Measure for Measure: The Relationship between Measures of Instructional Practice in Middle School English Language Arts and Teachers' Value-Added Scores

PAM GROSSMAN
Stanford University

SUSANNA LOEB
Stanford University

JULIE COHEN
Stanford University

JAMES WYCKOFF
University of Virginia

Over the past 2 years, educational policy makers have focused much of their attention on issues related to teacher effectiveness. The Obama administration has made teacher evaluation and teacher quality a central feature of many of its educational policies, including Race to the Top (RTTT), Investing in Innovation (i3), and the Teacher Incentive Fund (TIF) grants. In response, many states and school districts are developing measures of teacher effectiveness to reward, tenure, support, and fire teachers. In response to these policies, many observers are raising questions and concerns about the measures of teacher effectiveness that inform high-stakes personnel decisions. Unfortunately, we have little systematic knowledge regarding the properties of most of these measures. This article has two goals: to explore elements of instruction that may be associated with improved student achievement and to examine the domains of teaching skills that are identified in the literature as important to high-quality teaching but that may not be highly correlated with value-added measures of teacher effectiveness.

Over the past several years, policy makers have focused much of their attention on issues related to teacher effectiveness. The Obama administration has made

Electronically published March 20, 2013

American Journal of Education 119 (May 2013)

0195-6744/2013/11903-0005$10.00

teacher evaluation and teacher quality a central feature of many of its educational policies, including Race to the Top (RTTT), Investing in Innovation (i3), and the Teacher Incentive Fund (TIF) grants. In response, many states and school districts are developing measures of teacher effectiveness to use to reward, tenure, support, and fire teachers. For example, the District of Columbia Public Schools have recently received nationwide attention for firing 241 teachers, most of whom received very low scores on its teacher evaluation system, IMPACT. This represents a stark departure from just a few years ago when few school districts employed systematic approaches to measuring teacher effectiveness to inform personnel decisions.

In response to these policies, many observers are raising questions and concerns about the measures of teacher effectiveness that inform high-stakes personnel decisions. These measures include both value-added measures and classroom observation instruments, which have been used in conjunction with more traditional administrator evaluations. Value-added measures are receiving particular scrutiny in light of their use in highly visible debates regarding the effectiveness of individual teachers. While both value-added models (see Rivkin et al. 2005; Rockoff 2004; Sanders and Rivers 1996) and several different teacher observation protocols (Danielson 2007; La Paro et al. 2004; Pianta et al. 2006) have been employed for some time, we know relatively little about the properties of either value-added measures or observation protocols in the context of making judgments about the quality of teaching (see Goe et al. 2008; Hanushek and Rivkin 2010). To what extent do these measures capture the multidimensional qualities of effective teachers? To what extent do they measure similar or different dimensions of teacher effectiveness? What do we know about their validity and reliability? Unfortunately, we have little

---

PAM GROSSMAN is a professor of education and the faculty director of the Center to Support Excellence in Teaching at Stanford University. Her research focuses on teacher education, professional development, and the teaching of English. SUSANNA LOEB is a professor of education at Stanford University, faculty director of the Center for Education Policy Analysis, and a codirector of Policy Analysis for California Education (PACE). Her research addresses teacher policy, school leadership, and school finance. JULIE COHEN is a doctoral candidate in curriculum and teacher education at Stanford University. Her research focuses on the features of high-quality teaching across content areas and explores ways to develop teachers' use of instructional practices associated with student achievement gains. JAMES WYKOFF is a professor of education and the director of the Center for Education Policy and Workforce Competitiveness at the University of Virginia. His research focuses on issues of teacher labor markets, including teacher preparation, recruitment, assessment, and retention.

systematic knowledge regarding the properties of most of these measures or the extent to which these different measures are related to one another.

This article has two goals: to identify elements of instruction that are associated with teachers' impact on student achievement, as measured by value-added measures, and to examine the domains of teaching practice that are identified in the literature as important to high-quality teaching but are not necessarily highly correlated with value-added measures of teacher effectiveness. We start with the assumption that teaching quality is a multidimensional quality that is unlikely to be fully captured by any single measure. In addition, if different measures capture different facets of teaching quality, they may not, in fact, be highly correlated with each other.

In this exploratory study, we use several different measures to investigate the classroom practices that differentiate teachers with greater and lesser impact on student achievement, particularly in higher-poverty schools. Our goal is to better understand the relationship among these measures in a small, but carefully constructed, sample. In addition, this study explores the use of value-added analyses as a possible measure of teaching effectiveness by investigating whether value-added measures in fact reflect observable differences in instruction. In other words, are the kinds of value-added analyses used in teacher evaluation a reasonable signal for differences in instructional quality and teaching?

Background and Conceptual Framework

Employing multiple measures of teacher effectiveness offers a number of potential advantages to the use of any single measure. It is widely recognized that effective teachers possess a range of characteristics and skills that contribute directly and indirectly to improved student outcomes. As a result, any single measure of effectiveness is likely to be limited in its ability to capture this diversity of characteristics. For this reason, among others, school districts that are beginning to implement more rigorous teacher evaluation systems are typically using multiple measures of teacher evaluation, as is the case in New York City and Washington, DC, and other Race to the Top states. Typically, these multiple measures include at least some measure of student achievement as well as direct observations of classroom practice.

Many teacher evaluation systems now rely on some version of value-added modeling to look at teachers' impact on student achievement. Remarkable progress in quantifying teachers' impact on student achievement through the use of value-added has been made over the past several years, and we now understand much more clearly both the strengths and weaknesses of value-added methods (see Harris 2009; Ishii and Rivkin 2009; McCaffrey et al.

2009). Proponents argue that these measures better differentiate among teachers than existing evaluation tools (Harris 2009; Rockoff 2004). Moreover, recent work by Raj Chetty and colleagues (2011) finds that exposure to high-value-added teachers is associated with desirable long-run outcomes like college attendance, higher earnings, and a reduced probability of teen pregnancy. While this research suggests that value-added may be capturing meaningful differences among teachers that are associated with important outcomes, there are also well-documented problems with using value-added to evaluate individual teachers. The following are among the more troublesome issues (for a more detailed description of issues associated with value-added, see Hanushek and Rivkin [2010]):

- While tested outcomes are important, most would agree that effective teachers do more than just improve outcomes as measured on standardized achievement tests. Some aspects of good teaching, such as the ability to improve the quality of students' writing or to help students engage in rigorous academic discussions of complex texts, both important outcomes for English language arts (ELA), may not be captured by multiple-choice measures typical of standardized tests.
- Value-added estimates provide a ranking of teacher effectiveness through a methodology that is opaque to teachers and typically provide little guidance to teachers for how they can improve.
- Value-added estimates are unstable when based on relatively small numbers of students, thus requiring several classes of students to reduce measurement error.
- Value-added measures can be estimated for just those teachers in tested grades and subjects, typically mathematics and reading in grades 4–8.
- Empirically isolating the effect of individual teachers from other attributes of the school context, many of which are difficult to measure, is very complex.

For these reasons, among others, school districts have been eager to include classroom observations in efforts to measure teaching effectiveness. Teacher observation protocols have existed for some time, but only recently have they been considered in high-stakes teacher evaluations. These protocols, which address many of the issues raised above, can be a rigorous alternative or a complement to value-added measures. However, there is limited empirical evidence on whether classroom observations, when well implemented, can make valid distinctions among more and less effective teachers. Some studies suggest that value-added measures and observation protocols may be identifying similar groups of effective teachers. For example, Robert Pianta and his colleagues have looked at the relationship between teachers' scores on the Classroom Assessment Scoring System (CLASS) and students' social/emo-

tional and academic trajectories in elementary grades (Hamre and Pianta 2001; Pianta et al. 2008). In a recent paper, similar in spirit to the work reported here, Thomas Kane and colleagues (2010) explore the effectiveness of a sample of Cincinnati teachers as measured by both value-added growth in student achievement and a classroom observation protocol based on Charlotte Danielson's framework. They find that higher levels of performance on the observation protocol are associated with meaningful student achievement gains. In general, however, the associations between classroom observation and value-added measures are generally low to moderate (Bill and Melissa Gates Foundation 2012; Hill et al. 2011).

Yet the value of observation protocols may be not that they converge with value-added scores but that they provide a valuable complement to value-added measures of teaching effectiveness. Value-added measures are driven by assessments of student learning that are necessarily partial; such standardized tests may be less effective at measuring critical thinking, the ability to engage in rigorous academic discourse, or students' social and emotional development. Classroom observations can provide additional information about the interactions between teachers and students that are related to a wider range of student outcomes. For example, teachers' scores on the CLASS protocol have been linked to students' social and emotional development in preschool settings (Hamre and Pianta 2001). Classroom observation data can also be used diagnostically in ways that value-added measures cannot. Well-designed systems of classroom observation can not only identify more and less effective teachers but also provide information about the quality of teacher practice along multiple dimensions; observation protocols thus hold the potential to inform efforts to improve teaching. Identifying classroom practices associated with more effective teachers and then targeting these practices in teacher education and professional development provides a potential avenue for improving the quality of instruction for all students.

A number of current observation protocols have been designed to focus upon elements of classroom instruction that may be consistent across different grade levels and content areas, examining a series of features that could be considered generic elements of teaching. For instance, Danielson's (2007) *Enhancing Professional Practice: A Framework for Teaching* focuses on teacher preparation and knowledge, standards-based instruction, necessary material resources, and student and teacher relationships. Similarly, Pianta and his colleagues developed the Classroom Assessment Scoring System (CLASS) to assess instructional approaches, as well as the teacher-student and student-to-student interactions and the nature of the classroom environment (Pianta et al. 2006).

Other instruments have been designed to measure teachers' understanding of best practices and specific subject matter knowledge in content areas. Em-

phasizing the importance of making mathematics accessible to students, Heather Hill and her colleagues developed the Mathematical Quality of Instruction instrument (MQI) to assess the accuracy and richness of teachers' mathematical ideas, language, representations, and tasks (Hill 2005). In literacy, a number of observation protocols have been developed around comprehension, including systems developed by Barbara Taylor and colleagues (2003) at the University of Michigan's Center for the Improvement of Early Reading Achievement or the TEX-IN3 system (Hoffman et al. 2004).

None of the existing observation protocols, however, provide a way to observe across the many domains of ELA classrooms, particularly at the secondary level. The paucity of content-based observation approaches has been a persistent problem in efforts to develop assessments of teaching (Kennedy 2010). Indeed, Mary Kennedy argues: "Until recently, assessments have not attended to the intellectual substance of teaching: to the content actually presented, how that content is represented, and whether or how students are engaged with it. . . . Documenting the intellectual meaning of teaching events remains the elusive final frontier in performance assessment" (245–46). To that end, Pam Grossman and colleagues (2009) developed the Protocol for Language Arts Teaching Observation (PLATO), which builds on existing observation tools and research on effective teaching practices in ELA in an attempt to parse the different facets of teaching practice in secondary ELA classrooms.

In addition to classroom observations, teacher logs of instruction have been used to document the content of instruction and teachers' approaches to teaching that content (e.g., Rowan et al. 2004). Teacher logs can capture the variation in the content of instruction both across teachers and within teachers across different classes. Since the content to which students are exposed is likely to be related to their achievement, teacher logs of the content of instruction represent another important vantage point on instructional quality.

Method and Data

In this study, we observed the detailed teaching practices of New York City middle school ELA teachers who represented different levels of effectiveness as measured by value-added models to explore the relationships between observations of classroom practice and value-added as different measures of teaching effectiveness. We were interested both in whether we could identify classroom practices that were associated with teachers who were relatively more effective in predominately high-poverty schools and in whether value-added measures correlated with an independent measure of instructional quality. We also collected teaching logs to develop a richer and more comprehensive

picture of classroom practices. We describe each of these components in more detail below.

Sample Selection

To create the sample for this study, we first estimated the value-added of all New York City teachers who taught ELA to students in grades 6–8.[1] There are active debates concerning the best specification for estimating teacher effects. Because there is no consensus on the best approach, we chose to combine two measures. In particular, we use one estimate that models gains in student achievement as a function of student time varying characteristics (X), classroom characteristics (C), school characteristics (S), a teacher fixed effect (τ), student fixed effects (θ), and year (ν) and grade (ϕ) indicator variables (see model [1]). This strategy identifies value-added by comparing teachers who teach the same students, usually in different years.

$$A_{ijst} - A_{ijs(t-1)} = \beta_0 + X_{it}\beta_1 + C_{ijst}\beta_2 + S_{jst}\beta_3 + \tau_j + \theta_i + v_t + \phi_g + \varepsilon_{ijst}. \quad (1)$$

Our other estimate models gains on student controls, school controls, classroom controls, and year and grade indicator variables (see model [2]).

$$A_{ijst} - A_{ijs(t-1)} = \beta_0 + X_{it}\beta_1 + C_{ijst}\beta_2 + S_{jst}\beta_3 + \tau_j + v_t + \phi_g + \varepsilon_{ijst}. \quad (2)$$

The student controls include gender, race, eligibility for free lunch, prior year test scores in math and ELA, and English learner status, among other factors. Classroom variables include the aggregates of all the individual variables plus the standard deviations of the prior year test scores. The school variables include enrollment, the percent of both black and Hispanic students, the percent of English learners, and the school average expenditures per pupil. We shrink each measure of value-added using empirical Bayes techniques to adjust for estimation error in calculating value-added coefficients.[2] The data for the value-added analysis is based on student-level data for performance on the New York State achievement tests of grades 6–8 matched to each student's teacher, a rich set of student, classroom, and school controls for years 2001–7. (A more detailed discussion of the attributes of the data can be found in Boyd et al. [2008].)

We then divided teachers into quartiles based on each of these two estimated value-added measures. In particular, we identified teachers in their third year through sixth year of teaching who were in the second quartile of value-added performance on both measures or were in the fourth quartile (the top) of value-added performance on both measures. We employ teachers in their

third through sixth years of teaching, as by this point teachers have generally stabilized their teaching practices and have plateaued in their gains to value-added (Kane et al. 2006). In addition, we have multiple years of value-added data for these teachers, which increases the stability of the measure. For this analysis, we selected teachers in the second quartile rather than the lowest because we thought that there might be significant behavior management issues and relatively little instruction occurring in the lowest-quartile classrooms, based on our experiences observing in schools. Because PLATO is focused more on instruction and less on management issues, we believed we could learn more from the comparison of second and fourth quartile teachers. In addition, the differences between teachers in the top and bottom quartiles might have been striking enough that raters would ascertain a teacher's quartile during observations, potentially biasing their scoring of classroom practices.

Using these samples, we identified pairs of middle school teachers—at least one moderate-performing (second quartile) teacher and at least one high-performing (fourth quartile) teacher—teaching in the same school. If we were unable to recruit both a second quartile and a fourth quartile teacher from a school, we dropped the school from our sample. Ultimately, we selected 24 teachers across nine schools with at least two contrasting teachers from each school. Neither observers nor participants knew the value-added quartiles of specific teachers during any component of data collection. Teachers in our sample in the second quartile ($N = 13$) had value-added that averaged -0.08, while those in the fourth quartile ($N = 11$) averaged 0.19. For the subsequent analysis, we employ the value-added measure described in model (2) above.[3] The mean difference of 27% of a standard deviation of student achievement is large relative to the effect of most interventions intended to improve student achievement and reflects important differences in the estimated effectiveness of teachers in each group (Hill et al. 2007) and very large relative to educational interventions, which typically average from .03 to .10 standard deviations.

Table 1 describes the teachers in our sample and the larger population of grades 6–8 ELA teachers from which they were drawn. In many respects, the teachers from the two value-added groups (cols. 2 and 3) are similar, although the higher-performing teachers were somewhat more likely to be African-American, somewhat less likely to be New York City Teaching Fellows, and scored somewhat higher on the Liberal Arts and Sciences Test, the New York State general knowledge certification exam. The sample is similar to the larger population of middle school ELA teachers on many observable measures, for example, on gender, age, and performance on the general knowledge certification exam and the verbal portion of the SAT. However, in some respects they differ. There tend to be relatively more African-American teachers in our sample, with fewer white, Hispanic, and other teachers. The sample teachers are also more likely to have received their teacher preparation in a

TABLE 1

ELA Teachers in Grades 6–8, 2008, with 4 or 5 Years of Teaching Experience, Sample Teachers and All New York City Teachers

	SAMPLE MEANS			
VARIABLE	Low Value-Added (N = 13)	High Value-Added (N = 11)	Average (N = 24)	NEW YORK CITY MEAN (N = 3,777)
Female	.769	.909	.833	.831
Race/ethnicity:				
Black	.154	.273	.208	.142
Hispanic	.077	.000	.042	.097
White	.615	.455	.542	.695
Other	.154	.272	.208	.065
Year of birth	1975	1976	1975	1975
Pathway:				
College recommended	.538	.545	.542	.473
Individual evaluation	.077	.000	.042	.090
New York City teaching fellows	.308	.182	.250	.243
Teach for America	.000	.091	.042	.087
Temporary	.077	.000	.042	.009
Unknown	.000	.182	.083	.094
General Knowledge Certification Exam	249	261	255	258
SAT Math (N=13)	464	450	458	495
SAT Verbal (N=13)	506	492	501	509
Teacher value-added	−.081	.187	.041	~0

traditional preparation program than was true of the larger population of teachers.

Classroom Observation: Development of PLATO

A primary data source for this study involves two structured classroom observation protocols. In addition to PLATO, we used six elements from two domains of the CLASS (La Paro et al. 2004) to assess two of the more generic aspects of instruction we wanted to measure: emotional support and classroom organization. Within the domain of Emotional Support, we used the CLASS scales for Positive Climate, Negative Climate, and Regard for Adolescent Perspectives. Within the domain of Classroom Organization, we included the CLASS measures of Behavior Management and Productivity. We also included the CLASS measure of Student Engagement. All of our observers were trained and certified in the use of CLASS prior to entering the classrooms.[4]

We assessed interrater agreement on the instrument in the following manner. First, master coders identified target scores for each video. We then calculated the percentage of individual ratings that fell within one point of the target score. All observers completed training both in CLASS and in PLATO; CLASS ran its own training for teams in California and New York. PLATO agreement was assessed using at least five different videos of classroom instruction across three different content domains (writing, literature, and grammar). All observers achieved at least 80% agreement on both CLASS and PLATO before observing in the field.[5] We reassessed interrater agreement during the second wave of observations in New York. Fifteen percent of all instructional segments were double coded on both CLASS and PLATO, and we again achieved a minimum of 80% agreement on our ratings. (For details on the construction and attributes of PLATO, see Grossman et al. [2009].)

In addition to using the six elements from two domains of CLASS, we also developed our own structured observation protocol—the Protocol for Language Arts Teaching Observation (PLATO)—based on research on effective literacy instruction. This instrument is designed for observations of middle and high school English language arts classrooms. For this study, PLATO was initially designed to be used in conjunction with the CLASS instrument and therefore employed the same seven-point scale referenced to three levels (low, medium, high). For each element, we developed indicators of interactions that would receive low scores (1 and 2), medium scores (3, 4, and 5), and high scores (6, 7) by raters.[6] Researchers observed for 15-minute intervals and then coded that 15-minute segment of instruction according to both the PLATO elements and the six CLASS elements. In its first iteration, PLATO included 10 elements of effective English language arts instruction: Purpose, Intellectual Challenge, Representations of Content, Connections to Personal and Prior Knowledge, Models and Modeling, Explicit Strategy Instruction, Guided Practice, Feedback, Classroom Discourse, and Accommodations for English Learners (based on research on effective ELA instruction). We provide a brief description of these elements.

The element of Purpose derives from research that suggests that children learn better in classrooms in which the purposes and goals of their work are clearly articulated and the relationships between what they learn and broader goals are clear (Borko and Livington 1989; Smith and Feathers 1983). Intellectual Challenge was designed to focus upon the nature of the tasks and questions aimed to students and the degree to which these tasks were cognitively challenging for students and asked students to engage in demanding intellectual work (e.g., Newmann et al. 1998). We also wanted to capture the ways in which teachers made content accessible to students and contextualized that content in terms of students' prior or personal knowledge (Bransford and Johnson 1972; Lee 1995; Levin and Pressley 1981; Tharp and Gallimore

1988). The element of Representation of Content evaluates the teacher's disciplinary knowledge, and the accuracy and clarity of the representations of the content he or she made to students during the observed segment. Connections to Personal and Prior Knowledge looks at the extent to which the teacher connects new material to students' prior learning, and it captures the degree to which the teacher makes linkages to students' experiences to help them connect to the content.

Students also need examples of strong work in ELA, strategies to help them produce sophisticated readings and written texts, and structured and scaffolded opportunities to practice employing those strategies. Three additional elements—Modeling, Explicit Strategy Instruction, and Guided Practice—attempt to capture this triad of practices central to effective ELA instruction. When a teacher provides examples and models of what students are being asked to do, students have specific, concrete images of what their work can and should look like (Frederiksen and Collins 1989; Graham 2006; Hillocks 1995). Because research also suggests that teaching specific strategies that can be used flexibly across a range of ELA activities can enable students to be more successful, we included an element on Explicit Strategy Instruction (Beck and McKeown 2002; Greenleaf et al. 2001; Palinscar and Brown 1987). Guided Practice evaluates the level of support that a teacher provides in the segment observed as well as a teacher's capacity to check in with his or her students, evaluate their learning, and offer any needed support.

The element of Feedback is based upon a long line of research that suggests that feedback facilitates student learning (Kluger and DeNisi 1996; Thorndike [1931] 1968), particularly when it is specific and targeted (Sadler 1989; Sperling and Freedman 2001). This element captures when teachers elicit student ideas, probe student thinking, and have opportunities to address misconceptions. The Classroom Discourse element grows out of research on the nature of productive discourse that can promote learning in the classroom (Nystrand 1997; Nystrand and Gamoran 1991; Taylor et al. 2003). It assesses opportunities students have for conversations with the teacher or among peers, as well as whether the discourse is perfunctory and minimal, at the low end, or elaborated and purposeful, at the high end.

Finally, research suggests that teachers in most mainstreamed classrooms today are increasingly responsible for teaching English learners, and hence, must be able to respond both to their language needs as well as support their academic development (August and Hakuta 1997). The element of Accommodations for Language Learning captures the range of strategies and supports that a teacher might use to make a lesson accessible to nonnative speakers; this includes the teacher's taking into account individuals' levels of language proficiency, strategic use of primary language, grouping strategies, differenti-

ated materials and assessments, and the use of graphic organizers and visual displays.

Because content coverage can also be an important predictor of student achievement (Rowan et al. 2002), PLATO also includes checklists for the major content domains within English language arts, including reading, writing, literature, speaking/listening, and grammar and mechanics. For each segment of instruction, observers check the domain that was the focus of instruction and identify additional features related to that domain.[7] Finally, PLATO also captures the activity structures for every segment of instruction.

Data Collection

We observed teachers on 6 separate days during the spring of 2008. On each day, we observed teachers for at least 2 hours of instruction, generally in two different classes. The 6 days were divided into two waves of data collection, separated by between 2 and 6 weeks. In each class, we observed 15 minutes of instruction and then rated the teacher on the 10 PLATO elements and six CLASS dimensions for approximately 8 minutes. We then observed an additional 15 minutes and rated a second segment. Across the 6 days of instruction, we coded an average of 24 instructional segments per teacher.

To capture additional dimensions of classroom practice, including relationships among teachers and students, peer interactions in the classroom, curricular focus, and evidence of links to out-of-school literacy practices, we also included more open-ended observations. For this reason, we had two observers in the majority of the classrooms; one observer in the classroom took open-ended notes, while the second observer used PLATO and CLASS to score instruction.

Because we observed just 6 days of instruction for each teacher, we wanted to include a measure that captured a longer period of instruction. To supplement our direct observations, we included a teacher log, based, in large part, on the Study of Instructional Improvement's teacher log for ELA for middle school classrooms (Rowan et al. 2004). The log replicates the content categories of PLATO, allowing us to coordinate two of our measures of instruction. After a brief training in how to use the logs, we asked teachers to fill out this log for 15 consecutive days of instruction, or roughly 3 weeks.[8] Teachers began to fill out the logs on the days we began our observations, which allowed them to ask any questions once they began to use the log. With these logs, we are also able to assess the overlap between the observer's and the teacher's assessment of content coverage and activity structures for the days of observation.

Results

We look first at the overall picture of ELA instruction in these classrooms. Of the PLATO measures, on average, teachers scored highest on Purpose and lowest on Accommodations for English Language Learners (fig. 1).[9] The standard deviation across teachers is approximately 1.0 for all 10 PLATO elements. For the CLASS elements, the teachers in our sample scored highest on Behavior Management and lowest on Negative Climate, which is the one element constructed to represent a negative classroom behavior and which is reverse coded. The CLASS elements have somewhat higher variances across teachers than the PLATO elements (fig. 2).

Observational Measures and Value-Added

Our structured observation data suggest that there may be systematic differences between teachers in the two value-added quartiles (fig. 1). While this was an exploratory study with a small sample, we found a consistent pattern of teachers in the higher quartile of value-added scoring higher than teachers in the lower quartile on nearly all PLATO elements. In some cases the differences were relatively small, while in others the differences were larger. The small size of the sample (24 teachers) makes it difficult to statistically differentiate groups; however, *t*-tests of the differences across groups on each of the PLATO scores show that the two groups of teachers are statistically different on the element of Explicit Strategy Instruction ($p = .03$) and have relatively small *p*-values on the elements of Guided Practice ($p = .09$) and Intellectual Challenge ($p = .13$). The CLASS elements of Student Engagement ($p = .10$) and Negative Climate *p*-value (.12) also have relatively low *p*-values. While our sample size limits the extent to which we could find statistically significant differences between the groups, our exploratory results suggest that high-quartile teachers may be different from lower-quartile teachers on several aspects of their instruction.

Some differences in observed classroom practice are associated with meaningful differences in value-added. For example, a standard deviation improvement in PLATO's Explicit Strategy Instruction (ESI), for which the difference between the mean value-added of the two quartiles is significant at $p = .03$, is associated with an improvement in student achievement of 11% of a standard deviation.[10] Increasing Guided Practice ($p = .09$) by a standard deviation is associated with a 7% of a standard deviation increase in student achievement.[11] Although we cannot know if these are causal relationships, they do establish a credible hypothesis that readily measurable aspects of classroom practice may meaningfully improve student achievement.

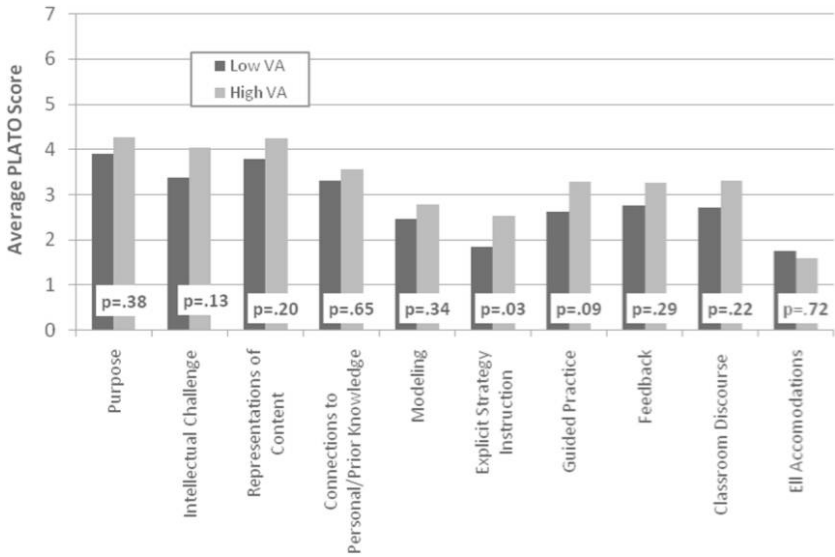

FIG. 1.—Average score by value-added group for each of the PLATO elements

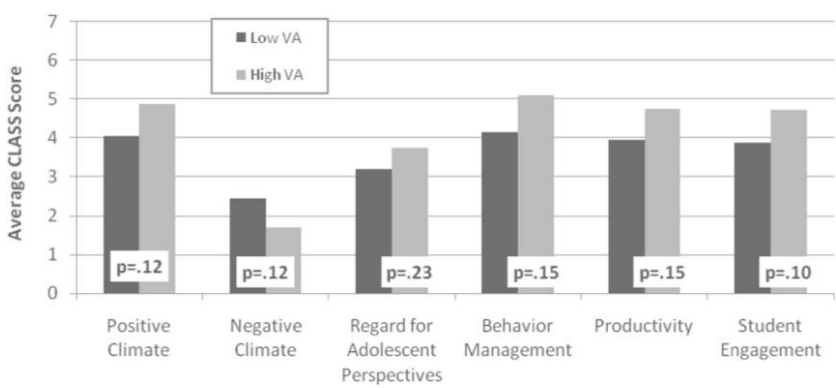

FIG. 2.—Average score by value-added group for each of the CLASS dimensions

While a few of the PLATO elements and CLASS dimensions appear to identify differences in value-added measures of teachers, they do not necessarily reflect discrete measures of instruction. The measures are highly correlated, particularly at the teacher level. Appendix B in the online version of this article provides data on teacher-level correlations across all the classroom observation measures. Within PLATO, particularly high correlations exist between Intellectual Challenge and Purpose (.91) and between Classroom Discourse and Feedback (.95).[12] Among the CLASS measures, Productivity, Student Engagement, and Behavior Management are all very highly correlated (at least .94). The high correlations make it difficult to assess whether particular instructional elements as captured by these scales are individually related to higher value-added scores or if they are tapping into a broader underlying dimension that characterizes the difference between the groups.

To look more carefully at the relationship between individual elements of instruction and value-added measures, we used the teacher's scores on PLATO elements and CLASS dimensions to predict that teacher's value-added quartile (table 2). Each column within each of the two vertical panels of the table represents a separate model in which one of the observational measures of instruction predicts the log odds of a teacher being in the high-value-added group. The numbers in parentheses represent the standard errors of these estimates. As an example, in the first specification, a teacher with a one unit higher score in Explicit Strategy instruction is 4.88 times more likely to be in the high-value-added group. This positive relationship between value-added and Explicit Strategy Instruction holds up to the inclusion of each of the other measures of instruction. None of the other measures have nearly as strong a relationship with value-added once Explicit Strategy Instruction is included in the model. This underscores that, in this study, Explicit Strategy Instruction is the dominant dimension that differentiates between high-quartile and low-quartile teachers.

As is clear from this analysis, the element of Explicit Strategy Instruction distinguishes the more effective teachers in our sample. To get a better sense of what such instruction looks like, we provide some examples from the field notes taken during open-ended observations. Teachers who score "high" (a score of 6 or 7) on Explicit Strategy Instruction provided students with very structured and specific ways to approach ELA activities. For example, one high-quartile teacher systematically broke down a newspaper article on "skinny jeans" to help students understand the features of effective journalism. She instructed them on how to compose a list of "4 Ws" (who, when, where, and what), how to use that list to create a focused lead, and then how to incorporate supporting details culled from graphic organizers. Students then wrote their own newspaper articles with an arsenal of specific strategies. This focus on how students could tackle ELA tasks was reflected in other high-quartile teach-

TABLE 2

Logit Results in Odds Ratios Predicting High-Value-Added Grouping

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Explicit strategy instruction	4.88 (.070)	3.31 (.146)	3.47 (.108)	5.91 (.058)	10.17 (.065)	3.16 (.171)	3.76 (.089)
Purpose	.76 (.643)						
Intellectual challenge		1.24 (.688)					
Representations of content			1.28 (.699)				
Connections to prior knowledge				.68 (.410)			
Modeling					.3 (.308)		
Guided practice						1.37 (.613)	
Feedback							1.06 (.891)
	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Explicit strategy instruction	3.54 (.099)	3.35 (.107)	3.06 (.138)	3.42 (.091)	3.37 (.11)	3.38 (.115)	3.35 (.117)
Classroom discourse	1.17 (.718)						
Positive climate		1.49 (.94)					
Negative climate			.611 (.405)				
Regard for adolescent perspectives				1.33 (.542)			
Behavior management					1.23 (.523)		
Productivity						1.25 (.577)	
Student Engagement							1.54 (.337)

ers' classrooms. These teachers made visible the often invisible processes requisite for successful, sophisticated literary analysis, reading comprehension, or writing.

Unfortunately, instances of effective strategy instruction are also quite rare. The mean score for Explicit Strategy Instruction is 2.1, and the modal score is 1. Even instructional segments that score high on other elements such as Intellectual Challenge score lower on Explicit Strategy Instruction. During a lesson in which students were asked "to anticipate an opponent's counterar-

gument in writing an editorial," undoubtedly an intellectually demanding activity, the teacher did not discuss how students might accomplish the lesson's goal when writing independently. While the vast majority of teachers in our sample provided students with directions for completing activities, they did not instruct them on the nuances of how to complete those activities effectively. The goal of many lessons was completion of the specific task rather than mastering a more broadly applicable skill or strategy.

The lack of strategy instruction discussed previously was particularly pronounced during writing instruction. Our qualitative data indicates a clear distinction between "writing instruction" and lessons during which students were asked to write. Unfortunately we saw few of the former, regardless of teacher quartile. The majority of what were coded as writing lessons were lessons in which students spent class time writing but were given little to no direction about how to structure their writing or strategies to improve their writing. This lack of structure was evident in lessons across the stages of the writing process. The majority of peer editing sessions involved students reading each other's writing without specific features on which they should focus or questions to use to guide the editing process. As a result, students often provided each other with general or vague feedback such as "I love it!" or "You could make this better." Often teachers told students to "work on" or revise a draft for an entire class period, with no specifications or guidance about how to structure their efforts. As a result, we saw numerous students simply typing or copying earlier drafts in neater handwriting in an attempt to revise.

Intellectual Challenge and Guided Practice were two other instructional elements that seem to have distinguished teachers in the higher-value-added quartile, though the differences in scores were only marginally significant. Our open-ended observation notes provide vivid illustrations of the range in the intellectual rigor of the instruction teachers provide. Many of the teachers in the lower-value-added quartile provided instructional activities involving students' writing instructions for what they do when they wake up in the morning or completing highly formulaic "bio poems" that required little more than filling in the blanks. In sharp contrast, teachers in the high quartile had students writing five-paragraph essays about *My Antonia*, generating alternative endings to short stories, or crafting speeches from the perspective of presidential candidates. In terms of Guided Practice, teachers in the lower-value-added quartile either did not provide opportunities for students to practice new skills in class or allotted time for students to work independently but did not provide support or "guidance" during class time. Often these teachers asked students to complete literature or writing assignments at home, without the benefit of teacher support and clarification. The teachers in the high quartile of value-added circulated during literature circles, answering student questions and clarifying their ideas or doing periodic whole class "check ins" as students

worked through stages of the writing process. In these instances, we find some evidence of a relationship between a teacher's value-added quartile and several facets of high-quality instruction that can be more systematically explored in future research. While these analyses are exploratory, they suggest directions for future research with larger samples of teachers.

Also of note here are the elements of instruction captured in PLATO that do not seem highly related to value-added achievement but which we have reason to believe may be important teaching skills. The element of Connections to Personal and Prior Knowledge provides a good example. In this study, the quality of teachers' connections to students' experiences does not seem highly associated with value-added measures. Yet, one goal of English courses is to help students make personal connections to the literature they read. Work on motivation in adolescent literacy suggests the importance of helping students see connections between their academic work and their worlds outside of school (e.g., Snow and Biancarosa 2003). There are several possible explanations for the fact that this measure is not highly associated with student achievement in this study. First, PLATO might not measure this dimension of teaching well. Alternatively, the outcomes of teachers' ability to make these kinds of connections are not easily captured on standardized achievement test. A better measure of this aspect of teaching might be students' motivation to read outside of school or their sense of identity as readers and writers, neither of which is assessed by standardized tests. Observation protocols may capture aspects of teaching that may be related to student outcomes that are not necessarily well measured by standardized assessments and therefore would not be highly associated with value-added measures. Classroom Discourse might represent another element that might not have a simple relationship to standardized test scores. The kinds of argumentation and interpretation that might be the result of sustained, high-quality discussion of a literary text, for example, might better be captured through assessments of students' analytic writing.

Teacher Logs

As described above, teachers in our sample also filled out logs of their daily activities. These logs were filled out daily over a period of 15 school days and thus capture a substantially longer time frame of instruction than the observations. While the logs provide another data point on the content of instruction in these classrooms, none of the content coverage measures we constructed from the log data statistically differentiate between teachers in the two quartiles of value-added; only reported time spent on research skills approaches statistical significance (see fig. 3). Interestingly, teachers in the lower-value-added

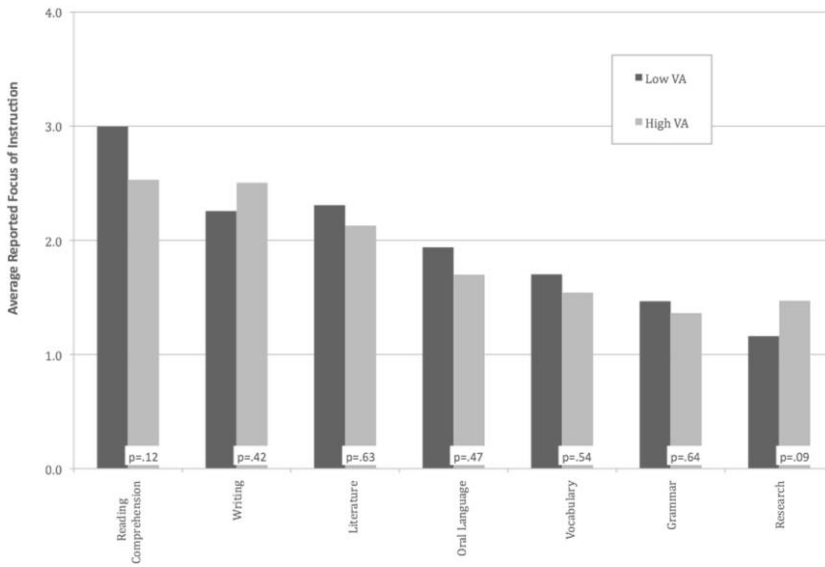

FIG. 3.—Content domain focus as reported in teacher logs by value-added group

quartile actually reported spending more time on reading instruction, while teachers in the higher-value-added quartile seemed to spread their time a bit more evenly across the content domains of ELA. While not significant in this study, these patterns may be worth exploring in larger-scale studies, especially in conjunction with more information about the content domains actually assessed in standardized assessments.

While content coverage does not seem to distinguish teachers in the two different quartiles, the logs indicate that high- and low-value-added teachers do seem to make different use of grouping structures in their instruction (see fig. 4). Although teachers in both quartiles reported using individual work time for approximately the same percentage of instructional time, high-value-added teachers reported using small groups far more than low-value-added teachers (36% vs. 16%), and they reported employing large-group instruction far less (26% vs. 44%).

Interestingly, teachers' self-reports of content coverage in the teacher logs differ from our structured observations, indicating that teachers perceive what they are doing differently from outside observers. For example, on days on which we observed, we noted one teacher teaching grammar over 30% of the time, while she reported teaching it only 7% of the time. This may be a result of the limited training provided to teachers on how to use the logs or

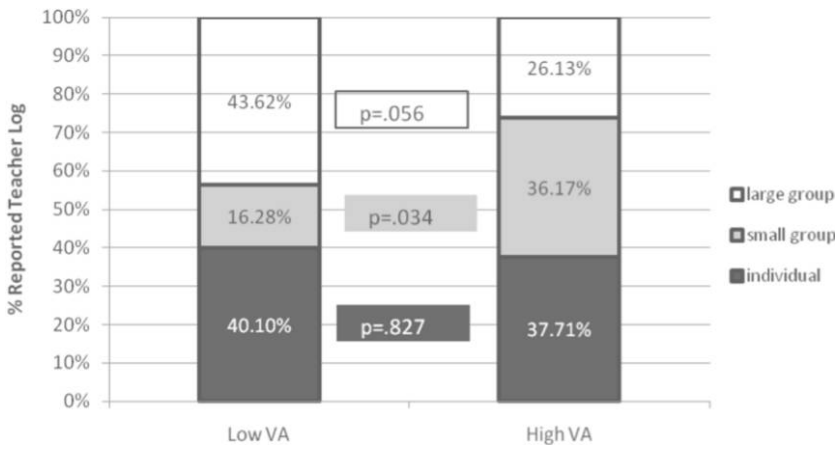

FIG. 4.—Use of small and large groups as reported in teacher logs by value-added group.

it may relate to different perceptions of observers and teachers on the content of instruction.

Implications

As a result of a substantial push by policy makers, the effort to quantify teacher effectiveness is progressing quickly. Many states and school districts feel a sense of urgency to employ measures of effectiveness to help shape the teaching workforce. Many of these efforts have been encouraged by federal grant programs, such as the RTTT, TIF, and i3, and by emerging work from the Gates Foundation on Measures of Effective Teaching project (Bill and Melissa Gates Foundation 2012). Critics suggest that practice is substantially outpacing our understanding of how to measure effective teaching. In a recent article on assessment of teacher quality in practice, Deborah Ball and Heather Hill (2009) commented: "The current enthusiasm for teacher quality requires caution. In the end, what matters is the quality of the instruction that students receive—that is, teaching quality. . . . However, given the underdevelopment of the field right now, we need to improve the precision with which we conceptualize and measure teacher quality. . . . We will have to delve into instruction and then map backward and forward to specific elements that we can use to predict instructional practice and its quality" (95–96).

This article is an attempt to better understand the relationship among

teacher value-added scores and multiple measures of teaching effectiveness across a range of teaching domains. Even with the small sample used in our analysis, we find some evidence that teachers in the high quartile of value-added have a different profile of instructional practices than teachers in the low quartile of value-added. Teachers in the high quartile score meaningfully higher than teachers in the lower quartile on a few elements of PLATO and CLASS, suggesting that effective teachers have better command over certain skills than less effective teachers. We also find that other instructional elements are not associated with gains in student learning as measured by value-added. Both sets of findings have important implications for future research and practice.

As is true of most exploratory research, our findings are correlational and do not imply causality. Teachers who effectively engage in explicit strategy instruction may contribute to increased achievement in their students for many other reasons. However, it is noteworthy that of all the elements of teachers' practices we observed in a double blind application of a rigorously designed observation protocol, teachers' use of Explicit Strategy Instruction is associated with the largest effect on value-added. Further research will need to confirm this finding. If, however, the quality of certain instructional practices, like Explicit Strategy Instruction, does differentiate teachers in terms of their impact on student achievement, we would also want to know if professional development can meaningfully improve the quality of quantity of teachers' use of explicit strategy instruction.

If we can identify classroom practices that are associated with student achievement, we may be in a much stronger position to improve the quality of teaching in ways that have demonstrable effects on students. Instructional practices such as explicit strategy instruction are teachable. If observation protocols can identify such practices, initial teacher preparation and other forms of professional development can focus their efforts on improving teachers' skills and practices in these areas.

Another important implication of this research is that some observed teaching practices are not associated with student achievement gains as measured by the New York standardized ELA exam. One possible explanation for this is that components of the observation protocol measure dimensions of teaching practice that do not affect student learning. A second possibility is that some aspects of instruction, such as the quality of classroom discourse, may be important in developing students' reasoning abilities and conceptual understanding of literature and writing, but these abilities may not be measured well by the tests used to construct value-added scores. In addition to developing multiple measures of instruction, we need to develop multiple measures of student outcomes to ensure that classroom instruction supports the development of a broad range of learning outcomes for students. Given the limited

slice of student learning that can be captured through current standardized assessments, we should be cautious of prematurely narrowing our definition of good teaching to only those elements that correlate with value-added measures.

Teacher observation tools like PLATO and CLASS will inevitably play an important role in the assessment and improvement of teacher quality. By identifying components of classroom practice that are related to student achievement, teacher preparation and professional development can systematically be improved. However, just like value-added measures, these tools are still far from perfect. For example, although many of the 16 elements used in this study appear to signal a higher-value-added teacher, they do not necessarily reflect different features of instruction. The elements are highly correlated. In addition, teacher observation protocols differ widely in their construction, and we know little regarding the effects of these differences on teacher evaluation.

Finally, we share the concern that none of the measures of teacher effectiveness are ready for the high-stakes uses to which they are being put. Advocates rightly suggest that these measures have important signals that result in the improvement of teaching quality. Critics counter that none of these measures will stand up to scrutiny when individuals fail to retain their job or receive a salary increment. Ultimately, before such measures are used to make high-stakes decisions about individual teachers, we need a much better understanding of how measures of effectiveness perform, both individually and in combination with other measures, and how they can best be employed to improve the quality of classroom teaching.

Notes

We would like to thank the Carnegie Corporation for its support of this work. We would also like to thank our collaborators Donald Boyd and Karen Hammerness, our project manager Sinead Mullen, and everyone who helped with this study, including Chandra Alston, Michelle Brown, Conra Gist, Sharon Greenberg, Hamilton Lankford, Dana McCormick, Rebecca Rufo-Tepper, and Rita Simpson-Vlach; it truly takes a village. Finally, we would like to thank the teachers who participated in this study, without whom this work would not have been possible.

1. These value-added estimates were used solely for research purposes. Neither teachers nor administrators in New York City were privy to these estimates nor were they used for any personnel decisions.

2. See app. A in the online version of this article for a brief discussion of the Empirical Bayes technique.

3. When we estimate a partial adjustment model identical to the model in eq. (2), we obtain estimates that are correlated with the model (2) estimates at .99 and produce quartile groupings exactly the same as those produced by model (2). For the subsequent

analysis, we used the quartiles of value-added, which are the same regardless of the model of value-added used.

4. Interrater agreement was assessed by CLASS researchers, who also trained the team. To achieve interrater agreement, raters needed to score at least 80% on all of the elements. Raters scored from between 81% and 85%.

5. Exact score match rates were notably lower, as would be expected with a seven-point scale with high-inference and multifaceted dimensions of teaching practice. In general, the exact match rates were in the range of 40% to 60%. The highest exact match rate was for Negative Climate (70%), and the lowest was for Regard for Adolescent Perspectives (31%). Some elements were easier for raters to achieve agreement using the exact and adjacent score criteria: Purpose (94%), Intellectual Challenge (85%), Modeling (86%), and Behavior Management (96%). Other elements proved more difficult for raters to score consistently with a master score: Classroom Discourse (63%), Strategy Instruction (62%), and Positive Climate (63%).

6. In subsequent versions of PLATO, we use a four-point scale and require an exact match for interrater agreement.

7. For example, if observers identify reading as the target domain, they also identify the nature of the text read (e.g., fiction or nonfiction), the focus of reading instruction (e.g., comprehension, decoding, metacognitive strategies, etc.), and the nature of in-class reading activity (e.g., independent reading, teacher reading aloud, etc.).

8. In retrospect, we should have spent more time training teachers to fill out the logs, as previous research emphasizes the importance of training to ensure high-quality responses (see Rowan et al. 2004). We provided a brief orientation during our first interview with teachers.

9. Means and standard deviations for each element are located in tables B1 and B2 in app. B in the online version of the article.

10. This estimate is based on an estimated coefficient of ESI of .09, and a standard deviation of ESI of 0.80, and a standard deviation of achievement of 1.0

11. Based on estimated coefficient and standard deviation of Guided Practice of .069 and 0.96, respectively, and a student achievement standard deviation of 1.0.

12. In subsequent versions of PLATO, we conducted factor analyses that show an underlying empirical and conceptual set of factors. The factors include Instructional Scaffolding, Cognitive and Disciplinary Demand of Talk and Tasks, and Representation and Use of Content.

References

August, Diane, and Kenji Hakuta, eds. 1997. *Improving Schooling for Language Minority Children: A Research Agenda.* Washington, DC: National Academies Press.

Ball, Deborah L., and Heather C. Hill. 2009. "Measuring Teacher Quality in Practice." In *Measurement Issues and Assessment for Teaching Quality*, ed. Drew H. Gitomer. Thousand Oaks, CA: Sage.

Beck, Isabel L., and Margaret G. McKeown. 2002. "Questioning the Author: Making Sense of Social Studies." *Educational Leadership* 60 (3): 44–47.

Bill and Melinda Gates Foundation. 2012. *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains: Initial Year 2 Findings from the Measures of Effective Teaching (MET) Project.* Seattle: Bill and Melinda Gates Foundation.

Borko, Hilda, and Carol Livingston. 1989. "Cognition and Improvisation: Differences

in Mathematics Instruction by Expert and Novice Teachers." *American Educational Research Journal* 26 (4): 473–98.

Boyd, Donald, Hamilton Lankford, Susanna Loeb, Jonah Rockoff, and James J. Wyckoff. 2008. "The Narrowing Gap in New York City Teacher Qualifications and Its Implications for Student Achievement in High-Poverty Schools." *Journal of Policy Analysis and Management*, 27 (4): 793–818.

Bransford, John, and Marcia K. Johnson. 1972. "Contextual Prerequisites for Understanding: Some Investigations of Comprehension and Recall." *Journal of Verbal Learning and Verbal Behavior* 11 (6): 717–26.

Carlin, Bradley P., and Thomas A. Louis. 1996. *Bayes and Empirical Bayes Methods for Data Analysis*. London: Chapman & Hall.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2011. "The Long-Run Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood." Working Paper No. 17699, National Bureau of Economic Research, Cambridge, MA.

Danielson, Charlotte. 2007. *Enhancing Professional Practice: A Framework for Teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.

Frederiksen, John R., and Allan Collins. 1989. "A Systems Approach to Educational Testing." *Educational Researcher* 18 (9): 27–32.

Goe, Laura, Courtney Bell, and Olivia Little. 2008. *Approaches to Evaluating Teacher Effectiveness: A Research Synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

Graham, Steven. 2006. "Strategy Instruction and the Teaching of Writing: A Meta-Analysis." In *Handbook of Writing Research*, ed. Charles A. MacArthur, Steve Graham, and Jill Fitzgerald. New York: Guilford.

Greenleaf, Cynthia, L., Ruth Schoenbach, Christine Cziko, and Faye L. Mueller. 2001. "Apprenticing Adolescent Readers to Academic Literacy." *Harvard Education Review* 71 (1): 79–130.

Grossman, Pam, Sharon Greenberg, Karen Hammerness, Julie Cohen, Chandra Alston, and Michelle Brown. 2009. "Development of the Protocol for Language Arts Teaching Observation (PLATO)." Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA, April.

Hamre, Bridget K., and Robert C. Pianta. 2001. "Early Teacher-Child Relationships and the Trajectory of Children's School Outcomes through Eighth Grade." *Child Development* 72 (2): 625–38.

Hanushek, Eric, and Steven Rivkin. 2010. *Using Value-Added Measures of Teacher Quality*. Brief No. 9. Washington, DC: CALDER.

Harris, Douglas. 2009. "Would Accountability Based on Teacher Value Added Be Smart Policy? An Examination of the Statistical Properties and Policy Alternatives." *Education Finance and Policy* 4 (4): 319–50.

Hill, Carolyn J., Howard S. Bloom, Alison R. Black, and Mark W. Lipsey. 2007. "Empirical Benchmarks for Interpreting Effect Sizes in Research." Working Papers on Research Methodology, MDRC, New York.

Hill, Heather. 2005. "Content across Communities: Validating Measures of Elementary Mathematics Instruction." *Educational Policy* 19 (3): 447–75.

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.

Hillocks, George. 1995. *Teaching Writing as Reflective Process*. New York: Teachers College Press.

Hoffman, James V., Misty Sailors, and Gerald R. Duffy. 2004. "The Effective Ele-

mentary Classroom Literacy Environment: Examining the Validity of the TEX-IN3 Observation System." *Journal of Literacy Research* 36 (3): 303–34.
Ishii, Jun, and Steven Rivkin. 2009. "Impediments to the Estimation of Teacher Value Added." *Education Finance and Policy* 4 (4): 520–36.
Kane, Thomas, Jonah Rockoff, and Douglas Staiger. 2006. "What Does Certification Tell Us about Teacher Effectiveness? Evidence from New York City." Working Paper No. 12155, National Bureau of Economic Research, Cambridge, MA.
Kane, Thomas, Eric S. Taylor, John Tyler, and Amy L. Wooten. 2010. "Identifying Effective Classroom Practices Using Student Achievement Data." Working Paper No. 15803, National Bureau of Economic Research, Cambridge, MA.
Kennedy, Mary M. 2010. "The Uncertain Relationship between Teacher Assessment and Teacher Quality." In *Teacher Assessment and the Quest for Teacher Quality: A Handbook*, ed. Mary Kennedy. San Francisco: Jossey-Bass.
Kluger, Avraham N., and Angelo DeNisi. 1996. "Effects of Feedback Intervention on Performance: A Historical Review, a Meta-analysis, and a Preliminary Feedback Intervention Theory." *Psychological Bulletin* 119 (2): 254–84.
La Paro, Karen M., Robert C. Pianta, and Megan Stuhlman. 2004. "The Classroom Assessment Scoring System: Findings from the Prekindergarten Year." *Elementary School Journal* 104 (5): 409–26.
Lee, Carol. 1995. "A Culturally Based Cognitive Apprenticeship: Teaching African-American High School Students Skills in Literary Interpretation." *Reading Research Quarterly* 30 (4): 608–28.
Levin, Joel R., and Michael Pressley. 1981. "Improving Children's Prose Comprehension: Selected Strategies That Seem to Succeed." In *Children's Prose Comprehension: Research and Practice*, ed. Carol M. Santa and Bernard L. Hayes. Newark, DE: International Reading Association.
McCaffrey, Daniel F., Tim R. Sass, J. R. Lockwood, and Kata Mihaly. 2009. "The Intertemporal Variability of Teacher Effect Estimates." *Education Finance and Policy* 4 (4): 572–606.
Morris, Carl N. 1983. "Practical Empirical Bayes Inference: Theory and Applications." *Journal of the American Statistical Association* 78 (381): 47–55.
Newmann, Fred M., George Lopez, and Anthony S. Bryk. 1998. *The Quality of Intellectual Work in Chicago Schools: A Baseline Report*. Chicago: Consortium for Chicago School Research.
Nystrand, Martin. 1997. *Opening Dialogue: Understanding the Dynamics of Language and Learning in the English Classroom*. New York: Teachers College Press.
Nystrand, Martin, and Adam Gamoran. 1991. "Instructional Discourse, Student Engagement, and Literature Achievement." *Research in the Teaching of English* 25 (3): 261–90.
Palinscar, Annemarie S., and Deborah A. Brown. 1987. "Enhancing Instructional Time through Attention to Metacognition." *Journal of Learning Disabilities* 20 (2): 66–75.
Pianta, Robert C., Jay Belsky, Nathan Vandergrift, Renate Houts, and Fred J. Morrison. 2008. "Classroom Effects on Children's Achievement Trajectories in Elementary School." *American Educational Research Journal* 45 (2): 365–97.
Pianta, Robert C., Bridget K. Hamre, Nancy J. Haynes, Susan Mintz, and Karen M. La Paro. 2006. *Classroom Assessment Scoring System (CLASS) Manual: Middle/Secondary Version Pilot*. Charlottesville: Curry School of Education, University of Virginia.
Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73 (2): 417–58.
Rockoff, Jonah. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *American Economic Review* 94 (2): 247–52.

Rowan, Brian, Eric Camburn, and Richard Correnti. 2004. "Using Teacher Logs to Measure the Enacted Curriculum: A Study of Literacy Teaching in Third-Grade Classrooms." *Elementary School Journal* 105 (1): 75–101.

Rowan, Brian, Richard Correnti, and Robert Miller. 2002. "What Large-Scale Survey Research Tells Us about Teacher Effects on Student Achievement: Insights from the Prospects Study of Elementary Schools." *Teachers College Record* 104 (8): 1525–67.

Sadler, D. Royce. 1989. "Formative Assessment and the Design of Instructional Systems." *Instructional Science* 18 (2): 119–44.

Sanders, William L., and June C. Rivers. 1996. "Research Project Report: Cumulative and Residual Effects of Teachers on Future Student Academic Achievement." University of Tennessee Value-Added Research and Assessment Center, http://www.cgp.upenn.edu/pdf/Sanders_Rivers-TVASS_teachereffects.pdf.

Smith, Frederick R., and Karen M. Feathers. 1983. "The Role of Reading in Content Classrooms: Assumption versus Reality." *Journal of Reading* 27 (3): 262–67.

Snow, Catharine, and Gina Biancarosa. 2003. *Adolescent Literacy Development among English Language Learners*. New York: Carnegie Corporation of New York.

Sperling, Melanie, and Sarah W. Freedman. 2001. "Research on Writing." In *Handbook of Research on Teaching*, 4th ed., ed. Virginia Richardson. Washington, DC: American Educational Research Association.

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, Roland G., and Ronald Gallimore. 1988. *Rousing Minds to Life: Teaching, Learning and Schooling in Social Context*. Cambridge: Cambridge University Press.

Thorndike, Edward L. (1931) 1968. *Human Learning*. New York: Century.