



*A Union of Professionals*

A Guide for Developing

# Multiple Measures

For Teacher Development  
and Evaluation



# Introduction

**CREATING EFFECTIVE AND EFFICIENT** teacher development and evaluation systems is one of the most important challenges facing America's schools. Some of the reasons for creating new systems or enhancing existing ones include:

- The critical relationship between teaching practice and student learning;
- The understanding that high-quality evaluation systems can identify teachers' strengths and weaknesses to guide professional development decisions that improve both teaching and learning;
- The fact that most current teacher evaluation systems are unable to accurately distinguish different levels of performance and provide meaningful guidance for improvement;
- The current structure of schools, which treats teachers as interchangeable cogs with little differentiation;
- Research that indicates that some of the traditional measures of teacher quality are not highly related to student achievement on standardized tests;
- The movement to include measures of student learning along with measures of teaching; and
- A call from stakeholders for differentiation of teacher performance for high-stakes decisions.

Given the importance of this effort, we must work collaboratively to create a development and evaluation system that can both effectively inform teacher development and ensure fair teacher accountability. It is important in creating a new teacher evaluation system that we not substitute the current inadequate evaluation system with another equally meaningless system where teachers' rankings are based, either solely or predominantly, on results from overly narrow standardized test scores or on scores from idiosyncratic unstandardized measures of student learning that lack rigor. Our challenge

*We must work collaboratively to create a development and evaluation system that can both effectively inform teacher development and ensure fair teacher accountability.*

is to create fair, rigorous and efficient teacher development and evaluation systems that not only define the standards of accomplished teaching and learning, but use multiple sources of data to measure them.

There are no simple ways to do this right; there are no quick and cheap solutions; there is no single instrument or procedure to fairly identify the effectiveness of teachers. Evaluating and developing teachers requires the use of well-constructed, multiple measures. And even after that, the final process for teacher evaluation will likely be “a marriage of insufficiencies.” The weakness of one measure will need to be supported by the strength of another. No single or predominant measure, instrument or assessment can provide a fair and accurate measure of the complexity of teaching and learning.

Good teacher development and evaluation systems also require professional judgment. Instrument scores and evaluation processes will provide evidence about the quality of teaching, but in the end, in a profession, there will and should be judgments made on the basis of evidence.

It is important to note that this brief represents the AFT’s best thinking on the issue of developing multiple measures for teacher development and evaluation at this point in time. It is an evolving piece of work that will be revised as new research becomes available about the use of student learning measures in teacher evaluation.

# Why Do We Need Multiple Measures?

**WHEN EVALUATING A COMPLEX ACTIVITY**—arguing a legal case, performing a surgical procedure or teaching a lesson—multiple measures examine diverse facets of the activity from many perspectives. These different perspectives are essential if we want to get an accurate representation of what we wish to assess. The sum of the various criteria that go into the assessment, as well as the multiple ways of looking at the criteria, allow us to make a considered judgment of a complex activity.

The importance of using multiple measures in educational evaluation is a well-established principle of testing and measurement. The National Council on Measurement in Education’s (NCME) Code of Professional Responsibilities states: “Persons who interpret, use and communicate assessment results have a professional responsibility to use multiple sources and types of relevant information about persons or programs whenever possible in making educational decisions” (Section 6.7).

Furthermore, The Standards for Educational and Psychological Testing, developed jointly by the American Educational Research Association (AERA), the American Psychological Association (APA) and the NCME, states: “In educational settings, a decision or characterization that will have major impact ... should not be made on the basis of a single test score. Other relevant information should be taken into account if it will enhance the overall validity of the decision” (Standard 13.7).

Because teaching is multifaceted, there are multiple standards for what makes an effective teacher. The evaluation of teacher effectiveness must assess performance on all of those standards. In particular, the assessment must look at a teacher’s knowledge of students, of how children learn, of subject matter, and of instructional and assessment techniques. It must

*In educational settings, a decision or characterization that will have major impact... should not be made on the basis of a single test score.*

also look at outcomes of student learning, student behavior and student engagement. To evaluate teacher effectiveness, we must gather data on all these areas and make a judgment based on all the evidence, even as we value some evidence more than others because it is more reliable, more valid and/or a more important aspect of teaching from the perspective of the teaching profession.

Thus, we weight the data, giving more value to some measures than others, but making our judgment based on all the evidence we have, without overemphasizing any single measure. In short, we use multiple measures to determine effectiveness. It is also worth emphasizing that the major reason for incorporating multiple measures in teacher evaluation systems is to create systems that provide useful and relevant information that can guide and enhance professional development throughout the year and throughout a teacher's career.

If this all sounds a bit confusing, it's because there are multiple meanings of multiple measures.

One meaning of multiple measures refers to the diverse information we use to make a judgment about something. For example, when buying a house, our decision is based on several factors. We might consider the price, the location, the size, the condition and the style. And we make our final judgment by weighing these various factors. The same is true when we think about evaluating a teacher—we have to look at more than one characteristic of teaching—instructional practice, student outcomes, contributions to the profession and the like. We must then determine what matters most as we put the various pieces (measures) together to come to a judgment.

A second meaning of multiple measures refers to the multiple ways we can measure a single piece of information we are using to make a judgment. For example, in deciding whether to buy a house and considering its location, we might look at its proximity to schools, to employment, to recreational facilities. Again, we would weigh the various pieces of information to determine which ones are more important. Is being close to school more important than being close to work? Is being close to public transportation more important than being close to school?

In teacher evaluation, each characteristic of importance (for example, quality of instruction, outcomes for students, contributions to the profession) can be measured in several different ways. We could collect evidence of student learning using standardized tests, pre- and post-tests in nontested subjects, student work and portfolios. We could collect evidence of instructional quality through classroom observations, analysis of teacher artifacts (including lesson plans and student assignments), and examination of grading practices and feedback to students. We could collect evidence of professional responsibility from administrator and/or supervisor reports,

and logs and documentation of professional activities. Evidence of student learning, evidence of instructional quality and evidence of professional responsibility are each used to measure different characteristics of teacher quality—as in the first definition of multiple measures discussed above. Standardized test scores, teacher-designed assessment results, student performances and student work are all different measures of the same characteristic: student learning—as described in our second view of multiple measures (see Figure 1, How To Use Multiple Measures To Evaluate Teachers).

FIGURE 1

## HOW TO USE MULTIPLE MEASURES TO EVALUATE TEACHERS

### Concept #1. Identify the multiple characteristics you want to include:

- **Evidence of growth on student learning**
- **Evidence of instructional quality**
- **Evidence of professional responsibility**

### Concept #2. Identify the multiple measures of each characteristic:

- **Evidence of growth on student learning**
  - Standardized tests
  - Student portfolios
  - Teacher-designed assessments
  - Student learning objectives
- **Evidence of instructional quality**
  - Classroom observations
  - Examination of teacher artifacts, including lesson plans and student assignments (e.g., cognitive demand)
- **Evidence of professional responsibility**
  - Administrator, supervisor reports
  - Logs and documentation of professional activities

\*\*THIS IS AN EXAMPLE AND NOT AN EXHAUSTIVE LIST OF EVIDENCE AND/OR CHARACTERISTICS.

# The Purpose of This Brief

**THIS BRIEF FOCUSES MAINLY** on one aspect of teacher evaluation: student learning. We identify several assessment criteria that must be considered when incorporating multiple measures of student learning, as a component, into a teacher evaluation system. We also address the considerations for combining the evidence into a single profile of student learning and teacher effectiveness.

## Assessment Criteria

There are technical, educational, political and administrative considerations that must be addressed when choosing instruments to demonstrate student outcomes as a result of teacher instruction. In this brief, we identify nine such factors: purpose, fairness, rigor, alignment, focus, growth, administrative capacity, cost and usefulness for teacher development.

**Purpose:** One of the basic rules of assessment is that tests should be used only for the purposes for which they were developed and/or validated (AERA/APA/NCME Standard 1.4). Many standardized tests for students were developed to determine student learning, not teacher effectiveness. Using those tests for purposes other than to measure what they were originally designed to measure (i.e., to evaluate teachers) may not only seriously undermine the validity of the instrument for determining teacher performance but also lead to “corruption” and/or unintended outcomes. For example, using the tests for teacher evaluation may result in narrowing the curriculum and learning opportunities for students. Furthermore, many instruments that are being considered for use in teacher evaluation are student diagnostic tools. These assessments provide valuable information to teachers about how well their students are doing, the progress they are making over time, and where to focus instructional efforts. Currently, these tools are “no stake” instruments we expect teachers to use routinely. When (or if) these tests are introduced as part of a high-stakes teacher evaluation

### NINE ASSESSMENT CRITERIA FOR EXAMINING MULTIPLE MEASURES OF STUDENT LEARNING

- ✓ Purpose
- ✓ Fairness
- ✓ Rigor
- ✓ Alignment
- ✓ Focus
- ✓ Growth
- ✓ Administrative capacity
- ✓ Cost
- ✓ Usefulness for teacher development



process, they must be validated for such a purpose, and safeguards must be built in to preserve the original intent of the tool—providing valuable diagnostic information about a student’s learning.

Purpose should be thought of as the overarching consideration that determines how the rest of the criteria, discussed below, should be evaluated. Validation needs to be conducted in light of the specific purpose for which the measure of student learning is being used, and other considerations, such as fairness, growth, alignment, depend on what the test is being used for.

**Fairness:** There are many ways to think about fairness (for example, educationally, psychometrically or politically). We have identified four critical aspects of fairness that must be taken into consideration.

The first aspect of fairness to consider is the concept of validity. While validity is a complex psychometric concept with specific definitions within the measurement community, we use it here in the construct validity sense: Do the instruments measure the thing they purport to measure, teaching? Is it reasonable to use the instruments for the proposed purposes? This is known as “consequential validity.” Many of the concepts we discuss in the fairness category and elsewhere (such as rigor) are related to issues of validity. Given that the intended audience for this brief is not experts in assessment or measurement, we have taken the liberty of simplifying our discussion of validity to make this concept and all its aspects more accessible to our readers.

The second aspect of fairness refers to reliability—that is, the degree to which the measure is free from error and provides “true” scores, as well as the degree to which the evidence the measure provides is repeatable and consistently reveals similar results.

Third, is the concept of standardization. How well are the instruments, processes and judgments standardized across teachers and subjects? Are the demands made on a teacher of biology comparable to those made on a foreign language teacher or a fifth-grade teacher? Are they similar for all biology teachers, all foreign language teachers, all fifth-grade teachers? While there should be room for professional judgment, we need to ensure that judgment is standardized in some ways across teachers and subjects so fairness can be achieved.

Finally, within the definition of fairness, we need to consider the concept of teacher influence and control, or the degree to which the student-learning measure is likely to be influenced by factors—both positive and negative—the teacher has no control over (for example, a scripted curriculum that is not aligned well with the state tests, poorly aligned or ineffective professional development, supportive colleagues who plan their lessons together, or conflicting school reforms.).

*Other measures such as unit tests and formative assessments may suffer from other perception problems. These must be carefully attended to if the system is to produce valid and useful results.*

Other concepts to consider relate to perceptions of fairness—or the extent to which the measure or analysis is perceived to be free of bias and an accurate portrayal of student learning. For example, many teachers distrust value-added methods (VAM) and perceive them to be unfair for four primary reasons: (1) They feel students' standardized test scores are not a good reflection of learning (or what is covered in the curriculum) and that focusing on the tests can lead to a distortion of the curriculum and the learning experience for students; (2) they distrust the capacity of the school data systems to accurately reflect the students they are teaching; (3) they believe VAM analyses of student achievement do not sufficiently take into account factors beyond the control of the teachers that affect growth—for example, students' prior level of performance, disruptive students in the class, the teaching and learning conditions of the school; and (4) if they are familiar with VAM analysis, they know that the results (as related to teacher effectiveness) tend to vary over time depending on the test, the method and the extent of data used. Indeed, teachers' concerns are reflected in many validity and reliability problems that researchers have identified with VAM procedures currently used. Though we give this one example of VAM, other measures such as unit tests and formative assessments may suffer from other perception problems. These must be carefully attended to if the system is to produce valid and useful results.

**Rigor:** There is concern by practitioners, parents and policymakers that the measures used to assess student learning often focus on superficial or low-level knowledge rather than on tasks that require integrating information and problem solving. While multiple-choice exams can be cognitively demanding, often these exams focus on recall of information rather than generative processes for developing and using information. As assessments get used more often for accountability, there has been a history of “dumbing them down,” either by making the items simpler or by lowering the cut scores. In examining the measures used in teacher evaluation, it is therefore necessary to assess the rigor of the learning assessed by the measures.

**Alignment:** In assessing the evidence of student learning for evaluating teachers, it is critical to understand how well the instrument used to measure learning is aligned to the student academic standards and the curriculum that is adopted in the classroom. All too often, the standardized tests are not aligned with the academic standards, the curriculum or the instruction. Teacher-made assessments and end-of-year exams can be expected to be more closely aligned to the curriculum than more distant or commercially developed instruments, although the standardized instruments may have stronger psychometric properties.

**Focus:** The measures we use to assess student learning, and the weights we give to those measures, indicate the importance we assign to evidence

derived from a measure. The more we value a measure for accountability, the more “privileged” it becomes as evidence of learning. In order to assess focus, we can ask: To what degree are the student-learning measures focused on narrow definitions of student learning or broader, more inclusive definitions? It is important to keep in mind the purposes of public education when we begin to think about how we define student learning. In the broadest sense, we can say that schools should be responsible for educating citizens to actively participate in a vibrant democracy, should seek to prepare future workers who can positively contribute to the national economy and compete in a global economy, and should help students realize their full potential.<sup>1</sup> Whichever measures of student learning (narrow or broad) are included in a teacher development and evaluation system will be the ones that are privileged as soon as we attach high stakes to them (for example, by tying the outcomes of a teacher’s evaluation to district staffing decisions). Which measures of student learning we privilege will have significant ramifications for students, their educational experiences and, ultimately, the future of our nation. Thus, focus should be given careful consideration.

**Growth:** In this instance, growth refers to whether a particular measure has the ability to show growth between two or more points in time, as defined in current federal legislation. While this is a somewhat simplistic definition of growth, many questions remain regarding how we make this definition operational. How much gain is good enough? Do we know whether it takes the same amount of effort to move up students who are on grade level, as it takes to move up students who are three grade levels behind? How do we measure growth when the content of the curriculum and tests change over time?<sup>2</sup> For example, how can growth be measured from 9th-grade biology to 10th-grade chemistry?

**Administrative capacity:** Different instruments make different demands on the system. For example, using standardized test scores requires an extensive test database and the capacity to ensure its accuracy, as well as technical knowledge about how to conduct VAM analyses. End-of-course exams require district or state capacity to develop and administer the tests. Portfolio assessment requires not only that comparable tasks be developed, but also that teachers be trained to assess the work reliably and accurately.

**Cost:** There are many costs—hidden and evident—associated with the collection of evidence of student learning for use in teacher evaluation. There are the costs of test or instrument development; the costs of implementation and data collection; the costs of scoring, whether by machine or trained assessors; and the costs of communicating the results to teachers and the community. These costs vary depending on the instrument. Standardized tests that are commercially developed may be cheap to administer and score, but require costly data banks and highly trained analysts, as well as

---

<sup>1</sup> In a discussion of the purposes of public schooling, David Labaree identifies three goals of education: “democratic equality, social efficiency and social mobility.” For further discussion of these goals, see D.F. Labaree, “Public Goods, Private Goods: The American Struggle over Educational Goals,” *American Educational Research Journal* 34(1), 39-81 (1997).

<sup>2</sup> Joseph Martineau refers to this phenomenon as “construct shift.” For more information on this issue, see, J. A. Martineau, “Distorting Value-Added: The Use of Longitudinal, Vertically Scaled Student Achievement Data for Growth-Based, Value-Added Accountability,” *Journal of Educational and Behavioral Statistics* 31(1), 35-62(2004).

considerable communication efforts to make the results understandable and useful to teachers and administrators.

**Usefulness for teacher development:** If the development of teacher capacity is a highly valued outcome from a teacher evaluation system, then it is necessary to understand the potential of each of the instruments to assist teachers in developing professionally. A VAM analysis can sort teachers, but it cannot by itself contribute to more effective teaching (although examination of student patterns might be useful for identifying where the curriculum needs to be enhanced or teacher practice needs improvement). Collaborative efforts at developing and implementing end-of-course assessments or rigorous student learning objectives, on the other hand, can be tied directly to professional activities to improve practice. Indeed, examining the standards, and coming to consensus on what constitutes unacceptable, average, above average and superior student performances can provide an opportunity for excellent professional development.

## **Strengths and Weaknesses of Tools for Assessing Student Learning**

There are many sources of evidence of student learning—student learning objectives, teacher-designed assessments, student portfolios and interim assessments, to name a few. It is important to reiterate that these various instruments were developed to evaluate student learning, not teacher effectiveness, and the validity of using these as instruments to evaluate teachers must be established. This should not preclude districts from beginning to use such instruments if educators are in favor of including them. However, it is critical the systems be piloted and studied before they become consequential.

It is impossible to make recommendations or draw definite conclusions on any type of measure without examining the specific assessment, tool or instrument. For example, a multiple-choice test can measure superficial aspects of rote learning or more complex aspects of knowledge. A portfolio may contain standardized performances from students or be a random collection of artifacts. One cannot adequately assess the strengths and weaknesses of a particular instrument or method without a fuller understanding of the context in which it is being used. Similarly, a cost-benefit analysis of the various instruments is dependent on values placed on teacher development, public credibility, psychometric merit, etc. We recommend that each measure under consideration for inclusion in a teacher development and evaluation system be examined using the nine criteria previously discussed, in order to build a more complete and accurate profile of student learning.

# Putting the Data Together for a Measure of Teacher Effectiveness

**WE CAN COMBINE THE DATA** from our multiple measures of student learning in various ways, depending on what we value most, how accurate we believe the measure to be, and how strong its relationship is to the various criteria discussed above. For example, we may give more weight to student portfolios than to student test scores, or vice versa, based on whether teacher perception or psychometric properties is more significant in our value system. In making such a determination, we may decide that a certain level of proficiency on each student-learning measure is essential, or we may decide that excellence on one measure (for example, student test scores as compared with student learning objectives) can compensate for shortcomings in another. We can then create a model to combine all the characteristics of student learning.

While there is no one right answer, careful consideration must be given to how evidence is combined and weighted to determine an overall score for student learning in a teacher evaluation system. We can then use the same process for creating a profile of teacher effectiveness by combining evidence of: (1) growth on student learning, (2) instructional quality and (3) professional responsibility (see Figure 2, Determining Teacher Effectiveness). In identifying the characteristics and the various ways they can be measured, and then weighing the evidence, the process must be transparent to the profession and the public. The evaluation must be fair to teachers and credible to the public. Most important, the data on which judgments are made must be valid and reliable. While the various pieces of evidence may be given different weight in the evaluation, the final judgment of teacher effectiveness is a complete picture based on ALL the evidence.

## DETERMINING TEACHER EFFECTIVENESS

---

### COMBINE ALL THE EVIDENCE.

---

*Assign value to each piece of evidence:*

---

- ✓ Determine how much weight to give each characteristic.
- ✓ Determine how much weight to give various measurements of each characteristic.

---

### CHOOSE A MODEL FOR COMBINING THE DATA TO MAKE A SUMMARY JUDGMENT.

---

- ✓ Use a conjunctive model, with a set score for each characteristic.
- ✓ Use a compensatory model, with strength in one area or characteristic compensating for weakness in another.
- ✓ Use a validation model, where the system identifies assessed discrepancies between characteristics and measures, and determines a procedure for how to handle those discrepancies
- ✓ Use a blended model, with conjunctive, compensatory and validation aspects.

# Credits and Acknowledgements

**THE AFT WOULD LIKE TO THANK** Courtney Bell, Research Scientist at ETS; Laura Hamilton, Senior Behavioral Scientist at the RAND Corporation; and Lindsay Clare Matsumura, Assistant Professor at the University of Pittsburgh's School of Education for reviewing an earlier version of this brief. The content of this brief is the responsibility of the AFT alone; it has not been endorsed by the individuals who reviewed the document or the organizations for which they work.

# Examining Multiple Measures

*Directions: Complete this worksheet for each measure examined.*

STUDENT LEARNING MEASURE:

## Assessment Criteria

Purpose	
Fairness	
Rigor	
Alignment	
Focus	
Growth	
Administrative capacity	
Cost	
Usefulness for teacher development	