

R E P O R T

---



# Incorporating Student Performance Measures into Teacher Evaluation Systems

Jennifer L. Steele, Laura S. Hamilton,  
Brian M. Stecher

Sponsored by the Center for American Progress

This work was sponsored by the Center for American Progress with support from the Bill and Melinda Gates Foundation. The research was conducted in RAND Education, a unit of the RAND Corporation.

The RAND Corporation is a nonprofit institution that helps improve policy and decisionmaking through research and analysis. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

**RAND**® is a registered trademark.

© Copyright 2010 RAND Corporation

Permission is given to duplicate this document for personal use only, as long as it is unaltered and complete. Copies may not be duplicated for commercial purposes. Unauthorized posting of RAND documents to a non-RAND website is prohibited. RAND documents are protected under copyright law. For information on reprint and linking permissions, please visit the RAND permissions page (<http://www.rand.org/publications/permissions.html>).

Published 2010 by the RAND Corporation  
1776 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138  
1200 South Hayes Street, Arlington, VA 22202-5050  
4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213-2665  
RAND URL: <http://www.rand.org/>  
To order RAND documents or to obtain additional information, contact  
Distribution Services: Telephone: (310) 451-7002;  
Fax: (310) 451-6915; Email: [order@rand.org](mailto:order@rand.org)

## Summary

---

### **The Use of Student Achievement to Evaluate Teachers Is Drawing Increasing Policy Attention**

In a growing effort to recognize and reward teachers for their contributions to students' learning, a number of states and districts are retooling their teacher evaluation systems to incorporate measures of student performance. This trend stems from evidence that teachers' evaluations and reward structures have not sufficiently distinguished teachers who are more effective at raising student achievement from those who are less effective (Toch & Rothman, 2008; Tucker, 1997; Weisberg et al., 2009). It has also likely been spurred by competitive federal grant programs, such as Race to the Top and the Teacher Incentive Fund, and by philanthropic efforts, such as the Bill and Melinda Gates Foundation's Empowering Effective Teachers Initiative, all of which encourage states and districts to enhance the way they recruit, evaluate, retain, develop, and reward teachers. Given strong empirical evidence that teachers are the most important school-based determinant of student achievement (Rivkin et al., 2005; Sanders & Horn, 1998; Sanders & Rivers, 1996), it seems increasingly imperative to many education advocates that teacher evaluations take account of teachers' effects on student learning (Chait & Miller, 2010; Gordon et al., 2006; Hershberg, 2005).

Meanwhile, improved longitudinal data systems and refinements to a class of statistical techniques known as *value-added models* have made it increasingly possible for educational systems to estimate teachers' impacts on student learning by holding constant a variety of student, school, and classroom characteristics. However, measuring teachers' performance based on their value-added estimates involves several challenges. First, despite recent advances in value-added modeling, in practice, most value-added systems have a number of limitations: The tests on which they are based tend to be incomplete measures of the constructs of interest, year-to-year scaling is often inadequate, and student-teacher links are generally incomplete—particularly for highly mobile students or in cases of team teaching (Baker et al., 2010; Corcoran, 2010; McCaffrey et al., 2003). Second, value-added estimates can be calculated only for teachers of subjects and grades that are tested at least annually, such as those administered under a state's accountability system. In most states, the tested grades and subjects are only those required by No Child Left Behind: math and reading in grades 3–8.

In light of these limitations, educational systems that are now attempting to incorporate student achievement gains into teacher evaluations face at least two important challenges: generating valid estimates of teachers' contributions to student learning and including teachers who do not teach subjects or grades that are tested annually. This report considers these chal-

lenges in terms of the kinds of student performance measures that educational systems might use to measure teachers' effectiveness in a variety of grades and subject areas.

## Considerations in Choosing Student Performance Measures to Evaluate Teachers

The report argues that policymakers should take particular measurement considerations into account when using student achievement data to inform teacher evaluations. Such considerations include score *reliability*, or the extent to which scores on an assessment are consistent over repeated measurements and are free of errors of measurement (AERA, APA, & NCME, 1999). We describe three reliability considerations in particular: the internal consistency of student assessment scores, the consistency of ratings among individuals scoring the assessments, and the consistency of teachers' value-added estimates generated from student assessment scores.

Policymakers should also consider evidence about the *validity* of inferences drawn from value-added estimates. Validity can be understood as the extent to which interpretations of scores are warranted by the evidence and theory supporting a particular use of that assessment (AERA, APA, & NCME, 1999). Validity depends in part on how educators respond to student assessments, on how well the assessments are aligned with the content in a given course, and on how well students' prior test scores account for their prior knowledge of newly tested content.

In addition, policymakers may wish to consider the extent to which student assessments are vertically scaled so that scores fall on a comparable scale from year to year. Vertically scaled tests can, in theory, be used to assess students' growth in knowledge in a given content area. In their absence, estimates of students' progress are based on their test performance relative to their peers in a given subject from year to year. However, vertical scaling is very challenging across a large number of grade levels and in cases where tested content is not closely aligned from one grade to the next (Martineau, 2006).

The report also discusses the merits and limitations of additional student performance measures that states or districts might use. Commercial interim assessments are relatively easy to administer consistently across a school system, but they are not typically designed for use in high-stakes teacher assessments, and attaching high-stakes use may undermine their utility in informing teachers' instructional decisions. Locally developed assessments have the potential to be well aligned with local curricula, but items need to be developed, administered, and scored in ways that promote high levels of consistency. Using aggregate student performance measures to evaluate teachers in nontested subjects or grades allows school systems to rely on existing measures but creates a two-tiered system in which some teachers are evaluated differently from others. In addition, policymakers must consider how teachers will be held accountable for students who receive instruction from multiple teachers in the same subject in a given year.

## How New Teacher Evaluation Systems Are Addressing Measurement Challenges

To describe how educational systems are beginning to address some of the aforementioned measurement challenges, the report presents profiles of two states and three districts that have begun or are planning to incorporate measures of student performance into their teacher evaluation systems. These are Denver, Colorado; Hillsborough County, Florida; the state of Tennessee; Washington, D.C.; and the state of Delaware. To identify these five, we collected information from the websites of systems incorporating some type of student performance measures into their teacher evaluations according to media reports, prior studies, and teacher-quality websites we reviewed. The five profiles describe the student assessments administered by these systems and how those assessments are or will eventually be included in teachers' evaluations. In addition, the profiles illustrate a few steps that systems are taking to promote the reliability and validity of teachers' value-added estimates, such as averaging teachers' estimates across multiple years and administering pretests that are closely aligned with end-of-course posttests. They also demonstrate how the systems evaluate teachers in nontested subjects and grades. Finally, we use the profiles to discuss how some of the systems assign teachers responsibility for students enrolled during only a portion of the school year.

## Policy Recommendations

The report offers five policy recommendations drawn from our literature review and case studies. The recommendations, which focus on approaches to consider when incorporating student achievement measures into teacher evaluation systems, are as follows:

- Create comprehensive evaluation systems that incorporate multiple measures of teacher effectiveness.
- Attend not only to the technical properties of student assessments but also to how the assessments are being used in high-stakes contexts.
- Promote consistency in the student performance measures that teachers are allowed to choose.
- Use multiple years of student achievement data in value-added estimation, and, where possible, use average teachers' value-added estimates across multiple years.
- Find ways to hold teachers accountable for students who are not included in their value-added estimates.

We conclude with the reminder that efforts to incorporate student performance into teacher evaluation systems will require experimentation, and that implementation will not always proceed as planned. In the midst of enhancing their evaluation systems, policymakers may benefit from attending to what other systems are doing and learning from their struggles and successes along the way.