Developing an Aggregate Metric of Teaching Practice for Use in Mediator Analysis

Valeriy Lazarev and Denis Newman (Empirical Education Inc.)
Pam Grossman (Stanford University)

Introduction

Efficacy studies often involve mediator analyses, e.g., a teacher development program targets teacher performance while the ultimate outcome is student achievement. Teaching practices affected by the program can be measured using an observational instrument (rubric) and included as mediator in the analyses of student outcomes.

Problems:

1) Rubrics consist of indicators measuring particular aspects (domains) of teaching. Domain scores are typically averaged to obtain a single composite metric. However, the relative contributions of each domain to student outcomes differ; some domains measure aspects of teaching that do not translate directly into observed student achievement; measurement error correlations between domain scores vary across domains.

2) Classroom observation scores are subjective estimates that use ordinal scale designed primarily to assess teaching practices, not student outcomes. Student outcomes are not necessarily a linear function of observation scores.

3) Estimation of the contribution of each single domain score to the student outcome may be impractical in an experimental study that is not powered to deal with an arbitrary number of teacher level covariates. It is desirable to have a single metric of teacher performance, calibrated on a large number of past observations, i.e. shown to be accurate and relevant.

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 PLATO rubric (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.

Data Collection and Analysis

The calibrating dataset was created in the framework of MET project. More than one thousand recorded lessons 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 on SAT9 test. 1502 complete observations were included in the analysis.

Analysis was performed using R package mgcv (Wood, 2006). The principal output of the procedure is a plot of the functional relationship between each domain score and value-added score, estimated degrees of freedom, 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 close to unity suggests that the relationship is linear, while higher order implies a non-linear relationship. 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.

Model

A complex intervention can have both direct and indirect effects on student achievement. We need therefore to estimate direct impact of a program on student outcomes, Θ, and the contribution of the improved teacher practices, which is due to the impact of the program, ϕ, on teacher performance, T:

$$Y_1 = \Theta + Y_0 + T(\phi; Z)\alpha + X\beta + \varepsilon,$$

where Y is the student outcome, Z is the vector of teacher characteristics (including teaching practice), and X is the vector of student characteristics.

T, the aggregate metric of teacher performance is a function of observed domain scores,

$T = \Sigma f_j(z_j)$, where each of $f_j(z_j)$ is an arbitrary smooth function chosen so as to maximize the correlation between T and a 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:

$$\hat{Y}_i = \Sigma f_{j\,i}(z_{j\,i}) + \varepsilon_i$$

Estimating this model using penalized spline smoothing (Wood, 2006) allows determining the true shape of the relationship between student outcomes and $f(z)$. Analysis of $f(z)$ allows finding a simple approximation for the aggregate teacher performance indicator, T, which can used be as a mediating variable in future studies.

Results

Our analysis revealed a variety of patterns of relationship between domains and available value-added measures of student achievement. Two domains had non-monotonic relationship to the outcome. Two more domains had monotonic relationships with a moderate degree of nonlinearity. Three remaining domains had strictly linear associations with the outcome. Removing domains with ambiguous (non-monotonic) relationship to the outcome and least significant monotonic components results in the specification of an optimal linear model, which explains only 20% less variation that the full generalized additive model. This model suggests that the composite score should include only three domains, with weights varying between by a factor of three (.03 vs. .10). A univariate regression of the outcome on the domain average has a much lower quality (less than 50% of variance explained by the full model), which suggests that using the latter in mediator analysis could result in substantial bias.

Generalized additive model

	Estimated degrees of freedom	*p* value
s (TIME)	1.00	0.20
s (SUI)	1.00	0.98
s (MDLG)	3.78	0.04
s (BEMT)	3.61	<.001
s (CLDI)	2.21	0.18
s (INCH)	2.95	0.53
s (RoC)	1.00	0.58
R^2	.053	

Optimal linear model

	Estimate	*p* value
Intercept	-0.52	<.001
TIME	0.03	0.18
BEMT	0.10	<.001
INCH	0.03	0.09
R^2	0.041	

Univariate linear model

	Estimate	*p* value
Intercept	-0.39	<.001
Average score	0.02	<.001
R^2	0.026