

## Abstract Title Page

**Title: How Non-Linearity and Grade-level Differences Complicate the Validation of Observation Protocols**

**Authors and Affiliations:**

Valeriy Lazarev, Empirical Education Inc.

Denis Newman, Empirical Education Inc.

## Abstract Body

### Background

Teacher evaluation is currently a major policy issue at all levels of the K-12 system driven in large part by current US Department of Education requirements. For research that addresses educational effectiveness, the generation of data about teachers by school systems is adding a new class of administrative data available to effectiveness studies both as outcome measures and as covariates to improve precision and to use as potential mediators in the analysis of impact on students. However, measures of teacher effectiveness being rolled on a large scale for purposes of evaluation are still immature. While questions are being raised in the high-stakes policy context, a similar need for more analysis must be satisfied before the resulting administrative data can be used confidently by researchers. This paper takes a step toward understanding one of the widely used frameworks for evaluating teaching practices through systematic classroom observations (Danielson, 2007, 2013). We focus on math instruction in grades four through eight. In examining the validity of the instrument, our focus is on the developmental changes in what can be described as effective student-teacher interaction as children progress from elementary to middle school.

This study came out of two strands of work representing multiple disciplines. First, we have been providing technical assistance to school systems that are concerned with the validity of the observational measures they are incorporating into teacher evaluation, where validity is often envisioned as a strong association between the teacher's score from a set of administrator's classroom observations and the teacher's effectiveness measured in terms of student growth (often value-added measures.) Second, we have begun using data from standard observation protocols in our experimental research where the strength of the association with achievement scores is important in using impact of classroom practice as a mediator of impact on student achievement. We have available the large corpus of data collected in the Gates Foundation's Measures of Effective Teaching (MET) project allowing us to delve deeply into the questions of the nature of the relationship between measures of classroom practice and measures of their students' achievement. Our work questions two assumptions that underlie existing analyses: first that protocols are "generic" (Danielson, 2007, pp. 21-23) in applying equally to teaching any content at any grade level and, second, that linear correlations are the appropriate characterization of the association between observation scores and measures of achievement.

In the view of the recent MET studies on which this work builds, development of a consistent teacher evaluation system requires that various evaluation instruments (observation rubrics and student surveys) can reveal the latent teacher effectiveness and are therefore correlated with a given student growth metric (e.g. value-added scores), which is the ultimate "output" of teaching practice. The MET reports provide a strong evidence of positive correlation between various metrics of teacher performance. The strength of this statistical association is however moderate (Kane and Staiger, 2012). An extensive and growing literature devoted to the analysis of robustness of value-added models (Ballou 2005; Jürges and Schneider, 2007; McCaffrey et al., 2008) suggest that the problem may lie with the sensitivity of value-added estimates with respect to model specification and measurement error. Some authors find that the value-added estimates are inherently volatile (Baker et al. 2010, Hill 2009).

Little attention has been drawn so far to the analysis of functional relationships between value-added and observation scores. The association between the two types of metrics is reported in terms of the linear correlation coefficient. Observation scores however are subjective estimates that use ordinal scale. Moreover, the goal of observation protocols is to catalog and evaluate teaching practices, not student outcomes. Observation scores are therefore not necessarily linear functions of underlying value-added, and correlation may not be an adequate measure of association. Some recent findings such as the striking asymmetry in the distribution of teacher scores reported by several states (see for example, Keesler and Howe, 2012, and Tennessee Department of Education, 2012) suggests that non-linearity in the relationship between observation and value-added scores may be preventing classroom observation instruments from differentiating among higher performing teachers. Furthermore, many observation instruments are expected to be equally effective in every classroom, which may be an unrealistically strong assumption given the variability of teaching practice across subjects and grade levels. To the best of our knowledge, no studies addressing these issues have been conducted.

### **Objective and Research Questions**

The main objective of this study is to explore the patterns of relationship between observational scores and value-added measures of teacher performance in math classrooms and the variation in these relationships across grade levels. While the MET analyses used a single composite score consisting of a simple average of the eight component scores of the protocol, in our work we treated each component separately since each measures a separately definable aspect of classroom practice. Specifically, across all the components, we pose the following questions:

- Do the relationships between observation scores of math teachers and their value-added scores tend to be non-linear?
- Is there a difference in the relationships associated with developmental changes as indicated by grade levels (four through eight)?
- To what extent the observed changes are associated with increasing departmentalization (in upper elementary and middle school)?

### **Setting:**

N/A

### **Population / Participants / Subjects:**

Data come from classrooms of teachers from six major urban school districts.

### **Intervention / Program / Practice:**

N/A

### **Significance / Novelty of study:**

This study questions two assumptions in prior work on the validity of observation protocols: first, such frameworks are generic and can be applied across grade levels and second, the relationships to achievement scores are linear. Our findings have several important methodological and policy implications, most importantly (1) analysis of teacher effectiveness should not be limited to the analysis of (linear) correlations, (2) not all components can be unambiguously associated with achievement gains, (3) classroom observation should be calibrated for use in particular types of teaching environments. Novel statistical methods used in this study are used increasingly in engineering and environmental studies, but are not well

known among educational researchers. Ultimately, this study helps develop a more rigorous approach to comparing different metrics of teacher performance and their correct interpretation and use in a policy context.

### **Statistical Model:**

Since the component scores can be an arbitrary function of the value-added score, our primary goal is to establish the “true” shape of such a function. We use therefore a nonparametric approach: estimation of a generalized additive model using penalized regression splines (Wood 2006). This approach allows determining the optimal degree of smoothing and therefore identifying the true shape of the relationship between the student growth metric  $x$  and the observational score  $y$ . Basic model used for exploratory analysis is:

$y_i = s(x_i) + \varepsilon_i$ , where  $x_i$  is the value-added score of teacher  $i$ ,  $s(x) = E(y|x)$  is a smooth function, and  $\varepsilon_i$  is the zero-mean random error term. Examples of estimated  $s$  functions are presented in Figures in Appendix B.

In order to answer the question about the association of the changing relationships with the departmentalization of teaching as opposed to gradual developmental change, we apply the following logic. If departmentalization is the main factor responsible for the different shapes of functional relationships in elementary and middle schools, then  $y_i = s_{nd}(x_i) + s_d(x_i) + \varepsilon_i$ , - separate functions are estimated for non-departmentalized and departmentalized teaching setting respectively. If the observed change is due to developmental factors, then the relationship is changing gradually with grade level. An appropriate model in this case is a two-dimensional smooth function allowing for arbitrary interactions with the grade-level:

$y_i = s_2(x_i, g_i) + \varepsilon_i$ , where  $g_i$  is grade level taught by the teacher. To analyze the relative contribution of the two types of factors, we estimate a single nested model:

$y_i = s_{nd}(x_i) + s_d(x_i) + s_2(x_i, g_i) + \varepsilon_i$  between Estimating separate functions of each grade level This model is identifiable due to the presence of a substantial number of elementary math classrooms observed in a departmentalized environment.

The resulting model is too complex to visualize, but for the purposes of our analysis, it is sufficient to compare statistical significance of smooth terms  $s$  included in the model. Results are presented in Table 1.

### **Applicability of Method:**

Methods used in this study are applicable for the analysis of teacher effectiveness and teacher evaluation systems that include observational and student growth components.

### **Research Design:**

N/A

### **Data Collection and Analysis:**

We used the dataset that was created in the framework of MET project. As part of this project, several thousand high-quality video recordings of upper-elementary and middle-school math and ELA lessons were scored by observers trained in the use of Framework for Teaching observation rubric (Danielson 2013) on eight components of two domains: “classroom environment” and “instruction.” The dataset also contained value-added scores for the teachers featured in the

videos calculated from the student performance on Balanced Assessment in Mathematics (BAM) and state assessments (see Kane et al., 2012, for details). We limit the analysis to the value-added scores based on BAM because of the uniformity of this assessment across the sites of this multi-state study and its closer alignment with common-core standards (MARS, 2013). The total number of usable data points was 1102, each representing a unique teacher. Grades levels represented in the sample are 4 through 8. About one quarter of elementary school math teachers, 95% of 6<sup>th</sup> grade teachers, and all 7<sup>th</sup> and 8<sup>th</sup> taught in a departmentalized setting.

Analysis was performed using *R* package *mgcv* and involved estimation of a generalized additive model using penalized regression splines (Wood 2006).

### **Findings / Results:**

First, we find that the relationships between many observation indicators (FFT component scores) and value-added scores are non-linear. Examples of two most frequent types of non-linear relationships are presented in Appendix B “ $\Gamma$ -relationship” (Figure 1) implies that the indicator can differentiate well between low to middle levels of teaching effectiveness but does not discriminate statistically between high-performing teachers. “ $\Pi$ -relationship” (Figure 2) implies that a component tends to evaluate deviations from the norm or expectations associated with the teachers in the middle of achievement distribution.

Second, we find that the shape of the relationship between observation scores and teacher value-added tend to differ between elementary and middle school. Figure 3 juxtaposes estimated functional relationships between teachers’ scores on “Managing student behavior” component and their BAM-based value-added scores. Elementary school relationship is characterized by a lower strength of association between the two metrics, which represents a general pattern that we identify in the data across the components. The middle-school relationship is a typical  $\Gamma$ -relationship suggesting that this instrument is effective only for the three lower quartiles of the teachers’ value-added score distribution but cannot differentiate among the highest performing math teachers. The pronounced differences between the elementary and middle level relationships are observed for all observational components.

Third, we find that developmental factors may have a much greater effect on the changing shape of the relationship between the observation and value added scores. Estimation of the nested model described earlier, shows that the grade-level interaction term is highly significant in each component model while the term associated with departmentalized teaching is typically not significant (see Table 1). The elementary-middle differences are therefore primarily associated with gradual change between grades rather with the specialization of math teaching.

### **Conclusions:**

The relationship between classroom observation instruments and teacher effectiveness is often presumed to be a simple linear mapping. We find that the relationship is complex: it is typically non-linear and the shape of this relationship appears to evolve with the grade level progression. Careful analysis of such relationships will improve our understanding of teacher effectiveness and result in more accurate teacher evaluation systems. Better evaluation systems will in turn generate more useful administrative data for use in effectiveness research.

## Appendices

### Appendix A. References

- Baker, E.L., Barton, P.E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R.L., Ravitch, D., Rothstein, R., Shavelson, R.J., & Shepard, L.A. (2010) "Problems with the use of student test scores to evaluate teachers," Economic Policy Institute, Briefing paper no. 278.
- 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.
- Danielson, C. (2007) *Enhancing Professional Practice: A Framework for Teaching*. 2nd ed. Alexandria, VA: Association for Supervision and Curriculum Development
- Danielson C. (2013) *The Framework for Teaching Evaluation Instrument*, Danielson Group
- 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.
- Hill, H. (2009). Evaluating value-added models: A measurement perspective. *Journal of Policy Analysis and Management* 28: 702–709.
- Jürges, H. & Schneider, K. (2007) "Fair Ranking of Teachers," *Empirical Economics*, 32, 411-431.
- 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
- Keesler V., & Howe, C. (2012) Understanding Educator Evaluations in Michigan: Results from Year 1 of Implementation, *Michigan Department of Education*. Retrieved from: [http://www.michigan.gov/documents/mde/Educator\\_Effectiveness\\_Ratings\\_Policy\\_Brief\\_403184\\_7.pdf](http://www.michigan.gov/documents/mde/Educator_Effectiveness_Ratings_Policy_Brief_403184_7.pdf)
- MARS (2013) Mathematics Assessment Resource Service (MARS): Balanced Assessment in Mathematics. Retrieved from: [http://mathshell.org/ba\\_mars.htm](http://mathshell.org/ba_mars.htm)
- 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](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

Figure 1. Relationship between observation scores and value-added scores. An example of “ $\Gamma$ -relationship.”

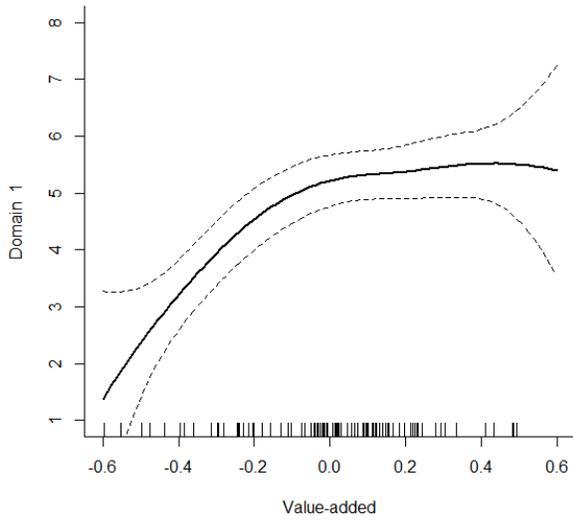


Figure 2. Relationship between observation scores and value-added scores. An example of “ $\Pi$ -relationship.”

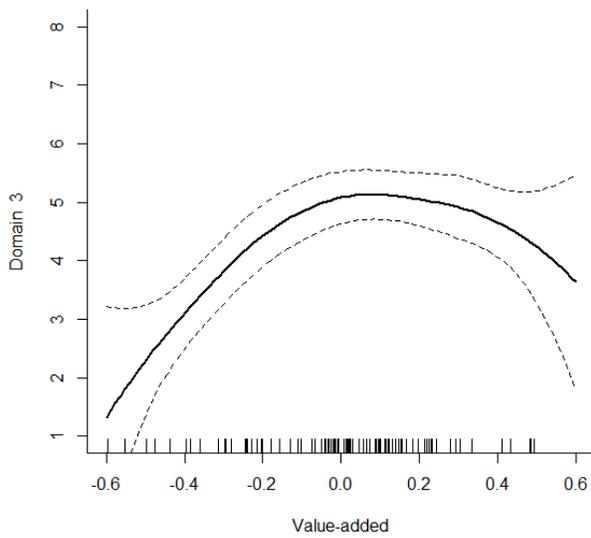
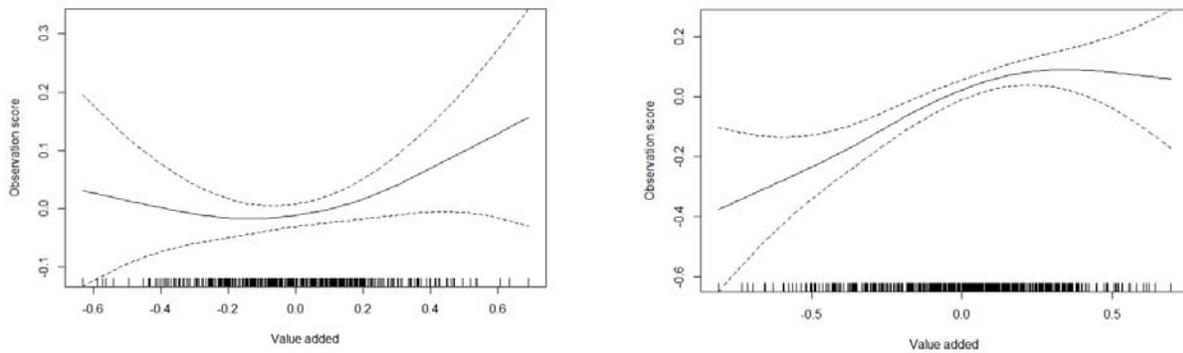


Figure 3. Changing relationship between observation scores and math value-added scores.



Left pane: grades 4 and 5. Right pane: grades 6-8. Observation scores are normalized.

Table 1. Significance of functional terms in component models

	R <sup>2</sup>	Statistical significance of terms		
		Common component, $s_{nd}$	Departmentalized teaching, $s_d$	Grade-level interaction, $s_2$
Creating an Environment of Respect and Rapport	0.096	-	-	++
Communicating With Students	0.074	-	-	++
Establishing a Culture for Learning	0.137	-	-	++
Engaging Students in Learning	0.097	++	-	++
Managing Classroom Procedures	0.073	++	+	++
Managing Student Behavior	0.079	-	+	++
Using Assessment in Instruction	0.067	-	-	++
Using Questioning and Discussion Techniques	0.050	-	-	++

+: p-level < 0.05      ++: p-level < 0.01