Multiple Ways to Use Multiple Measures for Teacher Accountability

Douglas N. Harris

Associate Professor of Economics
University Endowed Chair in Public Education
Director, Education Research Alliance for New Orleans

November 1, 2013

Widespread Agreement on Multiple Measures

- There is essentially no dispute that teachers should be evaluated using multiple measures
- While often unstated, the reasons include measurement problems with individual measures and the multiple goals of education
- But much less attention has been paid to what we do with multiple measures once we have them
- Weighted averages have become the default approach essentially by accident and it is important to consider the alternatives

Questions

1) What do we know about how to create and use multiple measures?
2) What more needs to be known on this issue?
3) How, and under what circumstances, does this issue impact the decisions and actions that districts can make on teacher evaluation?

- This brief is different from others in the CKN series because it’s less about summarizing research and more about policy design

Weighted Average: Description

- Assign some weight to each measure
  - Simplest approach is an average, i.e., with two measures assign 50% of weight to each
  - Might assign more weight to measures that capture more important elements of effectiveness or that are better measures of effectiveness
- This approach is encouraged by *Race to the Top*, *Teacher Incentive Fund* and more
- Also known as composites or indices

But Why Weighted Averages?

- It’s simple and easy to explain, and . . .
- It’s a common approach in daily life
  - Heat index (weighted avg of temperature & humidity)
  - Dow Jones Industrial Average (weighted avg of prices of major stocks)
  - College rankings (weighted avg of peer assessments, faculty-staff ratios, alumni giving, and more)
  - BCS College Football ratings (weighted avg of polls, win-loss, and more)
  - Consumer Reports (weighted average of quality indicators for products)

There are Alternatives

- Matrix
  - Place teachers into a category for each measure, but do not create an average
- Screening (two parts)
  - Use VA to “screen” for teachers where more intensive data collection is necessary, but not directly as part of the final decision
  - Use VA to screen for observers who may be poor judges of teacher performance
- More to come on all three approaches

Criteria

- We have also been evaluating these approaches too narrowly, focusing almost entirely on *accuracy* of classifications (validity and reliability)
- We also need to consider:
    - Cost (lower the cost the better)
    - Fairness (can apply it equally to all teachers)
    - Simplicity
- Important note: these criteria apply to the <u>decisions</u> resulting from the measures—the inferences made

In-Depth Discussion of the 3 Approaches for Using Multiple Measures

Weighted Averages: Applying Criteria

- Advantages
  - Simplicity (boils down to one measure)
- Disadvantage
  - Cost (it requires collecting all the data for every teacher)
  - Fairness (even at high cost, it cannot be applied to all teachers)
- Accuracy unclear compared with other methods

Matrix: Description

- Instead of combining the separate measures, place teachers in various boxes based on the combination of performance categories

- Allows more nuance

	Performance Measure A	
Performance Measure B	Low A – Low B	High A – Low B
	Low A – High B	High A – High B

Matrix: Applying Criteria

- Advantages:
  - Simplicity (in between weighted average and screening)
- Disadvantages:
  - Cost (still need all the measures)
  - Fairness (even at high cost still can’t apply the same method to all teachers)
- Accuracy unclear compared with other methods

Screening: Description

- Example: medicine
  - For many diseases, the first test is a low-cost “screener,” designed to identify anyone who might have disease
  - Those identified by screener as potential disease carriers take a more expensive “gold standard” test that gives a more definite determination
  - It is a *process* for collecting information and identifying diseases
  - They don’t create a weighted average of screener and gold standard test results

Two Screening Approaches

1) Use value-added and other past information to identify teachers for whom more (and better) performance data should be collected
2) Use value-added to identify weak classroom observers
    - Worry about <u>personal biases</u>, especially with educators in the same school are the observers
    - Worry about <u>inability</u> of some observers to identify effective teaching—letting VA drive the observation
    - The correlation between value-added and classroom observations could be a useful signal for these problems
    - Takes advantage of the strengths of value-added, while avoiding main weakness (reliability)

Screening: Applying Criteria

- Advantages:
  - Cost (don’t need to do as many classroom observations)
  - Fairness (final decision is based on observations, which are available for all teachers)
- Disadvantages:
  - Complexity
- Accuracy unclear compared with other methods
  - Screening observers could improve accuracy of non-VA measures
  - And still using all the information available
  - Again, main concern is accuracy of the performance categorization and related personnel decisions

Summary

	Weighted Average	Matrix	Screening
Accurate	?	?	?
Inexpensive	⬇	⬇	⬆
Fair	⬇	⬌	⬆
Simple	⬆	⬌	⬇

What if each measure captures a different element of effectiveness?

- So far, I have assumed that each measure captures the same element or aspect of effectiveness
- Probably more or less reasonable when using just value-added and classroom observations—both focus on classroom instruction
- But many schools consider contributions to school community, to take one other example
- This has implications for how we evaluate the three methods

Reconsidering Earlier Evaluations

- When measures capture different aspects of effectiveness, weighted average and matrix approach tend to make more sense
- With screening approach, it would make little sense to screen based on one element of effectiveness then use a gold standard test for a different element
  - Analogy: In Consumer's Reports, this would be like identifying cars that get good gas mileage in the first stage and collecting information about road handling only if the car got good mileage
- Bottom line: Cannot logically pick a method separately from the choice of measures

Approaches Not Mutually Exclusive

- Example 1: Could use a weighted average or matrix method as the first-stage screening process
- Example 2: What might seem like a weighted average usually includes an appeals process that includes more data collection (like screening)

Summary

- The most common way to use multiple measures in teacher accountability is through weighted averages of value-added with other gauges of teacher performance. This method has strengths and weaknesses.

- Policymakers should consider a wider range of options for using multiple measures.

- Because the main objective is to accurately classify teacher performance, most discussions of measures of teacher performance focus on validity and reliability. But fairness, simplicity, and cost should also be considered.

Summary (cont.)

- The matrix and screening methods are somewhat more complex than weighted averages, but they may be more accurate.

- The “screening” method is the least costly and fairest of the three options because it uses value-added measures to improve and streamline other forms of data collection, and it allows final decisions to be made based on the same criteria for all teachers.

- Ultimately, we should assess the method of using multiple measures based on how the options affect student learning, but the evidence does not yet exist to do that.