The test matters: The relationship between classroom observation scores and teacher value-added on multiple types of assessment

With the advent of *Race to the Top* and other policy initiatives, states and districts are literally racing to develop teacher evaluation systems that use multiple measures of teaching quality. Most of these systems include both some form of standardized classroom observations as well as value-added measures of student achievement gains. As states decide how to weight these different forms of evidence to create an overall composite evaluation of teachers, researchers have investigated the relationships among multiple indicators of teaching quality, generally finding modest to relatively low correlations between observational data and value-added measures (Bell et al., 2012; Identifying Reference, 2013; Strong, Gargani, & Hacifazlioğlu, 2011).

In some ways, these low correlations are not surprising, as student test scores and classroom observations represent quite different ways of measuring teaching. In fact, the rationale for including different measures in a composite evaluation is precisely because multiple measures are designed to capture a wider range of information about teaching. However, the low correlations may also reflect a lack of alignment between the goals of particular teaching practices captured by observation protocols and the kinds of student outcomes measured by many standardized tests. Different student assessments may be differentially sensitive to the more ambitious, cognitively demanding forms of teaching measured in classroom observation instruments (Popham, 2007; Yoon & Resnick, 2007).

The lack of a strong relationship between different measures of teacher effectiveness creates dilemmas for teachers and districts and a puzzle for the field. Districts who have invested in teacher evaluation systems with multiple measures face the challenge of helping teachers improve, when the measures may point in different directions. A student assessment of more basic skills may reward teachers who reduce the complexity of a task and coach students on multiple-choice responses; such a teacher is unlikely, however, to score well on an observation protocol that values intellectual challenge. Researchers, in turn, puzzle over the fact that the mechanism they expected to help explain student achievement—the quality of instruction—is so modestly related to the outcome. In any system of teacher evaluation, policymakers will want to consider the alignment between their theories of instruction, measures of teaching practice, and the outcomes measured by standardized tests.

To explore these issues, we examine how the relationships between one observation protocol, the Protocol for Language Arts Teaching Observation (PLATO), and value-added measures shift when different tests are used to assess student learning. We use a unique dataset to investigate whether the nature of the assessment influences the direction and strength of the association between observation scores and value-added scores. The Gates Foundation's Measures of Effective Teaching (MET) project assessed the teaching practices and the student achievement gains of thousands of teachers from six different districts located in different states. The database includes scores for students of MET teachers on both required state tests as well as supplemental assessments, including the SAT-9 open-ended in English language arts. The supplemental assessments were chosen to assess more cognitively complex learning outcomes, such as those being prioritized by the Common Core State Standards. While the state tests vary in content coverage and cognitive demand (Polikoff & Porter, 2012), several consist entirely of multiple-choice questions, where students are not asked to generate any written text. In contrast, the SAT-9 requires constructed responses, providing the potential for a different perspective on student achievement. In our analyses, we explore the following questions:

1) How does the overall relationship between scores of teaching practices and teacher value-added measures (VAM) differ depending upon the assessment used to construct the VAM?
2) What are the relationships between specific instructional elements and teacher value-added scores on different student assessments?

Background

Stability of Value-Added Across Assessment Type

Recent research has begun to investigate value-added teacher effects that were estimated using different achievement tests. Such studies have found moderate to low correlations between teachers' value added scores based on different student assessments, suggesting that a given teacher might be quite effective at improving student learning captured by one kind of test, but not necessarily effective given a different outcome measure. Lockwood, McCaffrey, Hamilton, Stecher, Le, and Martinez (2007), for example, found very low correlations among teacher effect coefficients when two different subscales of the Stanford Achievement Test (SAT-9) mathematics assessment—procedures and problem solving-- were used to compute teacher value-added. Particularly striking is their finding that the within-teacher variation in value-added across outcome measures is greater than the variation across teachers in their sample. While their data does not allow for analysis of classroom mechanisms that may or may explain these differential teacher effects, Lockwood et al. hypothesize that instruction may play a critical, even causal, role in these differences.

Papay (2011) performed a similar analysis with three different reading assessments, a high stakes state test along with the low stakes SAT-9 and Scholastic Reading Inventory (SRI), and found similarly low correlations in teacher value-added based on outcome measure. Both Lockwood et al. and Papay found that selection of achievement measure contributes substantially more to variation in teacher value-added coefficients then the specification of the value-added model, or the specific population of teachers against whom a teacher is compared. Harris and Sass (2009) found different relationships between National Board For Professional Teaching Standards and value-added coefficients in Florida, depending on the assessment used and whether or not the assessment was used for high stakes evaluation purposes.

These studies suggest the importance of attending to the features of assessments used in VAMs and analyzing the resulting differential relationships with other measures of teaching quality, including classroom observations. Little research to date has documented whether the variation in value-added based on assessment type is associated with differential relationships with specific classroom practices. The MET data, which include two different assessments, provide an initial opportunity to explore this issue.

PLATO's Theory of Instruction

While value-added measures imply that the hallmark of good teaching consists solely of gains in student achievement as measured by standardized assessments, classroom observation protocols often start with a theory of effective instruction that promotes the development of rich, conceptual understanding of subject matter (Seidel &

Shavelson, 2007). For example, Charlotte Danielson's Framework for Teaching (2007), a widely used observation measure, was built on the principles of constructivism that value student meaning making and active engagement. The Mathematical Quality of Instruction (MQI), developed by Heather Hill and her colleagues (2008), grew out of their work on the importance of teachers' mathematical knowledge for teaching and the value of precision in mathematical explanations.

In this article, we use the Protocol for Language Arts Teaching Observation (PLATO), one of the two subject-specific observation protocols used in the MET study, as one example of an observation protocol to investigate the larger issue of instructional sensitivity of different types of student assessments. PLATO was developed from research on high quality teaching in English Language Arts (ELA). The theory underlying PLATO articulates the importance of rigorous content and intellectually challenging tasks, the centrality of classroom discourse in developing sophisticated understanding of content and disciplinary skills, and the critical role of teachers in providing instructional scaffolding for students to help them succeed (see Identifying Reference, 2013 for a fuller description of the PLATO protocol).

There are four underlying factors in the full PLATO instrument, three of which were represented in the PLATO Prime instrument used for MET: instructional scaffolding, cognitive and disciplinary demand of classroom talk and tasks, and classroom environment. Figure 1 describes the structure of the PLATO Prime instrument and the elements used to code instruction.

The instructional scaffolding factor includes the PLATO elements of Modeling and Strategy Use and Instruction. Modeling measures whether a teacher visibly enacts a process, skill, or strategy as a guide for student work. At the low end, the teacher does not model or does so inaccurately or incompletely. At the high end, she accurately and completely models and decomposes or highlights specific features of the process in which students will engage. The element of Strategy Use and Instruction focuses on the teacher's use of flexible methods or approaches to skills and processes in reading, writing, or other areas of ELA. At the low end, the teacher refers to or prompts strategies without instruction. At the high end, the teacher provides detailed instruction including how, when, and why a strategy could be used for reading or writing.

The factor of Disciplinary Demand includes the PLATO elements of Intellectual Challenge and Classroom Discourse. The element of Intellectual Challenge focuses on whether teacher-provided activities or questions focus primarily on rote recall, at the low end, or promote high-level, analytic thinking at the high end. The element of Classroom Discourse focuses on opportunities for student talk and uptake of student ideas within those opportunities. At the low end, teachers or students do not have opportunities to engage in ELA-related talk or the talk is tightly teacher-directed with brief responses. At the high end, students have extended opportunity to engage in ELA-related discussion that includes explanation, clarification, or other high-level uptake of student ideas.

The factor of Classroom Environment includes the elements of Time Management and Behavior Management. Time Management focuses on the amount of time students are engaged in ELA-focused activity. At the low end, time may be lost due to disorganized transitions or materials. At the high end, instructional time is maximized. The element of Behavior Management focuses on the degree to which student behavior is appropriate to the task at hand and facilitates learning. At the low end, the classroom is disorderly and student behavior is a major impediment to learning. At the high end, the classroom is orderly and there are almost no instances of behavioral disruption that distract from student learning.

Prior research in ELA suggests that the practices highlighted in PLATO are designed to support students' growth as readers and writers. The research on which PLATO was built, for example, suggests that teachers who are more skilled at elaborating and drawing out student contributions during classroom discussions can support students' achievement as measured by writing assessments (e.g. Nystrand, 1997). Students who have teachers that provide time for discussion, elicit student ideas, and probe student thinking have opportunities to develop their ideas and arguments during classroom discussions (Nystrand, 1997; Nystrand & Gamoran, 1991; O'Connor & Michaels, 1993; Taylor et al., 2003; Tharp & Gallimore, 1991). However, many standardized assessments, particularly those featuring only multiple-choice questions of basic comprehension skills, may not be sensitive to the kinds of learning that result from such rich discussions. Nystrand and his colleagues (1997) for example, used student essays rather than multiple-choice tests to measure achievement gains, as argumentation skills and the ability to develop ideas are likely to be better captured on performance assessments in which students are required to write extended prose.

Similar to other observation protocols, PLATO also privileges instruction that is intellectually challenging and provides students with opportunities to engage in inferential thinking, interpretation, and meaning-making. Such opportunities support students to develop their own interpretations of complex text and to write critically. However, assessments vary in how well they capture interpretation and analysis; some state tests, for example may require students to engage in less challenging tasks, such as recall of vocabulary or grammar rules, or multiple choice questions that ask students for basic comprehension of information. For these reasons, it is plausible that some of the instructional practices highlighted in this instrument, and others like it, may be differentially aligned with different student outcome measures.

Data and Methods

Data

The analysis in this report uses data from the *Measures of Effective Teaching* study conducted by the Bill and Melinda Gates Foundation (Kane & Staiger, 2012). The larger sample from which this analysis was drawn is comprised of approximately 1,300 fourth through eighth grade mathematics and ELA teachers from six districts-- Charlotte-Mecklenberg, Dallas, Denver, Hillsborough County, New York City, and Memphis--who volunteered to have multiple lessons video recorded and analyzed using multiple classroom observation protocols, including PLATO. Teachers who volunteered for the study were similar to their peers in terms of their demographics, education level, and years of experience (Kane & Staiger, 2012). For this paper, we sampled all 4th through 8th grade ELA teachers for whom we had both outcome measures, resulting in 893 teachers and 3,362 video observations.

PLATO Scores. For the study, each teacher was videotaped on four occasions, across more than one class of students when possible. The videos were then coded by approximately 90 raters hired by the Educational Testing Service (ETS) specifically for PLATO; the vast majority (98 percent) of raters were current or former teachers with three or more years of experience (Kane & Staiger, 2012). All raters were trained in an online system created by the PLATO developers and administered by ETS. To become certified, raters had to achieve a 70 percent exact match with master-scored video segments. The MET study also included numerous procedural safeguards to ensure reliable scoring. In particular, raters were required to score frequent “calibration videos” to address potential rater drift. Throughout the study, scoring experts at ETS supported raters by helping answer questions and reconciling discrepant scores.

PLATO requires raters to score 15-minute segments of instruction. Raters assigned a score for each element on a scale of one (low) to four (high) for each segment. Scores are based on both the quality and quantity of the instruction. To analyze teacher practices, we aggregated scores across instructional segments to the teacher level, on the assumption that there is natural variation of instructional practices both within a lesson cycle and across lessons. These teacher level averages on the focal teaching practices were then standardized across the sample.

We used factor analysis on the six PLATO elements to examine their underlying structure and to identify latent constructs. Summarized in Appendix Table 1, our analyses revealed three underlying factors that were consistent with the instructional domains that PLATO had been designed to signal[1] (see Figure 1): 1) The Instructional Scaffolding Factor that primarily loaded on the modeling and strategy instruction elements, 2) The Disciplinary Demand Factor that loaded mostly on the intellectual challenge and classroom discourse elements, and 3) The Classroom Environment Factor that primarily loaded on the time management and behavior management elements.

---

[1] Only one component met the Kaiser criterion of having an initial eigenvalue of over 1.0. However, as illustrated in Appendix Figures 1, both Scree Plot and Horn’s Parallel Analyses indicated that three components should be retained. After rotating these three factors (using both orthoganol and oblique methods) for comprehensibility, they loaded on the PLATO elements in ways that reflected the instructional factors intended by the designers of PLATO, thus indicating conceptual consistency as well.

Value-Added Scores. Table 1 also summarizes the two value-added measures (VAM) we used for our analyses – one based upon the state assessments and the other based upon the SAT9 assessment. Because the VAM scores were standardized against the full sample of teachers, mean scores for our subsample are close to, but not exactly, zero. Following Kane and Staiger (2012), we convert value-added scores into months of schooling gained using a conversion factor that assumes nine months of learning to be equivalent to 0.25 standard deviations. There are a number of advantages to this approach: 1) coefficients are more interpretable; 2) comparisons with results from the larger MET study (Kane & Staiger, 2012) can be made directly; and 3) the same conversion can be applied across analyses in this paper[2]. A limitation of this approach is that it applies the same conversion across value-added measures even though each was estimated in student-level standard deviation units within grade and subject area for each test.[3] It is important that the reader bear in mind that "months of learning" units are relative units.

Data Analysis

The main purpose of this analysis is to examine the relationships between PLATO and two different ELA value-added measures—one based upon state tests and another based on SAT-9. We begin by generating lowess curves to provide a graphical representation for how these different teacher quality measures are related. First, we converted percentile ranks on average PLATO scores to Normal Curve Equivalents

[2] Values in units of "months of learning" can be divided by 36 in order to convert back to student-level standard deviations. This approach is consistent with that used in the MET report (Kane & Staiger, 2012)

[3] Although it might have been more appropriate to have used a separate conversion that reflected the tests and contexts specific to each measure, we did not have access to this information.

(NCE) to have an equal-interval scale[4]. Lowess curves use non-parametric, locally-weighted regression models to estimate teachers' value-added as a function of teachers' PLATO percentile rankings. Due to concerns that district and grade level differences might explain some of these observed relationships, we also use multiple linear regression models to account for these factors.

Our multiple linear regression models estimate teachers' value-added in ELA as a function of their PLATO scores, and take the general form:

$$VAM_{tdg} = \beta_1 (PLATO_{tdg}) + \delta_d + \Phi_g + \varepsilon_{tdg} \qquad [\textit{Equation 1}]$$

Here, the value-added (*VAM* based upon either state or SAT9 assessments) of teacher *t* in district *d* in grade *g* is estimated as a function of her *PLATO* score, an indicator (dummy) for her district (δ_d), an indicator (dummy) for her grade level (Φ_g), and an error term (ε_{tdg}). Each district used a different state assessment so we felt it was inappropriate to compare the value-added of teachers across districts. Thus, including district fixed effects (dummies) allows us to compare teachers only to peers within the same district. In addition, both VAM and PLATO scores vary by grade level[5]. Therefore, we include grade level fixed effects to compare teachers only to peers within the same grade level, ensuring that observed relationships between VAM and PLATO are not explained by grade level differences. In separate models, we also included a vector of school characteristics as controls[6]. Finally, we also reproduced our models but replaced PLATO

[4] Differences between the lowess curves in this analysis and those in the MET report result from our use of NCEs rather than percentiles (Kane and Staiger, 2012).

[5] For example, we find teachers' value added to be greater at lower grade levels. This phenomenon is likely due to the fact that VAM scores were standardized within grade level. Though it may be the case that teachers at lower grade levels are indeed more effective, we thought it was just as likely that it is easier for teachers to move (increase) achievement when students are younger.

[6] Due to page limit requirements, we do not report on these findings; results were similar and are available upon request from the authors. We decided not to focus on models with school controls for several reasons. First, about 15 percent of teachers had missing information on schools. Second, these same school

average scores with the three different PLATO factors (Disciplinary Demand, Instructional Scaffolding, and Classroom Environment) so we could test whether VAM scores are more sensitive to different domains of instruction as measured by PLATO.[7]

Findings

We found that the relationship between teachers' PLATO and VAM scores does vary depending upon the outcome measure used to construct value-added scores. We began our analysis by comparing correlation coefficients between PLATO and each of the VAM scores. Correlations were generally small, but PLATO was more highly correlated with SAT9 VAM scores (r=0.16) than with state VAM scores (r=0.09). Using a method for "comparing correlated correlation coefficients" developed by Meng, Rosenthal, & Rubin (1992), we found differences between these coefficients to be statistically significant (z=1.71, p=0.04). These results provided initial evidence that SAT9 VAMs may be more sensitive to the kinds of instruction measured by PLATO.

Figure 2 plots separately the relationships between PLATO and SAT9 VAM (blue) and between PLATO and state VAM (red) using locally weighted scatterplot smoothing (lowess curves). Both SAT9 and state VAMs increase with increasing PLATO scores, reflecting the positive correlation coefficients described above. However, these plots illustrate that differences exist, especially in the bottom half of the distribution. Among teachers with the lowest observational scores, SAT9 VAMs appear to be much more responsive to PLATO performance than are state VAMs. The difference between a 50^{th} and a 10^{th} percentile teacher on PLATO is estimated to be 10 months of learning using SAT9 VAM, but only 4 months using state VAM. This may suggest that PLATO is

characteristics were included in the generation of the value-added scores on the left hand side of the equation; thus, we were worried we might bias estimates by re-introducing these controls into our models.

more effective at distinguishing teacher effectiveness on SAT9, at least among individuals with low observational ratings. These results also indicate that the differences between correlation coefficients identified above are primarily among lower performing teachers.

To ensure that these relationships were not explained by district or grade level differences related to VAM construction, we used multiple regression to estimate teachers' VAM scores as a function of PLATO. We included fixed effects for district and grade level to compare teachers only to peers in the same district and grade level. The regression coefficient for the relationship between PLATO and SAT9 VAM was B=1.95 ($p<0.001$); the coefficient on state VAM was B=0.81 ($p<0.001$). These results indicate that teachers who are rated one standard deviation higher on PLATO produce, on average, an additional two months of learning on the SAT9 VAM as compared to one month of learning on the state VAM.

Because different VAM scores were constructed using different assessment measures and are standardized in different districts and grade levels, we do not believe the VAM scores, nor their corresponding regression coefficients, to be fully comparable.[8] Even so, these results provide further suggestive evidence that SAT-9 VAMs may be more sensitive than state VAMs to the kinds of instruction that PLATO measures. That trends were similar using other analysis methods—lowess curves and correlation comparisons—strengthens the case that real differences exist.

[8] In separate analyses, we decided to assume SAT-9 VAM and state VAM scores to be fully comparable. We used two approaches to test whether the coefficients on PLATO were significantly different between the SAT-9 and state VAM analyses: 1) Saving coefficients from our separate OLS regression and used Seemingly Unrelated Estimation with standard errors clustered at the teacher-level and 2) Using Seemingly Unrelated Regression to estimate the equations jointly. Both methods suggested that the PLATO coefficients on SAT-9 VAM were statistically greater than they were on state VAM.

As described in the introduction, PLATO intends to measure ambitious forms of instruction that emphasize, among other things, developing student ideas and arguments through intellectually challenging discussions. There is good reason then to believe that performance assessments designed to evaluate argumentation skills and the ability to develop ideas would do a better job of capturing the kinds of instruction that PLATO privileges as compared with assessments that tend to prioritize comprehension skills and multiple-choice questions. Thus, it seems reasonable that SAT-9 VAMs would have stronger relationships with PLATO scores than would state VAMs. We might also expect observational ratings on the particular instructional domains that target ambitious forms of instruction to be especially sensitive to which student outcome measure is used in the VAM. To test this, we examined the relationships between our two VAM measures and the different domains of instruction measured by PLATO.

Differences by Instructional Practice

Are SAT-9 VAMs and state VAMs sensitive to similar features of instruction as signaled by PLATO? Table 1 summarizes correlation coefficients of the relationships between both VAM scores and three PLATO factors measuring different domains of instruction. The results indicate that, across domains, SAT-9 VAM is more highly correlated with PLATO than state VAM. However, the only statistically significant difference is in the domain of Disciplinary Demand. As compared with state VAMs, SAT-9 VAMs have a significantly stronger relationship with PLATO scores on Disciplinary Demand; the two VAMs have statistically similar relationships with the two other PLATO domains.

Results from multiple linear regression models were consistent. Table 2 summarizes the results from regression models estimating VAM scores as a function of different PLATO factor scores; we include coefficients on PLATO average as well for comparison. Results in the columns on the left (Models 1 and 2) are from models using SAT-9 VAM as the outcome measure; results on the right (Models 3 and 4) are from models using state VAM as the outcome.

The findings indicate that teachers who excel in classroom management have better value-added regardless of which student assessment is used to estimate value-added scores. These results are consistent with prior evidence that teachers with more facility in classroom management have better student achievement gains (Kane et al., 2010). Although both VAM scores are significantly related to Classroom Environment, only the SAT-9 VAM is significantly related to Disciplinary Demand. This difference suggests that PLATO domains designed specifically to identify ambitious instructional practices are especially sensitive to which test is used to construct value-added scores. Assuming that the SAT-9 exam is better than district exams at assessing ambitious student outcomes, such as the capacity for critical thinking and reasoning, then it makes sense that teachers who teach in ways intending to promote these ambitious outcomes—those who excel on practices related to the Disciplinary Demand of instruction—would be better at raising student achievement on the SAT-9. [9]

[9] In separate analyses, we assumed SAT-9 VAM and state VAM scores to be fully comparable and tested to see whether the coefficients on each of the PLATO factors differed by VAM score. We used two approaches to test whether the coefficients on PLATO were significantly different between the SAT-9 and state VAM analyses: 1) Saving coefficients from our separate OLS regressions and used Seemingly Unrelated Estimation with standard errors clustered at the teacher-level and 2) Using Seemingly Unrelated Regression to estimate the equations jointly. Both methods indicated that the coefficients for Disciplinary Demand were significantly greater on SAT-9 VAM than on state VAM. Coefficients for Classroom Environment and for Instructional Scaffolding were also greater on SAT-9 VAM, but differences were not statistically significant at the $p<0.05$ level.

Discussion

This study suggests that researchers and policymakers need to pay careful attention to the assessments used to measure student achievement in designing teacher evaluation systems. Because classroom observation protocols used in the MET study had only modest relationships with value-added estimates on state assessments, it would be easy to dismiss the instruments or the practices they capture as less relevant for assessing teacher quality. However, a different picture emerges, for example, when comparing the relationship between PLATO scores and achievement gains on the SAT-9, a different type of assessment measuring different kinds of skills.

Our analysis suggests that classroom observation protocols may measure teaching practices that are differentially related to outcomes on standardized tests. The test used to measure outcomes really does matter in looking at the relationship between what teachers do in classrooms and what students learn. While these differences are likely consequential in terms of the relationship with teachers' classroom practice, little research to date has explored these relationships between outcome measures and classroom observations. However, our results suggest that if teachers are engaging in instructional practices designed to support students in generating text and ideas, this kind of instruction might not be captured on assessments that focus on less cognitively complex student outcomes.

As districts work to improve the alignment of different measures used in teacher evaluation systems, our results suggest that they think seriously about the student outcomes they value and how those outcomes are measured. These decisions can have serious consequences, especially when the student assessment becomes the standard against which all other evaluations are measured. Choosing a student assessment with low cognitive demand, for example, may mean that teachers who are enacting ambitious instruction in ways that support students to generate knowledge and high levels of analytic thinking will not be recognized for these contributions.

As we move towards the Common Core Standards that prioritize reasoning, argumentation, and generation of text, we need to both develop assessments that focus on these skills and then identify the teaching practices that move students towards meeting more ambitious standards. Perhaps the best indicator of whether or not students are learning to engage in rigorous academic discourse that prepares them for college classrooms, as identified in the Common Core State Standards, are actually classroom observations that capture the extent to which students are actually participating in such discussions. If we are not certain that our high-stakes assessments are accurately capturing the kind of learning we value, there may be good reason to include multiple sources of information about teaching quality that are sensitive to different facets of teaching and learning.

References

Bell, C.A., Gitomer, D.A., McCaffrey, D., Hamre, B., Pianta, R., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, *17*(2-3), 62-87.

Corcoran, S., Jennings, J.J., & Beveridge, A. (2011). Teacher effectiveness on high and low stakes test. *Unpublished manuscript.*

Danielson, C. (2007). *Enhancing professional practice: a framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.

Harris, D. N., & Sass, T. R. (2009). The effects of NBPTS-certified teachers on student achievement. Journal of Policy Analysis and Management, 28(1), 55–80.

Hill, Heather, Laura Kapitula, and Kristin Umland. 2011. "A Validity Argument Approach to Evaluating Teacher Value-Added Scores." *American Educational Research Journal* 49 (3): 794–831.

Kane, T.J. & Staiger, D.O. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Seattle, Washington: Bill and Melinda Gates Foundation.
Retrieved from:
http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf.

Kane, T.J., Taylor, E.S., Tyler, J.H., & Wooten, A.L. (2010). Identifying effective classroom practices using student achievement data. *Journal of Human Resources*, 46(3), 587-613.

Koretz, D. M. (2008). Limitations in the use of achievement tests as measures of educators'
productivity. The Journal of Human Resources, 37(4), 752–777.

La Paro, Karen M., Robert C. Pianta, and Megan Stuhlman. 2004. "The Classroom Assessment Scoring System: Findings from the Prekindergarten Year." *The Elementary School Journal* 104 (5): 409–426.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, 29, 4–1.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. Journal of Educational Measurement, 44(1), 47–67.

Nystrand, M. (1997). *Opening dialogue: Understanding the dynamics of language and learning in the English classroom*. New York: Teachers College Press.

Nystrand, M, & Gamoran, A. (1991). Instructional discourse, student engagement, and literature achievement. *Research in the Teaching of English, 25*, 261-290.

O'Connor, M.C. & Michaels, S. (1996). Shifting participant frameworks: Orchestrating thinking practices in group discussion. In D. Hicks (Ed.), *Discourse, Learning, and Schooling* (pp. 63- 103). New York: Cambridge University Press.

Papay, J. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcomes measures. *American Educational Research Journal, 48*(1), 163-193.

Polikoff, M. S. & Porter, A. C. (2012). *Surveys of Enacted Curriculum substudy of the Measures of Effective Teaching Project: Final report.* Seattle, WA: Bill and Melinda Gates Foundation.

Polikoff, M. S. (2010). Instructional sensitivity as a psychometric property of assessments. Educational Measurement: Issues and Practice, 29(4), 3-14.

Popham, J. W. (2007). Instructional insensitivity of tests: Accountability's dire drawback. Phi Delta Kappan, 89(2), 146–155.

Porter, A. C., Polikoff, M. S., Zeidner, T., & Smithson, J. (2008). The quality of content analyses of state student achievement tests and state content standards. Educational Measurement: Issues and Practice, 27(4), 2-14.

Seidel, T.S. & Shavelson, R. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454-499.

Shulman, L. (2009). Assessment of teaching or assessment for teaching? In D.H. Gitomer (Ed.), Measurement issues and assessment for teaching quality. Washington, DC: Sage Publications.

Strong, M., Gargani, J., & Hacifazlioğlu, Ö. (2011). Do we know a successful teacher when we see one? Experiments in the identification of effective teachers. *Journal of Teacher Education,62*(4), 1-16.

Taylor, Barbara M., P. David Pearson, Debra S. Peterson, and Michael C. Rodriguez. 2003. "Reading Growth in High-Poverty Classrooms: The Influence of Teacher Practices that Encourage Cognitive Engagement in Literacy Learning." *Elementary School Journal* 104 (1): 3–28.

Tharp, R.G., & Gallimore, R. (1991). *The instructional conversation: Teaching and learning in social activity*. Berkeley, CA: National Center for Research on Cultural Diversity and Second Language Learning.

Yoon, B., & Resnick, L. B. (1998). Instructional validity, opportunity to learn, and equity: New standards examinations for the California mathematics renaissance. Los Angeles, CA: Center for the Study of Evaluation.

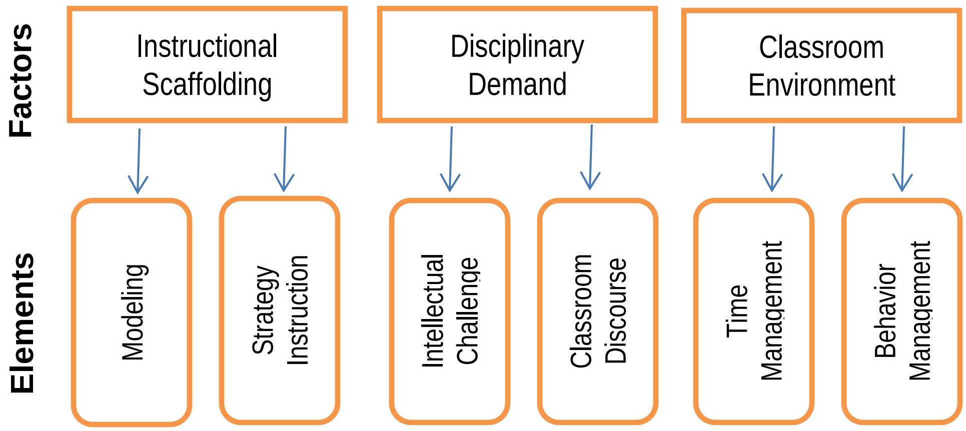


Figure 1- Structure of PLATO Observation Instrument

Figure 2: Teacher Value-Added (in months of learning units) as a Function of PLATO (all districts)

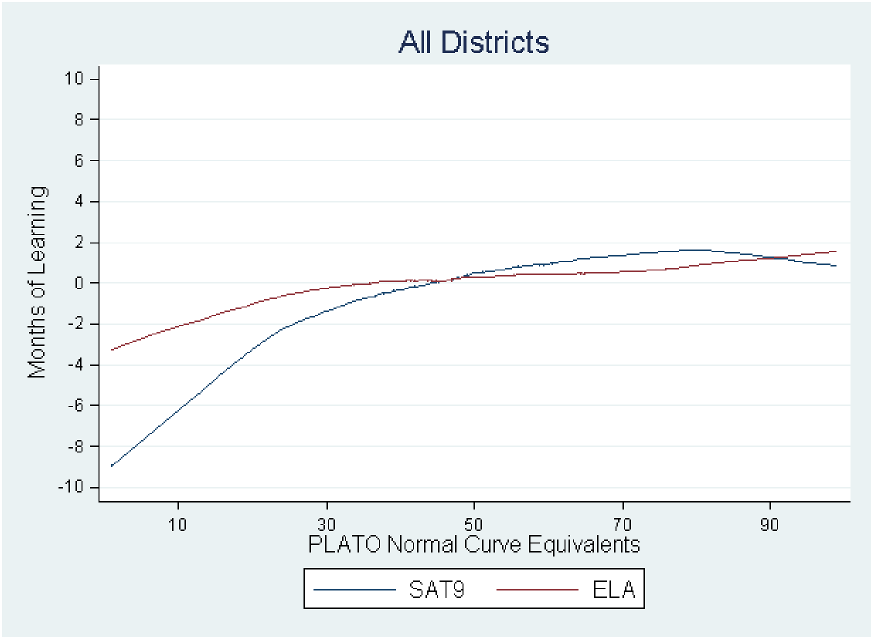


TABLE 1: Comparisons of Correlated Correlation Coefficients (all districts)

Measure	SAT9 VAM	State VAM	N	z-score	p-value
PLATO Average	0.156	0.089	893	1.715	0.043
Instructional Scaffolding	0.036	0.001	893	0.866	0.193
Disciplinary Demand	0.120	0.045	893	1.914	0.028
Classroom Environment	0.151	0.126	893	0.635	0.263

Z-scores were calculated using an approach developed by Meng, Rosenthal, & Rubin (1992) for comparing whether "correlated correlation coefficients" significantly differ from one another.

Table 2: Value-added as a function of PLATO (all districts)

	SAT-9 VAM		STATE ELA VAM	
	Model 1	***Model 2***	***Model 3***	***Model 4***
PLATO Average (standardized)	1.9498***		0.8054***	
	(0.434)		(0.230)	
Disciplinary Demand Factor		1.2870**		0.2427
		(0.488)		(0.264)
Classroom Environment Factor		2.1309***		1.1542***
		(0.511)		(0.265)
Instructional Scaffolding Factor		0.1604		0.1382
		(0.508)		(0.303)
Teachers	893	893	905	905
Grade Fixed Effects	x	x	x	x
District Fixed Effects	x	x	x	x

*** $p<0.001$, ** $p<0.01$, * $p<0.05$

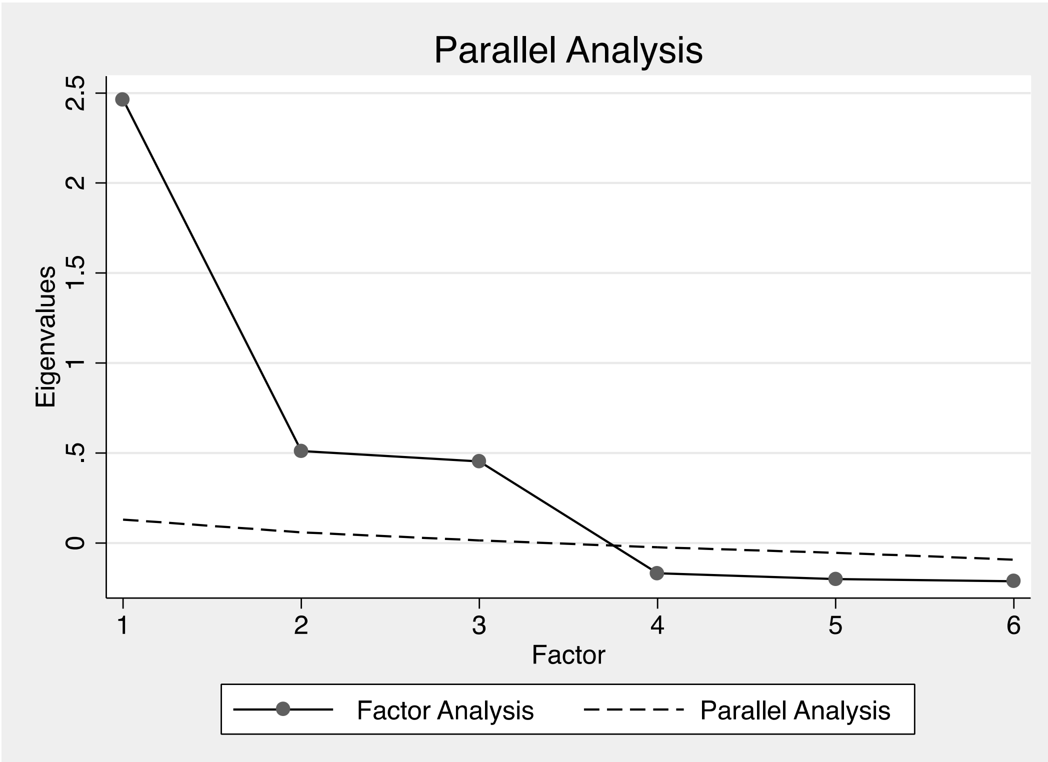
Appendices

Appendix Table 1: Factor loadings of PLATO instructional elements on unrotated and rotated factors

	ROTATED			UNROTATED		
PLATO Element	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 1**	**Factor 2**	**Factor 3**
Intellectual Challenge	0.7342	0.1907	0.1729	0.6867	-0.0446	-0.3631
Classroom Discourse	0.7289	0.1839	0.1390	0.6625	-0.0668	-0.3755
Behavior Management	0.2064	0.6834	0.0942	0.5581	-0.3599	0.2784
Modeling	0.1949	0.1661	0.6620	0.5505	0.4037	0.1948
Strategy Instruction	0.3281	0.1899	0.6588	0.6520	0.3749	0.1101
Time Management	0.3419	0.6839	0.2282	0.7165	-0.2665	0.2286

Note: The Disciplinary Demand, Classroom Environment, and Instructional Scaffolding Factors correspond with Factors 1, 2, and 3, respectively. Above rotated factors are from direct oblimin (oblique) rotation. We chose this standard, non-orthogonal approach because we assumed different domains of instruction would be correlated. We tried using verimax (orthogonal) rotation and results were similar.

Appendix Figure 1: Horn's Parallel Analysis



Note: Horn's Parallel Analysis suggests that all components above the dashed line be retained.