How Do Teacher Value-Added Measures Compare to Other Measures of Teacher Effectiveness?

Douglas N. Harris
Associate Professor of Economics
University Endowed Chair in Public Education

Carnegie Foundation webinar
October 31, 2012

New Teacher Evaluations and New Menu of Measures

Value-added to student achievement

Classroom observations (structured)

Principal evaluations (unstructured)

Student evaluations

Student Learning Objectives (SLOs)

Objective

A key goal for policymakers and practitioners is to develop a system that yields valid and reliable conclusions about teacher performance

Put differently, we want to minimize "misclassification" of teachers, e.g., putting high-performing teachers in the low-performing category

Comparing the various measures helps us identify the optimal mix of measures for making accurate performance judgments

Questions

What do we know about how alternative teacher effectiveness measures compare?

What more needs to be known on this issue?

What can't be resolved by empirical evidence on this issue?

How, and under what circumstances, does this issue impact the decisions and actions that districts can make on teacher evaluation?

How correlated are value-added measures with other measures?

Value-added measures (VA) are weakly correlated with teacher credentials

VA measures are correlated with almost all measures now on the table for the new breed of teacher evaluation

+0.12 to +0.34 correlation between value-added measures and classroom observation rubrics

+0.2 to +0.3 correlation between VA principal eval. (unstructured)

Importance Issue about Interpretation

A common reaction: these are weak correlations: and suggest value-added is not a very good measure

BUT it is important to be precise—what is a "weak" correlation?

Random error reduces the maximum possible correlation, possibly well below +1.0

Specifically, maximum is probably no higher than +0.7.

Other Reasons They Differ

- Reliability
  - See prior slide
  - Refers to the degree to which the measure is consistent when repeated.
  - If either measure is imperfectly reliable, then this reduces correlations
- Validity
  - Refers to the degree to which something measures what it claims to measure, at least on average
  - If two measures try to capture the same element of performance, and either is invalid, then correlation reduced
- Less obvious answer: The two chosen measures may not **intend** to measure the same thing
  - Implication: Differences between VA and other measures does not necessarily mean that either measure is wrong

Difficulty Determining Which Factors are Most Important

Reliability is easier to measure; only requires multiple observations over time under similar conditions

Direct test of validity requires a "true" measure of teacher performance as a comparison

However, no true measure exists, so we have to test validity indirectly

Some studies implicitly assume that value-added measures are correct

Other Approaches to Assessing Validity in VA

Simulations (evidence fairly positive for VA)

Experiments (evidence hard to interpret)

Testing the assumptions (evidence indicates problems with VA)

Other statistical tests (evidence is more supportive of VA)

Summary of Evidence

	Validity	Reliability
Value-added	Mixed evidence	Low-Modest Rel.
Classroom obs	N.A.	Modest Rel.
Principal evals	N.A.	N.A.
Student surveys	N.A.	N.A.
SLOs	N.A.	N.A.

What more needs to be known on this issue?

- Two main problems with existing evidence
  - Limited evidence about validity and reliability of most measures
    - Easy to (legitimately) criticize VA because there is so much more evidence
  - Evidence is detached from practice
  1. Evidence is about measures outside of high-stakes settings
  2. Numerous practical issues (e.g., applying common framework to all teachers)

What can't be resolved by empirical evidence on this issue?

- It is up to policymakers to decide what student outcomes we value and therefore what mix of measures of teacher effectiveness is optimal
- There does not seem to be complete agreement on what good teaching is
  - For example, some emphasize teacher performance in raising students' academic skills and others more concerned about motivating and engaging students
    - These skills are related but not the same
    - Classroom observations and student surveys are likely to be more effective measures of motivation/engagement

How does the evidence and discussion here impact decisions?

- Wide agreement on multiple measures, but idea only gets us so far
- It's probably not very useful to use measures that add no new information
  - Once we have 6 measures, is it likely that adding a 7th will lead to different performance classifications? Probably not.
  - Each measure is costly
- However, different measures may be useful for different purposes
  - Formative versus summative assessments
  - Classroom observations might be considered essential for formative feedback even if they classified teacher performance exactly the same way as other measures

Summary

While policymakers should consider the validity and reliability of all their measures, we know more about value-added than others.

Value-added measures are positively related to almost all other commonly accepted measures of teacher performance such as principal evaluations and classroom observations.

The correlations appear fairly weak, but this is due primarily to lack of reliability in essentially all measures.

Summary

The measures should yield different performance results because they are trying to measure different aspects of teaching

Using multiple measures can increase reliability; validity is also improved so long as the additional measures capture aspects of teaching we value.

Once we have two or three performance measures, the costs of more measures for accountability may not be justified. But additional formative assessments of teachers may still be worthwhile to help these teachers improve.