The New Paradigm for Teacher Evaluation – What Are The Implications for Improving Practice?
Tony Milanowski
Westat
Teacher Evaluation Moves from Back Burner to Barnburner

- Federal TIF, RTT & ESEA Waivers → a new paradigm for teacher evaluation: at least 3 levels of overall effectiveness as determined by a combination of observations of practice and student growth (with possible addition of other measures such as student surveys)

- Why Now?
  - New technology: “value-added” + standards-based practice evaluation
  - Value-added” studies of teacher effects → “teachers matter”
  - Disconnect between distribution of teacher ratings and student proficiency: 98% of teachers rated “proficient” but in many large districts only about half the students
  - Advocacy: “Widget Effect”, Gates MET Project
## Two Common Approaches to Combining Practice Ratings & Growth

<table>
<thead>
<tr>
<th>Component</th>
<th>Weight</th>
<th>Rating</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Growth</td>
<td>40%</td>
<td>3</td>
<td>120</td>
</tr>
<tr>
<td>Practice</td>
<td>40%</td>
<td>4</td>
<td>160</td>
</tr>
<tr>
<td>Student Survey</td>
<td>20%</td>
<td>3</td>
<td>60</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Overall Rating</th>
<th>100-175= 1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>176-225= 2</td>
</tr>
<tr>
<td></td>
<td>226-325= 3</td>
</tr>
<tr>
<td></td>
<td>326-400= 4</td>
</tr>
<tr>
<td></td>
<td>300</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Student Growth</th>
<th>Teaching Practice</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
</tr>
<tr>
<td>Low</td>
<td>4</td>
</tr>
<tr>
<td>Avg.</td>
<td>3</td>
</tr>
<tr>
<td>High</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Overall</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>100-175</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
What Are the Implications for Improving Performance?

- Student growth indicators + more rigorous practice evaluation should help to identify & remove poorest performers & recognize and learn from best performers
- Student growth could check the inherent tendency for leniency in practice ratings
- More differentiation + use of evaluation for stakes ➔ better assessment practice
  - More reliable, more credible feedback
  - Rubric ratings may help pinpoint areas in need of improvement
Adding Growth Allows More Differentiation of Performance

Observational Ratings From One TIF District

Combination of Practice Ratings & Growth Measures
To Make Use of the New Information, We Will Need:

- Disaggregation of data
  - Ratings of individual rubric components to pinpoint areas in need of improvement
- More than just feedback
- An explicit strategy to link evaluation ratings to professional development efforts.
- Sustained & directed PD strategies

What Strategies Are TIF Grantees Using?
National Institute for Excellence in Teaching - TAP

- 4-6 observations each year provide opportunity to demonstrate improvement
- Evaluators trained to give focused, specific feedback at post-observation conferences
- Video library illustrating practice at various rubric levels and examples of field-tested instructional strategies/lessons
- School week restructuring to provide time for small groups to develop, learn, & evaluate new strategies
- Individualized coaching/modeling by mentor eachers ling & coaching
Hillsborough, FL

- Individual PD plans updated quarterly & determined in part by evaluation system post conference feedback
- Courses specifically designed to improve performance on the evaluation rubric
  - Administrators can prescribe participation
- Each school has a teacher leader responsible for formative observation, feedback, modeling, helping choose PD and facilitating PLC’s
- Quarterly school meetings to discuss PD issues, obtain teachers’ views on PD needs & effectiveness
- Internet-based platform to share resources and access PD guide and video libraries of effective lessons
  - 8 video cameras for each school
Rutgers University + 4 New Jersey Districts

- School-based mentor teachers (1:24 ratio) supported by a central cadre of leadership teachers
- Formative instructional strategy & behavior management observation tool, including teacher self-rating, compliments main observation rubric (Framework for Teaching)
- Formative internet-based teacher logs track time on topics, content coverage, and basic instructional strategies so teachers can look back as well as set their own targets for practice change
- Self-paced, internet-based content library & PD modules keyed to specific competencies measured by evaluation process & personalized to teacher’s ratings
Example Teacher Log Reports

<table>
<thead>
<tr>
<th>Skills</th>
<th>% of Total Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1 Number/Operations</td>
<td>15.5%</td>
</tr>
<tr>
<td>S1C1P1 Compare/order</td>
<td>0.6% 1 hrs 15 mins</td>
</tr>
<tr>
<td>S1C1P2 Classify rational/irrational</td>
<td>1.3% 3 hrs 15 mins</td>
</tr>
<tr>
<td>S1C1P3 model read numbers</td>
<td>1.5% 3 hrs 15 mins</td>
</tr>
<tr>
<td>S1C1P4 model/solve absolute value</td>
<td>0.5% 1 hrs 10 mins</td>
</tr>
<tr>
<td>S1C2P1 Factors/multiples/prime</td>
<td>1.6% 3 hrs 30 mins</td>
</tr>
<tr>
<td>S1C2P2 Rational number effects</td>
<td>1.0% 2 hrs 10 mins</td>
</tr>
<tr>
<td>S1C2P3 Percent Inc., dec, simple interest</td>
<td>3.1% 6 hrs 45 mins</td>
</tr>
<tr>
<td>S1C2P4 Sci/scientific notation conver.</td>
<td>1.4% 3 hrs 5 mins</td>
</tr>
<tr>
<td>S1C2P5 Simplify expression</td>
<td>3.4% 7 hrs 20 mins</td>
</tr>
<tr>
<td>S1C3P1 Estimate</td>
<td>0.7% 1 hrs 35 mins</td>
</tr>
<tr>
<td>S1C3P2 Estimate on number line</td>
<td>0.7% 1 hrs 35 mins</td>
</tr>
<tr>
<td>S1C2P7 Data Analy., Prob., Discrete Math</td>
<td>18.5%</td>
</tr>
<tr>
<td>S2C1P1 Use displays, box/whisker, scatterplot</td>
<td>5.2% 11 hrs 12 mins</td>
</tr>
<tr>
<td>S2C1P2 Inferences, 2 data sets</td>
<td>1.2% 2 hrs 40 mins</td>
</tr>
<tr>
<td>S2C1P3 Summary/shape of distribution</td>
<td>1.2% 3 hrs 30 mins</td>
</tr>
<tr>
<td>S2C1P4 Bias, effective presentation</td>
<td>0.2% 3 hrs 30 mins</td>
</tr>
<tr>
<td>S2C1P5 Evaluate design</td>
<td>3.2% 6 hrs 30 mins</td>
</tr>
<tr>
<td>S2C2P1 Theoretical/experimental</td>
<td>3.2% 6 hrs 30 mins</td>
</tr>
<tr>
<td>S2C2P2 Compare outcome/prediction</td>
<td>2.3% 5 hrs 30 mins</td>
</tr>
<tr>
<td>S2C2P3 Sample space for dep/index</td>
<td>1.5% 3 hrs 15 mins</td>
</tr>
<tr>
<td>S2C3P1 Counting-order/presentation</td>
<td>2.4% 3 hrs 5 mins</td>
</tr>
<tr>
<td>S2C4P1 Counting-factorial notation</td>
<td>2.4% 3 hrs 5 mins</td>
</tr>
<tr>
<td>S2C4P1 Solve graph problems</td>
<td>2.4% 3 hrs 10 mins</td>
</tr>
</tbody>
</table>

Time Allocation by Domain

- S4 Geometry and Measurement: 11%
- Custom Skills/Activities: 30%
- S3 Patterns, Algebra, and Func.: 17%
- S2 Data Analy., Prob., Discrete Math: 19%
- S1 Number/Operations: 16%
- S5 Structure, Logic: 1%
- Time Not Available: 7%
Is something missing?

- Lots of good ideas on improving observed practice, but a lot less on improving student growth
  - Lots of districts give benchmark/interim assessments, but few seem to systematically use results to change teaching

• Formative student assessment is critical to teacher acceptance of the new paradigm because:
  - Feedback on student growth based on state tests comes back too late for teachers to do anything with it
  - The link between practice ratings and student growth is weaker than we’d like → getting better ratings does not ensure having higher growth
  - It may be easier & faster to improve growth by using diagnostic testing to guide practice change
Practice Ratings Are an Imperfect Predictor of Growth
Any Evidence That Evaluation Can Improve Performance?

- CPRE research on early adopters of “standards-based” evaluation (Framework for Teaching)
  - Promoted reflection on practice
  - Focused teachers on what was evaluated
  - Provided teachers with help on classroom management, ‘tips’, materials, solutions to specific problems
  - Most influence on new/junior teachers
  - At some sites, anecdotal evidence that poorer teachers left
Later Results in Cincinnati

- Veteran math teachers who went through the evaluation process had higher levels of value-added after the evaluation than before.
  - The average effect was 4.5 percentile points
  - [http://educationnext.org/files/ednext_20124_taylortyler.pdf](http://educationnext.org/files/ednext_20124_taylortyler.pdf)

- Long-term trend toward improved student achievement
  - Long term improvement in state performance index from 53 in 2000 to 87 in 2010-11
  - Above average student growth last 3 years
Standards-based Evaluation in Chicago

• More differentiation of evaluation ratings (70% in top two categories vs. 93% under old system)

• Principals reported teacher practice changes in planning, grouping, assessment, organization

• Principals’ conversations with teachers were more reflective, evidence-based, & focused on improvement, but many principals lacked the instructional coaching skills to have deep discussions of teaching practice

• Now being phased-in, with student growth, in ½ of Chicago schools

http://ccsr.uchicago.edu/sites/default/files/publications/Teacher%20Eval%20Report%20FINAL.pdf
Practice Improvement in TAP Schools
Source: NIET TAP Research Summary [http://www.tapsystem.org/publications/]

Observed Instructional Skills
National Cohort of Teachers: 2007-08, 2008-09, 2009-10

Average Points on 1-5 Scale

Average of TAP Instructional Domain indicators for teachers with data from all three years
n = 378 teachers
New Research on MyTeachingPartner

• Small randomized control trial of iterative web-based coaching process based on CLASS practice observation system, developed by Robert Pianta and colleagues at U. of Virginia

• Master teachers, trained in the CLASS system, provided individualized, regular, and systematic feedback to teachers in treatment group based on CLASS assessments; others got PD as usual

• Students from next year’s treatment group’s classes had net greater net gain (.22 standard deviations, 9 percentile points)

• Gain was also associated with improvement in practice as measured by CLASS, consistent with causal relationship

https://appam.confex.com/appam/2012/webprogram/Paper4012.html
For more on:

- **The New Paradigm**
  
  US Dept of Education:
  

- **Gates Measuring Effective Teaching**
  
  [http://www.gatesfoundation.org/college-ready-education/Pages/measures-effective-teaching-project.aspx](http://www.gatesfoundation.org/college-ready-education/Pages/measures-effective-teaching-project.aspx)

- **CPRE Research**
  

- **TAP**
  

- **Using student achievement growth to measure teacher performance**
  