - Student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

Why VAMs are better than status models (1)



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

Why VAMs are better than status models (2)



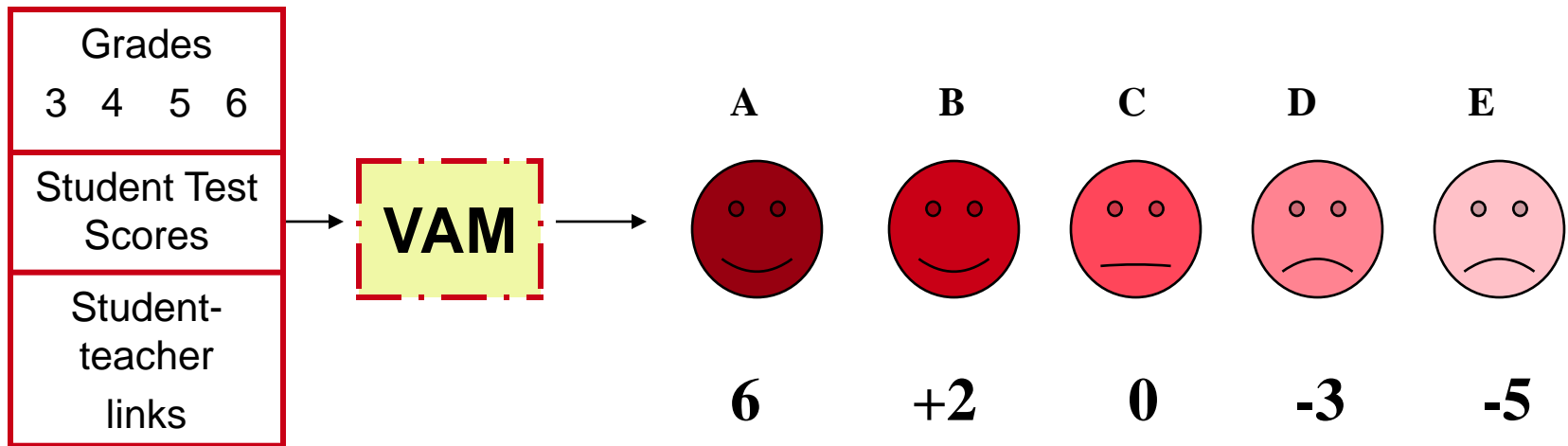
Slide courtesy of Doug Harris, Ph.D, University of Wisconsin-Madison

Interpreting VAM scores

- VAM produces an estimate of the average gain *in a particular class* in a particular year compared with the average gain in all similar classes in that year
- Some interpret this average as 100% “caused” by the teacher
- This requires ignoring the caveats and cautions issued by most researchers

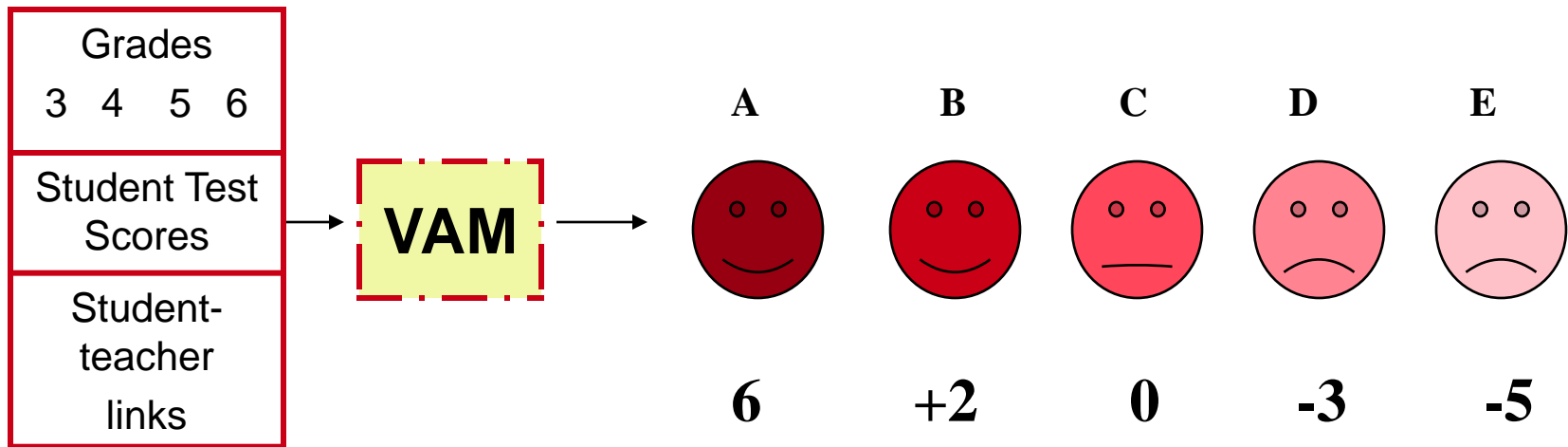
Teacher effects

Teacher Effects



Classroom effects

Classroom Effects



What the Research Says About Value-Added

- Researchers found that scores cannot be *solely* attributed to teachers' influence; VAMs provide a summary score of the “contribution of various factors toward growth in student achievement” (Goldhaber & Anthony, 2003, p. 38).

What the Research Says About Value-Added (Cont'd)

- Research does not find strong, consistent correlations between what teachers do in their classrooms (measured by observations) and value-added scores (Kimball, White, Milanowski, & Borman, 2004).

What the Research Says About Value-Added (Cont'd)

- Very little teacher effectiveness is explained by observable characteristics, and while teachers vary in their contribution to students' achievement score gains, researchers did not know what caused the variation (Rivkin, Hanushek, & Kain, 2005).

Concerns with value-added models

- Ceiling effect (depending on state test used)
- Non-random matching of students/teachers
- Context effects (classroom and school)
- Missing data patterns
- Horizontal and vertical test equating
- Student growth trajectories
- Influences on test scores *other than knowledge* (students' motivation, attitudes)

What VAMs can't tell you

- VAMs can't tell you **why** students in a particular classroom are scoring higher than expected
 - Maybe there is a narrow instructional focus on test content
 - Or maybe the classroom offers a rich, engaging curriculum that fosters deep student learning
- *How* teachers are getting results matters, not just the results themselves

VAMs don't measure most teachers

- About 69% of teachers (Prince et al., 2006) can't 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

Challenges for Special Education Teachers and ELL Specialists

➤ Challenges for SWD and ELLs

- Small student numbers
- Not all special educators and ELLs instruct students in tested subjects
- Teacher contribution to social and behavioral growth would not be factored into results
- Teachers working with students on alternate standards
- Little research exploring whether growth rates are comparable
- Little research on the use of accommodations & their impact related to teacher effects

Evaluating Teacher Effectiveness in Classrooms w/ Co-Teaching

- Majority of SWD are in the general education classroom
- Various co-teaching models make it difficult to evaluate teachers
 - For example, teachers as aides or working with small group of students
- Attributing value-added and non-value added scores to general ed, special ed, or ELL specialist

Is it possible to accurately measure teacher effectiveness with VAMs?

- We could get an estimate that would be close to the teacher's true contribution to student learning growth if
 - Teachers were randomly assigned to schools
 - Students were randomly assigned to teachers
- However, given differences in resources, school environment, etc., we would only know how the teacher did with **that** group of students in **that** setting

VAMs and teacher evaluation/assistance

- VAMs may be useful in identifying teachers who are not performing at acceptable levels in terms of student gains
 - *However, VAMs cannot be used to diagnose why the teacher is failing to meet student progress goals*
 - Additional information should be gathered in the classroom in order to properly diagnose problems
 - Teachers can then be provided with guidance and support to address specific needs

Getting to the bottom of poor VAM scores

- First incidence of poor VAM scores
 - Start after the beginning of the school year?
 - “Challenging” students assigned disproportionately to the teacher?
 - Personal/professional disruptions in teacher’s life?
- Consistently poor VAM scores
 - Use classroom observations diagnostically
 - Interview with teacher before/after observations
 - Examination of teacher assignments/student work

Non-VAM tests (accepted under Washington, DC's IMPACT evaluation system)

- DC Benchmark Assessment System (DC BAS)
- Dynamic Indicators of Basic Early Literacy Skills (DIBELS)
- Developmental Reading Assessment (DRA)
- Curriculum-based assessments (e.g., Everyday Mathematics)
- Unit tests from DCPS-approved textbooks
- Off-the-shelf standardized assessments that are aligned to the DCPS Content Standards
- Rigorous teacher-created assessments that are aligned to the DCPS Content Standards
- Rigorous portfolios of student work that are aligned to the DCPS Content Standards

Standardization is key

- Standardizing how curriculum- or classroom-based tests are given is key to ensuring that tests are “rigorous and comparable across classrooms”
- Ensure that tests meet district approval
- For subject-matter tests, ensure that
 - Tests are given on the same day, at the same time, for the same length of time, with supervision
 - Teachers agree to appropriate “test prep” rules

Evidence of growth in student learning

- Evidence is strongest when it is
 - **Standardized**, meaning that all teachers used the assessment in exactly the same way
 - Gave the assessment on the same day
 - Gave students a specific amount of time to complete the test
 - Used the same preparation/instructions prior to the test
 - Recorded/reported results accurately
 - **Valid**, meaning that it measures what is intended
 - Items (questions) accurately capture students' understanding and knowledge
 - Progress towards proficiency in a subject is captured because there are sufficient items to measure students at all levels
 - **Recorded**, meaning that student progress can be compared across classrooms and schools

Questions about Part 3?



Cincinnati study results

- Study by Kane et al. (2010) used teacher evaluation scores plus value-added scores
 - “...policies and programs that help a teacher get better on all eight ‘teaching practice’ and ‘classroom environment’ skills measured by TES will lead to student achievement gains” (p. 28)
 - “...helping teachers improve their ‘classroom environment’ management will likely also generate higher student achievement” (p. 28)
 - “...[adding] pedagogy that utilizes ‘questioning and discussion’ practices will generate higher reading achievement, but not higher math achievement” (p. 28)

Gates & Spencer Research (ETS)

- How do results from different measures compare?
 - Several multi-year research projects seek to answer that question
 - Projects involve ETS, the Institute for Social Research, RAND, the University of Virginia, and the University of Michigan
 - Numerous instruments (four types of classroom observations instruments), teacher assignment protocol, content knowledge measures, teacher self-efficacy measures

Basic Research Design (Gates, Spencer projects)

- Observe and video-record teacher multiple times in a classroom(s) during a school year
- Rate the classroom interactions using appropriate general and subject-specific rubrics
- Collect student assignments & resulting student work
- Assess teacher knowledge and beliefs (efficacy) through tests/surveys
- Estimate VAM scores for teachers
- Examine relationships within and across all measures

Instruments being used (Gates, Spencer)

- PLATO (ELA) – Pam Grossman - Stanford
- Math Quality Instruction (MQI) – Heather Hill, Harvard
- CLASS (observation instrument) – Bob Pianta, U of VA
- Framework for Teaching – Charlotte Danielson
- Intellectual Demand Assignment Protocol (IDAP) – Fred Newmann et al.
- Traditional Praxis measures – content knowledge
- Tests of *Knowledge For Teaching* measures from Deborah Ball, Heather Hill and colleagues
- Developing new measures of *Knowledge For Teaching* (collaboration with University of Michigan)
- Carol Dweck measures on views of intelligence

References/Resources

Coggshall, J. (2007). *Communication framework for measuring teacher quality and effectiveness: Bringing coherence to the conversation*. Washington, DC: National Comprehensive Center for Teacher Quality.

<http://www.tqsource.org/publications/NCCTQCommFramework.pdf>

Goe, L. (2007). *The link between teacher quality and student outcomes: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

<http://www.tqsource.org/publications/LinkBetweenTQandStudentOutcomes.pdf>

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

<http://www.tqsource.org/publications/EvaluatingTeachEffectiveness.pdf>

References/Resources (cont'd)

- Goldhaber, D., & Anthony, E. (2003). *Teacher quality and student achievement* (No. UDS-115). New York, NY: ERIC Clearinghouse on Urban Education.
- Holdheide, L., Goe, L., Croft, A., & Reschly, D. (2010). *Challenges in evaluating special education teachers and English Language Learner specialists*. Washington, DC: National Comprehensive Center for Teacher Quality.
- Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2010). *Identifying effective classroom practices using student achievement data*. Cambridge, MA: National Bureau of Economic Research.
<http://www.nber.org/papers/w15803>

References/Resources (cont'd)

- Kimball, S. M., White, B., Milanowski, A. T., & Borman, G. (2004). Examining the relationship between teacher evaluation and student assessment results in Washoe County. *Peabody Journal of Education*, 79(4), 54-78.
- Prince, C. D., Schuermann, P. J., Guthrie, J. W., Witham, P. J., Milanowski, A. T., & Thorn, C. A. (2006). *The other 69 percent: Fairly rewarding the performance of teachers of non-tested subjects and grades*. Washington, DC: U.S. Department of Education, Office of Elementary and Secondary Education.
- <http://www.cecr.ed.gov/guides/other69Percent.pdf>
- Race to the Top Application*
- <http://www2.ed.gov/programs/racetothetop/resources.html>
- Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2), 417 - 458.

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