Abstract Title Page

Title: Developing an aggregate metric of teaching practice for use in mediator analysis

Authors and Affiliations:

Valeriy Lazarev, Empirical Education Inc.
Denis Newman, Empirical Education Inc.
Pam Grossman, Stanford University

Abstract Body

Background / Context:

Efficacy studies of educational programs often involve mediator analyses aimed at testing empirically appropriate theories of action. In particular, in the studies of professional teacher development programs, the intervention targets presumably teacher performance while the ultimate outcome is the student achievement measured by a standardized test. In this case, teaching practices affected by the professional development can be measured using an observational instrument (rubric) and included as mediator in the analyses of student outcomes.

There are two obstacles to implementing this approach.

First, an observational rubric consists of a number of distinct domain indicators, each measuring a particular aspect of teaching practice. Domain indicators need to be aggregated to produce a single metric of teaching quality. A typical solution involves summation or averaging of domain scores to obtain a single metric. However, the relative contributions of each domain score to student outcomes in the short term may vary because some domains can be measuring aspects of teaching that do not translate directly into observed student achievement. Differences in the measurement error across domains and correlations between domain scores may also affect the contribution of each domain to the total.

Second, classroom observation scores are subjective estimates that use ordinal scale designed primarily to assess teaching practices, not student outcomes. Observation scores are therefore not necessarily related to student outcomes in a linear fashion. Some recent findings such as the asymmetry in the distribution of observation scores (Tennessee Department of Education, 2012) and the analyses performed in the framework of MET project (Kane and Staiger, 2012) point to substantial non-linearity in the relationship between aggregate observation and value-added scores. With a few exceptions (Grossman et al., 2010; Kane et al., 2010), little research has been conducted on the domain-level relationships between observational and value-added measures.

Purpose / Objective / Research Question / Focus of Study:

The main objective of this study is to develop a methodology for creating an optimal aggregate teacher performance metric from domain scores for use as mediators in the analyses of student outcomes. We use one particular observation rubric, PLATO (Grossman et al., 2010), to answer the question of how an aggregate teacher performance metric can be constructed from domain scores that would be best aligned with a selected measure of student achievement.

Setting:
N/A

Population / Participants / Subjects:
N/A

Intervention / Program / Practice:
N/A

Significance / Novelty of study:

Using the data of classroom observation in quantitative program evaluation requires developing aggregate performance metrics that combine domain scores in a way that maximizes the correlation between teacher performance indicators and student outcomes. Simple ad hoc approaches currently in use, such as summation of domain scores may result in inefficient and biased indicators that have low correlations with student achievement metrics. Whereas growing methodological literature on the measurement of teacher effectiveness tends to focus on the robustness of value-added models (Ballou 2005; Jürges and Schneider, 2007; McCaffrey et al., 2008; Harris, 2008; Braun et al, 2010) little attention has been paid to the statistical properties of efficient metrics based on classroom observation. This study intends to make a contribution to this area.

Statistical, Measurement, or Econometric Model:

In the context of an experimental study of a program that affects student achievement indirectly (such as a professional development program) or more generally a complex intervention that has both direct and indirect effects on student achievement, we may be interested in estimating the impact of the program, θ, on teacher performance, T, and the contribution of teacher performance, as we all other channels, on student outcomes: $Y_1 = \theta + Y_0 + T(\theta; Z)\alpha + X\beta + \varepsilon$, where Y is student outcome (e.g. test score), Z is the vector of teacher characteristics, and X is the vector of student characteristics. It is in principle possible to estimate the impact of on each component of teacher performance (observational rubric domain) and the contribution of each domain score to the student outcome (treated as representing a distinct teacher's skill, as in Kane et al 2010). In practice however, this task may be intractable and an experimental study may not be sufficiently powered to estimate a model where teacher performance using an instrument with an arbitrary number of domains. Moreover, domain of an observational rubric can be considered as partially complementary proxies for overall teacher quality rather than a distinct skill that has identifiable impact on student outcomes as measured by standardized tests. It is therefore desirable to have a single metric of teacher performance that has been calibrated (i.e. shown to correlate with the outcome of interest) on a large number of past observations.

We assume that such an aggregate metric is a function of its components (domain scores), $T = \Sigma f_j\ (x_j)$, where each of $f_j(x_j)$ is an arbitrary smooth function chosen so as to maximize the correlation between T and the student value-added metric based on the outcome of interest, $\hat{Y}$, in a sample of calibrating observations. It is therefore estimated from a generalized additive model: $\hat{Y}_i = \Sigma f_{j\,i}\ (x_{j\,i}) + \varepsilon_{\mathrm{i}}$

Estimating this model using penalized spline smoothing (Wood 2006) allows determining the optimal degree of smoothing and therefore the true shape of the relationship between student outcomes and $f(x)$, as well as identifying domains that do not contribute at a statistically significant level to the aggregate indicator either because they are unrelated to the outcome (at least in the short term) or because they are correlated with other domains. Analysis of the function $f(x)$ will generally allow finding a simple approximation for the aggregate teacher performance indicator, *T*, which can be subsequently calculated on the basis of classroom observations in a program evaluation study and used for mediator analysis.

Usefulness / Applicability of Method:

Our method is applicable to efficacy studies in which the intervention in question affects teacher performance and the mediating effect of the latter on student performance needs to be evaluated. It is assumed that teacher performance is assessed, in the framework of such a study, using a classroom observation instrument with multiple domain indicators.

Research Design:

N/A

Data Collection and Analysis:

The dataset we used to develop the methodology and to estimate the aggregate observational score for PLATO was created in the framework of MET project. As part of this project, several hundred high-quality video recordings of upper-elementary and middle-school lessons in ELA were scored by observers trained in the use of PLATO. The dataset also contained value-added scores for the teachers featured in the videos calculated from the student performance data (see Kane et al., 2012, for details). The total number of observations (scores) per domain was 1504.

Analysis was performed using *R* package *mgcv* (Wood 2006) and involved estimation of a generalized additive model. The principal output of the procedure (see example in Figure 1) was a plot of the functional relationship ("smooth") between each domain score and value-added score, estimated degrees of freedom for the smooth, proportion of explained dispersion, and other relevant statistics. Introspection of the plots together with assessing the estimated degrees of freedom allows making a decision about an appropriate parameterization of the relationship. The estimated degree of freedom of around unit suggests that the relationship is linear (possibly trivial if the significance of the estimate is low), while higher order implies that the relationship is non-linear. In some cases non-monotonic relationship (e.g. U-shaped) implies that a particular domain does not have an unambiguous effect on outcomes even though the relationship is technically significant. The analysis concluded with the estimation of a simple parametric approximation (linear regression) of the generalized additive model and determining if it is associated with a substantial loss of information.

Findings / Results:

Our preliminary analysis determined that only three out of six domains of PLATO used in the study were significantly associated with the value-added measure of student achievement. Of the three significant domains, two had linear relationship with the value added. For the only domain with non-linear relationship, log transformation appeared to be sufficient to linearize the relationship. In other words, the analysis showed that, as long as the aggregate teacher performance metric is meant to be used for research purpose in conjunction with student outcomes measured by standardized test scores (of a kind similar to used in the calibrating sample), only three domains have non-zero weights in the aggregation formula. The remaining three-element formula assigns weights of .03, .04, and .93 respectively to the remaining three domains. Correlation between the aggregate scores calculated according to the formula and the “naïve” score (sum of all domain scores with equal weights) is around 0.7, while the explained variance using the second method is about one half of that resulting from the calculated aggregate score. This suggests that the proposed method of aggregate score development has a potential to substantially increase the accuracy of the analysis.

Conclusions:

We have outlined a practically useful methodology of constructing aggregate teacher performance metrics for the purposes of program evaluation studies, in particular for mediator analyses.

The intermediate step of this process - estimation of a generalized additive model – is data intensive and requires large number of observations (thousands) in the calibrating sample to produce achieve sufficient power. This did not appear to be a problem in this study as most domains exhibited a linear relationship to the outcome. Developing an aggregate score for an arbitrarily complex observational instrument may require a much larger database of observations than is currently available.

Appendices

Appendix A. References

Ballou, D. (2005) "Value-added Assessment: Lessons from Tennessee," In R. Lissetz (Ed.), *Value Added Models in Education: Theory and Applications*. Maple Grove, MN: JAM Press.

Braun, H., Chudowsky, N., & Koenig, J. (2010) *Getting Value Out of Value-Added: Report of a Workshop*, Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Accountability; National Research Council.

Grossman, P., Loeb S., Cohen J., Hammerness K., Wyckoff J., Boyd D., & Lankford H. (2010) *The Relationship between Measures of Instructional Practice in Middle School English Language Arts and Teachers' Value-Added Scores*, CALDER, Working paper No. 45.

Harris, D. (2008) "The Policy Uses and Policy Validity of Value-Added and Other Teacher Quality Measures," In D. H. Gitomer (Ed.), *Measurement Issues and the Assessment for Teacher Quality*. Thousand Oaks, CA: SAGE Publications.

Kane, T., Taylor, E., Tyler, J., & Wooten, A. (2010) "Identifying Effective Classroom Practices Using Student Achievement Data," NBER Working Paper 15803.

Kane, T., & Staiger D.(2012), *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains*, Bill and Melinda Gates Foundation Research Paper

McCaffrey, D.F., Sass, T.R., Lockwood, J.R., & Mihaly, K. (2009). "The intertemporal variability of teacher effect estimates." *Education Finance and Policy* 4: 572–606.

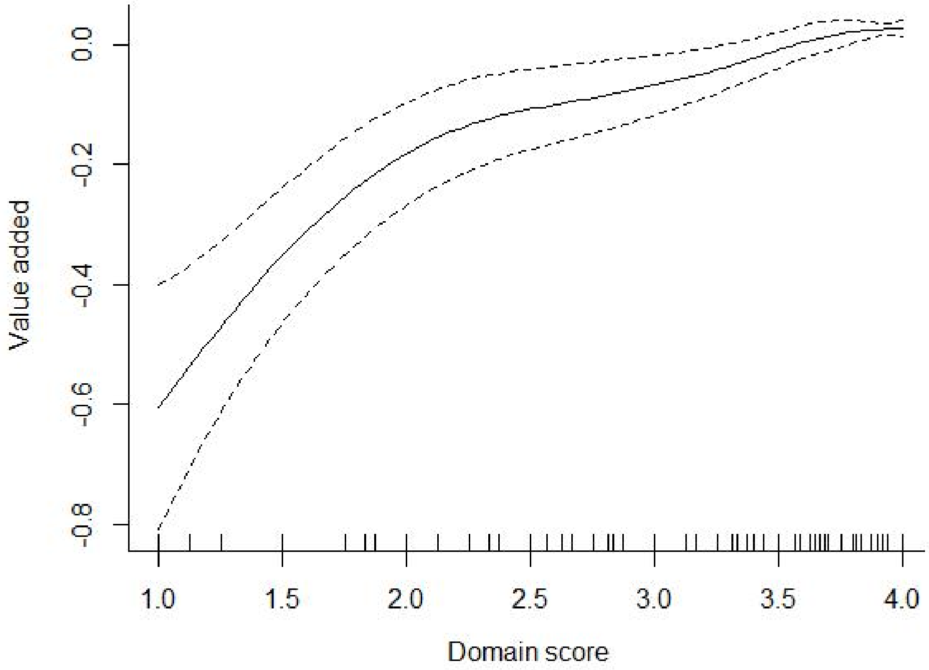
Tennessee Department of Education (2012), *Teacher Evaluation in Tennessee: A Report on Year 1 Implementation.* Retrieved from: http://www.tn.gov/education/doc/yr_1_tchr_eval_rpt.pdf

Wood, Simon (2006), *Generalized Additive Models: An Introduction with R*, Oxford: Taylor and Francis

Appendix B. Tables and Figures

Not included in page count.

Figure 1. Non-linear contribution of one of PLATO domains to the value-added. Plot and selected accompanying statistics (example).



```
Approximate significance of smooth terms:
                              edf      F        p-value
s(plato_AvgScrBEMT)  3.857  14.36   3.69e-13 ***
```