Empirical Education

EMPOWERING EDUCATORS FOR EVIDENCE-BASED DECISIONS

Can Multifactor Models of Teaching Improve Teacher Effectiveness Measures?

Valeriy Lazarev
Denis Newman
Empirical Education Inc.

NCLB waiver requirements have led to development of teacher evaluation systems, in which student growth is a significant component. Composite teacher evaluation scores commonly sum up the results of measurements made using several instruments. We hypothesize that, across different measures, there is more than one underlying factor and each measure can be decomposed into distinct factors. By performing factor analysis on the disaggregated evaluation data (observation components and survey items), we can identify several orthogonal factors, of which only one is associated with student test performance. We use teacher evaluation data collected by the Measures of Effective Teaching (MET) project as a model of state teacher evaluation system, which includes a value-added score as a measure of student performance, an observational rubric, and a student survey.

We find that one possible model of the latent data structure has three factors, of which only one is associated with value-added scores. This factor is also strongly associated with observation and survey items that deal with classroom control. The second factor is associated with such aspects of pedagogy as classroom dialog, questioning techniques, etc. The third factor is associated with items related to the notions of teacher sensitivity to students' well-being and includes mostly student survey items. Those factors can be interpreted as reflecting "effective," "constructive," and "positive" dimensions of teaching respectively. Each of them is an important and independent input into child development, while only first of them leads to achievement gains measurable in the short run.

The importance for policy of identifying orthogonal underlying factors is, first, that it provides precise knowledge of what exactly the evaluation system is measuring. This knowledge can be used by policy-makers to make an informed decision on how to combine the aspects of teaching into a single "teacher utility" function or, alternatively, if a multidimensional (matrix) evaluation system should be used. Second, it makes clear that teacher effectiveness consists of more than just the ability to promote student growth as measured by test scores. Additional factors may be weighted differently, for example, in identifying a teacher to become a mentor vs. to become a principal, which may require interpersonal capabilities unrelated to promoting student growth.

Third, evaluating teachers using several independent factor scores may help target resources (such as professional development) more accurately.

NCLB waiver requirements have led to development of teacher evaluation systems, in which a metric of yearly student growth is a significant component. In awarding waivers to states, Department of Education calls for educator evaluations, a substantial element of which is to be based on measures of student achievement. Once fully rolled out, these evaluation systems will be used in personnel decisions. It is generally assumed that an evaluation will consist of multiple measures. In particular, Race to the Top grant applications require states to design comprehensive evaluation systems with multiple measures of teacher performance. These measures often include – in addition to a test-based metric of student growth - observations by administrators, as well as surveys of parents, peers, and/or students. However DOE's focus on one component—student achievement—tends to put in the background what the other measures are measuring.

Recent empirical research has been focusing on metrics of student growth - value-added scores in particular – and their relationship to other metrics. An extensive set of recent teacher-evaluation studies conducted by the Measures of Effective Teaching (MET) project yielded a body of empirical evidence on the correlations among various teacher effectiveness metrics, including scores from several widely used classroom observation instruments, student surveys, and estimates of teachers' value-added contributions to student test achievement (Kane & Staiger, 2012). Conceptually, teacher effectiveness is viewed, if not explicitly defined, as one-dimensional: an evaluation system must be able to differentiate between high- and low-performing teachers and provide the resulting ranking for the purposes of allocation of professional development resources, personnel decisions, and ultimately to improve teaching through constructive feedback (Darling-Hammond, 2012; Grissom et al., 2013).

However recent studies, including the MET project, consistently show that the non-achievement measures (observations and surveys) have only limited correlation with the academic achievement measures, indicating that they may be measuring facets of teaching other than the teacher's ability to foster higher test scores. This leads us to ask: what are we measuring and what impact do factors, unrelated to short-term achievement gains, have on the education of our children, and how can these other factors be used in personnel decisions?

This paper will attempt to make the link from knowing what we are measuring in our evaluations to how we would use the information in personnel decisions. We view this link in terms of three steps, represented by the questions:

1. What are we measuring with observations and surveys? Can we identify a small set of underlying factors that are being measured?
2. What impact might these other factors have on the overall effectiveness of our education system?
3. How can the factors underlying the evaluation results be used in personnel decisions?

We present some preliminary results of our analysis that addresses the first question and discuss the approaches to the second and third questions.

1. What are we measuring with observations and surveys?

APPROACH

Adoption of an evaluation system relying on multiple metrics requires a method to combine several metrics produced by different instruments. If all metrics render same underlying concept, the task of producing a composite score is straightforward. An optimal composite would be a weighted sum of the component metrics, where the weights should reflect the relative reliability of each instrument. It appears as if current teacher evaluation systems follow this approach, by assuming that the teaching effectiveness is the only concept measured by all instruments employed in the system. Each instrument yields a single number—value-added score, observation score, etc.—and the composites (summative score) most commonly sum up these component scores. Measurements are therefore weighted at the level of instrument (see Figure 1 for schematic representation). In the absence of concrete findings to guide states and districts about how exactly component metrics might best be interpreted and combined (Rothstein & Mathis, 2013), the states tend to introduce ad hoc weighting schemas (e.g. student growth metric and observation should contribute each 50% to the total). Some recent studies of composite measures of teacher effectiveness have examined statistical foundations for compositing. For example, Hansen et al. (2013) and Mihaly et al. (2013) analyze properties of composite teacher scores created by assigning different weights to summative measures of student achievement, classroom observations, and student surveys, assuming that all three measure a single concept of teacher effectiveness.

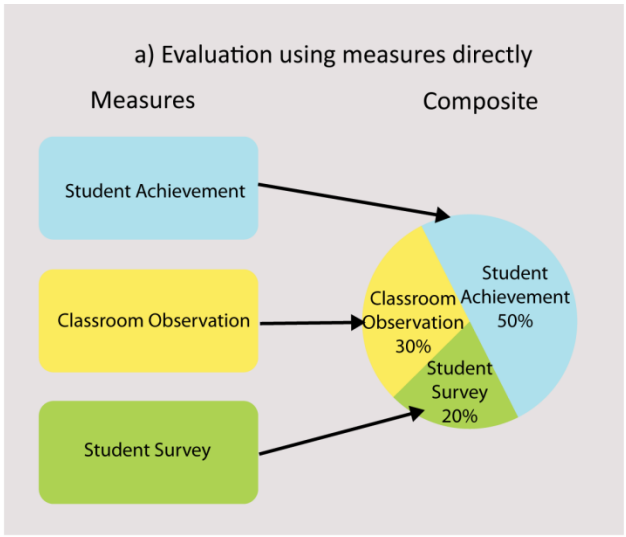
One important aspect of evaluation metrics that gets little attention in the studies of teacher evaluation systems is that they are, in their turn, multiple measures consisting of several elementary, separately scored items. Thus, classroom observation instruments consist of five or more components,[1] and surveys consist of dozens of items-questions. Moreover, the structure of cross-correlation between those items is complex. Our earlier analyses of MET data (Lazarev & Newman, 2013; Lazarev et al., 2013) showed that correlations between observational components and value-added scores vary widely. In particular, we found that only observational components associated with classroom and student behavior management are consistently correlated with value-added scores across grade levels and subjects. Kane and Staiger (2012) make a similar observation.[2] This suggests that complex instruments respond to different underlying concepts, and if so, summation of component metrics to obtain a single composite score may be an inadequate approach, which ultimately lowers the value of many kinds of personnel decisions made on the basis of teacher evaluations.

We take here an alternative approach. We hypothesize that, across different measures, there are more than one underlying factor. Only one of these factors may be associated with short-term achievement gains as measured by the test-based value added. Other factors may be correlated with distal outcomes that are impossible to measure synchronously with other metrics. Each of

[1] One of the widely used observation rubrics – FFT – has a total of 22 components.

[2] Their report however does not provide detailed results.

the n elementary measurement, v_i—an observational component or a survey item—is an imperfect instrument that may "pick signals" from one or more of the p factors, f_j (where $p << n$): $y_i = \sum_{j=1}^{p} \lambda_{ij} f_j$.[3]



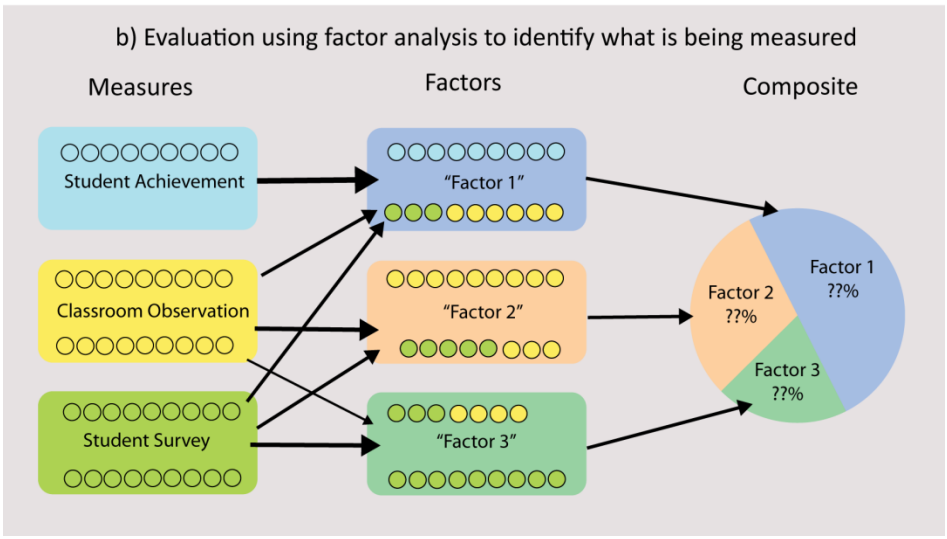


FIGURE 1. TWO APPROACHES TO COMBINING MULTIPLE MEASURES OF TEACHER PERFORMANCE

[3] In fact, some y_i can be a non-linear function of the underlying factors just as they exhibit non-linear relation to the value-added scores (Lazarev & Newman 2013). We will limit the analysis here to linear relationships for the purposes of tractability, although the existing methods of non-linear factor allow generalizing this model.

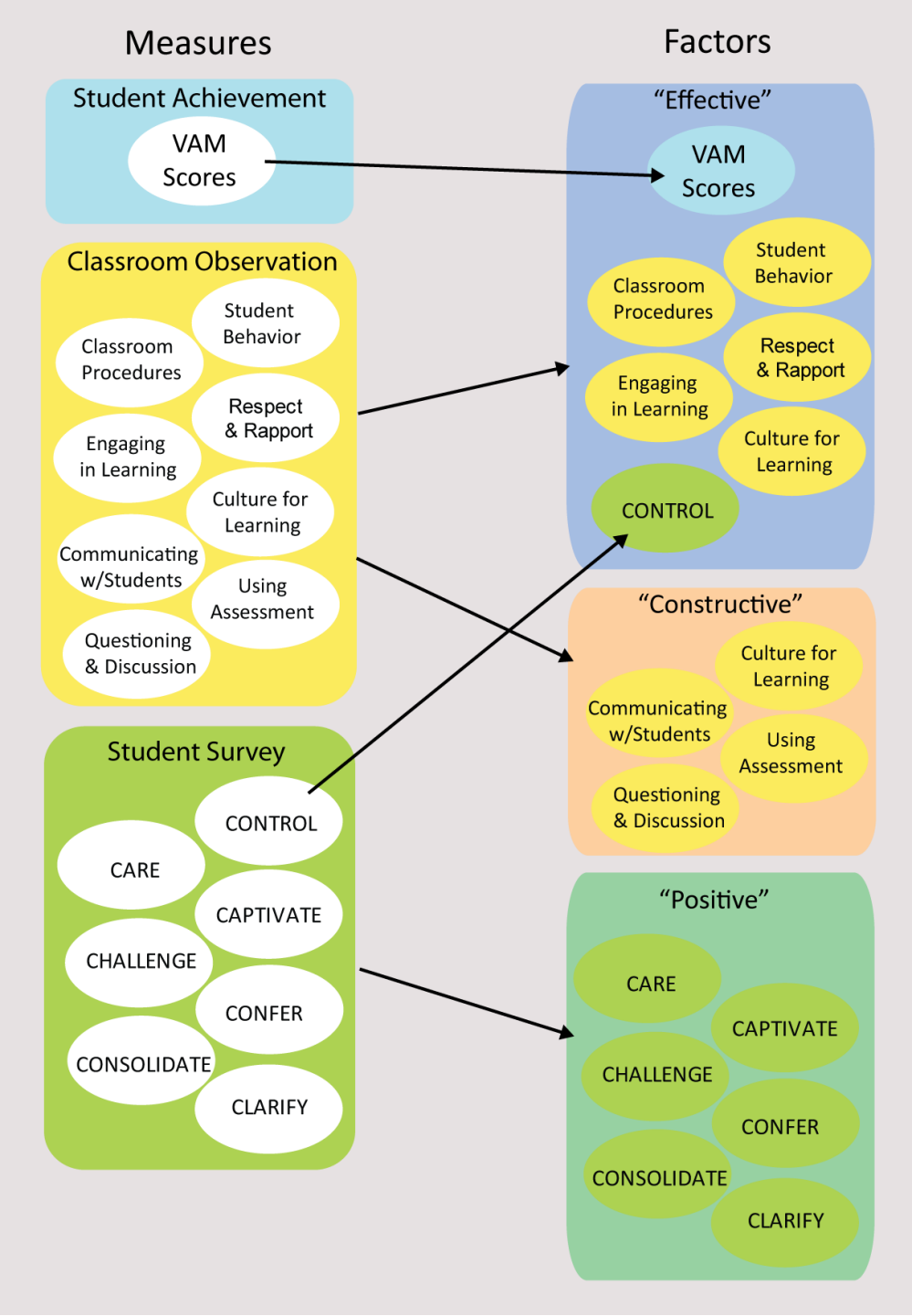


FIGURE 2. SCHEMATIC ILLUSTRATION OF FACTOR ANALYSIS RESULTS

As a result, measurements differ in the strength of their association with the student growth metric not because of the differences in reliability but because of the differences in the measured factors or combinations of factors. A natural analytic approach is then to apply factor analysis - a widely accepted method for analyzing a dataset with a large number of variables-

measurements, each of which may be a manifestation of a small number of latent underlying constructs (factors). Since we are particularly interested in identifying and interpreting factors that are *not* associated with value added, we will use a target rotation that produces the first factor that is uniquely associated with value added scores. This factor can be also expected to be associated with classroom management items. Once the additional factors are identified and interpreted, stakeholders can decide what weights to give each factor based on their comparative substantive importance (or social value). In other words, elementary measurements will be weighted at the level of measured concept rather than instrument (Figure 2).

DATA

We explore the idea outlined above using the data collected by the MET project—the largest existing corpus of teacher evaluation data collected in multiple large districts using the same set of instruments—student academic growth metric, observation rubric, and student survey. By design, the composition of this dataset resembles an output of teacher evaluation systems adopted by many states, with three instruments and multiple elementary measurements averaged to obtain component scores. Instead of limiting the analysis to a few aggregate scores for each teacher, we compiled a dataset with disaggregated measurements—survey items and observational components. This dataset includes, in addition to value added scores assigned to each teacher, 20 observable components of two generic observation rubrics[4]—8 of the Danielson Framework[5] and 12 of CLASS protocol—and 36 items of the Tripod student survey. These 36 items are categorized into seven broad characteristics of teacher performance as assessed by their students, the so called "7 Cs". These 7Cs categories include: Care, Clarify, Control, Challenge, Captivate, Confer, and Consolidate. Each category includes between three and eight yes/no questions. The dataset contains therefore a total of 57 variables—elementary measurements—for each teacher.

The MET project estimated two types of value-added models (VAM): one based on state test (distinct test in each of the five participating states) and another based on a study administered test (BAM for math and SAT9 for ELA). We use only the latter because the underlying tests are better aligned with Common Core and the same for all teachers in the dataset. We also limit our sample to middle school teachers (grades 6-8), which constitutes a majority of

[4] MET also used three subject-specific rubrics. We do not use those because they cannot be pooled together for the purposes of our analysis. Videos were scored by multiple teams of observers, so that most teachers have scores from several rubrics. We include in the dataset all teachers who have both CLASS and FFT scores.

[5] FFT has 22 components but only 8 of them are observable in the classroom, whereas the remaining 14 are based on administrator assessments of lesson plans, contribution to the school community, etc. Only the former eight components were observed and scored by the MET project.

records and cannot be pooled together with the elementary grades because of the differences in the composition of the survey.[6]

RESULTS

We follow a conventional approach in exploratory factor analysis. We start with a basic principal axes solution, determine the optimal number of factors, then perform an orthogonal rotation with known properties, and finally examine and interpret the rotated factor loadings, λ_{ij}. Using both scree and χ^2 tests, we find that a three factor model is adequate for the data.[7] In performing the target rotation of the factor structure, we impose a single constraint: only one factor should have a non-zero loading of the VAM score, i.e. only one row of the target matrix is defined.

Figure 2 illustrates main findings from our factor analysis. Although most loadings differ significantly from zero (see Table 1 in Appendix A), associations of variables-measurements with factors show clear patterns that are generally with findings mentioned earlier. We interpret the three factors based on the specific measurements that clustered together. We have labeled the three factors "Effective", "Constructive", and "Positive" dimensions of teaching.

1. Effective. One of the three factors is associated with student achievement as measured by the VAM scores. This factor is also associated with observational items reflecting teachers' skills in managing classroom and student behavior and following procedures. It is remarkable that among the student survey items, only questions relating to one "C", namely "Control", were associated with this factor. The Control category consisted of items also related to discipline and time management, such as "Our class stays busy and doesn't waste time."

2. Constructive. A second factor was associated with the classroom observational items reflecting mastery of such pedagogical devices as instructional dialog, feedback, and discussion. Constructivist pedagogical theory generally recommends building upon students' prior knowledge, active participation of students in their own learning, and teacher's role in scaffolding higher-order thinking and problem-solving. Thus, we identify these practices as "constructive". While Figure 2 only shows the FFT items, the CLASS observational items also show the same split between those associated with factor 1 and factor 2. It is important to note that specific items can be shared in some proportion between two or more factors. For example, the FFT component, "Establishing a Culture for Learning" is split approximately 50-50 between Effective and Constructive factors.

3. Positive. The third factor, also unrelated to VAM, consists primarily of all the remaining student survey items. The Tripod survey is heavily oriented to questions about the teacher's

---

[6] We have established that correlations between measurements differ between grade levels and that measurements, especially teacher observation and value added scores, are more closely interrelated in middle grades than in elementary grades (Lazarev & Newman, 2013).

[7] This does not mean that three-factor model is the only or the "best" possible model.

connection to students and students' positive feelings and perception of the teacher's empathy with the student's point of view, e.g., "My teacher knows when the class understands, and when we do not."

It is important to note the limitation of the MET data, which is that, because the observations were done with video of classrooms, it does not include the parts of the observational protocols that fall outside of the classroom such as supporting the school community, meeting with parents, and does not include peer or parental surveys that are also often part of teacher evaluations. We suspect that these could anchor a fourth factor associated with interpersonal skills and, perhaps, leadership. Analysis of data from well implemented large-scale teacher evaluations can help complete this picture but for now we have some intriguing results.

2. What impact might these other factors have on the overall effectiveness of our education system?

If our multiple measures are measuring factors that are not associated (or only weakly associated) with VAM scores, does that mean they have no particular value? This gets us to the second step of linking teacher evaluations with personnel decisions. We need to better understand the value of these factors for outcomes that we value for children's education. The VAM score is a measure of what the teacher accomplishes between fall and spring as measured by a standardized test. These and other calculations of a teacher's contribution to learning are given a high priority in most state teacher evaluation schemes. But there are many other outcomes, not always directly predicted by the VAM score, such as dropping out of school, persistence in applying for and getting into college, success in career, avoiding jail, and many other personal and social values that the school system may have an effect on. Although at this point, it is not even possible to measure such distal outcomes accurately, let alone predict how particular teacher characteristics will affect those outcomes, many school systems that have adopted multiple measures have at least a strong intuition that they are measuring something of value.

We can hypothesize linkages that our continuing research can explore. We can see that an efficiently run classroom is most strongly associated with test scores. The observational elements also suggest discipline and setting expectations. Teachers who score highly on this cluster of elements including generally doing well on the fall-to-spring achievement measure, may have an impact on student persistence, attendance, and socialization into schooling. We don't have to give an individual teacher a score on these kinds of outcomes to have a useful evaluation system. Understanding an association may be sufficient for seeing broader value in this factor we called "effective."

The second factor we identified as "constructive." We can speculate that this constructivism can have educational value beyond the test students take in spring. It may lead to a deeper understanding and seeking greater intellectual challenge that pays off in future educational choices such as, for example, following a STEM career. We don't know at this point that these

classroom practices have any positive effect or whether they have differential (positive or negative) impact depending on the student characteristic.

We saw that there was a third factor, which we labelled "positive" and is associated with what we might characterize as empathy, at least as perceived by students. This factor may increase a student's sense of belonging and value. Belonging uncertainty has been shown to increase student stress and undermine student engagement and motivation over time (Walton & Cohen, 2007) and we might speculate that it could have longer-term impacts on positive behavior, attendance, and staying in school. The student survey was the source of this third factor and the value of the surveys, beyond measuring classroom management that is associated with the "effective" factor (the "control" C), will depend on associating it with outcomes of value.

We pointed out that available data do not include observations made outside of classroom, as well as any kind of peer or stakeholder survey, which are also often elements of teacher evaluations. This makes this work illustrative but limited. It will be important to understand, for example, whether the "positive" factor is also associated with effective relationships with parents and peers. Alternatively, there may be a fourth factor associated with peer interactions and leadership. Exactly how these three or four factors will be weighted in the composite teacher scores depends ultimately on values that the stakeholders place on different types of outcomes. These weighting decisions however cannot be made without understanding what factors are actually measured by the multiple instruments of the evaluation systems and their components. Our analysis is an exploration of what appears to have been measured in the data compiled by the MET project and illustrates what can be done using data from any of the large-scale teacher evaluations currently being undertaken state and local education agencies.

3. How can evaluations be used in personnel decisions?

Turning to the question of how our measurements can be used in human capital considerations, we should note that creating a single composite teacher effectiveness score by weighting then adding up the multiple measures may not be an adequate approach to evaluation. Observations do not measure a single characteristic, and student surveys measure at least two quite distinct characteristics. Some items of both instruments contain information pertinent to test performance and some point to aspects of teaching that go beyond test outcomes. Observations and surveys about the teachers' abilities outside the classroom, not included in this study, may anchor an additional characteristic.

At this point, we can only speculate about the potential impact of characteristics beyond the factor we labeled as "effective." The approach that is being called for by the ED through its waivers as well as the rankings we see promoted by the National Council on Teacher Quality, centers teacher evaluations on student achievement. But we are seeing that the growth measured in the spring test plays a limited role in the set of characteristics emerging from a typical set of multiple measures. And the notion that these diverse characteristics are most usefully collapsed into a single aggregate score is questionable. It is questionable first because we don't know enough about the educational value of the underlying characteristics being measured. Second, it is not clear which characteristics (in what combination) should be considered in different kinds of decisions. At the very least, we should make an effort to

discover and explain the multiple facets of teaching and let the stakeholders make informed decisions.

What are the decisions in which teacher evaluations should be used? The usual suspects are raises, bonuses, layoffs, and terminations. And for these a "score" that rank orders all district teachers on a single scale can be useful. Weighting characteristics to favor the spring test scores is not unreasonable. But the same ranking would not be as useful in deciding which teacher to promote to vice principal or which to promote to math specialist. A leadership position may call for greater emphasis on empathy and consensus building. A specialist position may call for mastery of constructive techniques associated with the particular discipline.

We are beginning to be able to identify the set of characteristics that make up teaching. We can see that many of the characteristics highly valued by observational and survey measures are not major contributing factors to the spring test scores. They may nevertheless have great value to students longer term success as well as the success of the school as an organization, and measures of these characteristics may have great value in personnel decisions.

References

Cantrell, S., & Kane, T. J. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET Project's three-year study* (policy and practice brief). Seattle, WA: Bill & Melinda Gates Foundation.

Danielson Group. (2011). *The Framework for Teaching*. Retrieved April 22, 2013, from http://www.danielsongroup.org/article.aspx?page=frameworkforteaching

Darling-Hammond, L. (2012). *Creating a Comprehensive System for Evaluating and Supporting Effective Teaching*. Stanford, CA: Stanford Center for Opportunity Policy in Education.

Grissom, J. A., Loeb, S., & Master, B. (2013). Effective instructional time use for school leaders: Longitudinal evidence from observations of principals. *Educational Researcher, 42*(8), 433–444.

Hanushek, E., & Rivkin, S. (2010). Generalizations about the use of value-added measures of teacher quality. *American Economic Review, 100*, 267–271.

Kane, T., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains* (research report). Seattle, WA: Bill & Melinda Gates Foundation.

Lazarev, V., & Newman, D. (2013, September). *How non-linearity and grade-level differences complicate the validation of observation protocols*. Paper presented at the Fall 2013 Society for Research on Educational Effectiveness conference, Washington, DC.

Lazarev, V., Newman, D., & Grossman, P. (2013). *Developing an aggregate metric of teaching practice for use in mediator analysis.* Presentation at the Society for Research on Educational Effectiveness Spring 2013 Conference, Washington, DC. Retrieved April 22, 2013, from https://www.sree.org/download/files/5F_Lazarev_857.pdf

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

Rothstein, J., & Mathis, W. J. (2013). *Review of "Have We Identified Effective Teachers?" and "A Composite Estimator of Effective Teaching": Culminating findings from the Measures of Effective Teaching Project*. Boulder, CO: National Education Policy Center.

Walton, G. M., & Cohen, G. L. (2007). A question of belonging: race, social fit, and achievement. *Journal of Personality and Social Psychology, 92*(1), 82.

Appendix A. Results of Factor Analysis: Factor Loadings and Explained Variance

TABLE 1. RESULTS OF FACTOR ANALYSIS: FACTOR LOADINGS AND EXPLAINED VARIANCE

Measurement	Factor1	Factor2	Factor3
VAM	0.23		
Observation items – FFT			
FFT_CERR.Respect	0.49	-0.15	0.35
FFT_UQDT.Discussion	0.32		0.45
FFT_ECL.Culture.of.Learning	0.43	-0.14	0.42
FFT_MCP.Procedures	0.43	-0.21	0.27
FFT_CS.Communicating	0.38	-0.12	0.40
FFT_MSB.Manage.Behavior	0.52	-0.19	0.18
FFT_ESL.Engaging	0.41	-0.13	0.38
FFT_UAI.Assessment	0.36		0.38
Observation items - CLASS			
CLASS_Positive_Climate	0.41		0.62
CLASS_Negative_Climate	-0.48	0.16	-0.25
CLASS_Teacher_Sensitivity	0.36	-0.12	0.66
CLASS_Regard_for_Student_Persp	0.31		0.65
CLASS_Behavior_Management	0.54	-0.15	0.28
CLASS_Productivity	0.45	-0.13	0.39
CLASS_Instructional_Learning	0.37	-0.15	0.72
CLASS_Content_Understanding	0.34	-0.18	0.68
CLASS_Analysis_and_Problem_Solv	0.32	-0.14	0.67
CLASS_Quality_of_Feedback	0.35		0.77
CLASS_Instructional_Dialogue	0.32		0.77
CLASS_Student_Engagement	0.45		0.61
Student Survey Items ("7Cs")			
S_M_A10.CARE	0.45	0.72	0.21
S_M_B146.CARE	0.29	0.73	0.25
S_M_B34.CARE	0.46	0.71	0.23
S_M_B1.CLARIFY	0.44	0.75	0.18

TABLE 1. RESULTS OF FACTOR ANALYSIS: FACTOR LOADINGS AND EXPLAINED VARIANCE

Measurement	Factor1	Factor2	Factor3
S_M_B130.CLARIFY	0.47	0.68	0.15
S_M_B136.CLARIFY	-0.51	-0.35	
S_M_B17.CLARIFY	0.51	0.75	0.17
S_M_B80.CLARIFY	0.48	0.74	0.16
S_M_B112.CONTROL	0.86	0.12	-0.21
S_M_B113.CONTROL	0.76	-0.22	-0.30
S_M_B114.CONTROL	0.81		-0.23
S_M_B138.CONTROL	0.86		-0.29
S_M_B46.CONTROL	0.88	0.19	-0.18
S_M_B49.CONTROL	0.88	0.24	-0.13
S_M_B6.CONTROL	0.77	0.18	-0.13
S_M_B128.CHALLENGE	0.39	0.61	0.16
S_M_B133.CHALLENGE	0.37	0.52	0.17
S_M_B21.CHALLENGE	0.50	0.53	0.15
S_M_B36.CHALLENGE	0.59	0.55	0.16
S_M_B45.CHALLENGE	0.49	0.58	0.16
S_M_B59.CHALLENGE	0.42	0.55	0.20
S_M_B70.CHALLENGE	0.56	0.51	0.12
S_M_B90.CHALLENGE	0.54	0.64	0.16
S_M_B141.CAPTIVATE	0.47	0.57	0.13
S_M_B29.CAPTIVATE	0.52	0.71	0.18
S_M_B44.CAPTIVATE	0.49	0.73	0.20
S_M_B89.CAPTIVATE	0.55	0.67	0.16
S_M_B129.CONFER	0.51	0.49	0.19
S_M_B135.CONFER	0.39	0.47	0.12
S_M_B154.CONFER	0.40	0.75	0.25
S_M_B155.CONFER	0.46	0.61	0.20
S_M_A54.CONFER	0.50	0.70	0.20
S_M_B145.CONSOLIDATE	0.32	0.72	0.19
S_M_B147.CONSOLIDATE	0.46	0.74	0.18
S_M_B58.CONSOLIDATE	0.43	0.70	0.20

TABLE 1. RESULTS OF FACTOR ANALYSIS: FACTOR LOADINGS AND EXPLAINED VARIANCE

Measurement	**Factor1**	**Factor2**	**Factor3**
S_M_B83.CONSOLIDATE	0.43	0.71	0.20
Proportion of Total Variance	0.26	0.22	0.12
Cumulative Proportion of Total Variance	0.26	0.48	0.60

Reference for this paper: Lazarev, V., & Newman, D. (2014, March). *Can Multifactor Models of Teaching Improve Teacher Evaluation Systems?* Paper presented at the annual meeting for The Association for Education Finance and Policy, San Antonio, TX.