Teacher Evaluation Systems

Addressing Common Implementation Challenges

March 2013

Center on
GREAT TEACHERS & LEADERS
at American Institutes for Research

AIR
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 with the problem to do the work of solving it.
  - 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

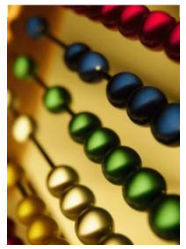
Brainstorming Activity

- 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 Assessments

Strengths of This Measure	Challenges for This Measure
▪ Already exist ▪ Teacher familiarity and use ▪ Not creating additional assessments/work ▪ Possibly formative in nature	▪ 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 Assessments

Strengths of This Measure	Challenges for This Measure
▪ 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.	▪ 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
▪ 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.	▪ 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.

What Are Student Learning Objectives?

- A goal that demonstrates a teacher's impact on student learning within a given interval of instruction.
- A measurable, long-term academic target written by an individual teacher or a teacher team.
- A process that allows teachers to demonstrate their impact on student learning within a given interval of instruction
- Student baseline data is collected
- Appropriate objectives are set for students
- Students are assessed at the end of the interval

High-Quality SLOs

Most SLOs include or address criteria like the following:

1. Baseline and trend data
2. Student population (general and special needs)
3. Interval of instruction
4. Standards and content
5. Assessment(s)
6. Growth target(s)
7. Rationale for growth target(s)

Student Learning Objectives

Strengths of This Measure

- Provide the opportunity to discuss teacher expectations and goals and reinforce teacher practices.
- Feedback from SLOs can provide detailed instructional goals for educator professional development plans.
- Flexible:
  - They can be tailored to specific grade levels, subjects, students, and individual teachers.
  - SLOs encourage collaboration among educators to set and achieve goals and provide educators with ownership over their evaluations

The Illinois Performance Evaluation Advisory Council (PEAC) decided to use SLOs in its model evaluation system for Type III assessments and recommends that districts use SLOs for Type III assessments.

Student Learning Objectives

Challenges for This Measure
▪ SLOs can be time-intensive to develop and evaluate while meeting requirements for rigor and comparability. ▪ SLOs require high-quality assessments, which may be difficult to identify or challenging to develop. ▪ SLOs require specific guidance to help educators define appropriate differentiated targets for students. ▪ SLOs can require a shift in school culture to support continuous improvement of educators.

State SLO Policies and Resources

State	SLO Policy	Resources
Georgia	• SLOs are set by the district. • The state reviews and approves the SLOs and provides supports for district implementation, including the development, review, and approval of SLO assessments. • The district creates course-level SLOs for all nontested grades and subjects, which teachers modify and use with specific targets for their individual classrooms.	Georgia SLO Operations Manual
Indiana	• Teachers develop SLOs, and administrators or groups of district leaders approve them. • The state provides trainings, webinars, and resources to assist with SLO development.	Indiana RISE handbook Indiana SLO Manual

State SLO Policies and Resources

State	SLO Policy	Resources
New York	• The state specifies the number and types of SLOs teachers must create based on teaching assignment and specifies which assessments are acceptable for the SLO. • Districts in New York can establish requirements or recommendations for assessments and rating scales to be used with SLOs to ensure consistency of expectations across schools.	New York Locally Selected Measures New York SLO Resources
Ohio	• Teachers develop SLOs, and administrators or groups of district leaders approve them. • The state plans to randomly audit SLOs at the local level.	Ohio SLO Process
Rhode Island	• Teachers develop SLOs, and administrators or groups of district leaders approve them.	Rhode Island SLO Materials

Additional SLO Examples and Resources

- Austin, Texas: Austin SLO Manual
- Denver, Colorado: Denver Schools Student Growth Objective and Monitoring Process
- Connecticut: Connecticut's System for Educator Evaluation and Development
- District of Columbia: DCPS IMPACT Guidebooks
- Louisiana: Louisiana COMPASS Teacher Evaluation Guidebook and Pointe Coupee Parish School System Student Learning Targets 2012–2013
- Maryland: Maryland Teacher and Principal Evaluation Guidebook
- Wisconsin: Wisconsin SLO Process Manual

Logistical Rules for Measuring Student Growth

- Which students are counted for a teacher's growth measure?
  - Attendance and continuous enrollment
  - Minimum number of student scores for validity
  - Grade advancement and retention
  - Missing assessment data
- How is teacher assignment, accounted for in student growth measures?
  - Teacher attendance and late assignment
  - Co-teaching scenarios

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

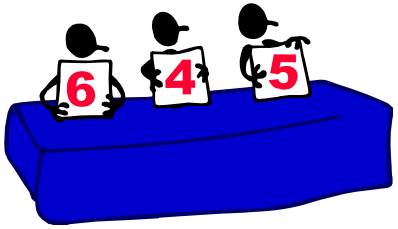
Include measures that will help teachers improve their practice:

- 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).

Rubric Design

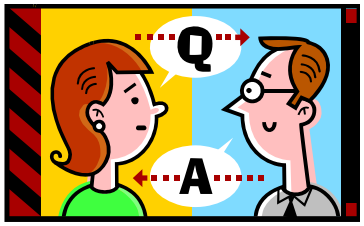
A Well Designed Rubric Can Improve…

- Scoring



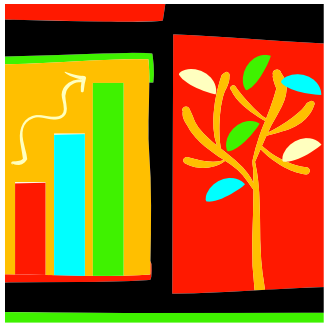
↘

- Feedback



↘

- Professional growth



Types of Rubrics

- Observation Rubric
  - Evidence from single or multiple observations
- Professional Practice Rubric
  - Observation evidence
  - Artifact Review
  - Other Measures (e.g. surveys)
- Evaluation Rubric
  - Professional Practice evidence
  - Student growth evidence
  - Goal-setting and/or professional development (if applicable)

Note: Rubrics may also include scoring metrics or notes

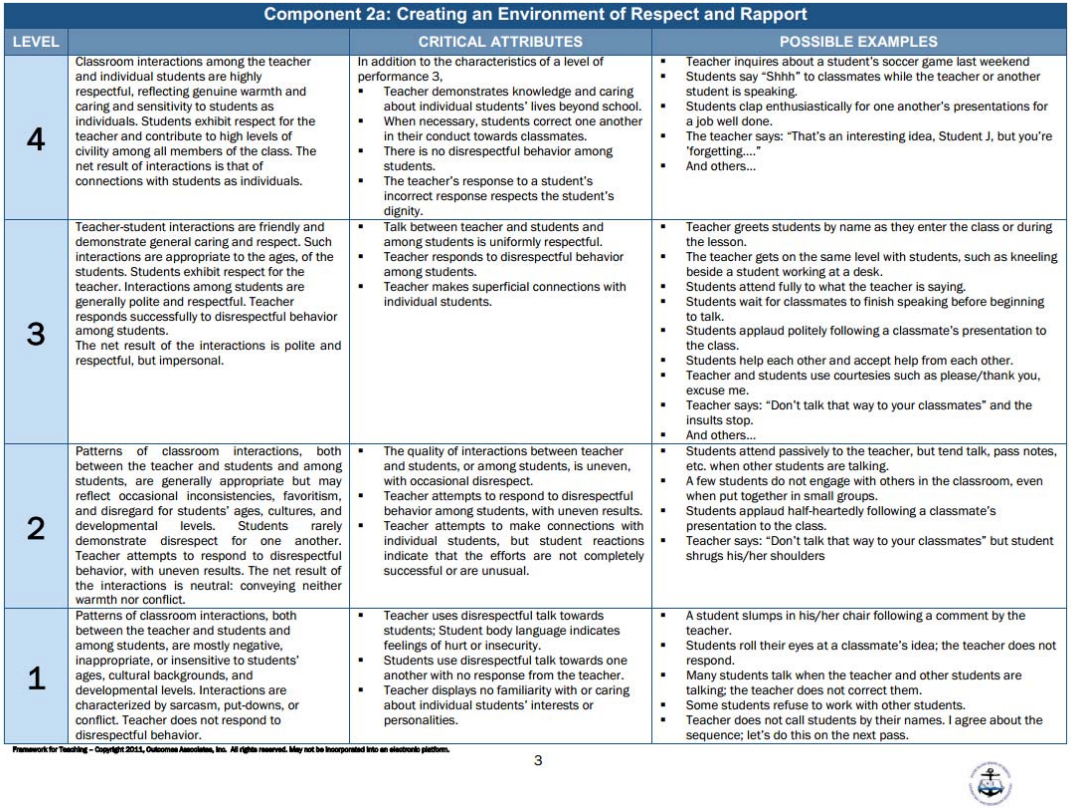
Ensuring High Quality Design

Common problems:

- A lack of vertical alignment
- Vague Language
- A lack of distinctness between domains and indicators

Can be addressed by:

- Stakeholder engagement
- High quality training
- Clear descriptions and examples
- Alignment to leading standards or other rubrics (InTASC, Danielson)
- Expert Review



Component 2a: Creating an Environment of Respect and Rapport			
LEVEL		CRITICAL ATTRIBUTES	POSSIBLE EXAMPLES
4	Classroom interactions among the teacher and individual students are highly respectful, reflecting genuine warmth and caring and sensitivity to students as individuals. Students exhibit respect for the teacher and contribute to high levels of civility among all members of the class. The net result of interactions is that of connections with students as individuals.	In addition to the characteristics of a level of performance 3, ▪ Teacher demonstrates knowledge and caring about individual students' lives beyond school. ▪ When necessary, students correct one another in their conduct towards classmates. ▪ There is no disrespectful behavior among students. ▪ The teacher's response to a student's incorrect response respects the student's dignity.	▪ Teacher inquires about a student's soccer game last weekend ▪ Students say "Shhh" to classmates while the teacher or another student is speaking. ▪ Students clap enthusiastically for one another's presentations for a job well done. ▪ The teacher says: "That's an interesting idea, Student J, but you're 'forgetting...." ▪ And others...
3	Teacher-student interactions are friendly and demonstrate general caring and respect. Such interactions are appropriate to the ages, of the students. Students exhibit respect for the teacher. Interactions among students are generally polite and respectful. Teacher responds successfully to disrespectful behavior among students. The net result of the interactions is polite and respectful, but impersonal.	▪ Talk between teacher and students and among students is uniformly respectful. ▪ Teacher responds to disrespectful behavior among students. ▪ Teacher makes superficial connections with individual students.	▪ Teacher greets students by name as they enter the class or during the lesson. ▪ The teacher gets on the same level with students, such as kneeling beside a student working at a desk. ▪ Students attend fully to what the teacher is saying. ▪ Students wait for classmates to finish speaking before beginning to talk. ▪ Students applaud politely following a classmate's presentation to the class. ▪ Students help each other and accept help from each other. ▪ Teacher and students use courtesies such as please/thank you, excuse me. ▪ Teacher says: "Don't talk that way to your classmates" and the insults stop. ▪ And others...
2	Patterns of classroom interactions, both between the teacher and students and among students, are generally appropriate but may reflect occasional inconsistencies, favoritism, and disregard for students' ages, cultures, and developmental levels. Students rarely demonstrate disrespect for one another. Teacher attempts to respond to disrespectful behavior, with uneven results. The net result of the interactions is neutral: conveying neither warmth nor conflict.	▪ The quality of interactions between teacher and students, or among students, is uneven, with occasional disrespect. ▪ Teacher attempts to respond to disrespectful behavior among students, with uneven results. ▪ Teacher attempts to make connections with individual students, but student reactions indicate that the efforts are not completely successful or are unusual.	▪ Students attend passively to the teacher, but tend talk, pass notes, etc. when other students are talking. ▪ A few students do not engage with others in the classroom, even when put together in small groups. ▪ Students applaud half-heartedly following a classmate's presentation to the class. ▪ Teacher says: "Don't talk that way to your classmates" but student shrugs his/her shoulders
1	Patterns of classroom interactions, both between the teacher and students and among students, are mostly negative, inappropriate, or insensitive to students' ages, cultural backgrounds, and developmental levels. Interactions are characterized by sarcasm, put-downs, or conflict. Teacher does not respond to disrespectful behavior.	▪ Teacher uses disrespectful talk towards students; Student body language indicates feelings of hurt or insecurity. ▪ Students use disrespectful talk towards one another with no response from the teacher. ▪ Teacher displays no familiarity with or caring about individual students' interests or personalities.	▪ A student slumps in his/her chair following a comment by the teacher. ▪ Students roll their eyes at a classmate's idea; the teacher does not respond. ▪ Many students talk when the teacher and other students are talking; the teacher does not correct them. ▪ Some students refuse to work with other students. ▪ Teacher does not call students by their names. I agree about the sequence; let's do this on the next pass.

3

Examples

- Rhode Island Professional Practice Rubric
  - Description of overall performance, indicators, and examples
  - Online platform for documentation
- Ohio Teacher Performance Evaluation Rubric
  - Evidence sources specified
  - Evidence documentation
- Colorado Rubric for Evaluation Teachers
  - Checklist across performance levels
  - Differentiation between observable and unobservable indicators

All have descriptions of performance at each level

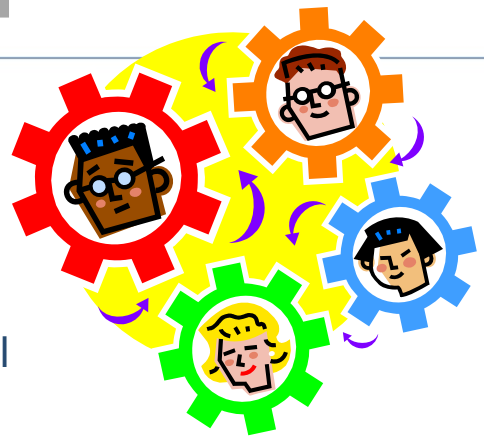
Personnel: Challenges and Solutions

The Role of the Evaluator

- At the heart of an evaluation system are personnel who can accurately assess teacher performance, communicate the results of that assessment to teachers, and help them plan for their professional growth.
- *Who the observers are is less important than whether they*
  - *Receive adequate training*
  - *Receive ongoing feedback on their performance*
  - *Periodically reassess and calibrate their observation skills*

The Role of the Evaluator

- Challenges for principals as evaluators:
  - Limited time
  - Limited background in certain subjects.
  - Lack of teaching experience
  - Limited skills and experience with conducting professional conversations and coaching



- Although the principal should conduct the final summative evaluation, other trained staff may observe teachers:
  - Principal
  - Vice-principal
  - Content-area administrators
  - Teacher leaders
  - Peer Observers (full-time or part-time)

Training Personnel to be Evaluators

- High-quality training builds trust.
- High-quality training needs to include:
  - Certification exams and calibration exercises
  - How to provide high-quality feedback
- Evaluation and feedback are a professional responsibility for principals.
  - What responsibilities can other staff assist with (i.e. observation, feedback)?
  - How do teachers and principals connect evaluation results to professional development and learning?

Interrater Reliability

Assessing Validity in Teacher Evaluation

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

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.

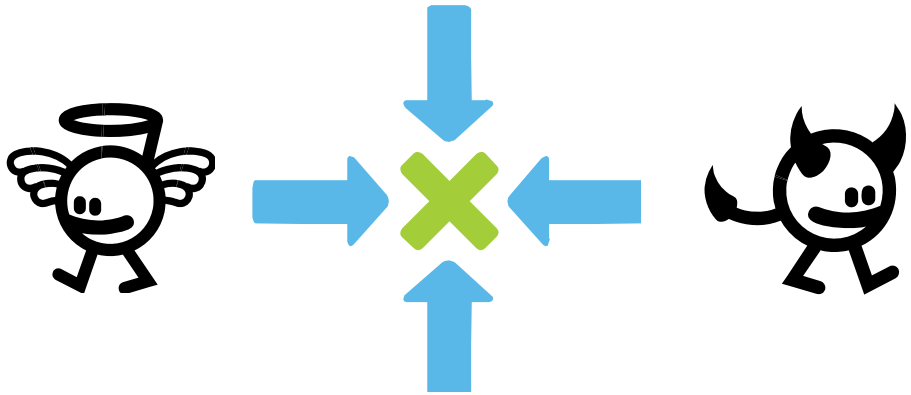
→ A teacher should get the same score no matter who observes him or her.

Interrater Reliability

- Interrater reliability is one element of an observational system:
  - Instruments
  - Raters
  - Scoring designs
- 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 or Severity
- Central tendency
- Halo or Horns effects



↘

High-quality training

Formal calibration and certification

Informal calibration through meetings

The Importance of Multiple Observers and Observations

- Using multiple observers—and multiple observations—improves the reliability of scores more than having longer observation periods.

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

Reliability Results With Various Combinations of Raters and Number of Lessons

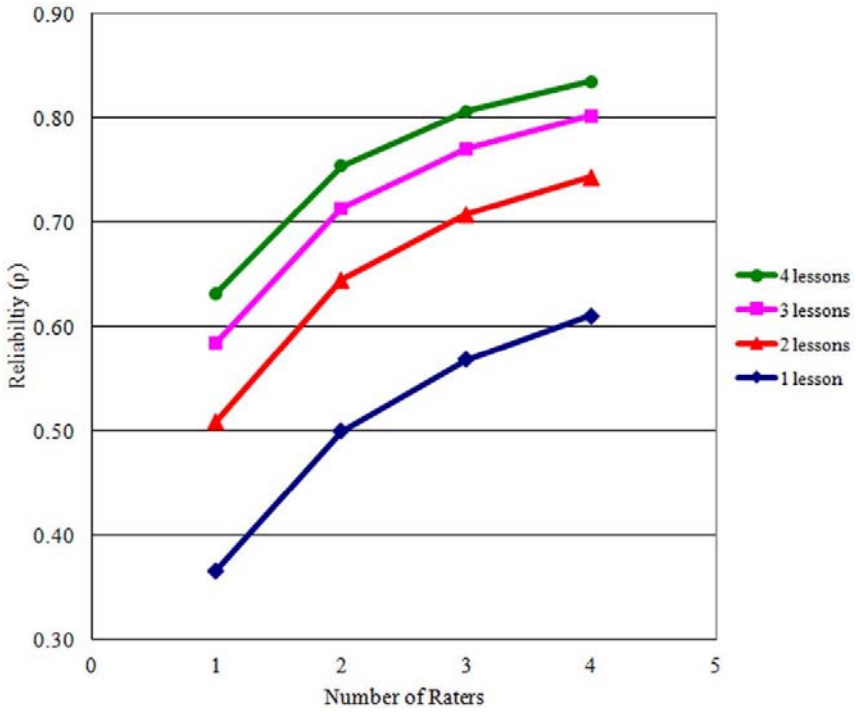


Figure 2. *Errors and Imprecision: The Reliability of Different Combinations of Raters and Lessons.* From Hill et al. (2012). Used with permission of author.

Lessons From the Measures of Effective Teaching (MET) Study

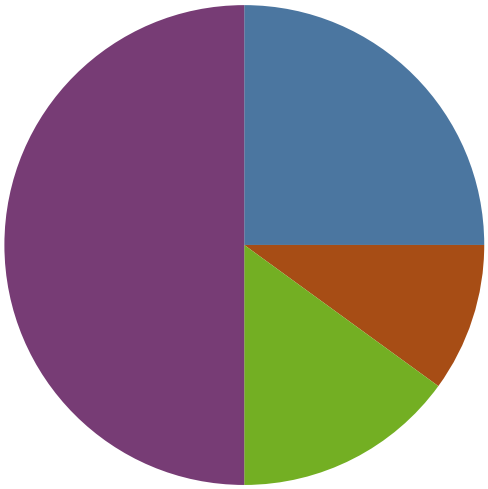
- Interrater reliability depends on factors beyond teacher quality, such as the consistency of classroom context, student demographics, and differences between lessons.
- Rater severity isn't an issue for the majority of observers but should be monitored
- The authors of the study recommend that
  - 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

- Evaluation 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



- 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

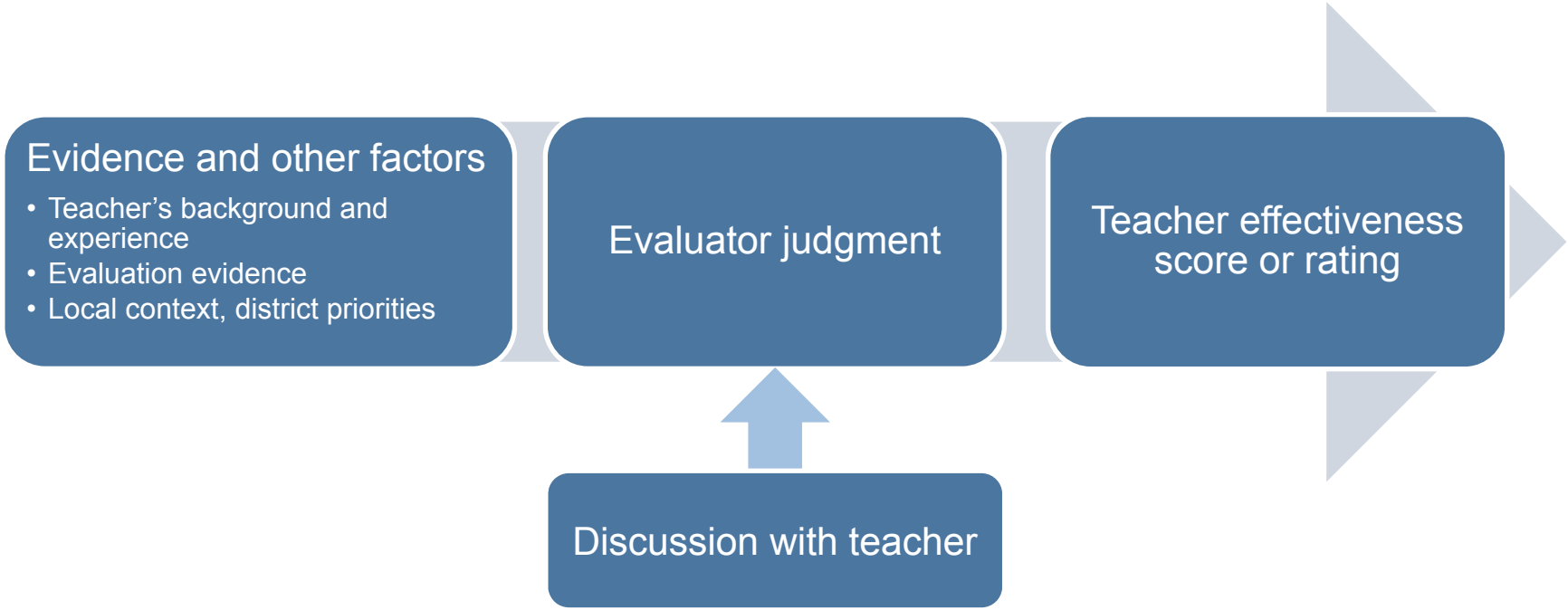
Profile Approach

- 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 will effect a teacher's final rating and the distribution of scores
- Using model performance data can help predict outcomes
- Ensuring that technical and policy needs and priorities are taken into account
- Ensuring the components and the overall system are valid

Lessons Learned

- Consider district policy priorities
- Use multiple measures of student growth when possible
- Use multiple observers of practice
- Recalibrate observers frequently
- Use real data to model summative scoring methods

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