

Teacher Evaluation Systems

Addressing Common Implementation Challenges

March 2013

Center on
GREAT TEACHERS & LEADERS
at American Institutes for Research ■



Adaptive or Technical Challenge?



“Indeed, the single most common source of leadership failure we’ve been able to identify...is that people, especially those in positions of authority, treat adaptive challenges like technical problems.”

—Heifetz & Linsky (2002), p. 14.

Technical Versus Adaptive Challenges

■ Technical Challenges

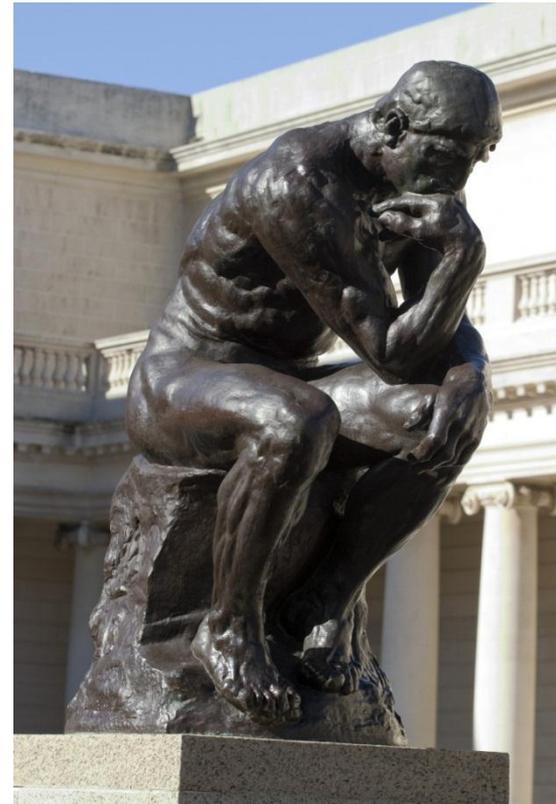
- Can be fixed by experts and by implementation of best practices.
- Are easy to identify and have solutions that can be implemented quickly.

■ Adaptive Challenges

- Require people to change their values, behaviors, and attitudes.
- Require people to learn new ways of doing business.
- Are often difficult to identify.
- Require people with the problem to do the work of solving it.
- Often require experiments, innovations, and new learning.
- Can take longer to implement.

Brainstorming Activity

- What are some technical challenges you are facing?
- What are some adaptive challenges you are facing?



Common Technical Challenges

- Growth measures for nontested grades and subjects
- Interrater reliability
- Combining evaluation measures for rating purposes



Growth Measures for Nontested Grades and Subjects

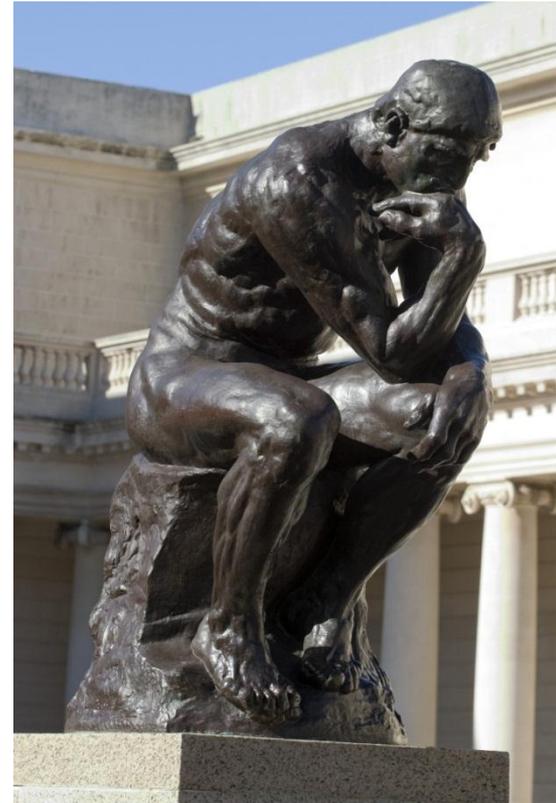
Center on

GREAT TEACHERS & LEADERS

at American Institutes for Research ■

Brainstorming Activity

- How is student growth currently measured in nontested subjects and grades?



Measures Must Meet State and/or Federal Requirements

1. Aligned with specific standards
2. Between two points in time
3. Comparable across classrooms

But this leaves plenty of options...



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

How Do We Measure Contributions to Learning Growth in the Following Cases?

- Teachers of nontested subjects (e.g., social studies, K–2, art, drama, band)
- Teachers of certain student populations and situations in which standardized test scores are not available or utilized
 - Teachers of students assessed on alternate assessments
 - Smaller teacher caseloads for some student groups (e.g., students with disabilities, English language learners)

Range of State and District Approaches

- Existing measures
- Rigorous new measures
- Portfolios/products/performance/projects
- Student learning objectives



Measures must be rigorous, between two points in time, and comparable across classrooms.

Existing Measures

Strengths of This Measure	Challenges for This Measure
<ul style="list-style-type: none">▪ Already exist▪ Teacher familiarity and use▪ Not creating additional assessments/work▪ Possibly formative in nature	<ul style="list-style-type: none">▪ Validity (whenever a measure is used in a way that was not intended)▪ Concern over content validity▪ Fidelity and standardization

Delaware, Tennessee, Rhode Island

- Assembled group of practitioners
- Tightly facilitated meetings
- Group recommended measures
- Expert panel approves measures

National RTI Center

- Progress monitoring tools
- Tiers I, II, and III
- <http://www.rti4success.org/chart/progressMonitoring/progressmonitoringtoolschart.htm>

New Measures

Strengths of This Measure	Challenges for This Measure
<ul style="list-style-type: none">▪ Tests can be made to match specific grade or subject standards.▪ Assessments can be created to meet standards of validity and reliability.▪ Same assessment can be given across district/teachers.	<ul style="list-style-type: none">▪ More tests!▪ Time and cost-intensive approach▪ Paper-and-pencil tests that may not be appropriate as the sole measure, particularly in subjects requiring students to demonstrate knowledge and skills (art, music, etc.)▪ Capacity to build valid and reliable assessments

Hillsborough County, Florida

- Race to the Top Grantee
- Pre- and postassessment for each course
- Scores averaged over three years to determine teacher effectiveness

Use Portfolio/Products/ Performance/Projects

Strengths of This Measure	Challenges for This Measure
<ul style="list-style-type: none">▪ Evidence of growth can be documented over time using performance rubrics.▪ Portfolios and projects can reflect skills and knowledge that are not readily measured by paper-and-pencil tests.	<ul style="list-style-type: none">▪ Training for interrater reliability▪ Logistical challenge for group raters▪ Ensuring rigor

- New York and Rhode Island districts participating in the AFT Innovation (i3) project
- As in Delaware, teachers identify existing measures already used in classrooms.
 - Must develop pretests to establish knowledge and skills students need prior to project.
 - Panel of experts and practitioners evaluate and approve measures.

Logistical Rules for Measuring Student Growth

- Which students are counted for a teacher's growth measure?
- How long does a student need to be in a teacher's class for the teacher to be expected to contribute to his or her growth?
- What portion of a teacher's students needs to be included in the growth measure?
- How is later teacher assignment, absence, transfer accounted for in student growth measures?
- Should students who skipped a grade or are held back be excluded?
- How do we attribute growth to teachers who share responsibility for students?
- What happens when assessment data is missing for a student or group of students?

Guidance

Identify ways to ensure that the measures are informative, accurate, and defensible.

- Validate measures through a process of determining factors to be measured, for what purpose, and how the evidence gathered addresses the need (Herman, Heritage, & Goldschmidt, 2011).
- Ensure rigor and high standards in expectations for students, especially college- and career-ready standards (e.g., see the [Rigor Rubric](#) that Austin [Texas] Independent School District uses).

Guidance

Include measures that will help teachers improve their practice:

- Motivate teachers to examine their practice.
- Give teachers opportunities to discuss the results with their peers and supervisors, fostering a collaborative environment.
- Provide specific guidance materials, including protocols and processes developed to help teachers understand the use of student achievement data for student growth measures (Goe & Holdheide, 2011).

Interrater Reliability

Measuring Validity in Teacher Evaluation

- Examine the relationship between teacher practice and student learning
 - Example
 - Teacher practice measure: classroom observation ratings
 - Student learning measure: teacher-level added value
- A valid teacher observation instrument
 - Low observation ratings = low value-added scores
 - High observation ratings = high value-added scores

Defining Reliability in Teacher Evaluation

A combination of

- Instruments
- Rater training and certification
- Scoring designs

Importance of Interrater Reliability

Even with a terrific observation instrument, the results are meaningless if observers are not trained to agree on evidence and scoring.

- A teacher should get the same score no matter who observes him or her.

Interrater Reliability

- Interrater reliability is one element of an observational system:
 - Instruments
 - Raters
 - Scoring designs
- Three types of variability may influence teacher scores:
 - Teachers
 - Lessons
 - Raters
- There is not a single right metric for interrater agreement.
- Generalizability studies can help can assist in the design of cost-efficient systems that produce reliable scores (Hill et al., 2012).

Obstacles to Rater Accuracy

- Rater bias
- Leniency
- Central tendency
- Halo or horns effects

The Importance of Training

- Classroom observations are valid only if they are also reliable, and reliability is highly dependent on training.
- Who the observers are is less important than whether they have been adequately trained and calibrated.
- High levels of interrater reliability requires high-quality training that includes
 - Initial training on instruments and processes
 - Certification and recertification exams
 - Calibration exercises

Key Training Elements to Ensure Interrater Agreement

- Guidance on how to interpret evidence, including time on task and importance
- Advice and practice on the physical demands of observation (i.e., handling materials, monitoring time)
- Guidance on how to include or exclude expertise about what the teacher should be doing
- Opportunities for observers to compare practice ratings with those of “master coders”

Certification and Calibration

- The *certification exam* should cover all grades and subjects that the observer will observe. There are a variety of ways to reduce the time burden of certification:
 - Include a knowledge assessment of the observation rubric.
 - Mix shorter videos of practice with longer, full-lesson videos of practice.
- The *calibration exam* should test the observer on a representative selection of skills and content to ensure continued accuracy in rating.
- Informal calibration through discussion forums where observers can share challenges and best practices can have a big impact.
- Certification and calibration exams are high-stakes.

The Importance of Multiple Observers and Observations

- Using multiple observers—and multiple observations—improves the reliability of scores more than having longer observation periods.
 - Using a mix of shorter and longer observations can decrease cost while maintaining increased reliability of scores.
 - Using a mix of internal and external (school-based and not) can help mediate scores that may be inflated by principals or teachers who have scored well previously; it is important, however, to note that the vast majority of scores are comparable between principals and external evaluators.
- Comparing student growth scores to overall practice scores can help determine the validity of the practice scores.

(Sources: Ho & Kane 2013; Sartin et al., 2011)

Reliability Results With Various Combinations of Raters and Number of Lessons

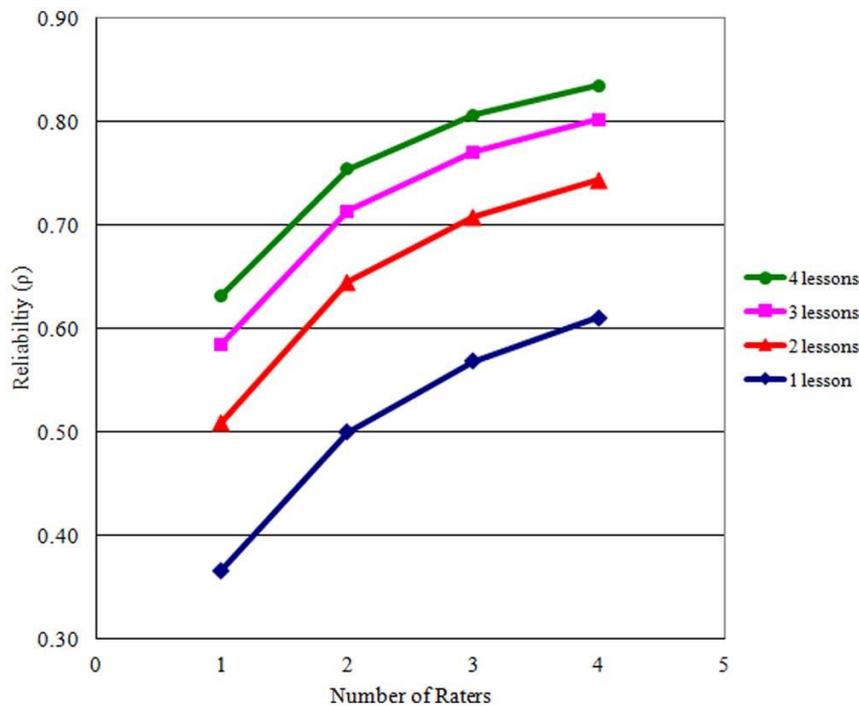


Figure 2. *Errors and Imprecision: The Reliability of Different Combinations of Raters and Lessons.* From Hill et al. (2012). Used with permission of author.

Lessons From the Measures of Effective Teaching (MET) Study

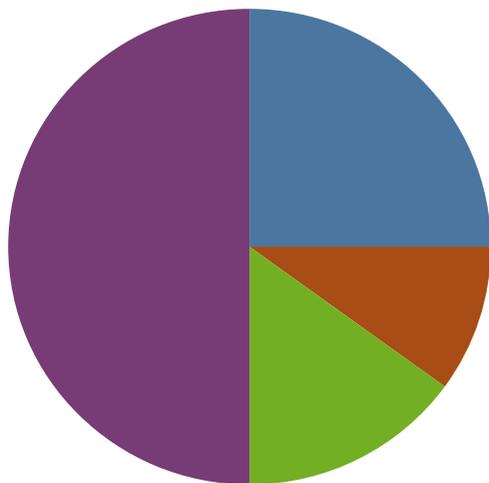
- The MET Study examined observers and interrater reliability in several districts implementing educator evaluations and found that
 - Interrater reliability depends on factors beyond teacher quality, such as the consistency of classroom context, student demographics, and differences between lessons.
 - Rater severity and course selection do not have a major impact on interrater reliability.
- The authors of the study recommend that
 - Observers undergo training and calibration prior to scoring.
 - Teachers be observed multiple times.
 - The district employ impartial observers from outside the school.

Lessons From the Measures of Effective Teaching (MET) Study

- The MET Study also examined what the most valid and reliable types of summative scores were. The MET Study found that outcomes were most valid when they combined
 - Student feedback (surveys)
 - Student learning (growth and/or achievement)
 - Observation
- The most valid way of combining these measures was to weight them comparably as part of a teacher's overall evaluation.

Combining Evaluation Measures for Rating Purposes

Numerical Approach



- Classroom observations
- Professionalism
- Professional goal setting
- Student growth

- Identify weight associated with each measure.
- Assign points to each measure and add or average together.
- Create and apply score ranges for each summative rating.

Metric	Indiv. Score	Weight	Final Rating
Classroom observations	88%	25%	0.22
Professional goal setting	90%	10%	0.09
Professionalism	76%	15%	0.11
Student growth	84%	50%	0.42
Summative teacher effectiveness score			0.84

Does Not Meet Standards	Partially Meets Standards	Meets Standards	Exceeds Standards
0.0–0.19	0.20–0.54	0.55–0.89	0.90–1.0

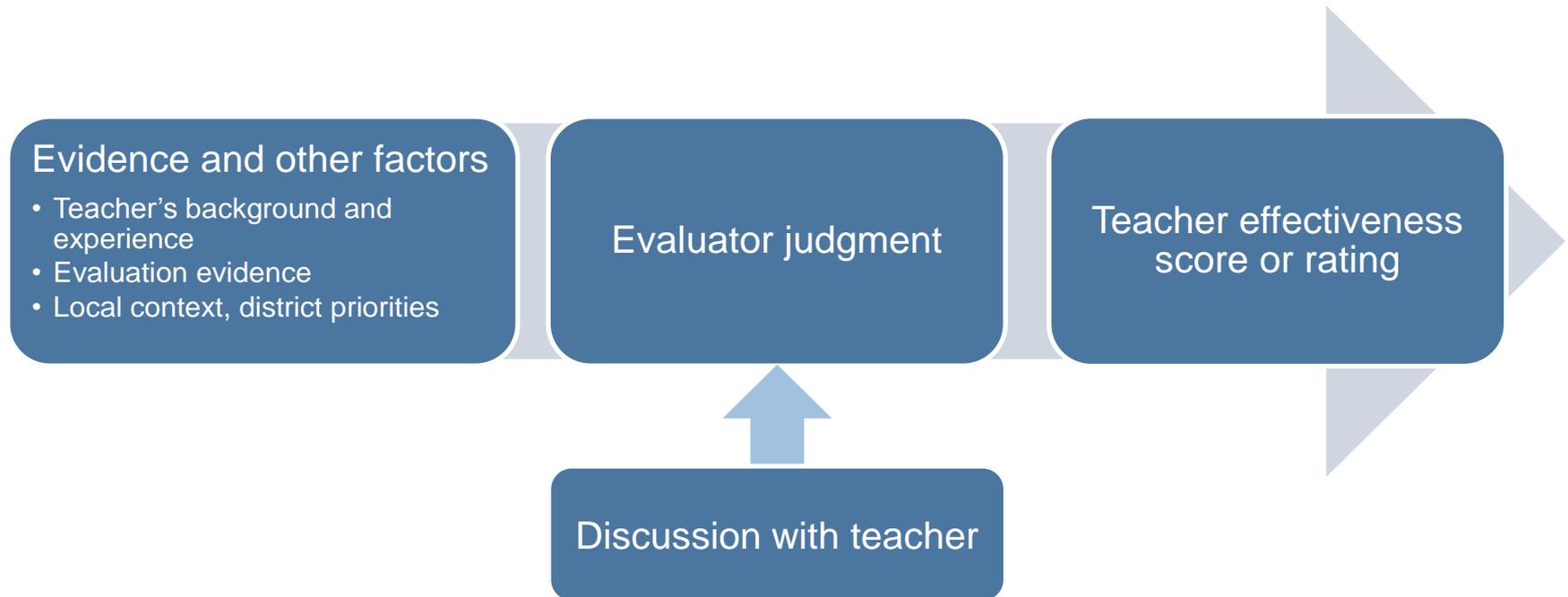
Profile Approach

- Gather and maintain evidence for multiple measures and rate educators separately on each measure.
- Combine results from disparate measures using a matrix, lookup table, or series of decision rules.

		Summative Professional Practice and Responsibility Rating				
		Distinguished	Accomplished	Proficient	Emerging	Unsatisfactory
Summative Student Growth Rating	4	Highly effective	Highly effective	Effective	Effective	Minimally effective
	3	Highly effective	Effective	Effective	Minimally effective	Ineffective
	2	Effective	Effective	Minimally effective	Minimally effective	Ineffective
	1	Minimally effective	Minimally effective	Minimally effective	Ineffective	Ineffective

Holistic Rating Approach

- Review the body of collected evidence and interpret it using the performance rubric to issue a single holistic rating for the educator.



Most Systems Use a Hybrid Approach

- Balances strengths and weaknesses of each pure approach.
- Incorporates stakeholder input and local context.
- Acknowledges the multiple levels of decision-making in rating performance.
- Breaks down the system into more easily communicated components.

Optional Implementation Rules

Minimum Competence Thresholds

- Create decision rules around minimum standards for some or all performance criteria that supersede other rules.
- Apply these rules to all or some educators (e.g., veteran, those nearing tenure).

Proficiency Progression

- Choose the performance criteria that are most critical for proficiency in the first year or phase.
- Increase minimum requirements year by year until desired proficiency standards are met.

Designing a Rating System and Setting Cut Scores

Considerations

- Where you set the bar and its effect on a teacher's final rating
- Model performance data
- Technical and policy considerations
- Ensuring the components and the overall system are valid

Closing Comments and Questions

- What do you need to know in order to move the work forward?
- What are your next steps for work?

References

- Donaldson, M. L. (2012). *Teachers' perspectives on evaluation reform*. Washington, DC: Center for American Progress. Retrieved from <http://www.americanprogress.org/wp-content/uploads/2012/12/TeacherPerspectives.pdf>
- Goe, L. & Holdheide, L. (2010). *Measuring teachers' contributions to student learning growth for nontested grades and subjects*. Naperville, IL: National Comprehensive Center for Teacher Quality. <http://www.tqsource.org/publications/MeasuringTeachersContributions.pdf>
- Heifetz, R. A., & Linsky, M. (2002). *Leadership on the line: Staying alive through the dangers of leading*. Cambridge, MA: Harvard Business School Press.
- Herman, J. L., Heritage, M., & Goldschmidt, P. (2011). *Developing and selecting assessments of student growth for use in teacher evaluation systems*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- 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/publications/when-rater-reliability-not-enough-teacher-observation-systems-and-case-g-study>
- Ho, A. D., & Kane, T. J. (2013). The reliability of classroom observations by school personnel. Seattle, WA: Bill and Melinda Gates Foundation. http://www.metproject.org/downloads/MET_Reliability%20of%20Classroom%20Observations_Research%20Paper.pdf

References

- Lachlan-Haché, L, Cushing, E., & Bivona, L. (2012). *Student learning objectives as measures of educator effectiveness: The basics*. Washington, DC: American Institutes for Research. Retrieved from http://educatortalent.org/inc/docs/SLOs_Measures_of_Educator_Effectiveness.pdf
- Lamb, L. M., & Schmitt, L. N. T. (2012). AISD REACH program update, 2010–11: Participant feedback. Austin, TX: Austin Independent School District Department of Research and Evaluation.
- 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>
- Slotnik, W. J., & Smith, M. D. (2004). *Catalyst for change: Pay for performance in Denver*. Community Training and Assistance Center. Retrieved from <http://www.ctacusa.com/PDFs/Rpt-CatalystChangeFull-2004.pdf>
- The New Teacher Project. (2012). *“MET” made simple: Building research-based teacher evaluations*. http://tntp.org/assets/documents/TNTP_METMadeSimple_2012.pdf

Robin Chait
Director, Teaching and Learning
Office of the State Superintendent
of Education
202-481-3783
robin.chait@dc.gov

Angela Minnici
Center on Great Teachers and
Leaders
Phone: 202-403-6321
aminnici@air.org
cjacques@air.org

1000 Thomas Jefferson Street NW
Washington, DC 20007-3835
877-322-8700
www.gtlcenter.org
gtlcenter@air.org