Listening. Learning. Leading.®

Educator Evaluation to Improve Teaching and Learning

Laura Goe, Ph.D.

Research Scientist, ETS

Understanding Teaching Quality Center

Presentation to

Nebraska Educator Evaluation Design Pilot Schools

October 24, 2012 ♦ Lincoln, NE

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
- Research Scientist in the Performance Research Group at ETS
- Sr. Research & Technical Assistance Expert for the federally-funded Center for Great Teachers and Leaders

The goal of teacher evaluation

*The **ultimate** goal of all teacher evaluation should be…*

TO IMPROVE TEACHING AND LEARNING

What's driving changes in teacher evaluation?

- 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) found that 90% of teachers were rated "good" or better in districts where students were failing at high levels

Chicago Study (Sartain et al., 2011)

- "The work in Chicago and across the country to improve evaluation was motivated by two main factors. **First, evaluation systems were failing to give teachers either meaningful feedback on their instructional practices or guidance about what is expected of them in the classroom. Second, traditional teacher evaluation systems were not differentiating among the best teachers, good teachers, and poor teachers."** (p. 1)

"Effective" vs. "Highly Qualified"

- The focus has shifted away from *ensuring **highly qualified** teachers in every classroom* to *ensuring **effective** teachers in every classroom*
- This shift is a result of numerous studies that show that qualifications provide a "floor" or "minimum" set of competencies but do not predict which teachers will be most successful at helping students learn

A succinct definition of teacher effectiveness

- Hunt (2009) stated that, "…the term "teacher effectiveness" is used broadly, to mean the collection of characteristics, competencies, and behaviors of teachers at all educational levels that enable students to reach desired outcomes, which may include the attainment of specific learning objectives as well as broader goals such as being able to solve problems, think critically, work collaboratively, and become effective citizens. (p. 1)

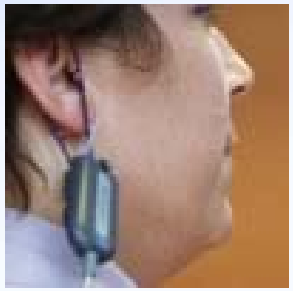
Evaluation for accountability *and* instructional improvement

- Effective evaluation relies on:
  - Clearly defined and communicated standards for performance
  - Quality tools for measuring and differentiating performance
  - Quality training on standards and tools
    - Evaluators should agree on what constitutes evidence of performance on standards
    - Evaluators should agree on what the evidence means in terms of a score

Listening. Learning. Leading.®

An Introduction to Measures

Measures: The right choice depends on *what you want to measure*



Measures and models: Definitions

- **Measures** are the instruments, assessments, protocols, rubrics, and tools that are used in determining teacher effectiveness
- **Models** are the state or district systems of teacher evaluation including all of the inputs and decision points (measures, instruments, processes, training, and scoring, etc.) that result in determinations about individual teachers' effectiveness

Rhode Island DOE Model: Multiple Measures include student learning

Student learning rating

+

Professional practice rating

+

Professional responsibilities rating

Final evaluation rating

The student learning rating is determined by a combination of different sources of evidence of student learning. These sources fall into three categories:

Category 1: Student growth on state standardized tests (e.g., NECAP, PARCC)

Category 2: Student growth on standardized district-wide tests (e.g., NWEA, AP exams, Stanford-10, ACCESS, etc.)

Category 3: Other local school-, administrator-, or teacher-selected measures of student performance

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
  - Student surveys such as Harvard's Tripod
  - Evidence binder/portfolio
- **Evidence of *professional responsibility***
  - Administrator/supervisor reports, parent surveys
  - Teacher reflection and self-reports, records of contributions

Measures that help teachers grow

- Measures that include protocols and processes that teachers can examine and comprehend
- Measures that are directly and explicitly aligned with teaching standards
- 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 portfolios)
- Measures that give teachers opportunities to discuss the results for formative purposes with evaluators, administrators, teacher learning communities, mentors, coaches, etc.
- Measures that are aligned with and used to inform professional growth and development offerings

Listening. Learning. Leading.®

Measures of Teaching Practice: Classroom Observations

Charlotte Danielson's Framework for Teaching Domains

Domain 1: Planning and Preparation 1a Demonstrating Knowledge of Content and Pedagogy 1b Demonstrating Knowledge of Students 1c Setting Instructional Outcomes 1d Demonstrating Knowledge of Resources 1e Designing Coherent Instruction 1f Designing Student Assessments	**Domain 2: Classroom Environment** 2a Creating an Environment of Respect and Rapport 2b Establishing a Culture for Learning 2c Managing Classroom Procedures 2d Managing Student Behavior 2e Organizing Physical Space
Domain 4: Professional Responsibilities 4a Reflecting on Teaching 4b Maintaining Accurate Records 4c Communicating with Families 4d Participating in a Professional Community 4e Growing and Developing Professionally 4f Showing Professionalism	**Domain 3: Instruction** 3a Communicating With Students 3b Using Questioning and Discussion Techniques 3c Engaging Students in Learning 3d Using Assessment in Instruction 3e Demonstrating Flexibility and Responsiveness

Why teachers generally value observations

- Observations are the traditional measure of teacher performance
- Teachers feel they have some control over the process and outcomes
- They report that having a conversation with the observation and receiving constructive feedback after the observation is greatly beneficial
- Evidence-centered discussions can help teachers improve instruction
- Peer evaluators often report that they learn new teaching techniques

Teacher observations: strengths and weaknesses

- Strengths
  - Great for teacher formative evaluation
  - Helps evaluator understand teachers' needs across school or across district
- Weaknesses
  - Only as good as the instruments and the observers
  - Considered "less objective"
  - Expensive to conduct (personnel time, training, calibrating)
  - Validity of observation results may vary with who is doing them, depending on how well trained and calibrated they are

Validity of classroom observations is highly dependent on training

- A teacher should get the same score no matter who observes him
  - This requires that all observers be trained on the instruments and processes
  - Occasional "calibrating" should be done; more often if there are discrepancies or new observers
  - Who the evaluators are matters less than the fact that they are trained to recognize evidence and score it consistently
- Teachers should also be trained on the observation forms and processes so they can participate actively and fully in the process

Reliability results when using different combinations of raters and lessons

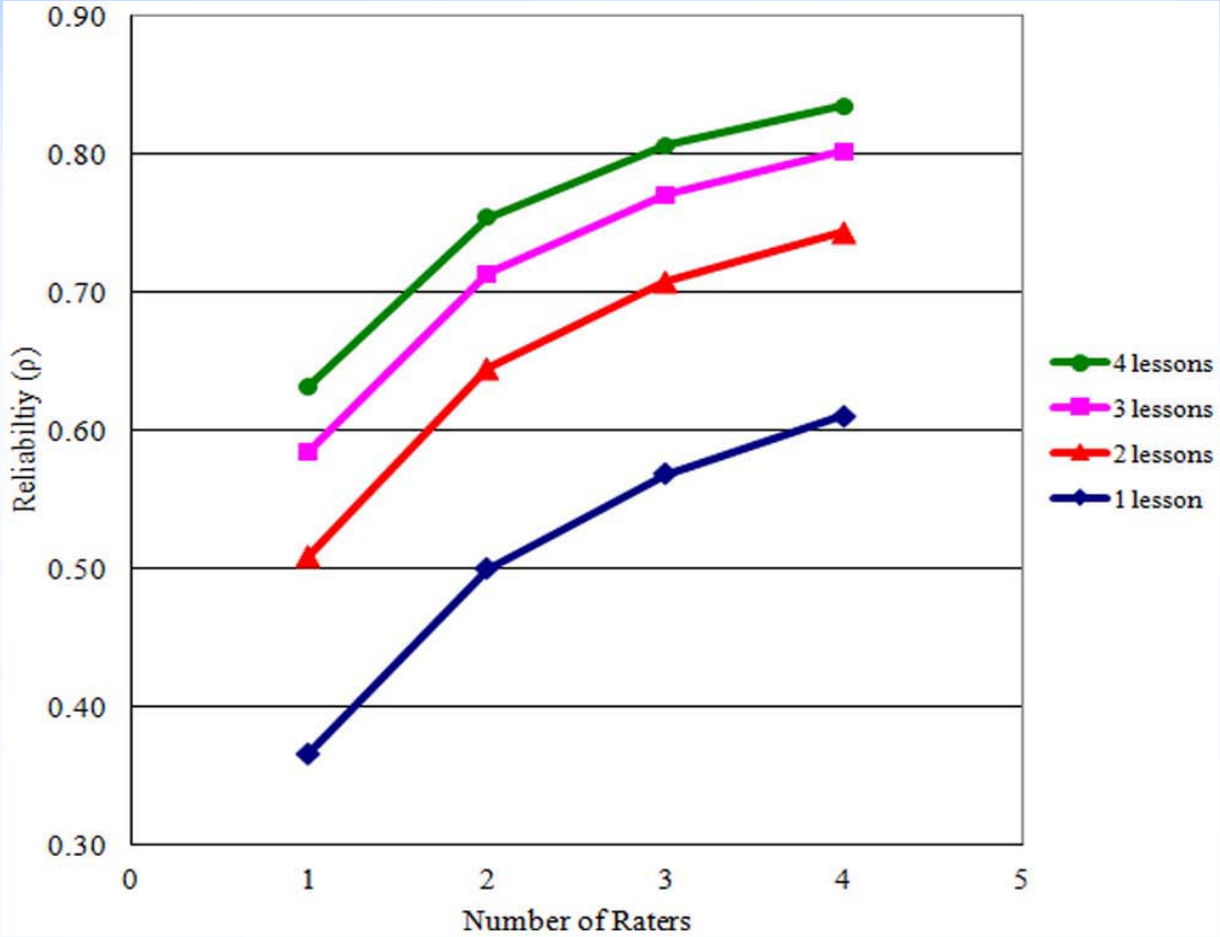


Figure 2. *Errors and Imprecision: the reliability of different combinations of raters and lessons.* From Hill et al., 2012 (see references list). Used with permission of author.

Formal vs. informal observations

- Formal observations are likely to be
  - Announced and scheduled in advance according to a pre-determined yearly schedule
  - Include pre- and post-conferences with review of lesson plans and artifacts
  - Last an entire class period
  - Result in a set of scores on multiple indicators
- Informal observations are likely to be
  - Unannounced, drop-in
  - Last less than an entire class period
  - Result in informal verbal or written feedback to the teacher, perhaps on only one indicator

Questions to ask about observations

- How many observations per year?
  - Vary by new vs. experience?
  - Vary by demonstrated competence?
  - Combination of formal and informal?
- Who should conduct the observations?
- Will multiple observers be required?
- How will they be trained?
  - Workshops? Online (video-based)?
- Will they need to be certified?

Validity of classroom observations is highly dependent on training

- *A teacher should get the same score no matter who observes him*
  - This requires that all observers be trained on the instruments and processes
  - Occasional "calibrating" should be done; more often if there are discrepancies or new observers
  - Who the evaluators are matters less than adequate training
  - Teachers should also be trained on the observation forms and processes

Listening. Learning. Leading.®

Measures of Teaching Practice: Student Surveys

Tripod Survey (1)

- **Harvard's Tripod Survey – the 7 C's**
  - Caring about students (nurturing productive relationships);
  - Controlling behavior (promoting cooperation and peer support);
  - Clarifying ideas and lessons (making success seem feasible);
  - Challenging students to work hard and think hard (pressing for effort and rigor);
  - Captivating students (making learning interesting and relevant);
  - Conferring (eliciting students' feedback and respecting their ideas);
  - Consolidating (connecting and integrating ideas to support learning)

Tripod Survey (2)

- Improved student performance depends on strengthening three legs of teaching practice: content, pedagogy, and relationships
- There are multiple versions: k-2, 3-5, 6-12
- Measures:
  - student engagement
  - school climate
  - home learning conditions
  - teaching effectiveness
  - youth culture
  - family demographics
- Takes 20-30 min
- There are English and Spanish versions
- Comes in paper form or in online version

Tripod Survey (3)

- Control (promoting cooperation and peer support) is the strongest correlate of value added gains
- *However, it is important to keep in mind that a good teacher achieves control by being good on the other dimensions*

Tripod Survey (4)

- Different combinations of the 7 C's predict different outcomes (student learning is one outcome)
- Using the data, you can determine what a teacher needs to focus on to improve important outcomes
- Besides student learning, other important outcomes include:
  - happiness
  - good behavior
  - healthy responses to social pressures
  - self-consciousness
  - engagement/effort
  - satisfaction

Listening. Learning. Leading.®

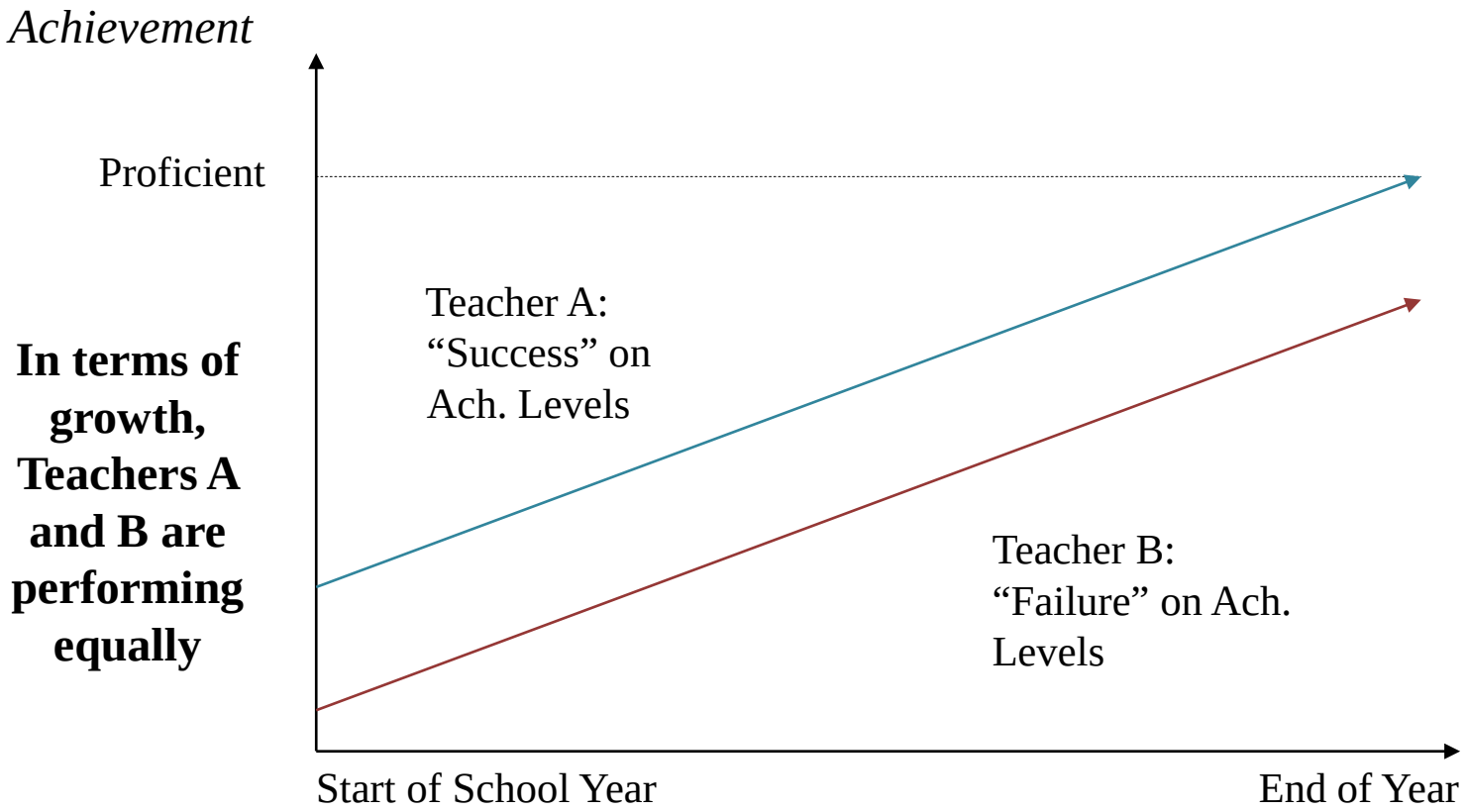
Measures of Teachers' Contributions to Student Learning Growth

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)

30

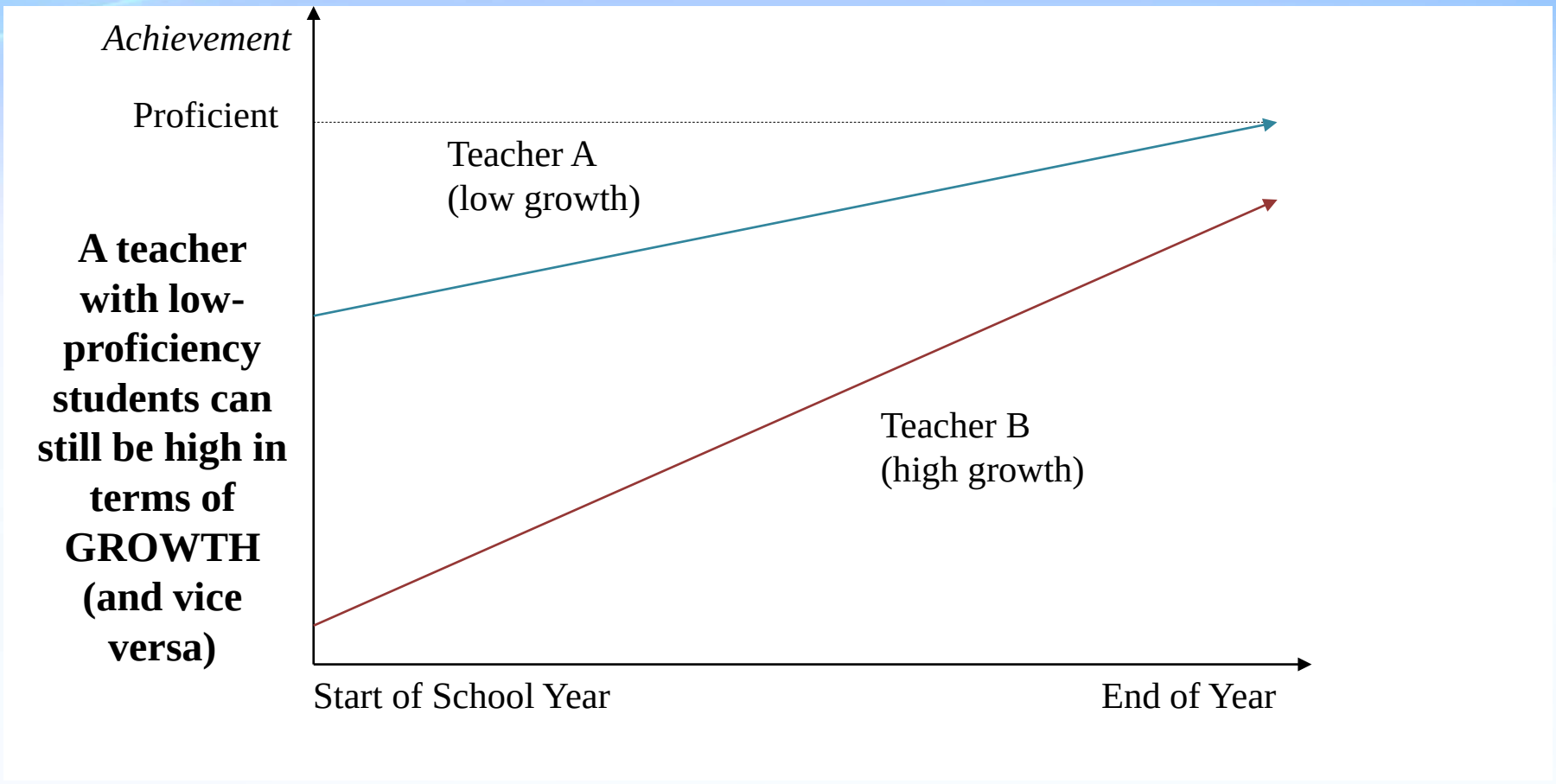
Growth vs. Proficiency Models



In terms of growth, Teachers A and B are performing equally

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

Growth vs. Proficiency Models



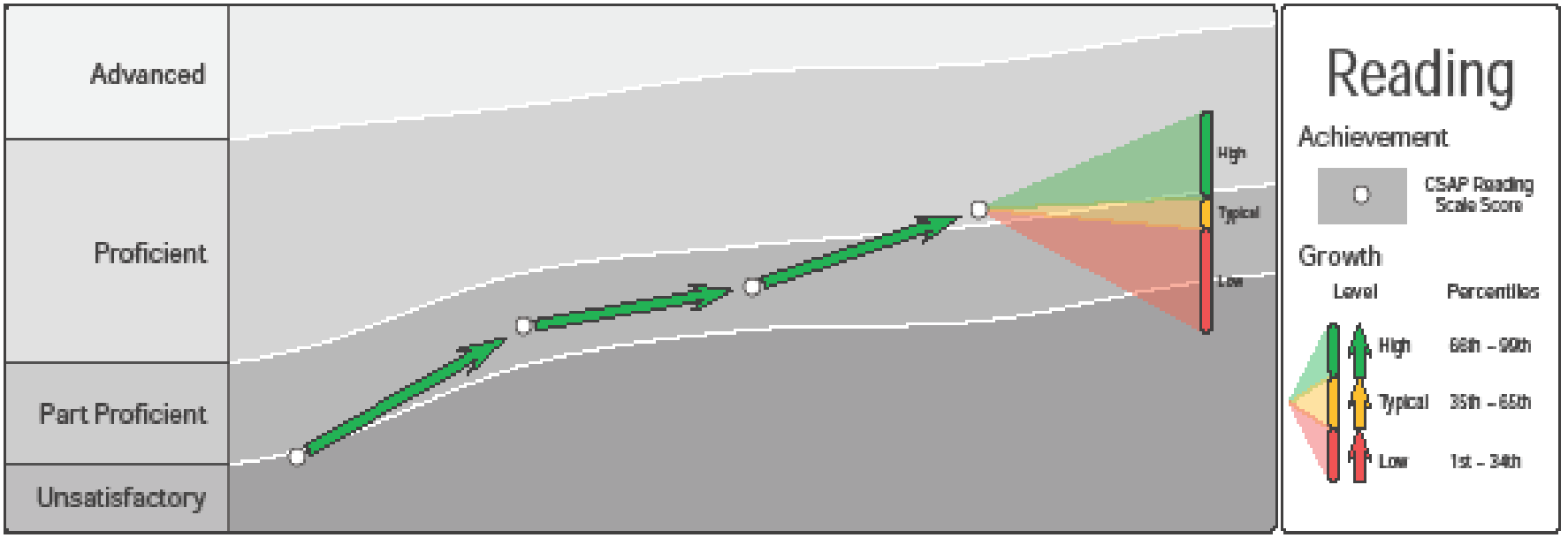
A teacher with low-proficiency students can still be high in terms of GROWTH (and vice versa)

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

Most popular growth models: Value-added and Colorado Growth Model

- EVAAS uses prior test scores to predict the next score for a student
  - Teachers' value-added is the difference between actual and predicted scores for a set of students http://www.sas.com/govedu/edu/k12/evaas/index.html
- Colorado Growth model
  - Betebenner 2008: Focus on "growth to proficiency"
  - Measures students against "academic peers" www.nciea.org

Linking student learning results to professional growth opportunities



	Grade 3 2006	Grade 4 2007	Grade 5 2008	Grade 6 2009	Next Year	
Scale Score Achievement Level	462 Unsatisfactory	539 Part Proficient	563 Part Proficient	609 Proficient		Achievement
Growth Percentile Growth Level		66 High	66 High	90 High		Growth

Slide courtesy of Damian Betebenner at www.nciea.org

Even if we could assess all students with standardized tests…

- Using only standardized test results to reflect teachers' contributions to student learning growth may capture just part of what we care about
  - Standardized tests may cover only about ½ the content standards (Polikoff et al, 2011)
- Good teachers also work to ensure that students can apply their knowledge
  - Applied knowledge may be more accurately measured with the 4 Ps (projects, portfolios, performances and products)

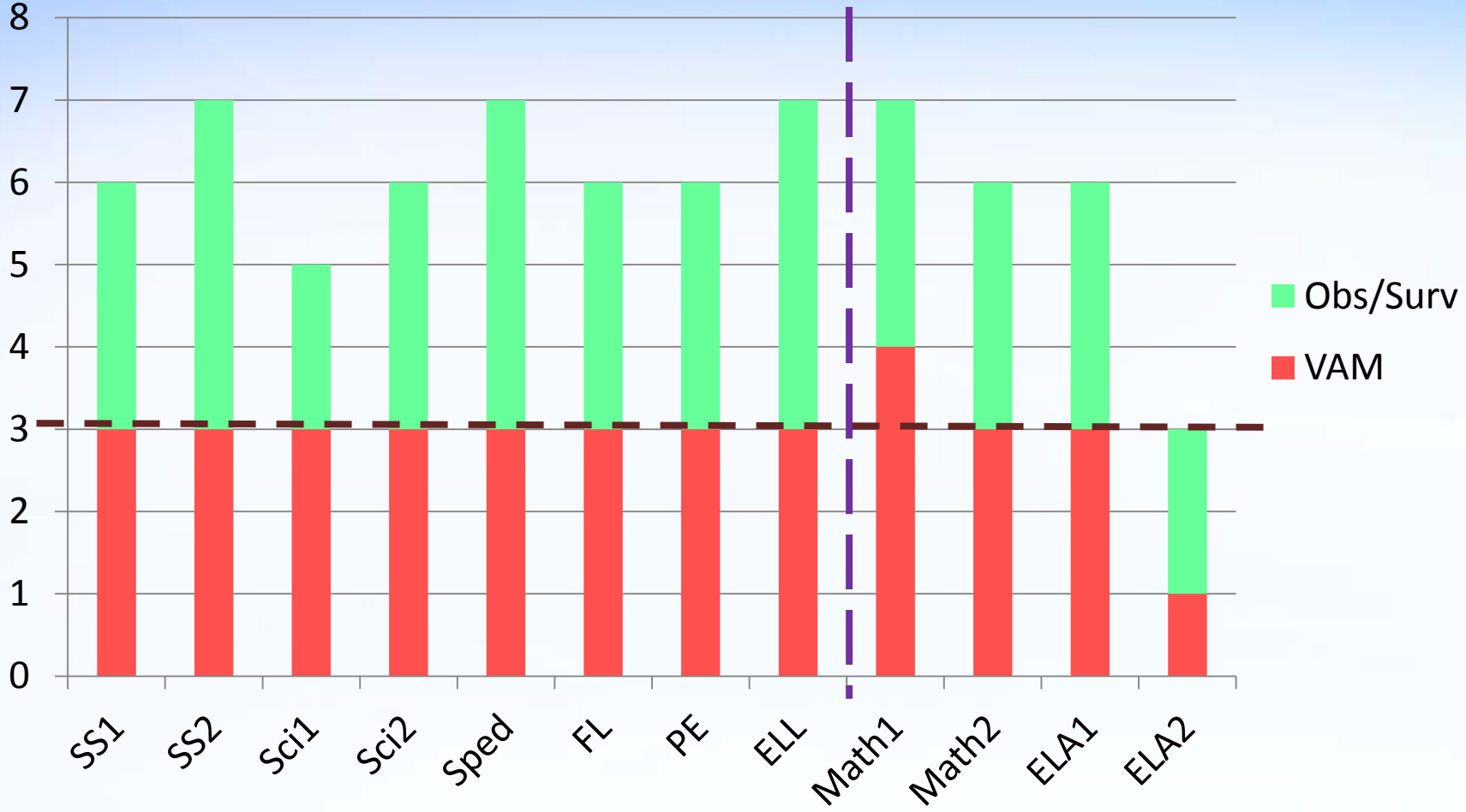
The 4 Ps (Projects, Performances, Products, Portfolios)

- Yes, they can be used to demonstrate teachers' contributions to student learning growth
- Here's the basic approach
  - Use a high-quality rubric to judge initial knowledge and skills required for mastery of the standard(s)
  - Use the same rubric to judge knowledge and skills at the end of a specific time period (unit, grading period, semester, year, etc.)

Measuring non-tested subject/grade teachers' contributions to student learning growth

Model	Description
Student learning objectives	Teachers assess students at beginning of year and set objectives then assesses again at end of year; principal or designee works with teacher, determines success
Subject & grade alike team models ("Ask a Teacher")	Teachers meet in grade-specific and/or subject-specific teams to consider and agree on appropriate measures that they will all use to determine their individual contributions to student learning growth
Colorado Content Collaboratives	Content experts (external) identify measures and groups of content teachers consider the measures from the perspective of classroom use; may not include pre- and post measures
Pre-and post-tests model	Identify or create pre- and post-tests for every grade and subject
School-wide value-added	Used in TAP Model; teachers in tested subjects & grades receive their own value-added score; all other teachers get the school-wide average

School-wide average growth used for non-tested subjects / grades



SLOs: Assessing students' current knowledge/skills (baseline data)

- Baseline data can be historic (found) or current (collected)
  - Historic (found) data includes all prior history on students' proficiency on specific standards
  - Current (collected) data includes all efforts made by the state, district, school, or teacher to establish students' current levels of proficiency
- Each has advantages and drawbacks

SLOs: Evidence of mastery

- Decide on the key standards you want your students to show proficiency in by the end of the course (semester/year)
- With colleagues in same subject/grade, ask yourselves the question: *How will we know that the students have mastered these standards?*
  - What is the evidence of mastery that you will be looking for?
  - How will you collect that evidence?
    - Assessments, 4Ps, etc.

SLOs: Measuring success

- High-quality assessments
- Rubrics will work best when performance, projects, portfolios or products are needed to show students' mastery of standards
  - Start by searching for rubrics on line
  - Use that as a starting place (you will probably need a more detailed one)
  - Work with colleagues to create a good rubric

SLOs: Scoring with rubrics

- Rubric-based assessments are best scored together with colleagues in the same grade/subject
- Start by selecting “anchor papers” at each level and discussing what qualities make them a 1, 2, etc.
- Try independently scoring sample papers/projects and discussing scores until agreement is reached

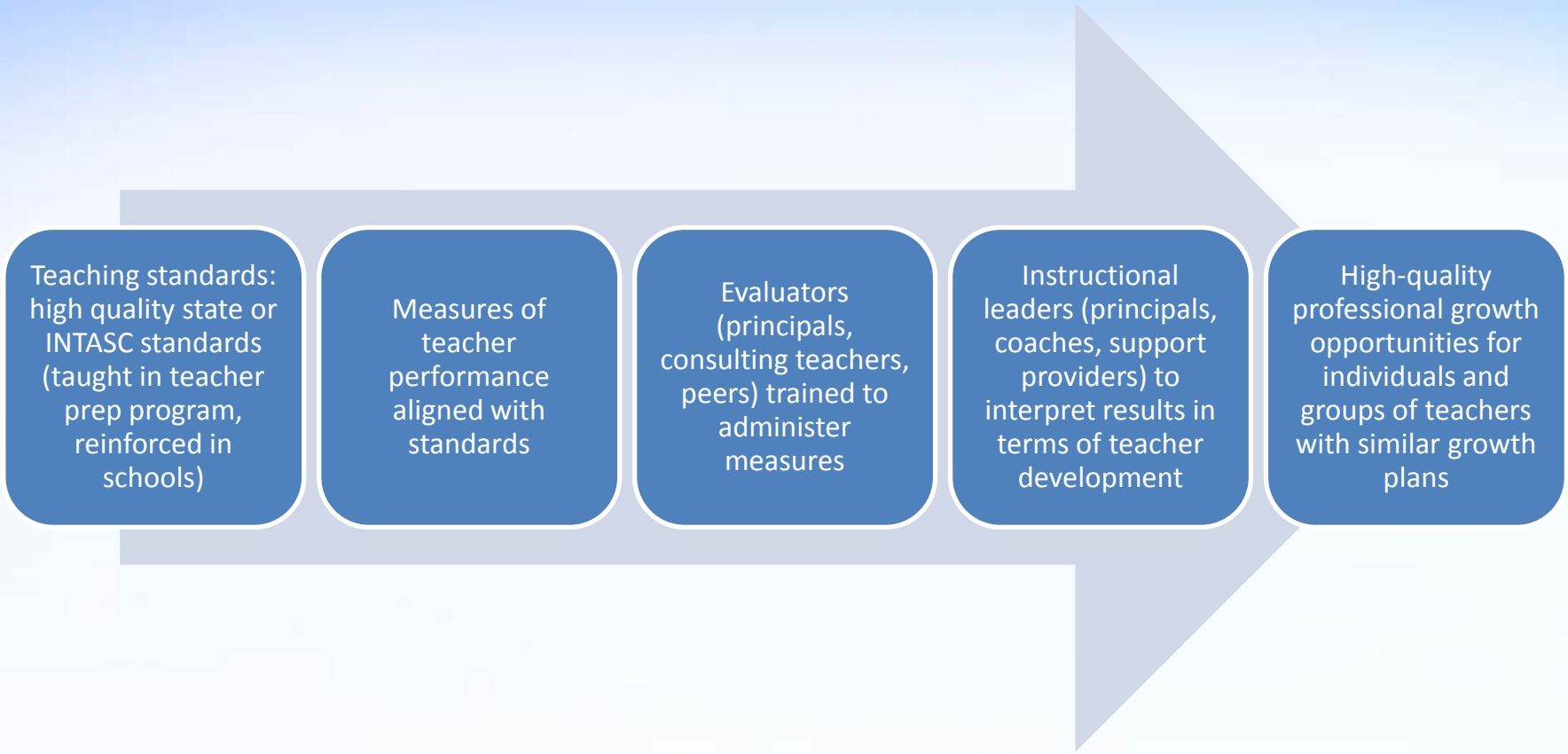
Options for scoring SLO results as part of teacher evaluation

1. Principal and teacher meet at several points including beginning and end of year to discuss results and principal makes holistic judgment
2. Principal considers results from all teachers in a subject/grade before determining scores for individual teachers
3. District receives documented results and determines scores for individual teachers

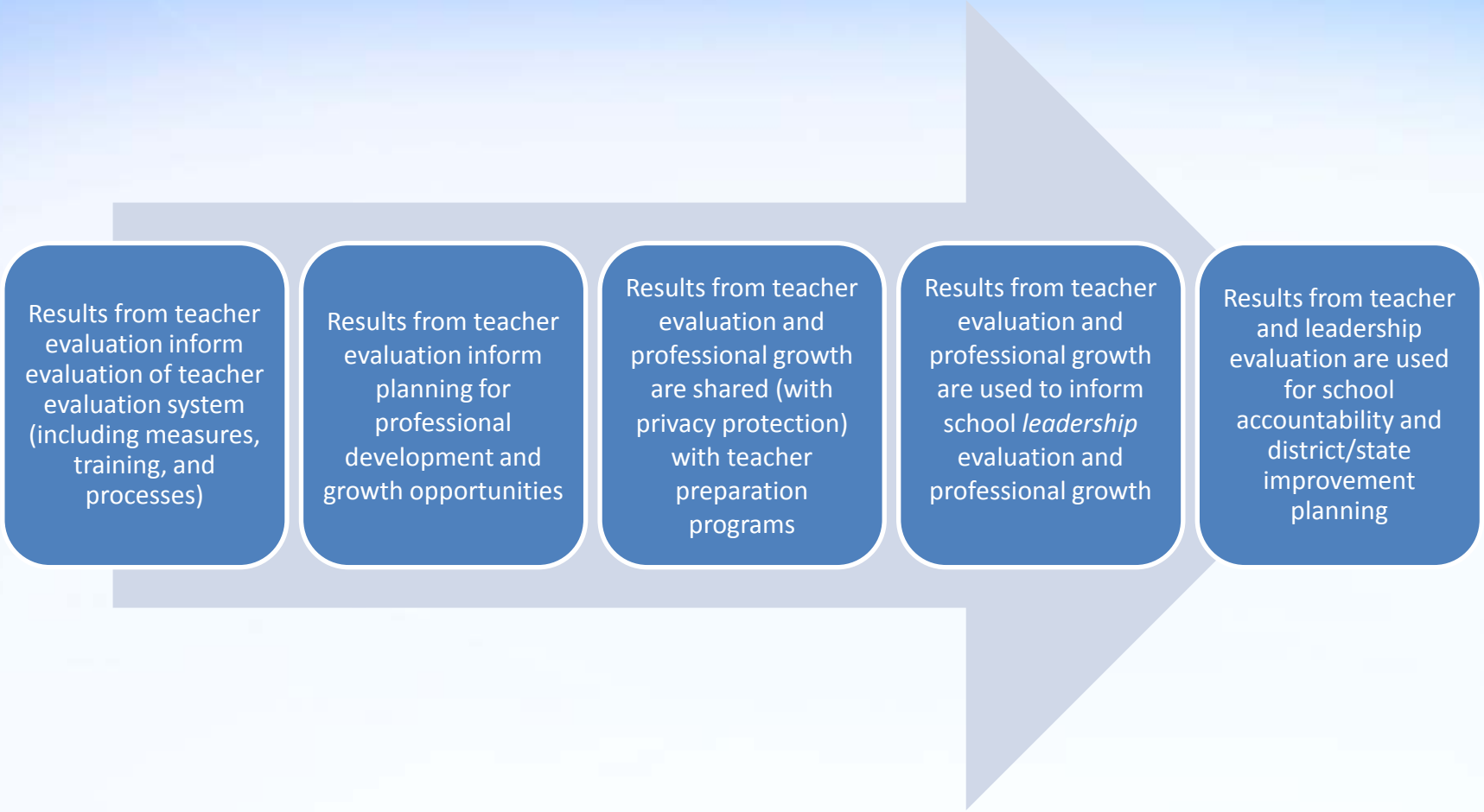
Listening. Learning. Leading.®

Using Evaluation Results to Inform Professional Growth

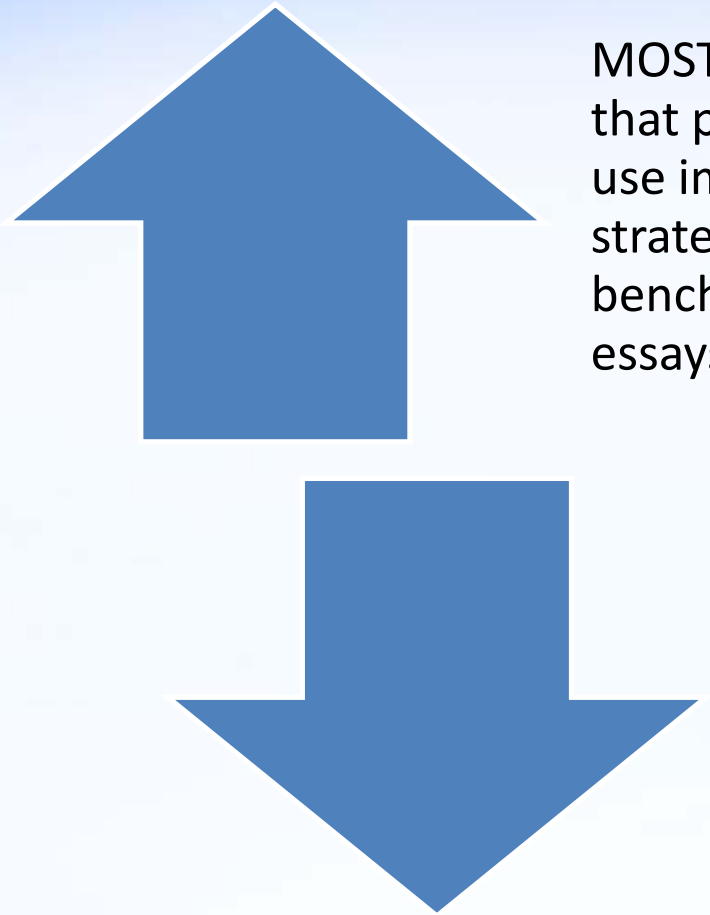
An aligned teacher evaluation system: Part I



An aligned teacher evaluation system: Part II



Using student learning outcomes to inform teacher professional growth



MOST helpful: Student assessments that provide information teachers can use immediately to adjust instructional strategies, such as results from benchmark or interim assessment or essays scores with rubrics

LEAST helpful: Student assessments that provide a snapshot of students' skills at a single point in time after most instruction is complete, such as last year's state standardized test results

A well-aligned evaluation system (Goe et al., 2012)

- Standards-based evidence will:
  - Form the basis for a professional growth plan
  - Give structure and consistency to coaching and mentoring by providing the basis for shared expectations and a common language
  - Provide a diagnostic approach to understanding inadequate student learning growth (i.e., determining which standards are not being met and considering how they might relate to student outcomes)
  - Offer a set of criteria to help principals, consulting teachers, mentors, and others identify areas in which teachers are successful and areas for improvement

Interpreting results for *alignment with teacher professional learning options*

- Different approach; not looking at "absolute gains"
- Requires ability to determine and/or link student outcomes to what likely happened *instructionally*
- Requires ability to "diagnose" instruction and recommend/and or provide appropriate professional growth opportunities
  - Individual coaching/feedback on instruction
  - Observing "master teachers"
  - Group professional development (when several teachers have similar needs)

Memphis professional development system: An *aligned* system

- Teaching and Learning Academy began April '96
- Nationally commended program intended to
  - "…provide a collegial place for teachers, teacher leaders and administrators to meet, study, and discuss application and implementation of learning…to impact student growth and development"
- Practitioners propose and develop courses
  - Responsive to school/district evaluation results
  - Offerings must be aligned with NSDC standards
  - ~300+ On-line and in-person courses, many topics

Principal Effectiveness: New Leaders for New Schools Definition

- "New Leaders for New Schools advocates for an evidence-based, three-pronged approach to defining principal effectiveness: 1) gains in student achievement, 2) increasing teacher effectiveness, and 3) taking effective leadership actions to reach these outcomes." http://www.newleaders.org/wp-content/uploads/2011/08/principal_effectiveness_nlns_overview.pdf

Vanderbilt Assessment of Leadership in Education (VAL-Ed)

High Standards for Student Learning		Sources of Evidence (Check Key Sources of Evidence): Reports from Others	Personal Observations	School Documents	School Projects or Activities	Other Sources	No Evidence	Effectiveness Rating (Circle One Number to Indicate How Effective): Ineffective	Minimally Effective	Satisfactorily Effective	Highly Effective	Outstandingly Effective
How effective is the principal at ensuring the school …												
Planning	1. plans rigorous growth targets in learning for all students.							1	2	3	4	5
	2. plans targets of faculty performance that emphasize improvement in student learning.							1	2	3	4	5

Resources

- Charlotte Danielson's Framework for Teaching http://www.danielsongroup.org/article.aspx?page=frameworkforteaching
- Memphis Professional Development system: http://www.mcsk12.net/aoti/pd/index.asp
- Common Core Standards rubrics from Elk Grove Unified School District (California) http://blogs.egusd.net/ccss/2012/01/12/ccss-aligned-rubrics-k-12/
- Harvard's Tripod Survey http://www.tripodproject.org/index.php/index/
- Vanderbilt Assessment of Leadership in Education http://www.valed.com/
  - Also see the VAL-Ed Powerpoint at http://peabody.vanderbilt.edu/Documents/pdf/LSI/VALED_AssessLCL.ppt

References

Glazerman, S., Goldhaber, D., Loeb, S., Raudenbush, S., Staiger, D. O., & Whitehurst, G. J. (2011). *Passing muster: Evaluating evaluation systems. Washington, DC: Brown Center on Education Policy at Brookings.* http://www.brookings.edu/reports/2011/0426_evaluating_teachers.aspx#

Goe, L., Biggers, K., & Croft, A. (2012). Linking teacher evaluation to professional development: Focusing on improving teaching and learning. Washington, DC: National Comprehensive Center for Teacher Quality. http://www.tqsource.org/publications/LinkingTeacherEval.pdf

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. Educational Researcher, 41(2), 56-64. http://scholar.harvard.edu/mkraft/files/hill_charalambous_kraft_2012._when_rater_reliability_is_not_enough_-_edr..pdf

Hunt, B. C. (2009). *Teacher effectiveness: A review of the international literature and its relevance for improving education in Latin America. Washington, DC: Partnership for Educational Revitalization in the Americas (PREAL).* http://preal.org/Archivos/Bajar.asp?Carpeta=Preal Working Papers&Archivo=Teacher Effectivenes.pdf

Polikoff, M. S. (2011). How well aligned are state assessments of student achievement with state content standards? *American Educational Research Journal, 48(4), 965-995.* http://aer.sagepub.com/content/48/4/965.abstract?rss=1

References (cont'd)

Prince, C. et al. (2008). The other 69 percent: *Fairly rewarding the performance of teachers of non-tested subjects and grades*. Center for Education Compensation Reform. http://www.cecr.ed.gov/guides/other69Percent.pdf

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417-458. http://www.econ.ucsb.edu/~jon/Econ230C/HanushekRivkin.pdf

Sartain, L., Stoelinga, S. R., & Brown, E. R. (2011). Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation. Chicago: Consortium on Chicago School Research at the University of Chicago. http://ccsr.uchicago.edu/sites/default/files/publications/Teacher%20Eval%20Report%20FINAL.pdf

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness.* Brooklyn, NY: The New Teacher Project. http://widgeteffect.org/downloads/TheWidgetEffect.pdf

Laura Goe, Ph.D.
609-619-1648
lgoe@ets.org
www.lauragoe.com
https://twitter.com/GoeLaura

Educational Testing Service
Rosedale Road, 02-T
Princeton, NJ 08541

Questions?