NBER WORKING PAPER SERIES

MEASURE FOR MEASURE:
EN MEASURES OF INSTRUCTIONAL PRACTICE IN MIDDLE SCHOOL ENGLISH LANGUAGE ARTS AND TEACHE

Pam Grossman
Susanna Loeb
Julia Cohen
Karen Hammerness
James Wyckoff
Donald Boyd
Hamilton Lankford

Working Paper 16015
http://www.nber.org/papers/w16015

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2010

We are grateful to the New York City Department of Education and the New York State Education Department for the data employed in this paper. We appreciate financial support from the Carnegie Corporation and the National Center for the Analysis of Longitudinal Data in Education Research (CALDER). CALDER is supported by IES Grant R305A060018 to the Urban Institute. The views expressed in the paper are solely those of the authors and may not reflect those of the funders. Any errors are attributable to the authors. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Measure for Measure: The relationship between measures of instructional practice in middle school English Language Arts and teachers' value-added scores
Pam Grossman, Susanna Loeb, Julia Cohen, Karen Hammerness, James Wyckoff, Donald Boyd, and Hamilton Lankford
NBER Working Paper No. 16015
May 2010
JEL No. I21

ABSTRACT

Even as research has begun to document that teachers matter, there is less certainty about what attributes of teachers make the most difference in raising student achievement. Numerous studies have estimated the relationship between teachers' characteristics, such as work experience and academic performance, and their value-added to student achievement; but, few have explored whether instructional practices predict student test score gains. In this study, we ask what classroom practices, if any, differentiate teachers with high impact on student achievement in middle school English Language Arts from those with lower impact. In so doing, the study also explores to what extent value-added measures signal differences in instructional quality. Even with the small sample used in our analysis, we find consistent evidence that high value-added teachers have a different profile of instructional practices than do low value-added teachers. Teachers in the fourth (top) quartile according to value-added scores score higher than second-quartile teachers on all 16 elements of instruction that we measured, and the differences are statistically significant for a subset of practices including explicit strategy instruction.

Pam Grossman
School of Education
Stanford University
Stanford, CA 94305
Pamg@stanford.edu

Susanna Loeb
524 CERAS, 520 Galvez Mall
Stanford University
Stanford, CA 94305
and NBER
sloeb@stanford.edu

Julia Cohen
School of Education
Stanford University
Stanford, CA 94305
juco@stanford.edu

Karen Hammerness
School of Education
Stanford University
Stanford, CA 94305
hammerness@optonline.net

James Wyckoff
Curry School of Education
University of Virginia
P.O. Box 400277
Charlottesville, VA 22904-4277
wyckoff@virginia.edu

Donald Boyd
The Center for Policy Research
University of Albany
135 Western Ave.
Albany, NY 12222
boydd@rockinst.org

Hamilton Lankford
School of Education, ED 317
University at Albany
State University of New York
Albany, NY 12222
hamp@albany.edu

Even as research has begun to document that teachers matter, there is less certainty about what attributes of teachers actually make the most difference in raising student achievement. Our own work and the work of others suggest that differences in teacher preparation may account for some of these differences, particularly in the first year of teaching, while other work identifies teacher attributes, including certification, the selectivity of teachers' undergraduate institutions, and teachers' scores on tests of general knowledge and verbal ability as factors that may be related to student achievement gains (Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2009; Kane, Rockoff, &Staiger, 2005; see Rice, 2003 for a review). The emphasis on teacher characteristics and preparation, however, obscures the importance of instruction within the classroom. Teachers' classroom practices are likely to be the mechanism by which teachers affect students. In addition, even if certain teacher attributes make a difference, such as having attended a more selective college, such findings may have limited utility for policy makers and teacher educators. Identifying classroom practices associated with high student achievement gains, and then targeting these practices in teacher education and professional development, provides a potential avenue for improving the quality of instruction for all students.

While our first goal is to identify those elements of instruction that lead to improvement in student learning, we are equally interested in assessing whether value-added measures capture instructional differences among teachers. Measures of value-added to student test score gains are becoming more and more common tools for measuring teacher effectiveness (c. f. Sanders & Rivers, 1996; Rivkin, Hanushek and Kain, 2005; Rockoff, 2004). Yet, researchers have debated the validity of these measures, questioning whether value-added measures reflect the characteristics of the students in the classroom more than the contribution of teachers to student achievement (Rothstein, 2007). This paper describes the pilot study of a larger effort to uncover the extent to which value-added measures reflect differences in instruction, as well to identify the classroom practices that are more characteristic of more effective teachers. We focus on middle school English Language Arts instruction and test performance in New York City. To measure instruction, we used a structured observation protocol that combines new measures of instruction with measures that have been used elsewhere (LaParo, Pianta, & Stuhlman, 2004; Pianta et al., 2006) in addition to collecting additional measures of instruction, including teacher logs and student work. In what follows we describe the differences in instruction between teachers with high value-added scores and those with moderately low value-added scores.

In this study, we built on our earlier work on pathways into teaching and on our data base on New York City teachers to investigate the classroom practices associated with teachers whose students make above average gains in English/Language Arts, particularly in more challenging schools. We ask what classroom practices, if any, differentiate between teachers with high impact on student achievement in middle school English/Language Arts, as measured through value-added analyses, from teachers with lower low value-added scores? In addition, this study explores whether these kinds of value-added analyses represent a reasonable signal for differences in instructional quality and teacher effectiveness; are value-added measures, in fact, associated with observable differences in instruction?

In order to investigate the classroom practices of middle school English/Language Arts teachers, we also need tools for classroom observation that are able to capture differences in instruction that might differentiate more effective teachers. In this study, we combined the use of several dimensions of the CLASS (Pianta, Hamre, Haynes, Mintz, &La Paro, 2006), a widely used classroom observation system, with a new observation protocol we developed specifically for secondary English/Language Arts instruction. In this paper, we describe the origins of this instrument, its elements, and its use in a study of middle school ELA instruction in New York City. We also discuss the relationship between the practices targeted in our instrument and value-added assessments of teacher quality. Our goals include both identifying those elements of instruction associated with improvement in student learning and investigating empirically the extent to which value-added measures capture differences in instructional quality among teachers.

Framework for Instructional Interactions in English/Language Arts

Our study focuses specifically on instruction in English/Language Arts, as there is growing concern about adolescent literacy in this country. While students in elementary schools have shown gains in literacy, as measured by the National Assessment of Educational Progress, students at the 8th and 12th grade levels have demonstrated relatively few gains in reading (http://nces.ed.gov/nationsreportcard). Even more disturbing, the achievement gap between white and non-white students, and between higher and lower SES students, seems to be increasing (c.f. Snow & Biancarosa, 2003). Though a number of curricular and instructional programs have been aimed at improving literacy (see Snow et al. for an overview of 12 such

programs), we still know relatively little about the instructional practices that support literacy achievement during middle school, particularly in urban schools. While we know that teachers in other content areas help develop students' academic literacy, English/Language Arts teachers are the teachers who are most explicitly charged with developing students' literacy skills. We focus on instruction in ELA classrooms where we can look at relationships between instruction and student achievement in reading and writing.

Much of the existing research on middle schools has focused on the importance of school organization (e.g. blocked scheduling; interdisciplinary teams), opportunities for student choice (electives and exploratory classes) and on the creation of strong adult-student relationships through structures such as advisories (c.f. Carnegie Corporation, 1989). While school-level structures can certainly impact student experiences and a variety of outcomes, we are most interested in classroom-level instructional interactions and how these interactions predict students' academic achievement.

We focus on middle school because so little research has focused on instruction at this level. Yet we believe that middle school is a consequential time in students' academic lives (Carnegie Corporation, 1989) and that the quality of instruction students receive during this time affects their chances at success in high school and beyond. While there is a growing consensus on effective approaches to early literacy instruction (Snow, Burns & Griffin, 1998), there is much less agreement about effective literacy instruction for secondary school students, including middle grades.

Prior research has suggested multiple dimensions that characterize classroom interaction including the intellectual challenge of tasks assigned to students (Newmann, Lopez, & Bryk, 1998), the quality of instructional conversation, including teachers' uptake and elaboration of student ideas (Nystrand, 1997; O'Connor & Michaels, 1993); and representations of content, including the use of analogies or examples (Leinhardt, 2004). In addition, measures of classroom practice must capture how those teaching English/Language Arts attend specifically to the needs of English learners in their classrooms as such students make up an increasing percentage of students in classrooms nationwide (e.g. Short & Fitzsimmons, 2007). Such accommodations might include more scaffolding of tasks, support for academic language development, and targeted vocabulary instruction (Short & Fitzsimmons, 2007). Measures of classroom practice must therefore be sensitive to a range of interactions around instruction,

including the nature of the task, the focus of instruction, the features of classroom discourse, the types of accommodations provided, and the quality of feedback given to students.

While some practices associated with higher student gains in English/Language Arts may be inherently subject-specific, such as comprehension instruction, other more generic factors may also contribute to teacher effectiveness at the middle school level. Student motivation and engagement, for example, have been strongly associated with student achievement in literacy, particularly at the secondary level (Guthrie & Wigfield, 2000). A long history of research on teaching also suggests that effective teachers may be better at capturing more time for academic instruction (Denham & Lieberman, 1980) and keeping students focused on their tasks than less effective teachers. Effective teachers may have more efficient routines for transitions between activities, and better classroom management that result in more time for instruction.

Given the multifaceted nature of effective teaching, our instrument takes many of these potential dimensions of instruction into account, while focusing on ELA specific instructional interactions. While some aspects of classroom practice are likely to be general, cutting across subject areas and grade levels, other features of instruction are almost certainly domain-specific, requiring subject-specific measures (Stodolsky, 1988). One of the challenges of creating tools for measuring instruction that can be useful across a wide range of classrooms lies in addressing this balance between measuring both generic and more domain-specific elements of classroom interaction (Grosman & McDonald, 2008).

Review of Observation Protocols

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 area, examining a series of features that could be considered generic elements of teaching. For instance, Danielson's (1996/2007) *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 nature of the classroom environment (Pianta et al., 2004). CLASS focuses upon a number of key instructional dimensions that can be examined in different subject areas, such as the degree to which a teacher

helps the student understand ideas within a disciplines and the broader framework of the domain, and the degree to which a teacher provides students opportunities for analysis and higher order thinking skills, and the quality of feedback a teacher provides for students. The protocol also captures the emotional climate of the classroom and the teacher's attempts to address adolescent needs and perspectives.

Other instruments have been designed to measure teachers' understanding of best practices and/or specific subject matter knowledge in content areas. Emphasizing the importance of making mathematics accessible to students, 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 et al., 2008). In literacy, a number of observation protocols have been developed around comprehension, including systems developed by Taylor et al. (2005) at the Center for the Improvement of Early Reading Achievement out of the University of Michigan. The TEX-IN3 system (Hoffman et al., 2004) focuses specifically on the presence and use of texts in classrooms.

None of the existing observation protocols, however, provides a way to observe across the many domains of ELA classrooms, particularly at the secondary level. Both of the aforementioned literacy protocols focus specifically on one element of English Language Arts instruction—reading—and were designed for elementary classrooms. The paucity of discipline-based observation approaches has been a persistent problem in efforts to develop assessments of teaching (Kennedy, in press). Indeed, as 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 and how students are engaged with it…Documenting the intellectual meaning of teaching events remains the illusive final frontier in performance assessment" (p. 21). To that end, the PLATO instrument 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.

Development of PLATO

The Protocol for Language Arts Teaching Observation (PLATO) is based on research on effective literacy instruction at the secondary level and is designed for observations of middle and high school English/Language Arts classrooms. In its initial version, which we use for the

study presented here, PLATO was designed to supplement our use of the CLASS instrument and therefore employed the same 7-point scale referenced to 3 levels (low, medium, high).(FN-2) For each element, we developed indicators of interactions that would receive low (scores of 1 and 2), medium (scores of 3, 4, and 5), and high scores (6, 7) by raters (see attached instrument for full details). Researchers observed for 15-minute intervals, and then coded that 15-minute segment of instruction, according to both the PLATO elements and six CLASS elements described in more detail below. (FN-3)

In its first iteration, PLATO included ten elements of effective English Language Arts instruction:

- clarity of purpose of the lesson
- level of intellectual challenge in both teacher questions and tasks assigned to students
- representations of content
- connections to both personal and prior knowledge
- use of models and modeling of both high quality work and strategies for reading and writing
- presence of explicit strategy instruction in reading and writing
- use of guided practice in the classroom
- quality of feedback offered to students by both teachers and peers
- qualities of classroom discourse, including teachers' response and elaboration of student ideas
- accommodations for English Learners

The element of Purpose derives from research that suggests that children who learn 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 & Livington, 1989; Smith & Feathers, 1983). The element "Intellectual Challenge" was designed to focus upon the nature of the task and the degree to which it represents a developmentally appropriate stretch or reach for the specific students (Newmann, Lopez, & Bryk, 1998; see also Vygotsky, 1978). We also wanted to capture the ways in which teachers made content accessible to students and/or contextualized that content in terms of students prior or personal knowledge. Based upon research that suggests that teachers' content knowledge—and the ways in which they represent content to children—may affect the ways in which students learn (Lee, 1995, 2007; Stevenson

& Stigler, 2002), we included the element "Representation of Content" in order to evaluate the teachers' disciplinary knowledge, and the accuracy of the representations of the content she or he made to students during the observed segment. When teachers connect new material to students' personal experiences and prior learning, students are more likely to develop a deeper understanding and make their own connections (Bransford & Johnson, 1972; Tharp & Gallimore, 1988; Levin & Pressley, 1981). Connections to personal and/or cultural experiences are particularly relevant for ELA instruction, in which students may be asked to make such connections to a literary text. To that end, the element "Connections to Prior/Personal Knowledge" evaluates the degree to which teachers make these linkages.

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 (Frederickson & Collins, 1989; Graham, 2006; Hillocks, 1992; Knudsen, 1991). The element "Models/Modeling" captures whether there are models of quality work available in the classroom to guide student work and whether or not they are analyzed. It also assesses the extent to which a teacher names and then models specific meta-cognitive strategies or skills that she wants students to use, such as being able to figure out the meaning of a word using context clues, being able to write a strong opening paragraph for an essay, or how to edit someone else's writing. Because research also suggests 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 & McKeown, 2002; Greenleaf, Schoepenhauer, Cziko, & Mueller, 2001; Palinscar & Brown, 1987). The element of "Guided Practice," based upon research upon the role of practice with support (Vygotsky, 1978; see also Lave and Wegner,1991), evaluates the level of support that a teacher provides in the segment observed as well as a teachers' capacity to check in with his or her students about their programs, evaluate their learning, and offer any needed support.

The opportunities for extended discussion as well as the ways in which student ideas are responded to and clarified are equally important aspects of an ELA classroom. The "Feedback"

element is based upon a long line of research that suggests that feedback facilitates student learning (Thorndike, 1931/1968; Kluger & DeNisi, 1996), particularly when it is specific and targeted (Sadler, 1989; Sperling & 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 & Gamoran, 1991; O'Connor & Michaels, 1993; Taylor, Pearson, Peterson, & Rodriguez, 2003) as well as upon research on the typical discourse pattern of "initiation-response-evaluation", and a tendency to pose lower-order, rote or recall questions that do not require analytic thinking (Cazden, 2001; Mehan, 1979). Thus the "Classroom Discourse" element 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. Building upon work that suggests that introducing academic language in classrooms can help bridge students' home discourse with the language used in school (Delpit, 1988; Schleppegrell, 2004), this element also focuses upon whether or not the teacher introduces specific ELA language and concepts and the degree to which he or she supports students in using those terms.

Finally, building upon work that 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 & Hakuta, 1997), we developed the dimension of "Accommodations for language learning" in order to capture the range of strategies and supports that a teacher might use to make a lesson accessible to non-native speakers. Accommodations we wanted to capture included teachers taking into account individuals' levels of language proficiency, strategic use of primary language, grouping strategies, differentiated materials and assessments, as well as graphic organizers and visual displays.

Because content coverage can also be an important predictor of student achievement (Rowan, Correnti, & Miller, 2002), we wanted to capture different aspects of teachers' curricular focus during instruction. For this reason, PLATO also includes checklists for the major content domains within English/Language Arts including reading, writing, literature, speaking/listening, and grammar & mechanics. For each segment of instruction, observers check the domain that was the focus of instruction, and identify additional features related to that domain. For

example, if observers identify reading as the target domain, they also identify the nature of the text read (e.g. fiction or non-fiction), 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.).

Research Design

Sampling procedures

In order to select teachers for this pilot study, we began by estimating the value-added to student achievement of all New York City teachers teaching sixth, seventh or eighth grade classes that take the English Language Arts exam. 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 used one estimate that includes student effects and models gains in student achievement as a function of student fixed-effects, student time-varying characteristics, (such as whether a student changes schools), school characteristics, classroom characteristics, and year and grade indicator variables. This strategy identifies value-added by comparing teachers who teach the same students, usually in different years. Our other estimate includes student controls, school controls, classroom controls and year and grade indicator variables. 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 emprical Bayes techniques to adjust for estimation error in calculating value-added.

We then divided teachers into quartiles based on each of these two estimated value-added measures. In particular, we identified teachers in their third through fifth 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.[1] Using these samples, we

[1] We selected teachers in the second quartile rather than the lowest because we thought that there might be relatively little instruction occurring in the lowest-quartile classrooms and hence less to be learned. In addition, the differences between teachers in the top and bottom quartiles might

identified matched pairs of middle-school teachers – one moderate-performing (2nd quartile) and one high-performing (4th quartile) – teaching in the same school. We choose 12 pairs, 24 teachers, for the pilot. Neither observers nor participants knew the value added quartiles of specific teachers during any component of data collection.

We focused on third through fifth year teachers because we have been following the cohort of teachers who were in their fourth year at the time of this study since they entered teaching in 2004; from our earlier work, we know a significant amount about their preparation, entry into teaching, perceptions of school context, and early experiences in the classroom. In addition, research on teaching also suggests that by the fourth or fifth year, teachers have developed a more stable set of instructional practices, and many teachers have reached a stage where they begin to plateau in their impact on student achievement (e.g. Boyd et. al., 2006; Kane et al, 2006). However, there were not enough teachers from this single cohort that met our stringent requirements (value-added and matched in schools), so we included one cohort with an additional year of teaching experience and one cohort with one year less of teaching experience.

Table 1 describes the teachers in our sample. 83 percent of the teachers are female. Approximately 21 percent of the teachers are black, four percent are Hispanic and 54 percent are white. Half of the teachers entered teaching through traditional teacher education programs (college recommending) and a quarter through the largest alternative route in NYC, the NYC Teaching Fellows. The remainder of the teachers entered through a variety of pathways including Individual Evaluation, Teach for America, and the Temporary license process.

Data Sources

The primary data source for this study is structured classroom observation. In addition to PLATO, described above, we used 6 elements from two domains the Classroom Assessment Scoring System (CLASS) (La Para, Pianta & Stuhlman, 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 measured Positive Climate, Negative Climate, and Regard for Adolescent Perspectives. Within the domain of Classroom Organization, we included Behavior Management and Productivity. We also included the CLASS measure of Student

---

have been so striking that they might have alerted raters to a teacher's quartile during observations, potentially biasing their scoring of classroom.

Engagement. All of our observers were trained and certified in the use of CLASS prior to entering the classrooms.

We assessed inter-rater reliability on PLATO consistent with the procedures used for CLASS. We first identified target scores for each video, based on master coders, and then calculated the percentage of ratings by individuals that fell within 1 point of the target score. All observers completed training both in CLASS and in PLATO. Reliability for using PLATO was assessed using at least 5 different videos of classroom instruction across three different content domains (writing, literature, and grammar). All observers achieved at least 80% reliability on both CLASS and PLATO before observing in the field. We re-assessed inter-rater reliability during the second wave of observation and again achieved a minimum of 80% agreement on our ratings.

We observed teachers on six separate days during the spring of 2007. On each day, we observed teachers for at least two hours of instruction, generally in two different classes. The six days were divided into two waves of data collection, separated by between two to six weeks. As mentioned above, observers did not know the quartile of the teachers they were observing.

We worried, however, that using structured observation instruments alone might cause us to miss other aspects of classroom practice not included on the protocols that may in fact distinguish the highly effective teachers. To capture additional dimensions of classroom practice, including relationships among teachers and students, peer interactions in the classroom, curricular focus, 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 classrooms; one observer in the classroom took open-ended notes while the second observer used PLATO and CLASS.

To gain insights over an extended period, we supplemented six days of direct classroom observations with teacher logs based, in large part, on the Study of Instructional Improvement's teacher log for ELA for middle school classrooms (Rowan, Camburn, & Correnti, 2004). The log replicates the content categories of the PLATO, allowing us to coordinate two of our measures of instruction. We asked teachers to fill this log out for 15 consecutive days of instruction, or roughly three weeks. 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 observer's and teacher's assessment of content coverage and activity structures for the days of observation.

In addition to the log, we also collected both assignments and student work samples to get another vantage point on instructional practice. We adapted the protocols for collecting and analyzing student work developed by Newmann and his colleagues (see Table 2). Because prior research suggests more challenging assignments are positively correlated with higher student performance (Newmann, Lopez, & Bryk, 1998), we collected copies of assignments teachers designated as challenging and typical for their classroom. For each assignment, we collected the work from students that teachers considered high, average, and low performing (2 students from each group). For each teacher, we had 12 pieces of student work. We then coded qualities of the both the assignment and of the written work, using protocols adapted from Newmann and his colleagues.

Findings

The distribution of scores across observations varies for both the PLATO and the CLASS elements (See Figures 1 and 2).[2] On average, teachers received lower scores for PLATO elements across the board than on CLASS elements (see Tables 3 and 4). This result confirms other studies using CLASS alone, in which teachers tend to score higher on the domains of Emotional Climate and Classroom Organization than they do on the CLASS domain of Instructional Support. Of the PLATO measures, on average, the teachers scored highest on Purpose and lowest on ELL Accommodation. The standard deviation across teachers is approximately 1.0 for all ten 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 is reverse coded. The CLASS elements have somewhat higher variances across teachers than the PLATO elements.

[2] Of the PLATO elements, Purpose, Intellectual Challenge and Representations of Content have approximately bell-shaped distributions. Connections to Prior/Personal Knowledge has an approximately uniform distribution, except that few observations achieved the very top score. The remaining measures are positively skewed across the observations. Among the CLASS measures, Positive Climate, Adolescent Perspectives, Productivity, Student Engagement are approximately bell-shaped, while Negative Climate is positively skewed and Behavior Management is negatively skewed. See Figures 1 and 2 for distributions.

Our structured observation data suggests systematic differences between teachers in the two value-added quartiles. Figure 3 illustrates the average PLATO scores for teachers by their value-added group. Across all PLATO elements, the high value-added teachers (top quartile) scored higher on average than the low value-added teachers (2nd quartile). This difference is evident even though observers were unaware of the value-added quartile of their subjects when scoring instructional practice. Figure 4 illustrates the average scores for teachers for CLASS dimensions by their value-added group. 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 shows that the groups are statistically different on the element of Explicit Strategy Instruction ($p=0.03$) and close to statistically different on the elements of Guided Practice ($p=0.09$) and Intellectual Challenge ($p=0.13$).

The high value-added teachers also received more positive scores on each of the six CLASS elements, with the exception of Negative Climate, which is lower among this group (see Figure 4). However, only the difference in Student Engagement across the groups is significant at the .05 level, although Negative Climate is close ($p= 0.13$).

Explicit Strategy Instruction

One of the most striking findings was the association between a teacher's score on the Explicit Strategy Instruction element and his/her value-added quartile. However, many elements of PLATO are positively correlated with each other (see Tables 5 and 6), potentially confounding our understanding of the relationship between instructional elements and value added quartiles. In order to provide a somewhat clearer picture of these relationships, Table 4 provides the results of simple logits predicting value-added group as a function of PLATO scores on multiple dimensions. This set of analyses, in particular, assesses the extent to which the relationship between Explicit Strategy Instruction and teachers' value-added scores holds up to the inclusion of other measures of instruction. Each column within each of the two vertical panels of the table represents a separate model in which Explicit Strategy Instruction and one other observational measure of instruction predicts the log odds of a teacher being in the high value-added group. 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 every one 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.

As is clear from this analysis, Explicit Strategy Instruction, in particular, appears to distinguish the more effective teachers in our sample. To get a better sense of what such instruction looks like, we provide several examples from the field notes taken during open-ended observations. Teachers who scored "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 teachers' classrooms. One teacher carefully broke down the steps involved in using context clues to determine the meaning of unknown words (reading the sentence before and after the sentence at hand, substituting in multiple words etc.). Another taught students how to identify the passive voice in their own writing and then change it to the active voice. Thus these teachers made visible the often invisible process requisite for successful, sophisticated literary analysis, reading comprehension, or writing.

Content Domain Findings

In addition to collecting information on instructional practice, the observers recorded the content domain(s) covered in the lesson. Lesson segments could be coded for multiple content domains. On average 31.2 percent of the observations focused on reading, compared with 43.3 percent on writing, 38.5 percent on literature and 19.8 percent on speaking and listening. During reading instruction, we observed that the majority of the reading texts were fiction, the primary instructional focus was on comprehension rather than decoding or evaluation, and in about 23 percent of the observations, reading was done in class, either independently or in small or whole class groups.

We also noted differences in content coverage based on a teacher's value-added quartile. Within our sample, high quartile teachers are more likely to teach across content domains and are more likely to focus on writing and speaking, while low value-added teachers are more likely to focus their instruction on reading and literature. On a more detailed note, the low value-added teachers are more likely to use class time for individual reading and reading aloud, while the high value-added teachers are more likely spend time on pre-writing and having the students make oral presentations. None of these differences are statistically significant at the .05 level, which is perhaps not surprising for such a small sample.

One possibility is that the differences that we see in instruction between high and low value-added teachers (as shown above in Figures 3 and 4) is driven by differences in the content domains. Perhaps, for example, all teachers teach better when they teach writing than reading and the higher scores on PLATO and CLASS are simply driven by high value-added teachers teaching more writing. In fact, this is not the case. If anything, instructional quality appears lower in writing than in other observations. Across almost all of the 16 elements, the writing observations have less positive scores. The differences are statistically significant for Representations of Content, Connections to Personal/ Prior Knowledge, Feedback, Positive Climate, Negative Climate, Adolescent Perspectives, Behavior Management, and Productivity. Only in the element of Modeling, does writing instruction appear better than instruction in other content areas.

In contrast to lower scores during writing instruction, teachers in both quartiles scored significantly higher on the majority of elements during literature instruction. These differences are significant across the board except for modeling, explicit strategy instruction, and positive and negative climate. There were few differences in instructional quality as measured by PLATO and CLASS between observed segments of teaching reading and teaching other content domains.

Looking Across Measures: Examination of Teacher Logs and Student Work

In addition to analyzing differences in instruction, as measured by PLATO and CLASS elements, we also examined teacher logs and student work.

Teacher Logs. The teacher logs indicate some differences between teachers in the two quartiles in content coverage and grouping strategies. Mirroring the observational results, high

value-added teachers in our sample focused less on reading and more on writing and research skills than low value-added teachers. While only the reading and research differences are even close to statistical significant at usual levels (see Figure 5), the sample is small and the similarity in results between the logs and the observations suggest a trend worth further exploration. The logs also indicate differences between how teachers in the two quartiles use grouping structures (see Figure 6). Though teachers in both quartiles use independent work time for approximately the same percentage of instructional time, there were statistically significant differences between the frequencies with which teachers in the two quartiles used small group versus whole class instruction. In particular, high value-added teachers use small groups far more than low value-added teachers (36 percent compared with 16 percent), and they use large groups far less (26 percent compared with 44 percent.

Student Work: Finally, we draw on student work. We coded the work for feedback, construction of knowledge, extended writing, disciplined inquiry, value beyond school, representations of content, and scaffolding. Across all measures but one, high value-added teachers score higher than low-value added teachers, although none of the differences are statistically significant.

Beyond determining the instructional features that predicted teacher's value-added quartile, we are interested in the relationships between the different measures of instruction. Do different measures, such as structured observations, self-reported teacher logs, and student work assignments and samples pick up similar or different aspects of instruction? We found several interesting connections between the PLATO/CLASS observations and student work data. For example, teachers who provide models or rubrics to illustrate 'good' work and define what constitutes quality in a specific lesson or domain are more likely to also give students feedback on how to improve their ($r= 0.44$, $p< .05$). In a way, modeling sets the stage for the feedback teachers provide when the assignment is complete. If students have greater clarity on a teacher's expectations for their work, the teacher can tie his/her feedback to the rubric or model he/she has provided and is thus better to able to give constructive, targeted feedback on that work.

Modeling is also positively correlated with students' opportunity to construct knowledge, whether or not the assignment requires basic recall of facts or asks students to provide rationales for their analyses and interpretations ($r= 0.44$, $p< .05$). This suggests that teachers who provide a clearer sense of their expectations may be better able to push students for higher levels of

analysis and interpretation perhaps because the assignment/instruction is better structured and goal-directed. There is also the possibility, however, that teachers who provide assignments that only ask for basic recall of facts do not see the need for models or rubrics because they are looking for right or wrong answers rather than more elaborated written work that could be stronger or weaker on several different dimensions.

Interestingly, scaffolding for student work assignments is significantly correlated at the .1 level with several PLATO dimensions including feedback ($r=.39$, $p<.1$), classroom discourse ($r=.37$, $p< .1$), and modeling ($r= .37$, $p< .1$). Scaffolding looks at whether or not the assignment provides "structural supports, such as graphic organizers, multiple drafts, etc., that support further investigation and revision." It seems logical that teachers who are clearer about their expectations for an assignment, indicated by a higher score on modeling, would be better able to provide appropriate scaffolds to complete that assignment. In the same way, teachers who engage in "frequent back and forth exchanges with students about their work" and provide "specific and timely feedback that acknowledges what students did well and problems/incomplete understanding" (PLATO rubric) are the same teachers who scaffold the completion of assignments, as multiple drafts provide multiple opportunities for teacher feedback.

While our analyses of the teacher logs and student work samples demonstrate some connections across measures, they also highlight important differences among the measures. Teachers' self-reports of content coverage in the teacher logs differ significantly from our structured observations, indicating teachers perceive what they are doing differently from outside observers. For example, on days in which we observed, we noted one teacher teaching grammar over 30% of the time, and she reported teaching it only 7% of the time. While there are some interesting links between PLATO elements and features of the assignments teachers provide for students, as noted above, we observe surprisingly few connections between the features and/or quality of student work samples and PLATO or CLASS elements. This may be a result of the fact that the work samples represent the end product of instruction, while classroom observations represent the process leading up to the product. In this way, the two measures provide very different lenses on instruction. Moreover different measures collapse or disaggregate different instructional elements making comparisons across measures potentially more challenging.

Discussion

PLATO and value-added measures

This paper describes a pilot study to assess the relationship between classroom instruction and measures of teachers' value-added to student achievement based on students' test score performance. Even with the small sample used in our analysis, we find consistent evidence that high value-added teachers have a different profile of instructional practices than low value-added teachers. Teachers in the fourth (top) quartile according to value-added scores score higher than second quartile teachers on all 16 elements of instruction that we observed, and these differences are close to statistically significant in certain elements such as intellectual challenge and guided practice.

Our open-ended observation notes provide vivid illustrations of the range in the intellectual rigor of the instruction teachers provide. Instructional segments that score "low" (scores of 1 or 2) on intellectual involve 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, instruction that scored high on Intellectual Challenge (scores of 6 or 7) included students writing five paragraph essays about My Antonia, generating alternative endings to short stories, or crafting speeches from the perspective of presidential candidates. Instruction that scores low on Guided Practice either does not provide opportunities for students to practice new skills in class or allotted time for students to work independently but without sufficient support or "guidance" during class time. Lesson segments that score high on Guided Practice involve teachers circulating during literature circles answering student questions and clarifying their ideas or doing periodic whole class "check ins" as students work through stages of the writing process. In these elements, we find a clear relationship between a teacher's value-added quartile and several facets of high quality instruction.

Why is strategy instruction essential? What does it look like in practice?

The dimension of explicit strategy instruction strongly differentiates teachers at the different levels of value-added and confirms research in the area of reading comprehension on the importance of strategy instruction (Pearson & Fielding, 1991). In order to be successful on standardized assessments, the primary tool for determining a teacher's value-added score, students must consistently employ strategies for interpreting literary text, making a compelling

argument, or analyzing grammatical errors. When students understand when and how to use specific strategies, as well as why they are useful, they may be better able to use such strategies on less familiar tasks or material. In contrast, students who write editorials without a broader understanding of how to craft persuasive prose or build an effective paragraph may encounter more difficulties on the open-ended writing questions included on the majority of standardized assessments.

Unfortunately, instances of effective strategy instruction are rare, as the positively skewed distribution of scores indicates. The mean score for Explicit Strategy Instruction, 2.1, is the lowest of all the PLATO elements except ELL accommodations, and the modal score is a 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 "to anticipate an opponent's counter argument 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. In fact it is during intellectually rigorous activities just beyond students' intellectual "comfort zone" that teachers most need to equip students with specific strategies. However, the vast majority of teachers provided students with directions for completing activities, but they did not instruct them on the nuances of how to complete those activities effectively. In literature circles, students were often told to analyze a character's actions or determine the meaning of unknown words without any discussion of the strategies that would enable them to do so. Similarly, teachers highlighted the features of cinquains or editorials but did not teach students how they might approach different types of writing based on those features. Thus the goal of many lessons was completion of the specific task rather than mastering a more broadly applicable skill or strategy.

What makes writing instruction more problematic and literature less so?

High value-added teachers were also more likely to teach writing than their less effective colleagues. Yet, writing instruction seems to provide another challenge for teachers. Across quartiles, teachers' scores on both PLATO and CLASS elements went down during writing lessons. Interestingly the only PLATO element on which teachers on average scored higher during writing instruction was modeling. Writing provides the opportunity to generate a concrete model or exemplar (student work, published pieces, or teacher's own writing). Moreover

providing a model for a writing strategy (brainstorming, organizing etc.) may be more familiar for many middle school teachers than modeling reading strategies that involve more metacognition (what a teacher is thinking when he/she is reading a text or interpreting literature).

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. For example, at the conclusion to an introductory lesson on the features of poetry that highlighted a number of academic terms including simile, onomatopeia, and stanza, a teacher instructed students to "do a pre-write in poetic form." After students asked number of questions about the specifics of the assignment, she instructed them to, "write whatever [they] want about poetry- turn to a clean page in [their] journal and title it, 'Free Write: Poetry.'" A significant number of the writing assignments consisted of little more than "do nows" written on the board at the start of class.

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.

Not only were the PLATO dimensions of practice lower in writing lessons; all of the social-emotional CLASS elements scored significantly lower during writing lessons, other than Negative Climate, which was significantly higher. As noted, the observed writing lessons tended to be less structured, with students spending a great deal of time working independently. This lack of structure might make writing lessons more difficult to manage and keep productive. When students were supposed to be writing independently, we noted the teacher's focus was heavily on behavior management. Regardless of the effectiveness of these efforts, highly visible behavior management, such as circulating with a notebook grading students on behavior, seemed

to demand the vast majority of the teacher's time during writing lessons. We saw few instances in which teachers meaningfully conferenced with students about their writing; management seemed to take precedence.

In sharp contrast, teachers' scores were significantly higher scores on seven of the ten PLATO elements during literature instruction. Literature lessons tended to be more structured around comprehension questions or reading texts aloud and analyzing as they went. This inherent structure may have facilitated teachers' ability to engage in higher-level discourse with students, make connections to students' experiences and prior knowledge, and demonstrate their own content knowledge through the use of examples and/or analogies. In addition, literature lessons involved more whole class and small group instruction, which may have made them easier to manage and keep on task, while writing lessons often had students working independently.

Perhaps the general instructional challenges during writing instruction result from the fact that most secondary ELA teachers have degrees in English literature, and are thus more confident and competent with content related to literature rather than writing. English majors may be more familiar and hence more comfortable discussing theme or character in a novel than explaining the intricacies of persuasive rhetoric and teaching students how employ those techniques in their own work. Yet, New York City has invested considerable resources around the teaching of writing in recent years, so these findings regarding the challenges of writing instruction remain puzzling.

Refining the Instrument and Future Research

The improvement of classroom instruction is at the heart of improving outcomes for all students. Yet, researchers and professional educators have been hampered by the lack of common tools for measuring classroom practice. In a recent article on assessment of teacher quality in practice, Ball and 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."

Our efforts to develop a tool for classroom observations for secondary English Language Arts represent a beginning contribution to this effort.

Ultimately, we hope to create a tool that is not only useful for research on teaching, but can be used for teacher development as well. By identifying components of classroom practice that are related to student achievement, we hope to contribute to the preparation of future English teachers and the ongoing development of practicing teachers. This tool might also prove useful in clinical supervision of English teachers, as it provides a common language for discussing and analyzing classroom practice. In a field that still lacks a technical language to describe practice (Lortie, 1975), tools that help coordinate how we view classrooms and calibrate our sense of qualities of practice are sorely needed.

The tool is still far from perfect. For example, in our effort to reach across the domains of ELA, we ended up with more generic elements than we might have originally imagined. In addition, while each of the 16 elements used in this study appears to signal a higher value-added teacher, they do not necessarily reflect different features of instruction. The elements are highly correlated, particularly at the teacher level. Because of these high correlations, we cannot necessarily conclude that it is one or more particular elements that drive the relationship between the observations of instruction and value added.

Based on this pilot study, we have significantly revised this instrument, developed broader instructional factors within PLATO, collapsing, disaggregating, and eliminating elements that are highly correlated with one another, and changed the scoring from a 7 point to a 4 point scale. We recently completed using this revised version of PLATO in a follow-up study of 177 teachers in New York City middle schools to see if these initial findings hold up.

Overall, our study provides evidence that value-added measures do more than measure the characteristics of students that arrive in a teachers' classroom. They also seem to be capturing important differences in the quality of instruction. This said, the small sample size of this pilot study limits our ability to identify with precision the instructional practices that most directly impact student achievement gains. Nonetheless, our findings regarding the impact of Explicit Strategy Instruction suggests that future research can begin to develop measures of teaching that can identify the attributes of instruction that can make a difference to student achievement. It is also possible that the impact some aspects of instruction, such as the quality of classroom discourse that may be important in developing students reasoning abilities and

conceptual understanding of literature and writing, 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.

References

August, D., & Hakuta, K. (Eds.). (1997). *Improving schooling for language minority children: A research agenda.* Washington, DC: National Academic Press.

Ball, D. (1993). With an eye on the mathematical horizon: Dilemmas of teaching elementary school mathematics. *The Elementary School Journal*, *93*(4) pp. 373-397.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education, 59*(5), 389-407.

Ball, D. L., & Hill, H. C. (2009). Measuring teacher quality in practice. In D. H. Gitomer, (Ed.), *Measurement issues and assessment for teaching quality*. Thousand Oaks, CA: Sage Publications.

Barton, E. P. (2006). *What jobs require: Literacy, education, and training, 1940–2006. Policy Information Report.* Princeton: Policy Information Center, Educational Testing Service.

Beck, I. L., & McKeown, M. G. (2002). Questioning the author: Making sense of social studies. *Educational Leadership, 30,* 44-47.

Borko, H., Stecher, B., Alonzo, M., Moncure, A., Shannon, C., & McClam, S. (2005). Artifact packages for characterizing classroom practice: A pilot study. *Educational Assessment, 10*, 73-104.

Borko, H., Stecherm B., & Kuffner, K. (2007). *Using artifacts to characterize reform-oriented instruction: The Scoop notebook and rating guide. CSE Technical Report 707.* Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Borko, H., & Livingston, C. (1989). Cognition and improvisation: Differences in mathematics instruction by expert and novice teachers. *American Educational Research Journal, 26*, 473-498.

Boyd, D., Grossman, P., Lankford, H., Loeb, S. & Wyckoff, J. (2009). Teacher Preparation and Student Achievement. *Education Evaluation and Policy Analysis 31*(4).

Boyd, D., Grossman, P., Lankford, H., Loeb, S. & Wyckoff, J. (2006). How Changes in Entry Requirements Alter the Teacher Workforce and Affect Student Achievement" *Education Finance and Policy 1*(2).

Bransford, J., & Johnson, M. K. (1972). Contextual prerequisites for understanding: some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior. 11*(6): 717-726.

Brophy, J., & Good, T. (1974). *Teacher-student relationships: Causes and consequences.* New York: Holt, Rinehart and Winston.

Brophy, J., & Good, T. (2000). *Looking in classrooms*. New York: Longman.

Bryk, A., Kerbow, D., Pinnell, G.S., Rodgers, E., Hung, C., Scharer, P.L., Fountas, I., Dexter, E. (2008). Measuring change in the practice of literacy teachers. Under review.

Camburn, E. & Barnes, C.A. (2004). Assessing the validity of a language arts instruction log through triangulation. *The Elementary School Journal, 105,* 49-73

Cazden, C. (2001). *Classroom discourse: The language of teaching*. New York: Heinemann.

CIERA Update. Retrieved March 23, 2009 from http://www.ciera.org/index.html.

Cuban, L. (2007). Hugging the middle teaching in an era of testing and accountability, 1980–2005. *Education Policy Analysis Archives, 15*, 1-29.

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

Delpit, L. (1988). The Silenced Dialogue: Power and Pedagogy in Educating Other People's Children. *Harvard Educational Review, 58,* 280-298.

Denham, C. & Lieberman, A. (Eds.). (1980). *Time to learn.* Washington, DC: National Institute of Education.

Durkin, D. (1978-79). What classroom observations reveal about reading comprehension instruction. *Reading Research Quarterly, 15,* 481-533.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher 18,* 27-32.

Graham, S. (2006). Strategy instruction and the teaching of writing: A meta-analysis. In C. MacArthur, S. Graham, & J. Fitzgerald, (Eds.), *Handbook of writing research* (pp. 187–207). New York: Guilford.

Greenleaf, C. L., Schoenbach, R., Cziko, C., & Mueller, F. L. (2001). Apprenticing adolescent readers to academic literacy. *Harvard Education Review, 71,* 79-129.

Guthrie, J. T., & Wigfield, A. (2000). Engagement and motivation in reading. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr. (Eds.) *Handbook of reading research, Volume III* (403-422). Malwah, NJ: Lawrence Erlbaum Associates.

Hamre, B. K., & Pianta, R. C. (2001). Early teacher-child relationships and the trajectory of children's school outcomes through eighth grade. *Child Development 72* (2), 625–638.

Harms, T., Clifford, R. M., & Cryer, D. (2005). *Early childhood environment rating scale (Rev. ed).* New York: Teachers College Press.

Heneman, H.G. III., Milanowski, A., Kimball, S. M., & Odden, A. (2006). *Standards-based teacher evaluation as a foundation for knowledge- and skill-based pay. CPRE Policy*

Briefs, RB-45. Philadelphia: University of Pennsylvania, Consortium for Policy Research in Education.

Hill, H. (2005). Content across communities: Validating measures of elementary mathematics instruction. *Educational Policy, 19*, 447-475.

Hillocks, G. (2000). *Teaching writing as reflective process*. New York: Teachers College Press.

Hoffman**,** J.V., Sailors, M., & Duffy, G. (2004). The effective elementary classroom literacy environment: examining the validity of the TEX-IN3 observation system" *Journal of Literacy Research, 36,* 303-334.

Hoffman, J. V., Roller, C. Maloch, B., Sailors, M., Duffy, G., & Beretvas, S. N. (2005). Teachers' preparation to teach reading and their experiences and practices in the first three years of teaching. *Elementary School Journal, 105*, 267-288.

Horizon Research, Inc. (2005–06). *Core Evaluation Manual: Classroom Observation Protocol.* September 2005.

Kennedy, M. M. (in press). Recognizing good teaching when we see it. To appear in *Handbook of Teacher Assessment and Teacher Quality*. M. Kennedy (Ed.) San Francisco, Jossey-Bass.

Kluger, A. N. & A. DeNisi (1996). Effects of feedback intervention on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119,* 254-284.

Knudson, R. E. (1991). Effects of instructional strategies, grade, and sex on students' persuasive writing. *Journal of Experimental Education, 59,* 141–152.

Lave, J. & E. Wenger (1991). *Situated Learning: Legitimate peripheral participation*. Cambridge, UK; New York; Melbourne, Australia: Cambridge University Press.

Lee, C. (2007). *The role of culture in academic literacies: Conducting our blooming in the midst of the whirlwind*. New York: Teachers College Press.

Lee, C. (1995). A culturally based cognitive apprenticeship: Teaching African American high school students skills in literary interpretation. *Reading Research Quarterly 30*, 608-628.

Levin, J. R. & M. Pressley (1981). Improving children's prose comprehension: Selected strategies that seem to succeed. In C. M. Santa & B. L. Haes, (Eds.), *Children's prose comprehension: Research and practice*, (pp. 44-71). Newark, DE: International Reading Association**.**

Lortie, D. (1975). *Schoolteacher: A sociological study*. Chicago, University of Chicago Press.

Mathes, P.G. & Torgesen, J.K. (1998). All children can learn to read: critical care for the prevention of reading failure." *Peabody Journal of Education*, Vol. 73, No. 3/4.

Mehan, H. (1979). "What time is it Denise?" Some observations on the organization and consequences of asking known information questions in classroom discourse. *Theory into Practice, 18,* 285-292.

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.

Palinscar, A. S. (2003). Collaborative approaches to comprehension instruction. In A. S. Sweet & C. E. Snow (Eds.), *Rethinking reading comprehension* (pp. 99-114). New York: Guilford Press.

Palinscar, A. & A. Brown (1987). Enhancing instructional time through attention to metacognition. *Journal of Learning Disabilities, 20,* 66-75.

Pearson, P. D. & Fielding, L.. (1991). Comprehension instruction. In Barr, Pearson & Kamil (Eds.), *Handbook of reading research,* (pp. 815-860). New York: MacMillan Press.

Pianta, R. C., LaParo, K. M., Stuhlman, M. (2004). The Classroom Assessment Scoring System: Findings from the Prekindergarten Year, *The Elementary School Journal, 104,* 409-426 .

Pianta, R. C., Hamre, B. K., Haynes, N. J, Mintz, S., La Paro, K. M. (2006). *CLASS Classroom Assessment Scoring System: Manual Middle Secondary Version Pilot*, June 2006.

Pianta, R., Belsky, J., Houts, R., Morrison, F., & the NICHD ECCRN. (2007). Opportunities to Learn in America's Elementary Classrooms. *Science, 315,* 1795–1796.

Piburn, M. & Sawada, D. (2000). *Reformed Teaching Observation Protocol (RTOP): Reference manual ACEPT technical report no. IN00-3*. Arizona Collaborative for Excellence in the Preparation of Teachers.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73,* 417-458.

Rockoff, J. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review, Papers and Proceedings, 94,* 247-252.

Rowan, B., Correnti, R., & 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,* 1525-1567.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems.*Instructional Science, 18,* 119-144.

Saginor, N. (2008). *Diagnostic classroom observation: moving beyond best practice*. Thousand Oaks, CA, Corwin Press.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee value-added assessment system (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8,* 299-311.

Sanders, W. L., & Rivers, J. C. (1996). *Research project report: cumulative and residual effects of teachers on future student academic achievement*. University of Tennessee Value-Added Research and Assessment Center. Retrieved from http://www.mdk12.org/practices/ensure/tva/tva_2.html

Schleppegrell, M. (2004). *The Language of Schooling: A functional linguistics perspective.* Mahwah, NJ: Lawrence Erlbaum Associates.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4-14.

Sizer, T. R. (1984). *Horace's compromise : the dilemma of the American high school :*

the first report from a study of high schools, co-sponsored by the National Association of Secondary School Principals and the Commission on Educational Issues of the National Association of Independent Schools. Boston: Houghton Mifflin.

Smith, F. R., & Feathers, K. M. (1983). The role of reading in content classrooms: Assumption vs. reality. *Journal of Reading*, 27: 262-267.

Snow, C. E., & Biancarosa, G. (2003). *Adolescent literacy and the achievement gap: What do we know and where do we go from here? Report of the Adolescent Literacy Funders Meetin*g. New York: Carnegie Corporation of New York. Retrieved from http://www.all4ed.org/resources/CarnegieAdolescentLiteracyReport.pdf.

Snow, C. E., Burns, S. M., & Griffin, P. (Eds.) (1998). *Preventing reading difficulties in young children.* Washington DC: National Academy Press.

Sperling, M., & Freedman, S. W. (2001). Research on Writing. *Handbook of research on teaching*, 4th Edition (pp. 370-389). Washington DC: American Educational Research Association.

Stein, M. K., & Matsumura, L. C. (2009). Measuring instruction for teacher learning. In D. Gitomer, (Ed.), *Measurement issues and assessment for teaching quality* (pp.179-206). Thousand Oaks, CA., Sage Publications.

Sterbinsky, A., & Ross, S. (2003). School observation measure reliability study. Memphis: Center for Research in Education Policy.

Taylor, B. M., Pearson, D. P., Peterson, D. S., Rodriguez M. C. (2003). Reading growth in high-poverty classrooms: The influence of teacher practices that encourage cognitive engagement in literacy learning. *Elementary School Journal, 104*, 3-28.

Taylor, B. M., Pearson, D. P., Peterson, D. S., Rodriguez M.C. (2005). The CIERA School Change Framework: An evidence-based approach to professional development and school reading improvement. *Elementary School Journal*. *104,* 3-28. Reading Research Quarterly Vol. 40, No. 1 January/February/March 2005, (pp. 40–69).

Tharp, R. G. & R. Gallimore (1988). *Rousing Minds to Life: Teaching, learning and schooling in social context.* Cambridge, UK, Cambridge University Press.

Thorndike, E. L. (1931/1968). *Human Learning*. New York: The Century Co.

Vygotsky, L. S. (1978). *Mind in Society*. Cambridge, England: Cambridge University Press.

Weade, G., & Evertson, C.M. (1991). On what can be learned by observing teaching. *Theory into Practice, 30*, Educational Evaluation: An Evolving Field (Winter, 1991), pp. 37-45.

Figures and Tables

Table 1: Teachers in the Sample

Variable	Mean	Variable	Mean
Female	0.833	Pathway – College Recommended	0.542
Race/Ethnicity - Black	0.208	Pathway – Individual Evaluation	0.042
Race/Ethnicity - Hispanic	0.041	Pathway – NYC Teaching Fellows	0.250
Race/Ethnicity - White	0.541	Pathway – Teach For America	0.042
Year of Birth	1975.1 (5.5)	Pathway - Temporary	0.042
General Knowledge Exam	254.6 (24.2)	SAT Math (n=13)	458(70.7)
		SAT Verbal (n=13)	501(77.5)

Table 2: Measures of assignments, adopted from the Consortium for Chicago School Reform's Manual of Scoring Tasks and Student Work in Writing, 1998.

	1	2	3	4	
Construction of Knowledge	Assignment asks for simple recall of basic facts.	Assignment asks for students to recall facts and give their interpretation.	Assignment asks students to interpret or analyze and push for rationales *or* an understanding of different perspectives.	Assignment asks students to give rationales for their analysis and interpretations, *including* an understanding of different perspectives.	*Insufficient Evidence*
Extended Writing	Assignment asks for minimal writing, such as multiple-choice or fill in the blank	Assignment asks students to produce individual sentences.	Assignment asks students to produce paragraphs	Assignment asks students to produce connected paragraphs	*Insufficient Evidence*
Disciplined Inquiry: Elaborated Written Communication *The student work creates an argument, makes a generalization, or draws a conclusion and supports it with evidence.*	The assignment requires no elaborated written communication.	The student is asked to either: A) create an argument, make a generalization, or to draw a conclusion or B) to provide evidence supporting such an elaboration.	The student is asked to draw conclusions or to make generalizations ***or*** arguments.	The student is asked to draw conclusions or to make generalizations ***or*** arguments *and* to support them with evidence or illustration.	*Insufficient Evidence*
Value Beyond School	Assignment does not appear to be connect to adolescent lives	Assignment is connected to student interests, but does not support learning for future life	The skill ***or*** content of the assignment is connected to students' lives and helps prepare them for issues they may face beyond school	The skill ***and*** content of the assignment is connected to students' lives and helps prepare them for issues they may face beyond school	*Insufficient Evidence*
Representations of Content: Depth and Accuracy	Assignment misrepresents the content being studied.	Assignment does not cover all aspects of the content being studied.	Assignment covers all aspects at a surface level and doesn't provide opportunities for students to probe.	Assignment covers all relevant aspects of the content being studied ***and*** provides opportunities for students to probe in depth into the content.	*Insufficient Evidence*
Scaffolding	Assignment lacks necessary supports to help students be successful.	Assignment offers few supports in the way of graphic organizers ***or*** the use of multiple drafts.	Assignment offers structural supports, such as graphic organizers, multiple drafts, etc.	Assignment offers structural supports, such as graphic organizers, multiple drafts, etc., that support further investigation and revision.	

Table 3: Average PLATO Scores Across Teachers (out of 7 point scale)

Element	Mean	Std. Dev.
Purpose	4.067	(0.99)
Intellectual Challenge	3.671	(1.05)
Representations of Content	4.002	(0.85)
Connections to Prior/Personal Knowledge	3.425	(1.25)
Modeling	2.597	(0.81)
Explicit Strategy Instruction	2.151	(0.80)
Guided Practice	2.918	(0.96)
Feedback	2.984	(1.14)
Classroom Discourse	2.985	(1.15)
ELL Accommodation	1.679	(1.01)

Table 4: Average CLASS Scores Across Teachers (on a 7 point scale)

Dimension	Mean	Std. Dev.
Positive Climate	4.417	(1.31)
Negative Climate	2.109	(1.14)
Regard for Adolescent Perspectives	3.449	(1.08)
Behavior Management	4.591	(1.60)
Productivity	4.308	(1.34)
Student Engagement	4.267	(1.26)

Table 5: Effect of Explicit Strategy Instruction and another PLATO Element or CLASS Dimension on the Odds Ratios Predicting High Value-Added Quartile

Explicit Strategy Instruction	4.88 (0.070)	3.31 (0.146)	3.47 (0.108)	5.91 (0.058)	10.17 (0.065)	3.16 (0.171)	3.76 (0.089)
Purpose	0.76 (0.643)						
Intellectual Challenge		1.24 (0.688)					
Representations of Content			1.28 (0.699)				
Connections to Prior and Personal Knowledge				0.68 (0.410)			
Modeling					0.33 (0.308)		
Guided Practice						1.37 (0.613)	
Feedback							1.06 (0.891)
Explicit Strategy Instruction	3.54 (0.099)	3.35 (0.107)	3.06 (0.138)	3.42 (0.091)	3.37 (0.11)	3.38 (0.115)	3.35 (0.117)
Classroom Discourse	1.17 (0.718)						
Positive Climate		1.49 (0.94)					
Negative Climate			0.611 (0.405)				
Regard for Adolescent Perspectives				1.33 (0.542)			
Behavior Management					1.23 (0.523)		
Productivity						1.25 (0.577)	
Student Engagement							1.54 (0.337)

Figure 1: Distribution of Scores for each of the PLATO Elements across Value Added Quartiles

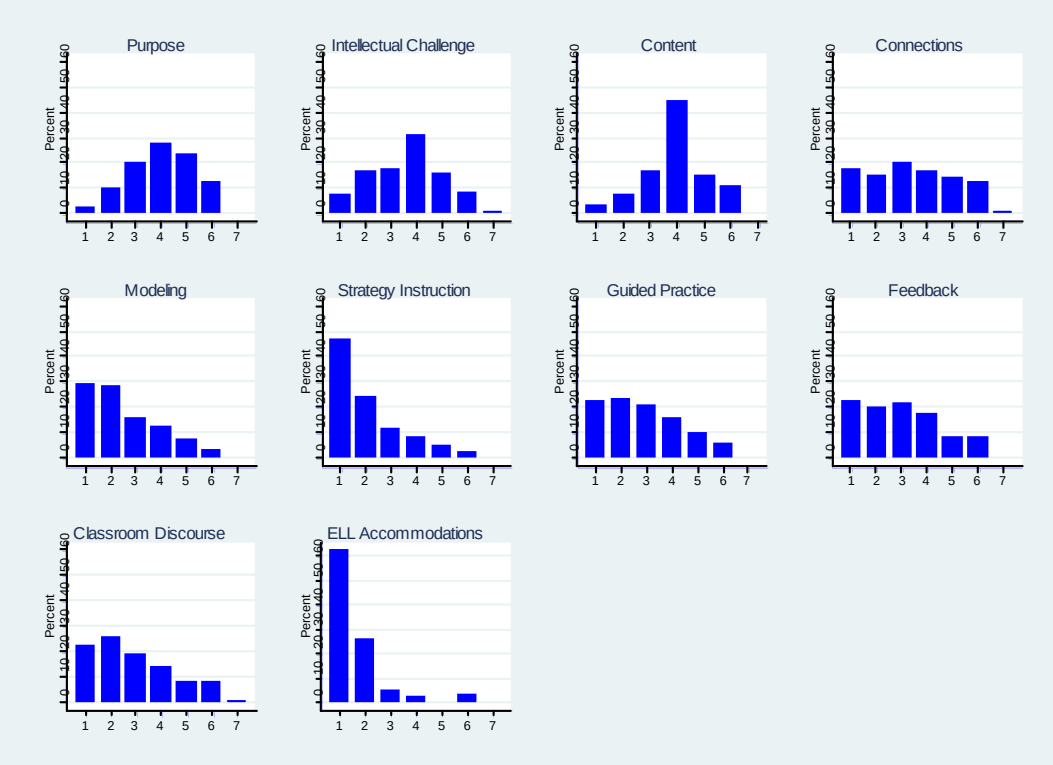


Figure 2: Distribution of Scores for each of the CLASS Dimensions across Value Added Quartiles

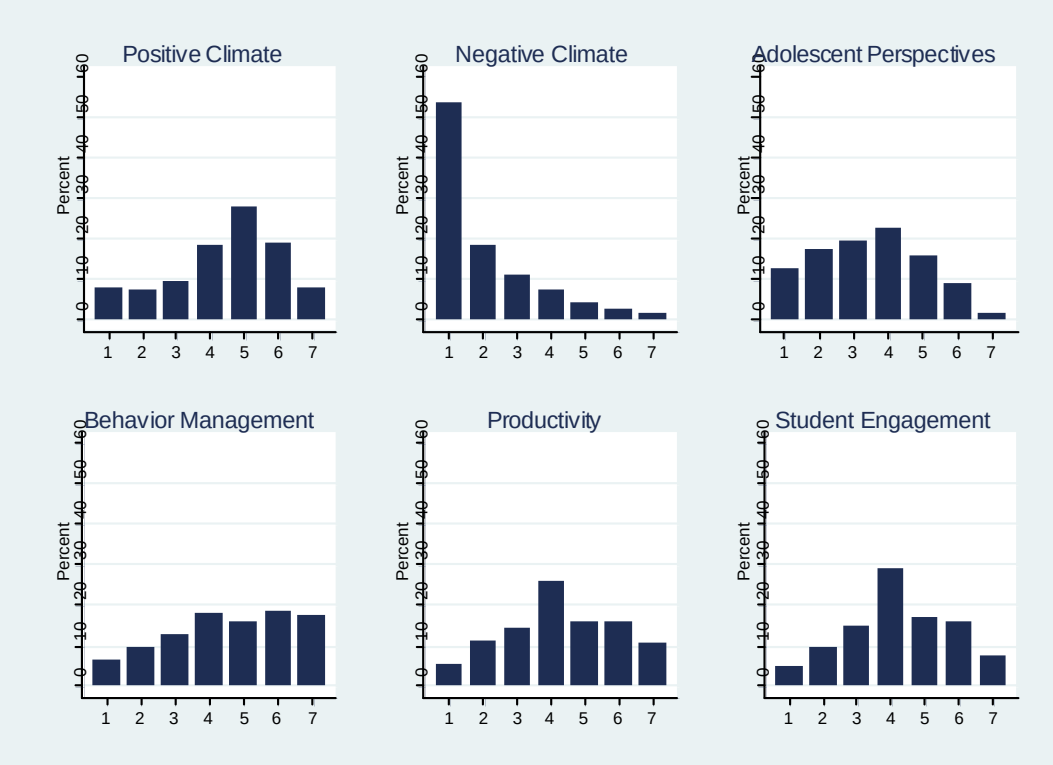


Figure 3: Average Score by Value Added Quartile for each of the PLATO Elements

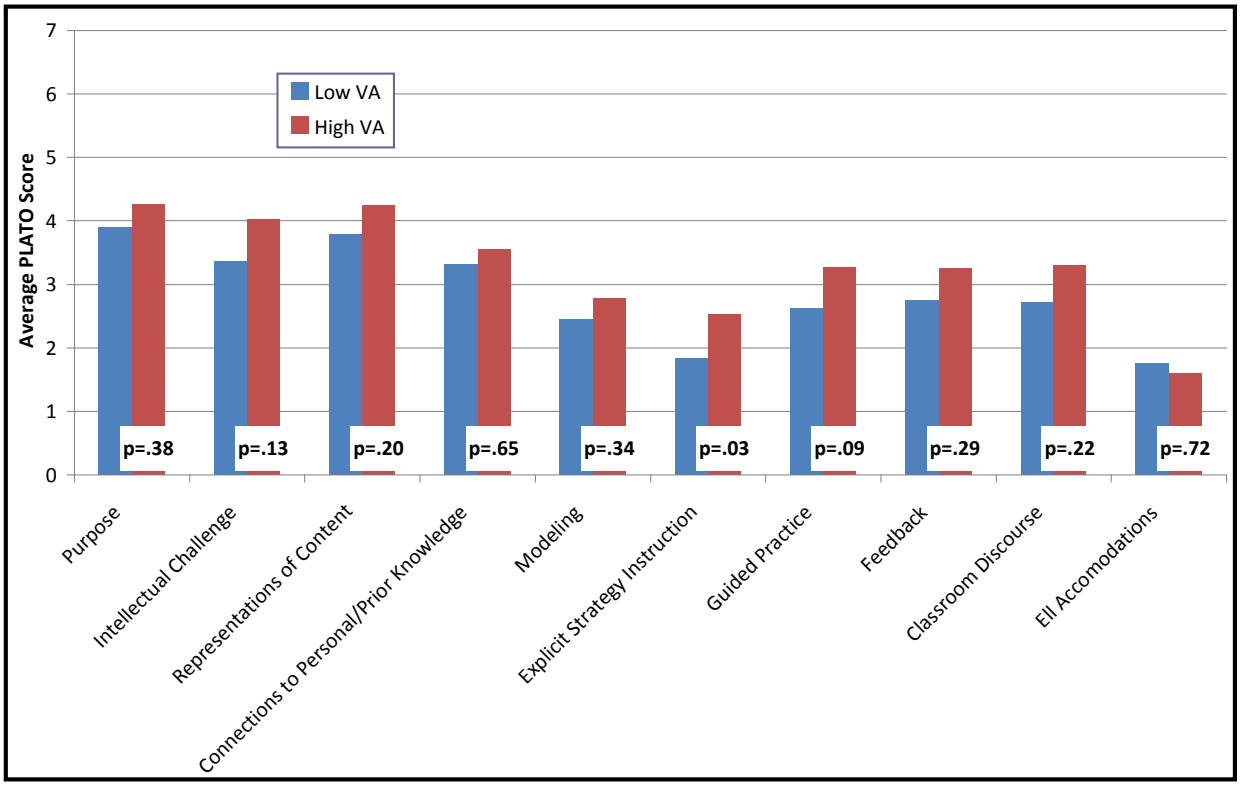


Figure 4: Average Score by Value Added Quartile for each of the CLASS Dimensions

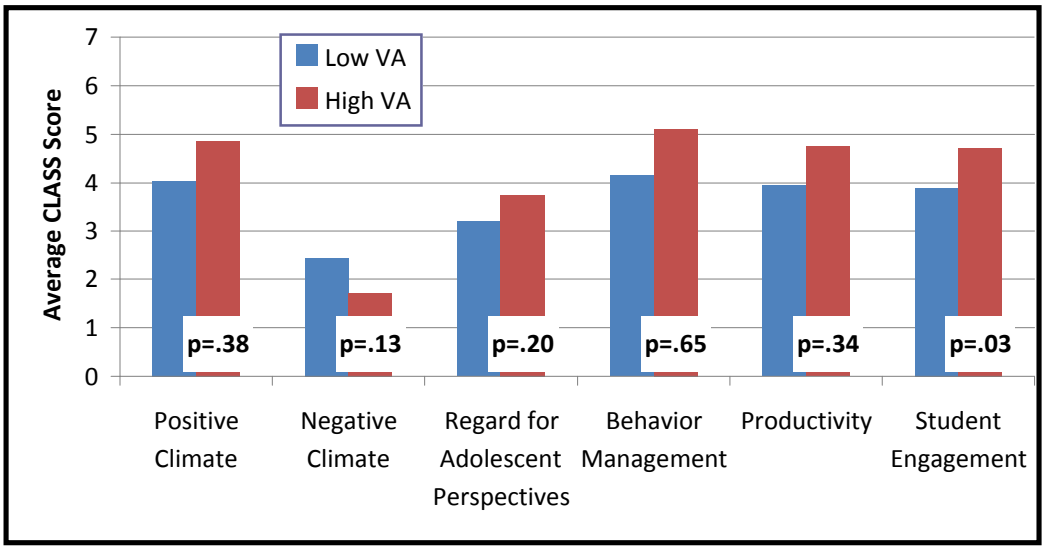
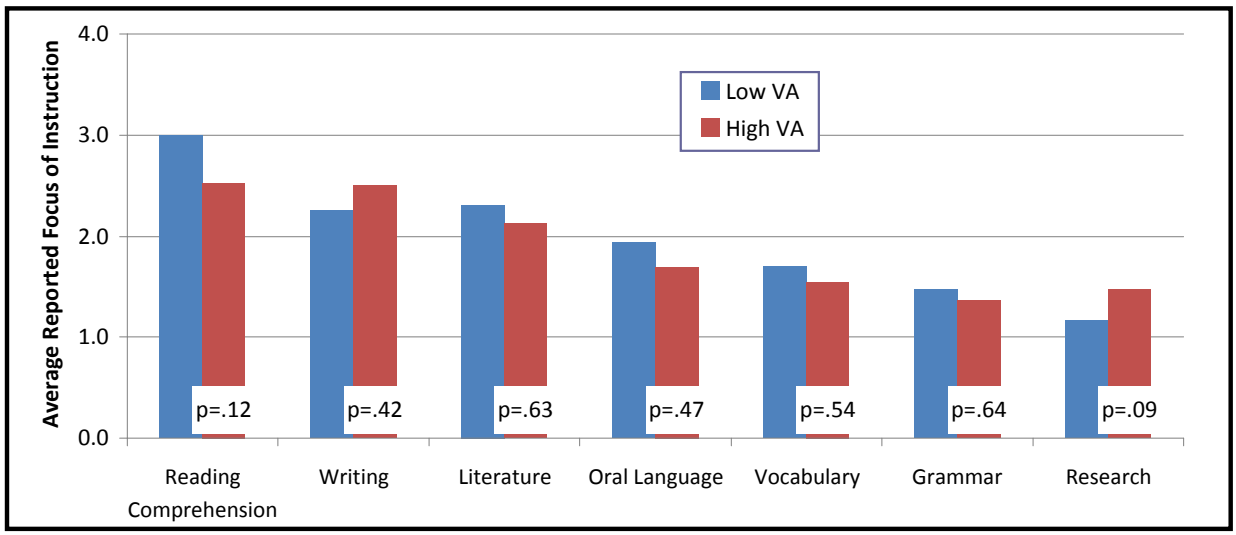


Figure 5: Content Domain Focus As Reported in Teacher Logs By Value Added Quartile



Focus of Instruction Input in Teacher Logs
1= not taught at all, 2= touched on briefly, 3= minor focus of instruction, 4= major focus of instruction

Figure 6: Use of Small and Large Groups As Reported in Teacher Logs By Value Added Quartile