Measuring Teacher Effectiveness in Untested Subjects and Grades

Laura Goe, Ph.D.

Council of Chief State Schools Officers
State Consortium on Educator Effectiveness

Webinar ♦ February 8, 2011
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 Performance Research Group 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

➢ Educational Testing Service
  • Technical assistance, research, dissemination
The goal of teacher evaluation

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

TO IMPROVE TEACHING AND LEARNING
Federal priorities (August 2010)

From “Race to the Top” and reiterated in the August 5, 2010 Federal Register (Vol. 75, No. 150) “Secretary’s Priorities for Discretionary Grant Programs”

• Teachers should be evaluated using state standardized tests where possible

• For non-tested subjects, other measures (including pre- and post-tests) can be used but must be “rigorous and comparable across classrooms” and must be “between two points in time”

• Multiple measures should be used, such as multiple classroom evaluations
Evaluation Models

- Austin, TX
- Delaware
- Georgia
- Hillsborough, FL
- Rhode Island
- TAP (Teacher Advancement Program)
- Washington, DC
Evaluation System Models

Austin (Student Learning Objectives)
http://www.austinisd.org/inside/initiatives/compensation/slos.phtm

Delaware Model
http://www.doe.k12.de.us/csa/dpasii/student_growth/default.shtml

Georgia’s CLASS Keys
Rubric:
p=6CC6799F8C1371F6B59CF81E4ECD54E63F615CF1D9441A92E28BFA2A0AB27E3E&Type=D
Evaluation System Models (cont’d)

Hillsborough, Florida
http://communication.sdhc.k12.fl.us/empoweringteachers/?page_id=317

Rhode Island Model
http://www.ride.ri.gov/educatorquality/EducatorEvaluation/Docs/Working%20Group%

Teacher Advancement Program
http://www.tapsystem.org/

Washington DC IMPACT Guidebooks
http://www.dc.gov/DCPS/In+the+Classroom/Ensuring+Teacher+Success/IMPACT+(Performance+Assessment)/IMPACT+Guidebooks
Questions to ask about models

- Are they “rigorous and comparable across classrooms”?
- Do they show student learning growth “between two points in time”?
- Are they based on grade level and subject standards?
- Do they allow teachers from all subjects and grades (not just 4-8 math & ELA) to be evaluated with evidence of student learning growth?
Student Learning Objectives:

- Teachers determine two SLOs for the semester/year
- One SLO must address all students, other may be targeted
- Use broad array of assessments
- Assess student needs more directly
- Align classroom, campus, and district expectations
- Aligned to state standards/campus improvement plans
- Based on multiple sources of student data
- Assessed with pre and post assessment
- Targets of student growth
- Peer collaboration
# Rubric for student learning objectives

<table>
<thead>
<tr>
<th>APPROVAL</th>
<th>NEEDS REVISION</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LEVEL 4</strong></td>
<td><strong>LEVEL 3</strong></td>
</tr>
<tr>
<td>Indicates expert use of data, rigorous goal-setting based on student and community strengths and needs, incorporates appropriate valid/reliable assessments, considers teacher’s own prior performance, and demonstrates alignment with campus improvement goals.</td>
<td>Indicates good use of data and acceptable level of rigor for students, considers teacher’s prior performance, and utilizes appropriate assessments, but is not aligned with broader campus goals and does not incorporate the strengths of students or the school community.</td>
</tr>
</tbody>
</table>

### Needs Assessment/Rationale: What are the needs?

- Campus data are reviewed for areas of strength and need (within subject area, within grade level, within student group, examining the Campus Improvement Plan (CIP), etc.).
- Classroom data are reviewed for areas of strength and need (by subject area, by student group, by concepts/skills/behavior).
- Campus needs and strengths are not incorporated.

### Learning Content/Context and Student Group: What and who is targeted?

- Targets specific academic concepts, skills, or behaviors based on TEKS/TAKS Objective
- Targets the needs of the identified population
- Considers demonstrated strengths of identified population, as well as classroom & school community
- Targets year-long (or semester-long) concepts, skills, or behaviors
- Supports goals of the Campus Improvement Plan (CIP)
- Classroom data are reviewed for areas of need, but needs of specific student groups are not examined in depth, and strengths are not identified.
- Classroom needs and strengths are not incorporated.
<table>
<thead>
<tr>
<th>Learning Objective: What will students learn?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Based on the identified student needs.</td>
</tr>
<tr>
<td>Supports goals of the CIP.</td>
</tr>
<tr>
<td>Is rigorous.</td>
</tr>
<tr>
<td>Is a good example of ongoing, reflexive practice.</td>
</tr>
<tr>
<td>Provides clear focus for instruction and assessment.</td>
</tr>
<tr>
<td>Is measurable.</td>
</tr>
<tr>
<td>Reflects strengths of students and school community.</td>
</tr>
<tr>
<td>Does not relate to goals of CIP.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Outcome Assessment: How will you know whether they learned it?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aligns with the targeted learning content area.</td>
</tr>
<tr>
<td>Relationship with learning objective is apparent.</td>
</tr>
<tr>
<td>Has been demonstrated as reliable and valid for targeted students.</td>
</tr>
<tr>
<td>Follows guidelines for appropriate assessments.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Performance Target: What is your goal for student achievement?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicts performance based on past performance of students when available.</td>
</tr>
<tr>
<td>Is a rigorous expectation for students.</td>
</tr>
<tr>
<td>Is a rigorous expectation for teachers, based on past performance.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rigor: How rigorous is your SLO?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content is challenging, complex and progressively deepens knowledge of core content</td>
</tr>
<tr>
<td>Content is thought-provoking requiring high thinking demand</td>
</tr>
<tr>
<td>Requires analytical thinking and active use of knowledge</td>
</tr>
<tr>
<td>Content is relevant to life/ experiences.</td>
</tr>
</tbody>
</table>
SLO Model Strengths/Weaknesses

**Strengths**

- Teachers take an active role in determining student learning goals
- Good professional growth opportunity for teachers
- If objectives are of high-quality and teachers plan instruction to meet them, students should benefit

**Weaknesses**

- Heavily dependent on administrator understanding and time commitment to supervision
- Not “comparable across classrooms” because teachers set the objectives and they will vary widely
- “Rigor” dependent on evaluators’ understanding and/or having an appropriate rubric
“Rhode Island Model” is another example of an SLO Model

- Under consideration, not yet implemented
  - Teachers measure student growth by setting student academic goals aligned to standards
  - Principals, during the goal setting process, will confer with teachers to establish each goal’s degree of ambition and select the appropriate assessments for measuring progress against the goals
  - Teacher evaluation will be based on students’ progress on the established goals, as determined by an end-of-the-year principal review of the pre-determined assessments and their results
The “Rhode Island Model”

The Rhode Island Model (RI Model)

1. Impact on student learning
2. Professional Practice (including content knowledge)
3. Professional Responsibilities

“...each teacher’s Student Learning (SL) rating will be determined by a combination of state-wide standardized tests, district-selected standardized tests, and local school-based measures of student learning whenever possible.”
Category 1: Student growth on state standardized tests that are developed and/or scored by RIDE

Category 2: Student performance (as measured by growth) on standardized district-wide tests that are developed and/or scored by either the district or by an external party but not by RIDE (e.g., NWEA, AP exams, Stanford-10, ACCESS, etc.)

Category 3: Other, more subjective measures of student performance (growth measures and others, as appropriate) that would likely be developed and/or scored at the district- or school-level (e.g., student performance on school- or teacher-selected assessments, administrator review of student work, attainment of student learning goals that are developed and approved by both teacher and evaluator, etc.)
Rhode Island DOE Model: Framework for Applying Multiple Measures of Student Learning

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
“Rhode Island Model”: Student Learning Group Guiding Principles

• “Not all teachers’ impact on student learning will be measured by the same mix of assessments, and the mix of assessments used for any given teacher group may vary from year to year.”

Teacher A (5th grade)

Category 1 (growth on NECAP) + Category 2 (e.g., growth on NWEA) + Category 3 (e.g., principal review of student work over a six month span) = Teacher A’s student learning rating

Teacher B (11th grade English)

Category 2 (e.g., AP English exam) + Category 3 (e.g., joint review of critical essay portfolio) = Teacher B’s student learning rating

Teacher C (middle school art)

This teacher may use several category 3 assessments
"Rhode Island Model" Strengths and Weaknesses

Strengths

- Includes teachers in evaluation of student learning (outside of standardized tests)
- Teachers will benefit from having assessment of student learning at the classroom level

Weaknesses

- Heavily administrator/evaluator driven process
- Teachers can weigh in on assessments, but do not determine student growth
Teacher Advancement Program (TAP) Model

➢ TAP requires that teachers in tested subjects be evaluated with value-added models

➢ All teachers are observed in their classrooms (using a Charlotte Danielson type instrument) at least three times per year by different observers (usually one administrator and two teachers who have been appointed to the role)

➢ Teacher effectiveness (for performance awards) determined by combination of value-added and observations

➢ Teachers in non-tested subjects are given the school-wide average for their value-added component, which is combined with their observation scores
TAP strengths/weaknesses

Strengths

- Value-added becomes everyone’s responsibility, which should encourage efforts from teachers in non-tested subjects to support teachers in tested subjects
- Multiple yearly observations should be more informative and produce more reliable information about practice
- Professional development aligned with results is required

Weaknesses

- Concerns about “fairness” when only a few teachers’ student achievement and progress toward learning goals “counts”
- Tells you nothing about how teachers in other subjects are performing in terms of student learning growth (grades are not always good indicators)
IMPACT sorts teachers into groups that are evaluated differently

- Group 1: general ed teachers for whom value-added data can be generated
- Group 2: general ed teachers for whom value-added data cannot be generated
- Group 3: special education teachers
- Group 4: non-itinerant English Language Learner (ELL) teachers and bilingual teachers
- Group 5: itinerant ELL teachers
- Etc...
## Score comparison for Groups 1 & 2

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Group 1 (tested subjects)</th>
<th>Group 2 (non-tested subjects)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher value-added (based on test scores)</td>
<td>50%</td>
<td>0%</td>
</tr>
<tr>
<td>Teacher-assessed student achievement (based on non-VAM assessments)</td>
<td>0%</td>
<td>10%</td>
</tr>
<tr>
<td>Teacher and Learning Framework (observations)</td>
<td>35%</td>
<td>75%</td>
</tr>
<tr>
<td>Commitment to School Community</td>
<td>10%</td>
<td>10%</td>
</tr>
<tr>
<td>School Wide Value-Added</td>
<td>5%</td>
<td>5%</td>
</tr>
</tbody>
</table>
Group 2 assessment rubric

- 3 “cycles” of data collected & averaged/year

- Highest level of rubric:
  - “Teacher has at least 1 high-quality source of evidence (i.e., one that is rigorous and reliable) demonstrating that approximately 90% or more of her/his students are on track to make significant learning growth (i.e., at least a year’s worth) towards mastery of the DCPS content standards over the course of the year.”
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
DC IMPACT Strengths & Weaknesses

Strengths

- Uses multiples measures to assess effectiveness
- Permits the use of many types of assessment for students in non-tested subjects and grades
- Includes what is important in the system (in order to encourage specific teacher behaviors)

Weaknesses

- No multiple measures of student learning growth for teachers in tested subjects and grades
- Huge differences in how teachers are measured
Georgia KEYS

STUDENT ACHIEVEMENT - “Annual teacher evaluations shall as a minimum take into consideration the following: (1) the role of the teacher in meeting the school’s student achievement goals, including the academic gains of students assigned to the teacher.” Georgia Code 20-2-210 (b) (1) and (a)

“In making a determination of the academic gains of the students assigned to a teacher, evaluators should make every effort to have available and to utilize the results of a wide range of student achievement assessments, including those utilized by the teacher, set by the local board of education, or required under this article.” Georgia Code 20-2-210 (b) (1) and (c)

*Student Achievement Teacher Standard 1:* The teacher has a positive impact on student learning and academic achievement.

SA 1.1 Students taught by the teacher demonstrate the Georgia Performance Standard (GPS) related academic achievement progress on measures of student learning including state-mandated achievement tests or other measures as determined by the school district (e.g., teacher-developed assessments, department or district common assessments, benchmark tests, student work samples, portfolios, etc.).

<table>
<thead>
<tr>
<th></th>
<th>Not Evident</th>
<th>Emerging</th>
<th>Proficient</th>
<th>Exemplary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuum of Improvement</td>
<td>No quantifiable evidence exists that student achievement has increased, based on pre- and post-assessments using measures identified by the school district.</td>
<td>Quantifiable evidence exists that student achievement has increased, but has not met the established benchmark identified by the school district.</td>
<td>Quantifiable evidence exists that student achievement has met the benchmark based on pre- and post-assessments using measures identified by the school district.</td>
<td>Quantifiable evidence exists that student achievement has exceeded the benchmarks based on multiple measures of student learning including pre-and post-measures identified by the school district and also includes data from multiple measures of student learning.</td>
</tr>
</tbody>
</table>
# Georgia KEYS for Non-tested subjects

SA 1.2 Students taught by the teacher of content areas not addressed by the Georgia Performance Standards (GPS) demonstrate academic achievement progress on measures of student learning as determined by the school district (e.g., teacher-developed assessments, department or district common assessments, benchmark tests, student work samples, portfolios, etc.).

<table>
<thead>
<tr>
<th>Continuum of Improvement</th>
<th>Not Evident</th>
<th>Emerging</th>
<th>Proficient</th>
<th>Exemplary</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No quantifiable evidence exists that student achievement has increased, based on pre- and post-assessments using measures identified by the school district.</td>
<td>Quantifiable evidence exists that student achievement has increased, but has not met the benchmarks based on pre- and post-assessments using measures identified by the school district.</td>
<td>Quantifiable evidence exists that student achievement has met the benchmarks based on pre- and post-assessments using measures identified by the school district.</td>
<td>Quantifiable evidence exists that student achievement has exceeded the benchmarks based on multiple measures of student learning including pre- and post-assessments identified by the school district.</td>
</tr>
</tbody>
</table>

### Teacher Generated Performance Standards

This option is similar to Example 3 in SA 1.1. A district-wide group of teachers could collaborate to determine proficiency or progress standards for a given subject. The type of assessment would depend on the skills and knowledge that students are expected to master. Art and music classes, for example, may require students to demonstrate skills through performance. Art students might be required to amass a portfolio that exhibits progress and eventual mastery of certain skills. Band students may be required to make recordings or give live performances.

### Certification Based Assessment

Students in some fields, such as career and technical education, can seek certification that they have mastered certain skills. These certification tests may have been developed by national associations, state boards, or private companies. Districts may choose to adopt some of these tests as assessments of proficiency for their own coursework. This strategy has the advantage of holding students to a recognized standard and allowing for comparisons to students outside the district. Drawbacks may include the monetary cost of testing and the challenge of finding tests that are representative of course content.

### National Standards

Some subjects may be covered by standards set by a national organization. For example, physical education students may be assessed using the President’s Physical Fitness Test. Students that achieve passing scores may be considered proficient, and progress can be measured across multiple testing periods. In addition, information from sporting associations may be used to assess students’ knowledge of the rules and strategies of various sports.
Georgia KEYS Strengths & Weaknesses

Strengths

• Rubric for measuring teacher contribution is easy to understand

• Includes examples of multiple measures of student learning for all teachers, including those in tested grades and subjects

Weaknesses

• Rubric (including observation and other information) is about 100 pages long

• Might be a challenge to implement
Stated goal is to evaluate every teacher’s effectiveness with student achievement growth, even teachers in non-tested subjects and grades.

Undertaking to create pre- and post-assessments for all subjects and grades.

Expanding state standardized tests and using value-added to evaluate more teachers.

Part of a multiple measures system.
Hillsborough Strengths & Weaknesses

Strengths
• Teacher and union involvement in evaluation system decisions
• Teachers may be able to recommend tests they are already using
• All teachers included, not just tested subjects

Weaknesses
• Very expensive to create tests for all grades and subjects
• Takes teachers out of the assessing/scoring/improving instruction loop
Delaware Model

- Standardized test will be used as part of teachers’ scores in some grades/subjects
- “Group alike” teachers, meeting with facilitators, determine which assessments, rubrics, processes can be used in their subjects/grades (multiple measures)
- Assessments must focus on standards, be given in a “standardized” way, i.e., giving pre-test on same day, for same length of time, with same preparation
- Teachers recommend assessments to the state for approval
- Teachers/groups of teachers take primary responsibility for determining student growth
- State will monitor how assessments are “working”
Delaware Model: Strengths & Weaknesses

Strengths

- Teacher-driven process (assumes teachers are the experts in assessing their students’ learning growth)
- Great professional growth opportunity as teachers work together across schools to determine assessments, score student work, etc.

Weaknesses

- Validity issues (how the assessments are given and scored, teacher training to score, etc.)
- Time must be built in for teachers to work together on scoring (particularly for rubric-based assessments)
Final thoughts

- We are not very good at predicting which sets of teacher qualifications, characteristics, and practices will result in the best student outcomes
  - Once in the classroom, multiple measures of teacher performance and student outcomes can help determine effectiveness

- There is not enough research yet to say which model and combination of measures will provide the most accurate and useful information about teacher effectiveness
  - Focus on models and measures that may help districts, schools, and teachers improve performance
Models and measures should provide useful information about effectiveness

Those models that yield actionable information are most likely to contribute to improvements in teacher practice

- Standardized tests scores provide little information about how to change practice
- Teacher practice linked to multiple student outcomes is most actionable
  - Teachers benefit from knowing how their specific practices resulted in student learning
  - Thus, create opportunities for teachers to examine outcomes in light of practice
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