A Composite Estimator of Effective Teaching

Kata Mihaly
J.R. Lockwood
Daniel McCaffrey
Douglas Staiger

October 3, 2012

Measuring Teacher Performance

- There is general agreement that teachers matter and that teachers vary in their effectiveness
- Although the term "teacher effectiveness" is not easy to define, there is a desire to have effective teachers in classrooms
- Starting in late 2000's in the U.S. there was a consensus that existing evaluation systems were failing to accomplish this goal:
  - Typically infrequent observations
  - "Widget Effect" (Weisberg et al. 2009)
  - Salary schedule based on experience and degree earned

Policy Initiatives

- A number of policy initiatives jump started reform
- Race to the Top: a competitive grant program that rewards States for satisfying certain educational policy standards
- Primary criterion for RTTT funding is demonstrating the ability to improve teacher and principal effectiveness
- "Design and implement rigorous, transparent, and fair evaluation systems for teachers and principals that (a) differentiate effectiveness using multiple rating categories that take into account data on student growth as a significant factor, and (b) are designed and developed with teacher and principal involvement"

Multiple Measures

- As a result, 32 States and DC have revised evaluation systems
- States are collecting multiple measures of teaching:
  - student learning (e.g., value-added, student growth percentiles)
  - classroom observations (e.g., Danielson's Framework for Teaching or state developed protocols)
  - student surveys
  - student learning objectives, or examples of student work
- Policymakers and practitioners want a single "effectiveness" score to support decision making in multiple areas: professional development, promotion, retention, tenure, compensation and removal decisions
- States are creating ad-hoc algorithms for combining measures

Creating a Combined Measure

- Goe (2011) provides a toolkit for creating a composite:

1. Stakeholders and experts define the underlying concepts of interest (target criterion)
2. Identify indicators related to the target criterion
3. Specify the rules for weighting and combining the measures
4. Decide cutoff rules and proficiency rating levels

- How can statistical methods be used to improve this process?
  - Can they provide a method for calculating weights?
  - Can combined measure improve prediction of target criteria when target is measured?
  - Can combined measure improve prediction of an unmeasured target criterion?

Defining Terms and Notation

- Target Criterion: η_i
- Indicator: $Y_{ik} = \varphi_{ik} + \varepsilon_{ik}$
  - Sum of "true score" or stable component and measurement error
- Examples:
  - Value Added from one year is average career value added plus error due to performance measure, classes taught and students from a particular year
  - Score on teacher observation protocol from one year is teachers stable level of teaching performance plus error due to the lessons observed, the rater, and maybe the section of students selected for the observations
- Assume that indicator is unbiased for its stable component

Defining Concepts and Notation (cont.)

- Target criterion can be
  - Stable component of indicator
  - Combination of stable components
  - Unobserved measure
- Vector of indicators (Y_i), vector of stable components (φ_i), and vector of measurement errors (ε_i).
- Assume expected values of φ_i and ε_i are μ and 0.
- Variance-covariance matrices for φ_i and ε_i are A and E_i
- Measurement error can vary with teachers and depend on how data were collected
  - Value added and student surveys
  - Teacher observation protocols

The Role of Expert Judgment

- Because teacher effectiveness is unobserved, first step in forming composite requires expert guidance to identify target criteria: "what is teacher effectiveness" Goe (2011)
- If there are multiple criteria, experts need to consider how to combine these concepts theoretically
- Define λ as vector of value weights on the stable components:

$$\eta_i = \lambda' \varphi$$

- These weights cannot be determined empirically, because they depend on the target criterion.
- For a given set of value weights, optimal statistical weights can be used to form best predictors of concept of interest

Optimal Weights

- Ideally, teacher effectiveness equals the weighted sum of the stable components, and these weights are known
- Under this ideal scenario we can define the "optimal" weights for combining measures
  - By optimal, we mean the weights that make the composite measure most correlated with the ideal effectiveness
- Optimal predictor of the target criterion is

$$\widetilde{\eta}_i = \omega_i{}' Y_i$$

- Predictor minimizes mean square error and maximizes the correlation between the predictor and the target criterion.

Optimal Weights (cont.)

- Solution :

$$E[\eta_i|Y_i] = \lambda'\mu + \lambda' A(A + E_i)^{-1}(Y_i - \mu)$$

- Optimal weights are a function of the matrix

$$\lambda A(A + E_i)^{-1}$$

- Goodness of Fit Statistic:

$$Corr(\tilde{\eta}_i \eta) = \sqrt{\frac{\lambda' A(A + E_i)^{-1} A\lambda}{\lambda' A\lambda}}$$

Simple Example

- Consider two stable components with corresponding indicators
- Weighting matrix is 2 x 2
  - First row contains optimal weights w_{11i} and w_{12i} for predicting the first stable components
  - Second row contains optimal weights w_{21i} and w_{22i} for predicting the second stable component
- Let $R_i, i = 1,2$ be the reliability of component i and ρ equal the correlation between the two stable components
- Then:

$$w_{11} = \frac{R_1 - R_1 R_2 \rho^2}{1 - R_1 R_2 \rho^2} \qquad w_{12} = \frac{(1 - R_1) R_2 \rho}{1 - R_1 R_2 \rho^2}$$

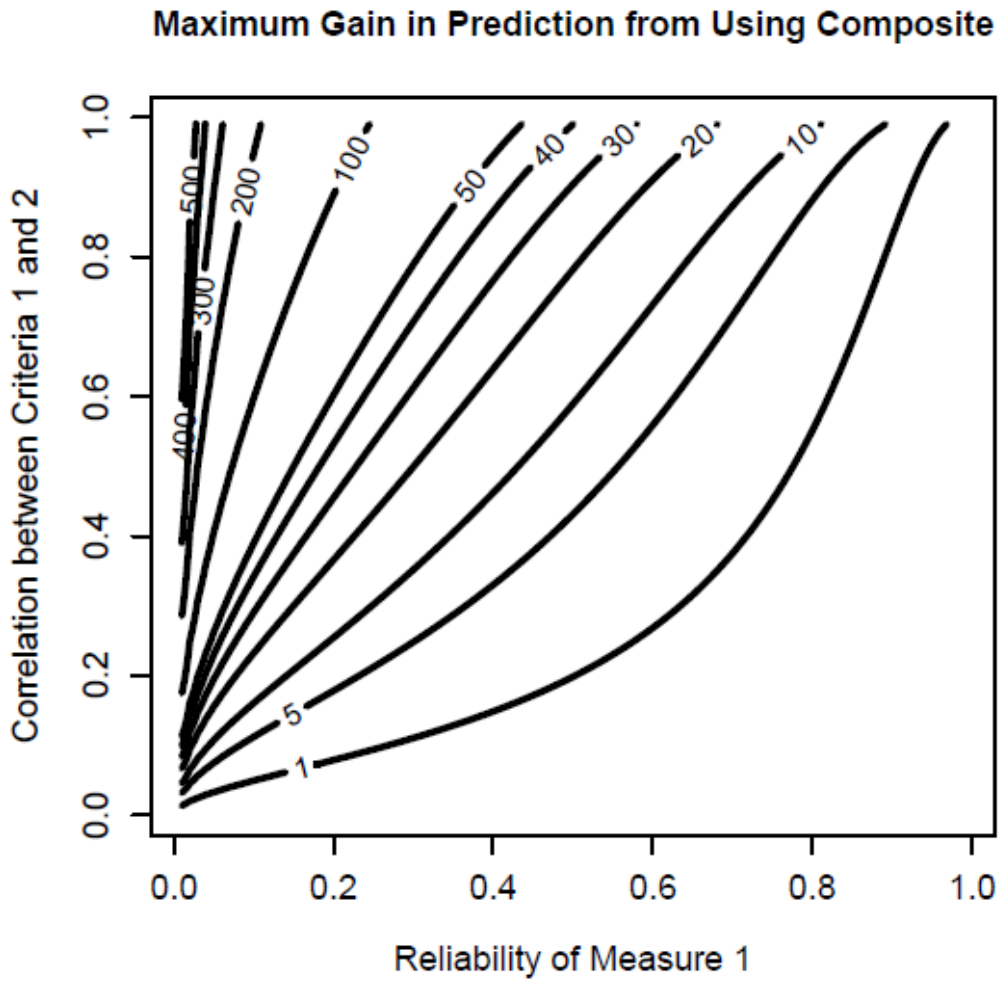
Simple Example

- Weights are functions of the reliability of the indicators and the correlation between the stable components
- As the reliability of the indicator approaches 1, that indicator receives all of the weight
- As the reliability of the indicator approaches 0, it gets none of the weight
- If the correlation between the stable components is high, then the weights are more even
- The fit statistic is also a function of the reliabilities and the correlation between the stable components
  - Increases with reliability
  - Conditional on reliability, increases with correlation

Evidence from earlier work

- Previous work has shown that the reliability of measures that are used or being considered is relatively high
  - MET first report, reliabilities range from 0.3 - 0.7
- Also, other studies have shown that the correlation between indicators that are collected by schools is relatively low
  - Bell et. al (2012) show that correlations between measures is low (0.17 – 0.30)
- However, correlation may depend on types of measures
  - The correlation of scores on two observation protocols may be higher than the correlation of either score with value added

Assume target is stable component of single indicator



Consequences of optimal weighting to predict stable components

- Compared to just using the target, optimally weighted composite yields little gain
- Optimal composite puts a significant percentage of the weight on the target unless:
  - Reliability of other measures is very low

    AND
  - Correlation of target and other component very high
- As a result, for any given set of value weights the optimal predictor of the weighted sum of the stable components is highly correlated with the value-weighted sum of the indicators.

Consequences for Unobserved Outcomes

- Unobserved target:
  - target criterion of teacher that is of interest but not clearly articulated or is not measured frequently
- Likely more correlated with a composite measure than any individual indicator or optimal predictor of any indicator
- If predicting unobserved target is of interest, a composite that equally weights each component is the best option
- If stakeholders can identify outcomes that are expected to be close to the unobserved target then it should receives somewhat more weight, but not too much

Extensions

- Allow for stable components to be broken into:
  - Common component
  - Mode component
  - Unique component
- How do the weights change when the teacher effectiveness measures is:
  - Same mode as one of the observed measures
  - Not in the same mode as any of the observed measures
- Use data from the MET project to examine optimal weights and ability to predict composite
  - Report to be released in early January


RAND
EDUCATION