Empirical Education
EMPOWERING EDUCATORS FOR EVIDENCE-BASED DECISIONS

Developing Composite Metrics of Teaching Practice for Mediator Analysis of Program Impact

Val Lazarev, Empirical Education Inc.

Denis Newman, Empirical Education Inc.

Cite this paper: Lazarev, V., & Newman, D. (2014). Developing Composite Metrics of Teaching Practice for Mediator Analysis of Program Impact.

An earlier version of this paper was presented at the Spring 2013 Conference of the Society for Research on Educational Effectiveness (SREE), Washington, DC.

Background

Efficacy studies of educational programs often involve mediator analyses aimed at testing empirically appropriate theories of action. In particular, in the studies of professional development programs, the intervention targets primarily teachers' pedagogical sills and content knowledge, while the ultimate outcome is the student achievement measured by a standardized test. In this case, teaching practices affected by the professional development program in question can be measured in the process of classroom observation, as well as using one or more additional instruments including peer and stakeholder surveys. Using an observation instrument-rubric has a number of advantages. One such advantage is the immediacy of observation: it can be used to measure teacher practice at any moment in time, while surveys typically rely on a cumulative experience of survey takers. Another advantage is that a certain level of reliability can be achieved and maintained through rater training, certification, and calibration, which is seldom an option with surveys. Classroom observation can therefore produce series of reasonably accurate data points (scale scores) that could be included as mediator variables in the analyses of student outcomes.

Researchers may face two types of problems when implementing observation-based approach to measuring the quality of a professional development program. First, an observation rubric consists of a number of items, which measure distinct observable aspect of teaching practice. These components need to be aggregated to produce a single metric or several composite scores, depending on the dimensionality of the underlying concept of teaching quality. Without some sort of aggregation, the task of mediation analysis using all rubric components can be intractable.[1] A typical solution involves summation or averaging of element scores to obtain a single metric, which is often referred to as *the* observation score. However, the relative contributions of item scores to student outcomes in the short term can vary because some rubric items can be measuring aspects of teaching that do not translate directly into higher student

[1] Most observation rubrics currently in use by researchers and administrators consist of at least one dozen separate items (called components, elements, or standards) grouped into several domains.

scores in the nearest testing period. In addition, it is practically impossible to develop item definitions and scoring guidelines that would guarantee independence of each item and equal distribution of measurement error across items. Differences in the measurement error across items and correlations among item scores also affect the contribution of each item to the total variance of observation data.

Second, observation item scores are ordinal ratings, designed primarily to assess teaching practices, rather than measure student outcomes in a precise way. Observation scores are therefore not necessarily related to student test scores and test-based teacher value-added in a linear fashion, even if they reflect meaningful inputs into student outcomes. Some recent findings such as the asymmetry in the distribution of observation scores (Tennessee Department of Education, 2012; <other REFs>) and the analyses performed in the framework of Measure of Effective Teaching (MET) project (Kane and Staiger, 2012) suggest possible non-linearity in the relationship between observation and value-added scores. With few exceptions (Grossman et al., 2010; Kane et al., 2010), little research has been conducted on the item-level relationships between observation and student outcomes or value-added measures of teacher effectiveness.

Current Study

The main objective of this study is to develop a methodology for creating summative teacher performance metrics from item scores for use as mediators in the analyses of student outcomes. We use one particular observation rubric, PLATO (PLATO, 2014), to answer the question of how a summative teacher performance metric, aligned with a selected measure of student achievement, can be constructed from item scores.

The main requirement for a summative metric in the context of mediator analysis is that it combines item scores in a way that maximizes the correlation between the set of teacher performance indicators (rubric items) and student outcomes. A metric that is poorly aligned with the outcomes variable(s) is statistically inefficient: it may prevent detecting a mediating effect even when sample size is expected to be sufficient to detect the effect on student outcomes. Simple ad hoc approaches currently in use, such as summation of item scores, result in 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) or optimal combinations of component scores in multi-measure teacher evaluation systems (Hansen et al, 2013; Mihaly et al, 2013) little attention has been paid to the statistical properties of observation metrics per se. In this study, we develop several alternative composite metrics of teacher performance and compare their statistical properties.

Statistical Model

In the context of an experimental study of a program that affects student achievement indirectly, such as a professional development program, 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 possible covariates, on student outcomes:

$$Y_1 = Y_0\alpha + X\beta + T(\theta; Z)\,\gamma + \varepsilon, \qquad (1)$$

where $Y_{1,0}$ are student outcomes (e.g. test score) before and after the intervention, θ is the treatment indicator, X is the vector of student characteristics, Z is the vector of teacher characteristics, α, β, and γ are coefficients to be estimated. Teacher performance may be a single holistic score or a vector of observable teacher performance aspects (observation rubric item scores), so that $T = \sum_{j=1}^{N} T_j$, $N \geq 1$.

In this specification, teacher performance function, $T(\theta; Z)$, allows for any shape of relationship between its inputs and the performance metric(s), as well as differential impact of each component of teacher performance. For practical purposes, some simplifying assumption about $T(\theta; Z)$ have to be made. One approach could be to treat every observation rubric item as a direct measure of a distinct teacher's skill, as in Kane et al (2010), and to estimate the mediating relationship for each item score independently. In practice however, this task may be intractable as an experimental study may not be sufficiently powered to estimate a model with a large number of mediators-observation rubric items. Moreover, high correlations among items (component scores) of observation rubrics reported in many recent studies (e.g. Chaplin et al., 2014; Lazarev et al., 2014) suggest that such items can be considered as partially complementary proxies for overall teacher quality rather than indicators of distinct skills with unique contributions to student outcomes. It is therefore desirable to have a single metric of teacher performance that efficiently aggregates the teacher performance information inherent in individual observation items.

Rearranging the terms in equation (1) and noting that $Y_1 - Y_0\alpha - X\beta = \hat{Y}$ is the definition of teacher value-added (denoted hereafter $\hat{Y}$) obtains:

$$T(Z) \sim \hat{Y} \tag{2}$$

This implies that, in the research context considered here, an empirical approximation of the teacher performance function, T, needs to be correlated with the teacher value-added metric derived from the outcome of interest. In other words, we need to calibrate an observation instrument applying a regression technique to extant observation and student outcome data.

In practical applications, we can assume that the empirical summative metric of teacher performance is a sum of observation components - item scores, s_j - or some transformation of those scores, possibly non-linear:

$$T = \Sigma f_j\,(s_j) \tag{3}$$

A representation of T needs to be chosen so as to maximize the correlation between T and the value-added metric $\hat{Y}$ in a sample of calibrating observations. In the following, we consider several alternative approaches to constructing transformation functions, f_j, and compare properties of the resulting summative metrics.

The most straightforward and frequently used approach is to use a simple sum of item scores, or their average (which is statistically equivalent), and a perform a simple (univariate) regression analysis:

$$\hat{Y}_i = \Sigma\, s_{ji} + \varepsilon_\mathrm{i} \tag{4}$$

Technically, this involves substituting all functions f_j in (3) with an identity function: $f_j = I$.

Another approach is to perform a linear regression analysis[2] and obtain an approximation of the form:

$$\hat{Y}_i = \Sigma\, s_{ji}b_{ji} + \varepsilon_i \qquad (5)$$

A version of the approach above, applicable in a situation where more than one distinct dimension of teacher quality can be identified, involves performing factor analysis first and using factor scores, ϕ_k in the subsequent regression analysis:

$$\hat{Y}_i = \Sigma\, \phi_{ki}\, b_{ki} + \varepsilon_i, \text{ where k } \in(1,K),\ K < N \qquad (6)$$

A linear regression model can be extended to a polynomial regression by inclusion of powers (squares, cubes, etc.) of items scores in an attempt to model inherent non-linearity:

$$\hat{Y}_i = \Sigma\, P_{ji}\,(s_{ji}) + \varepsilon_i \qquad (7)$$

Finally, generalized additive function approach allows defining the most accurate and complete representation of T. Under this approach, each of the $f_j(s_j)$ terms in (3) is an arbitrary smooth function (transformation) of item scores. Estimate a generalized additive model

$$\hat{Y}_i = \Sigma f_{ji}\,(s_{ji}) + \varepsilon_i \qquad (8)$$

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 $f(s)$. It also allows identifying items that do not contribute at a statistically significant level to the summative indicator either because they are measured with too much statistical noise, "crowded out" due to multiple correlation with other items, or unrelated to the outcome of interest in principle, at least in the short run.

[2] More generally, it could be a mixed model if available data allows modeling rater and/or school effects.

Analysis of functional terms $f_j\ (s_j)$ may help understand the relationship between teacher performance and student outcomes. However, non-parametric nature of these terms makes it difficult to use them for prediction. In practice, most non-parametic smooth curves can be approximated by polynomials and the generalized additive model can be used to define an optimal polynomial or linear regression model by relying on the *estimated degrees of freedom* (e.d.f.)—a measure of non-linearity—reported for each term in the generalized model.

Each of the approximations obtained by methods (4) – (8) can be used to calculate composite scores for use in mediator analysis.

Data Collection and Analysis

The dataset we used to develop the methodology and to estimate the summative observation 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. Eight PLATO components-elements (out of 13 defined by the protocol) were used in coding (Table 1). 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 item was 1504.

TABLE 1. PLATO ELEMENTS USED IN THE STUDY

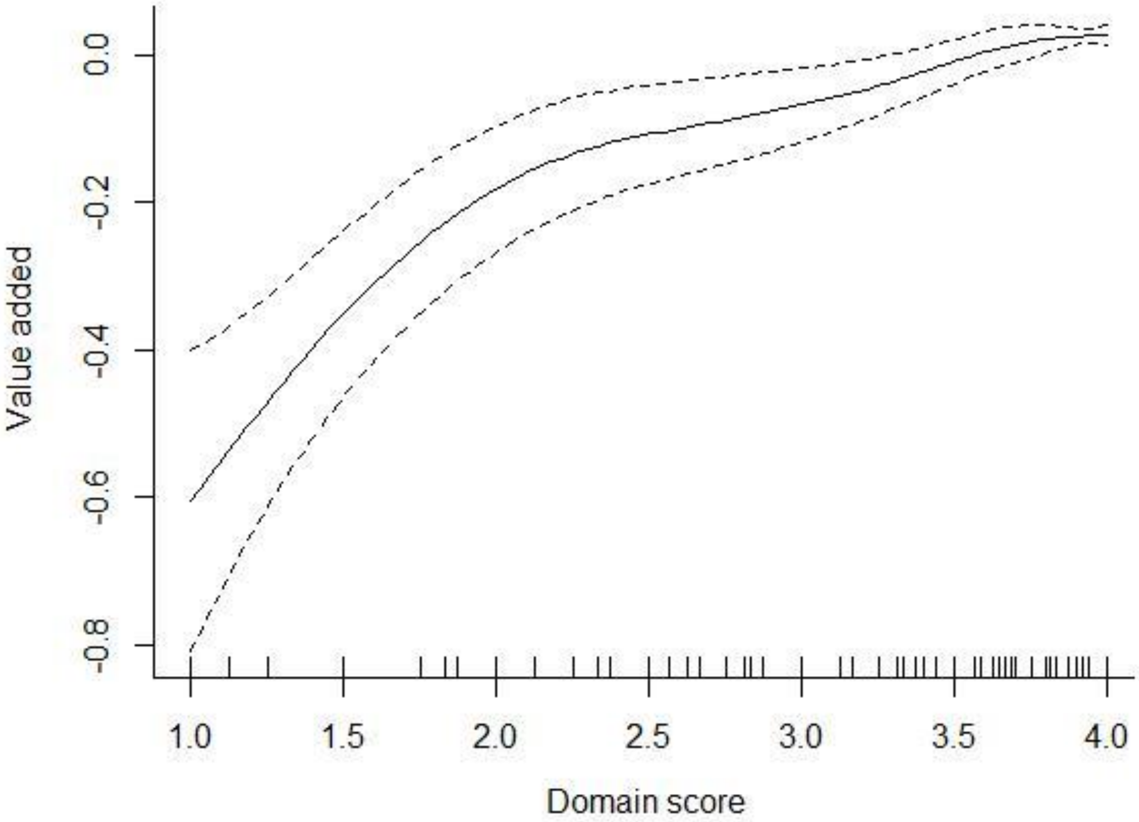
Element	Abbreviated name	Captured aspects of instruction
Representation of Content	RoC	Richness, accuracy, and clarity of the teacher's explanations and examples
Intellectual Challenge	INCH	Intellectual rigor of the activities in which students engage.
Modeling/ Use of Models	MDLG	Degree to which a teacher visibly enacts strategies, skills, and processes targeted in the lesson to guide students' work before or while they complete the task, the extent to which they are analyzed or not, and whether they are used to illustrate for students what constitutes good work on a given task.

TABLE 1. PLATO ELEMENTS USED IN THE STUDY

Element	Abbreviated name	Captured aspects of instruction
Strategy Use and Instruction	SUI	Teacher's ability to teach ELA strategies that can be used flexibly and independently.
Classroom Discourse	CLDI	Opportunity for and quality of student conversations with the teacher and among peers
Behavior Management	BEMT	Degree to which behavior management facilitates academic work.
Time Management	TIME	How well-paced and efficient tasks and transitions are in the classroom.

Analysis was performed using *R* package *mgcv* (Wood, 2006) and involved estimation of a generalized additive model. The principal output of this procedure (see an example in Figure 1) is a plot of the functional relationship (“smooth”) between two quantities (in this case, an observation item score and the value-added score) together with the confidence bands, estimated degrees of freedom (e.d.f.) 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 is a measure of non-linearity. If its value is close to one, then the relationship is linear or possibly non-existent if the statistical significance of the estimate is low), while higher degrees imply that the relationship is non-linear. In some cases non-monotonic relationship (e.g. U-shaped) implies that a particular item does not have an unambiguous effect on outcomes even though the relationship is technically significant. Analysis of an estimated generalized additive model allows developing a simpler parametric approximation by replacing non-parametric “smooths” with linear terms or low-order polynomials based on the e.d.f. values, thus moving from a model of type (8) to an adequately specified model of type (5) or (7). These types of models are easier to interpret and can be used to calculate a composite observation score.

FIGURE 1. ESTIMATED RELATIONSHIP BETWEEN AN OBSERVATION RUBRIC ITEM SCORE AND THE VALUE-ADDED. PLOT OF A "SMOOTH" WITH THE CONFIDENCE BANDS (DOTTED LINES) AND ACCOMPANYING STATISTICS. AN EXAMPLE.



Approximate significance of smooth terms:

	e.d.f.	F-value	*p* value
f(Domain_score)	3.857	14.36	3.69e-13 ***

Findings

In the following, we present the results of estimation of models (4)-(8) and compare the models in terms of goodness of fit (R^2). The simplest model (4)—a univariate regression of value-added score against observation total (sum of all observation item scores) presented in Table 2—provides a useful baseline, with the R^2 of .026. A complete linear regression (Table 3) has a much higher R^2 of .042. However most terms in it are not significant. Limiting the model only to significant terms results in a model with only a marginally smaller R^2 of 0.040 (Table 4).[3]

TABLE 2. REGRESSION OF THE TEACHER VALUE ADDED ON PLATO AVERAGE SCORE

	Regression coefficient	*p* value
PLATO score (Average of 7 components)	0.024	<.01
Constant	-0.395	<.01
R^2	=0.026	

TABLE 3. CONTRIBUTIONS OF PLATO COMPONENT SCORE TO THE TEACHER VALUE ADDED (LINEAR REGRESSION)

	Regression coefficient	*p* value
TIME	0.024	0.221
SUI	-0.002	0.926
MDLG	0.005	0.790
BEMT	0.098	0.000
CLDI	-0.007	0.766
INCH	0.036	0.140
RoC	0.048	0.556
Constant	-0.564	<.01
R^2	= 0.042	

[3] In fact, adjusted R^2 of the reduced model is higher than that of the complete because of the smaller number of terms: .038 vs. .036.

TABLE 4. CONTRIBUTIONS OF PLATO COMPONENT SCORE TO THE TEACHER VALUE ADDED. LINEAR REGRESSION, REDUCED

	Regression coefficient	*p* value
BEMT	0.113	<.01
INCH	0.039	0.031
Constant	-0.510	<.01
R^2	= 0.040	

A linear regression model with factor scores instead of raw item scores (Table 5) performs not much better than the baseline model (R^2= .028). This is a result of the incomplete set of items in the data and limited proportion of common variation accounted for the model.[4]

TABLE 5. CONTRIBUTIONS OF PLATO FACTOR SCORE TO THE TEACHER VALUE ADDED (LINEAR REGRESSION) AND CORRESPONDING FACTOR MODEL

Regression model					Factor loadings				
	Coefficient	*p* value	TIME	BEMT	INCH	RoC	CLDI	MDLG	SUI
Factor 1	0.055	<.01	1.019	0.619					
Factor 2	0.041	<.01						0.595	1.017
Factor 3	-0.043	.051			0.925	0.106	0.723		
Constant	-0.001	.87							
R^2	= 0.028								

[4] Factor model used here recovers two of the four theoretical factors identified by the developers of PLATO: Instructional Scaffolding (includes SUI and MDLG), and Classroom Environment (based on BEMT and TIME). Three remaining items present in the MET data are associated with a single factor, with RoC largely unexplained by the model (uniqueness = .98).See <AERA paper mentioned by Lindsay> on the factor structure of PLATO.

Our most advanced results based on the generalized additive function approach—type (8) model—are summarized in Table 6 and Figure 2.

TABLE 6. CONTRIBUTIONS OF SMOOTH TRANSFORMATIONS OF PLATO COMPONENT SCORE TO THE TEACHER VALUE ADDED. NON-PARAMETRIC (GENERALIZED ADDITIVE) MODEL

	Estimated degrees of freedom	*p* value
f(TIME)	1.00	0.20
f(SUI)	1.00	0.98
f(MDLG)	3.78	0.04
f(BEMT)	3.61	<.01
f(CLDI)	2.21	0.18
f(INCH)	2.95	0.53
f(RoC)	1.00	0.58
R^2	=0.053	

Note. "Estimated degrees of freedom" is a measure of linearity of relationship, with 1 corresponding to strictly linear relationship.

Graphs in Figure 2 reveal a variety of patterns of relationship between component scores and teacher value-added, similar to what we have reported elsewhere using other observation instruments (Lazarev & Newman, 2013). Two components—MDLG and CLDI—exhibit non-monotonic relationship to the outcome. One of them—Modeling (MDLG)—has a significant inverted U-shaped relationship to the outcome (ignoring the range above ~3.0 where only a few observations are located). This pattern of relationship suggests that teachers' performance has an optimum in the middle, whereas deviations towards either tail of the distribution of MDLG scores are associated with poorer student performance. Two more items—BEMT and INCH—have monotonic relationships with a moderate degree of non-linearity. Three remaining domains—Roc, TIME, and SUI—had strictly linear but insignificant associations with the outcome (signified by e.d.f. of 1 and high p values in Table 6). Thus model is the strongest of all with R2 of 0.053, probably a maximum of explained variation that can be achieved by any model.

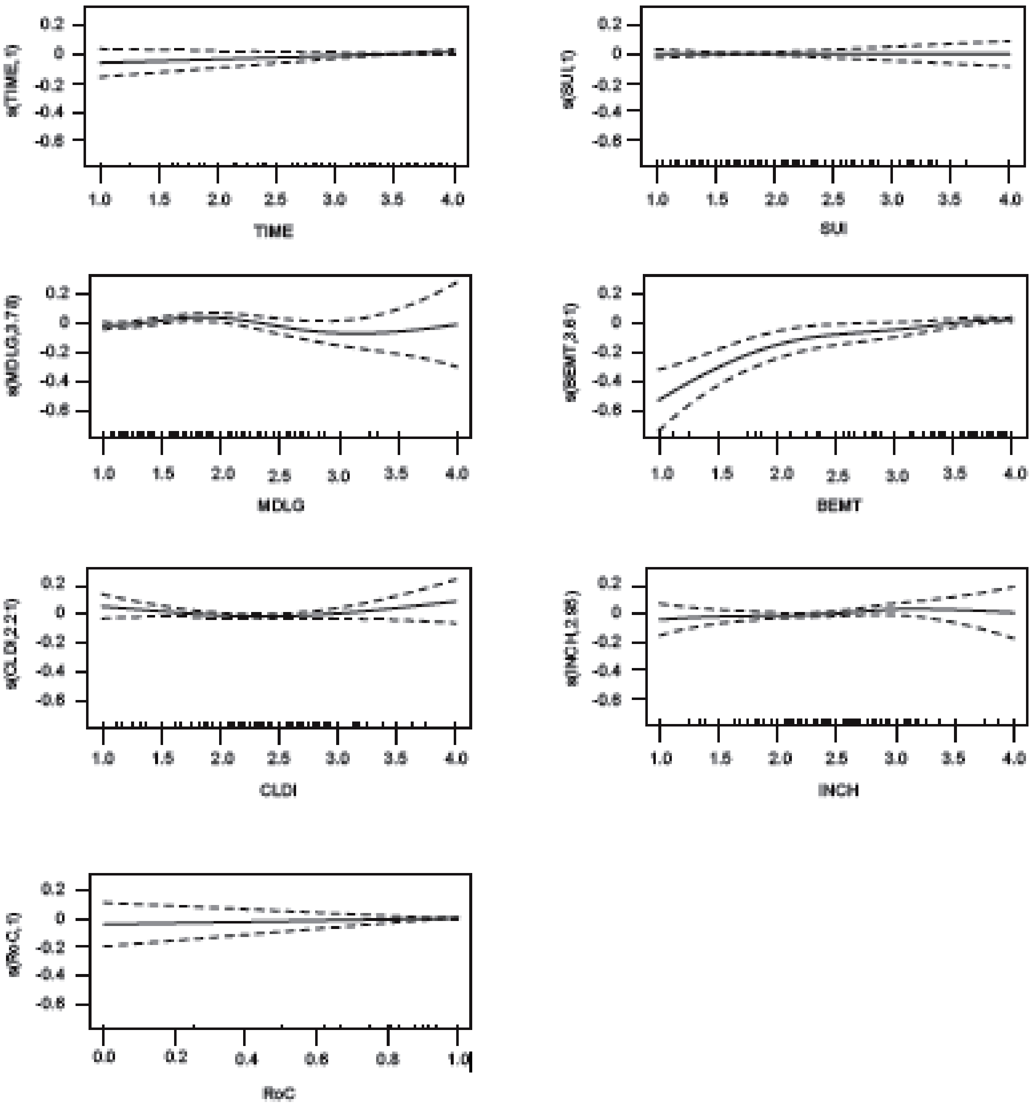


FIGURE 2. SHAPES OF RELATIONSHIPS BETWEEN PLATO COMPONENT SCORES AND TEACHER VALUE-ADDED

Analysis of this model allows specifying an efficient polynomial model without much exploratory analysis. First, we remove items that are insignificant in the model: Roc, TIME, and

SUI. Then for each remaining item, we include item itself and its powers up to rounded e.d.f. For example, e.d.f. for MDLG is 3.78, which is approximately four; we therefore include square, cube, and the fourth degree of MDLG into the model. After initial estimation, the model is refined by the elimination of statistically insignificant power terms. The resulting model (Table 7) has R2 of 0.051, which makes it almost as good as the generalized additive model thanks to its ability to reproduce the complexities of the relationships between observations cores and student outcomes. At the same time, this is a parametric model that can be easily used to calculate composite scores from data.

TABLE 7. CONTRIBUTIONS OF PLATO COMPONENT SCORE TO THE TEACHER VALUE ADDED. PARAMETRIC MODEL (POLYNOMIAL REGRESSION)

	Regression coefficient	*p* value
BEMT	0.448	<.01
BEMT^2	-0.054	<.01
INCH	0.037	0.12
CLDI	-0.279	<.01
CLDI^2	0.060	0.01
MDLG^2	0.258	<.01
MDLG^3	-0.150	<.01
MDLG^4	0.023	<.01
Constant	-0.841	<.01
R^2	=0.051	

Discussion

We have outlined a number of approaches to constructing summative teacher performance metrics for the purposes of program evaluation studies, in particular for mediator analyses.

The intermediate steps involving estimation of a generalized additive model or factor analysis are fairly data intensive and requires large number of observations (hundreds to thousands) in the calibrating sample. Once an optimal metric is constructed the sample size requirements for a program evaluation study are moderate. Proposed methods of summative score development help to increase the accuracy of the analysis substantially and lower the sample size requirements. The sample size is approximately inversely proportionate to the effect size, and the effect sizes (*R*) in the best models reported here are almost twice as large as those of the baseline model with the sum of item scores. The decision, as to which model to use, lies with the researcher in a particular study, and it has to do mostly with the tradeoff between complexity and accuracy. This decision is particularly difficult in case when a non-monotonic relationship is identified: on one hand, ignoring such a relationship would result in a biased and inefficient model. On the other hand, using an observation score as a mediator of the program effect that has itself an ambiguous relationship to the outcome of choice, creates problems for the interpretation of the results. In a simpler case of monotonic, albeit non-linear relationships, a choice in favor of a more complex model can be clearly justified by the benefit of a higher model quality.

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.

Chaplin, D., Gill, B,. Thompkins, A., & Miller, H. (2014). Professional practice, student surveys, and value- added: Multiple measures of teacher effectiveness in the Pittsburgh Public Schools (REL 2014–024). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic.

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.

Hansen M., Lemke M., & Sorensen N. (2013). *Combining Multiple Performance Measures*. Washington, DC: American Institutes for Research.

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

Lazarev, V. & Newman, D. (2013). *How Non-Linearity and Grade-Level Differences Complicate the Validation of Observation Protocols*. Paper presented at the Fall 2013 SREE conference, Washington, DC, September 2013.

Lazarev, V., Newman, D., & Sharp, A. (2014). *Combining classroom observations with other measures of educator effectiveness in Arizona’s pilot teacher evaluation model* (REL 2014-050). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National

Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory West.

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.

Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). *A Composite Estimator of Effective Teaching*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved April 22, 2013, from http://metproject.org/downloads/MET_Composite_Estimator_of_Effective_Teaching_Research_Paper.pdf

PLATO. (2014). Description of the Thirteen Elements. Retrieved from: http://platorubric.stanford.edu/Elements.html

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, S. (2006). Generalized Additive Models: An Introduction with R. Oxford: Taylor and Francis