Journal of Education for Students Placed at Risk (JESPAR)
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hjsp20

What Works to Improve Student Literacy Achievement? An Examination of Instructional Practices in a Balanced Literacy Approach

Catherine Bitter [a], Jennifer O'Day [a], Paul Gubbins [a] & Miguel Socias [a]

[a] American Institutes for Research (AIR),
Published online: 10 Feb 2009.

To cite this article: Catherine Bitter, Jennifer O'Day, Paul Gubbins & Miguel Socias (2009) What Works to Improve Student Literacy Achievement? An Examination of Instructional Practices in a Balanced Literacy Approach, Journal of Education for Students Placed at Risk (JESPAR), 14:1, 17-44, DOI: 10.1080/10824660802715403

To link to this article: http://dx.doi.org/10.1080/10824660802715403

Journal of Education for Students Placed at Risk, 14:17–44, 2009

ISSN: 1082-4669 print/1532-7671 online
DOI: 10.1080/10824660802715403

Routledge
Taylor & Francis Group

What Works to Improve Student Literacy Achievement? An Examination of Instructional Practices in a Balanced Literacy Approach

Catherine Bitter, Jennifer O'Day, Paul Gubbins, and Miguel Socias

American Institutes for Research (AIR)

A core assumption of the San Diego City Schools (SDCS) reform effort was that improved instructional practices, aligned with a balanced literacy approach, would be effective in improving student outcomes. This article explores this hypothesis by presenting findings from an analysis of classroom instruction data collected in 101 classrooms in 9 high-poverty elementary schools. Data were collected using a literacy observation tool adapted from prior research. The study found a prevalent focus on reading comprehension instruction and on students' active engagement in making meaning from text. Teachers' use of higher-level questions and discussion about text were substantially higher than that found by a prior study using the same instrument in similar classrooms elsewhere. Hierarchical Linear Modeling analyses of instruction and student outcome data indicate that teacher practices related to the higher-level meaning of text, writing instruction, and strategies for accountable talk were associated with growth in students' reading comprehension.

Literacy is the foundation for success in school and, in today's society, for success in life. Helping children become independent readers and writers is, thus, a central goal not only of elementary educators, but also of district and school improvement efforts across the nation. Even before the passage of the No Child Left Behind (NCLB) Act of 2001, many urban districts were focusing on strategies to improve reading achievement and reduce learning gaps among the students in their charge. Since NCLB, with its stringent accountability measures and massive Reading First initiative, that focus has only sharpened. But what are the best strategies for achieving this goal, particularly at scale? And what needs to change in classroom instruction to foster reading growth for all students? This article addresses the second of these questions, using fine-grained observation data on classroom literacy instruction linked to student achievement collected in the course of a 3-year study of the implementation and effects of San Diego's instructional reforms under the leadership of Alan Bersin and Anthony Alvarado.

San Diego is an appropriate venue for an investigation of instructional effects in literacy. At the heart of the San Diego reform effort was the assumption that, to improve student learning, teachers' instructional practice must improve. San Diego City Schools (SDCS) put this assumption into practice by systemically implementing specific instructional strategies aligned with a

Request for reprint should be sent to: Jennifer O'Day, American Institutes for Research, 1070 Arastradero Road, Suite 200, Palo Alto, CA 94304. E-mail: jo'day@air.org.

balanced literacy framework. San Diego's underlying theory of change hypothesized that these instructional practices would be effective in improving student outcomes.

As part of our larger theory-based evaluation of San Diego's reform efforts, we sought to directly test this hypothesized link between instruction and student outcomes. Most other studies of San Diego's reforms had considered instruction only marginally, through the reports of school staff, while focusing more closely on other aspects of the reform, such as instructional leadership and teachers' reactions to the reforms. Prior studies had not closely examined the instructional practices of San Diego teachers and how these practices have influenced students' achievement. Starting this study several years into the reform allowed us (a) to observe which aspects of the reform had taken hold in the classroom, and (b) to assess the effects these practices had on students' performance. Two primary research questions guided our investigation of literacy instruction:

- To what extent was classroom literacy instruction consistent with the instructional approach that San Diego hypothesized as effective?
- To what extent are these literacy instructional practices associated with increased student literacy achievement?

Although our conceptual framework for the evaluation was based on the core elements of San Diego's instructional reform and relationships among these elements, we utilized constructs identified in prior research and selected an instrument that had been used and validated in prior studies. This link to prior research enabled us to compare our findings to those of studies in other contexts and generalize our findings based on the theory underlying the reform effort.

In this article, we explore the underlying assumptions and approach of the literacy reforms in San Diego and outline our methodology for addressing our central research questions. We provide evidence that teachers' literacy instruction did, indeed, reflect the tenets of San Diego's balanced literacy reforms and that several of these instructional practices were significantly associated with student growth in reading comprehension.

LITERACY INSTRUCTION IN SAN DIEGO: THE RESEARCH BASE

To ground our study in the theory of change in San Diego, we first had to define San Diego's approach to instruction in relation to the literature base on literacy instruction. According to Cohen and Ball (1999), the heart of instruction takes place in the classroom and is situated in the interaction of teachers and students around content or instructional materials. Although instructional content is an important component of the instructional process and one that we explored through other aspects of our study (e.g., surveys and interviews), this article focuses on our efforts to measure primarily the pedagogical practices relevant to literacy. That is, we focus here on the ways in which teachers interacted with their students around content, rather than on the content itself. This approach to the study of instruction views both students and teachers as active participants in the learning process (e.g., Cohen, 1996).

San Diego's balanced literacy approach was multifaceted and supported by prior research. First, the San Diego approach heavily emphasized reading comprehension and students' creation of meaning through active interaction with text. The importance of student engagement with text has been recognized in earlier studies. Several researchers have found that readers who are

cognitively engaged in literacy activities have better comprehension and reading achievement than those who are less engaged (Campbell, Voelkl, & Donahue, 1997; Cunningham & Stanovich, 1998). One reason for this is that meaning exists not in the text itself, but rather must be created by the reader interacting with that text (RAND, 2002; Resnick, 1987). Meaning creation has also been a theme in similar research on mathematics instruction and learning (Schoenfeld, 1988, 1998).

The conception of students as active participants in the learning process was central to the literacy framework employed by SDCS. One manifestation of this was the emphasis on promoting *accountable talk* among students to motivate student engagement, facilitate connections between the text and students' prior knowledge, and engender meaningful interplay among readers. Accountable talk is an approach to engagement with text that seeks to foster student responsibility; interactive learning; and sustained, idea-focused, evidence-based discourse.

Another strategy valued by SDCS to foster student responsibility in learning was the incorporation of methods that engage students in meaning creation at multiple levels of their literacy competence. This approach aims to gradually release responsibility to students by utilizing instructional methods that range from a high level of teacher control and modeling (e.g., through a teacher reading aloud) to a high level of student independence (e.g., through independent reading). Additional instructional strategies that support this release of responsibility included shared reading, in which the teacher and students, together, read text at a level of cognitive demand that is higher than students could tackle on their own, and guided reading, in which the teacher and a small group of students closely read a text, focusing on an area of need for these students.

The need for students to actively engage in instructional activities for deep learning to take place has important implications for the role of teachers in the instructional process. As students are expected to take more and more responsibility for their own learning, the teacher's role is to encourage and scaffold the learning process. Teachers thus become facilitators and collaborators in learning, rather than its controllers. Empirical research has lent support to this conception of teaching for the promotion of literacy competence. Duffy and colleagues (1987) and Roehler and Duffy (1984) found, for example, that effective teachers of reading used modeling and explicit instruction in comprehension strategies, and Chinn, Anderson, and Waggoner (2001) observed that literacy instruction emphasizing collaborative reasoning resulted in greater student engagement and higher-level thinking. In a study of 140 high-poverty classrooms, Knapp and his colleagues (1995) discovered that teachers of reading who stressed meaning creation and higher-level thinking skills (in addition to lower-level skills) were more effective in increasing student achievement than others. This conception of the teacher as a facilitator of the learning process is also consistent with other research on learning (e.g., the National Research Council panel on How People Learn; Bransford, Brown, & Cocking, 2000) and with the literacy and mathematics frameworks developed and used by SDCS.

METHODS

To measure elementary literacy instruction in San Diego, we conducted an intensive longitudinal classroom observation study across 2 years. Coupled with interviews to tap into teacher beliefs about literacy instruction, their professional development experiences, and the school context, we considered the observation-based study to be a strong approach to capturing the depth of

implementation of the SDCS literacy framework and to linking variations in that implementation to variations in student achievement.

Observation Protocol

To measure instruction through observations, we built on a protocol used by Taylor, Pearson, Peterson, and Rodriguez (2003) in their Center for the Improvement of Early Reading Achievement (CIERA) School Change study. This study examined instruction in nine high-poverty elementary schools participating in a reading improvement program and measured the effects of specific instructional practices on students' reading achievement.

We chose a protocol that had been used and validated in a prior study for several reasons. First, the scheme had been used effectively to link instructional practices to outcomes, which was the relationship we were most interested in measuring. In addition, using this scheme provided a point of comparison for results and allowed us to generalize our findings beyond San Diego. In examining San Diego's reforms, it was important for us to consider how different (or not different) instruction in San Diego was relative to similar districts and schools. Taylor et al.'s study (2003) included a sample of high-poverty schools from across the country, and thus served as a reasonable, though rough, comparison.[1]

We chose to base our observation study specifically on the work of Taylor et al. (2003) in part because they grounded their instrument in the literature on literacy learning and teaching discussed earlier. These authors posited a framework that delineates four dimensions of literacy instruction to maximize cognitive engagement in literacy learning:

1. supporting higher-level thinking (in both talk and writing about text);
2. encouraging independent use of word recognition and comprehension strategies during reading activities (both instruction and text reading);
3. using a student support stance (in contrast to a teacher-directed stance); and
4. promoting active, as opposed to passive, involvement in literacy activities. (pp. 5–6)

These dimensions are consistent with SDCS's balanced literacy approach, and thus we reasoned that this tool would enable us to test key aspects of the district's literacy framework and its hypothesized relationship with student outcomes. For example, this tool included codes to describe activities related to making meaning from text. It included one code to denote *lower-level* comprehension instruction (i.e., questioning, discussion, and writing that focuses on understanding information explicitly stated in the text) and another for *higher-level* comprehension instruction (i.e., instruction that requires students to go beyond what the author has explicitly written to interpret or apply the text's meaning). In other words, teachers' lower-level comprehension questions focus on the literal meaning of text, and higher-level questions ask students to generalize, evaluate, or aesthetically respond to a text.

In addition to specific instructional activities, the scheme included codes to describe teacher and student interactions. Taylor et al. (2003) designated certain interaction styles as reflecting a

[1]In these schools (Taylor et al. 2003), 70% to 95% of students qualified for subsidized lunch. In addition, across the schools, 2% to 68% of students were nonnative English speakers. These demographics were similar to those of our San Diego sample, although we had slightly more variation in poverty levels and higher percentages of English learners overall (as described in Data Collection Procedures).

''student support stance'' (e.g., coaching; p. 6), meaning that these interaction styles provide support for students to engage actively with the text and to help students transfer relevant skills or knowledge from one literacy encounter to another. For instance, if a teacher asked, ''What do you predict will happen in this story?'' coaching might involve the provision of prompts to enable students to form predictions in a meaningful way, such as, ''What do you see on the cover of the book?'' or ''What does the title mean to you?'' Other codes representing student support and active response included teachers' modeling or providing feedback, as well as students' active reading and discussion.

Beyond the dimensions and codes that are well aligned with San Diego's theory of literacy instruction, the observation scheme also captured a wide range of other literacy practices commonly found in classrooms across the country. In other words, rather than measuring only, or mainly, those activities hypothesized to be effective by the San Diego approach, the coding scheme also captured practices emphasized in other approaches to literacy instruction, and, as a result, was unbiased toward any particular theory of literacy instruction. For example, in addition to comprehension-related codes, the scheme included four different codes for phonics instruction, as well as codes for teaching phonemic awareness, sight words, or grammatical structures. With respect to interaction patterns, the scheme included codes to capture more teacher-directed instruction, such as telling, as well as those for coaching or conversation. We considered telling to occur when teachers simply gave information to students. To illustrate, an example of telling would be a teacher explaining to students, ''The book you are going to read is a nonfiction book about sea creatures. It has a glossary at the end to help you with words you do not understand.''

An important benefit of the Taylor et al. (2003) instrument was that it had been used effectively to detect consistent significant relationships between specific aspects of literacy instruction (notably higher-level questioning about the meaning of text and teachers' scaffolding of student learning) and achievement gains in literacy (both reading and writing). With such a fine-grained and detailed technique, the scheme could capture variation in practices that mattered across teachers (again, positively or negatively) and relate them to outcomes.

Finally, we chose this scheme because the coding focused on the observable behavior of teachers and students, and was thus fairly low on the inferential scale. Validity and reliability were further enhanced by combining qualitative scripting of what took place in the classroom and quantitative coding of the same data. This allowed both the data collector and others to monitor coding accuracy and reliability across observers.

Our refined observation coding scheme for observed literacy instructional behaviors contained over 80 codes divided into 7 categories. The categories of codes included:

Level 1: Person providing instruction during the specified time segment (e.g., classroom teacher, aide);
Level 2: Instructional groupings used (e.g., small group, pairs, whole group);
Level 3: Major academic areas covered (e.g., reading, writing);
Level 4: Materials used (e.g., nonfiction text, big book, worksheet, overhead projector);
Level 5: Specific literacy instructional activities observed (e.g., use of questions or discussion about the higher- or lower-level meaning of text, comprehension strategies, writing, vocabulary instruction, phonics instruction, etc.);
Level 6: Teacher–student interaction styles (e.g., coaching, modeling, telling, reading aloud, discussion); and
Level 7: Expected student responses (e.g., oral response, listening, reading).

Codes at Levels 5 through 7 captured the core features of instructional practice and interaction style. Specifically, Level 5 indicated the detailed activity in which students were engaged. Levels 6 and 7, together, specified in detail the mode of interaction between teacher and students.[2]

Although the Taylor et al. (2003) coding scheme was a good fit with our work in San Diego, we made some modifications to reflect all aspects of literacy instruction in the district and to improve the usability of certain codes. Specifically, we added codes to capture activities related to English language development not covered by other codes, as well as activities related to writing strategies and writing content. In addition, we clarified several codes based on input from researchers involved in the CIERA study and feedback from our own data collectors. To allow for the highest possible comparability with the prior research, we limited significant changes to existing codes (Bitter, O'Day, Gubbins, Socias, & Holtzman, 2007).

In addition, to capture the levels of accountable talk in the classrooms, we developed a set of eight dimensions based on careful review of the literature on accountable talk and of pilot observations in SDCS classrooms. These dimensions included measures of classroom control (who controls classroom conversation, who selected the topic, and the teacher's role), student response (whether students respond to each other, whether students elaborate on what others have said), and content of talk (whether the discussion is focused on ideas, rather than recall; whether students' contributions are accurate and relevant; and whether the teacher presses for clarification, explanation, and evidence). These measures were then incorporated into the data collection template.

Despite its benefits, the use of this data collection tool involved several trade-offs. First and foremost, an underlying assumption of the coding scheme was that instructional quality can be measured by the frequency of practices that may be associated, on average, with greater growth in students' literacy achievement. What this means is that data collectors simply recorded whether any given practice occurred during an observed 5-min segment of instruction (approximately 12 segments per observation). The data were then aggregated to the teacher level to estimate the percentage of a given teacher's instruction that included the relevant practice. For example, a classroom observer would record a code for phonics instruction if, during the specified segment of the lesson, teachers and students engaged in that activity. The data collector would not evaluate how well those practices were delivered. Subsequent analysis might address the question, ''Did students with more phonics instruction have higher achievement?'' but could not answer, ''Was the phonics instruction itself good (along some measure of 'good')?'' or whether the quality of practice was associated with achievement. We chose to use a nonevaluative approach to the observation data both to enhance interrater reliability and to encourage teacher participation in the study. We believe the tool was successful on both fronts. Nevertheless, the fact that we do not distinguish other aspects of quality relevant to the specified literacy activities must be considered when interpreting the results.

Data Collection Approaches and Challenges

Data Collection Procedures

This literacy study was conducted within our broader study of San Diego's reforms that included intensive case studies of nine schools in SDCS. The sample included primarily high-

[2]The full list of codes is available upon request from the authors.

poverty schools (with between 61% and 100% of students qualifying for free or reduced-price lunch)[3] and had relatively large percentages of English learners (ELs; between 25% and 79% of students).[4] These characteristics were comparable to those of the CIERA study schools.

We collected data on literacy instruction using the observation protocol five times over the course of 2 years. In the 2004–2005 academic year, we observed classrooms in the fall, winter, and spring. In 2005–2006, we conducted an additional two follow-up observations, once in the fall and once in the early spring. In total, we conducted observations of 100 teachers across the nine case study schools in 2004–2005.[5] In 2005–2006, we observed 106 teachers. To include a representative sample of teachers for this study, we randomly chose two teachers at each grade level and asked them to participate. If they declined, we randomly picked replacements. In addition to being observed, we asked these teachers, as part of the broader study, to participate in interviews (1 hr each visit), fill out a survey in the spring of 2005, and complete monthly logs of their professional development experiences.

Observations were conducted by 14 data collectors. During each visit, our data collectors observed literacy instruction for 90 min in each classroom. The observation tool combined repeated intervals of qualitative recording of teacher–student interactions with immediate coding of the observed session along the seven dimensions of the coding scheme. Using a data collection template, observers took running notes of classroom activities in 5-min segments. They tried to record classroom conversations as close to verbatim as possible. After each 5-min segment of note-taking, the data collectors took 2 min to clean up their notes and start coding the segment using the observation coding scheme. Data collectors also recorded the number of students on-task at the end of each 5-min segment. In an average 90-min observation, observers recorded and coded approximately 12 5-min segments.

After every three 5-min observation segments, data collectors coded the extent to which they observed accountable talk in the preceding three segments. Using a common rubric, data collectors coded each dimension of accountable talk on a scale from 0 to 2, with 0 indicating the absence of the relevant accountable talk practice, 1 indicating scaffolded presence of the dimension, and 2 reflecting a high level of the dimension in classroom interaction during the specified period. Instances in which a question was not applicable (e.g., when students were reading independently) were coded as 0 for no accountable talk. Each accountable talk summary reflects approximately 20 min of instruction, and levels of accountable talk were recorded approximately four times per observation.

Ensuring Interrater Reliability

Over half of the data collectors were retired elementary literacy teachers in the San Diego area, some of whom had been literacy coaches, literacy specialists, or administrators in the

[3]In four of the case study schools, 100% of students qualified for free or reduced-price lunch.

[4]California uses the term *English learner* (EL) to designate students who speak a language other than English and who are not yet proficient in English. This designation is equivalent to *English language learner* (ELL) or *limited English proficient* (LEP) used in other jurisdictions.

[5]Ninety-one teachers were observed three times during the year; nine were observed only twice. One teacher moved from 5th grade in the fall to kindergarten in the winter and spring; she is treated as two separate teachers for analysis purposes, giving us an *n* of 101. Observations that were not in English (i.e., from biliteracy classes) were eliminated from these analyses.

district. Others were elementary literacy teachers on leave or graduate students in education with literacy teaching experience.

Although we chose a scheme that was highly objective and less inferential or evaluative, the protocol still required that all of our data collectors code a given practice in the same way over time. To ensure that the data collectors coded the observations reliably, we provided several phases of training. Training was designed to apply theories of adult learning and effective professional learning environments by incorporating active learning techniques, collaborative work among the data collectors, and ongoing feedback. Despite the training, maintaining a high level of interrater reliability throughout the study was a challenge due to the complexity of the tool, the multiple coding levels, and the large number of coders with different levels of background knowledge and experiences as researchers and classroom teachers.

To measure interrater reliability, each data collector individually coded three 5-min segments of instruction using videos of classroom instruction. We compared the coding from each data collector to a standard set of codes developed by our two literacy consultants experienced with the CIERA observation tool. For comparability, we approached the testing of interrater reliability in a similar way as Taylor et al. (2003): We reviewed the coding, provided feedback, and asked the observer to code the video clips a second time. Through this process, like Taylor et al. (2003), we were able to reach a minimum of 80% reliability at all levels of codes.[6]

Observers with fewer than 80% of their codes matching the standard codes at Levels 5 through 7 in any data collection season participated in one-on-one consultations with trainers. In addition, to maintain quality control throughout the year, the research team incorporated a review process for each data collection effort.

Assessing Student Learning Outcomes

The identification of valid, specific measures of students' literacy skills was essential for the analysis of relationships among literacy instructional activities and student achievement. Our most valid outcome measure of reading comprehension was the Degrees of Reading Power (DRP) assessment, a district-adopted reading comprehension assessment for Grades 4 through 8. The DRP measures how well students understand increasingly difficult texts. The test consists of nonfiction passages with missing words. Students must select the correct words to fill in the blanks from sets of multiple-choice options. Because we needed a consistent measure across grade levels, our nine case study schools administered the DRP at Grades 2 and 3 in addition to Grades 4 and 5 in the spring of 2005 and in fall and spring of 2005–2006.

We also used data from the Standardized Testing and Reporting assessments in California, including the California Standards Test (CST). The English-language arts (ELA) CST is administered to students in Grades 2 through 11 and measures performance of students based on the California content standards. We obtained individual student-level results, linked across years. Because the overall ELA score measures a wide range of literacy skills such as reading comprehension, grammar, and phonics, we also ran analyses using CST strand scores, focusing on those

[6]Each observer also individually coded three 5-min segments of instruction using detailed scripts of classroom instruction in San Diego. The average percentage of codes matching the standard for each level was: 100% (Level 1), 94% (Level 2), 100% (Level 3), 80% (Level 4), 89% (Level 5), 81% (Level 6), and 98% (Level 7).

most pertinent to the San Diego literacy reforms: the reading comprehension strand, literary response strand, and word analysis/vocabulary strand.

Finally, to capture outcomes in the primary grades, we used data from the district-adopted Developmental Reading Assessment (DRA), an individually administered reading assessment for Grades K–3 given three times per year. The DRA consists of a series of leveled books and recording sheets that allow teachers to ''determine students' reading accuracy, fluency, and comprehension levels'' (San Diego Unified School District, 2007).

RESULTS

As we have already discussed, our literacy study addressed two primary goals: (a) to describe literacy instruction in San Diego, and (b) to determine which specific practices, if any, were associated with increased student achievement. In this section, we provide illustrative findings relevant to these two goals from the 2004–2005 school year, our most complete and extensive year of data collection. We first discuss our findings from descriptive analyses of instruction in San Diego to create a picture of what classroom instruction looked like in our case study schools. Next, we provide results from our analyses of 2004–2005 literacy code and student achievement data to identify certain practices that showed a significant relationship with students' literacy skills. In each section, we compare our results to the findings of the Taylor et al. (2003) study to show how our results fit in and add to the current knowledge about effective literacy instruction.

To What Extent is Observed Instruction Consistent With San Diego's Theory?

We conducted descriptive analyses to quantify the occurrence of each of the literacy practices in the participating teachers' classrooms. Given the wide range and specificity of the codes, we focus here on a subset of the findings most relevant to reading comprehension and a student-support interaction style, important focuses of San Diego's approach to literacy instruction.

Descriptive Analysis Techniques

To analyze the literacy observation data for each year of data collection, we calculated across all three observations in 2004–2005, for each teacher, the proportion of 5-min observation segments in which the observer coded a given literacy instructional activity or interaction style. We then calculated averages across all of the teachers.[7] For several measures, we compared our case study results with descriptive statistics from the CIERA study.[8]

For each of the eight accountable talk measures, the values (from 0 to 2) were averaged across all of a teacher's segments in a year, providing an average accountable talk score for each teacher for each year. We then averaged the means across teachers to generate the figures presented

[7]For instance, if we observed a total of 30 5-min segments for a teacher in a given year and saw *coaching/scaffolding* in 10 of these segments, then the *coaching/scaffolding* proportion for that teacher equaled 10/30, or .333. Then, for each code or activity, we averaged together the proportions for all of the teachers.

[8]We were unable to test for significant differences between the two sets of data because we did not have access to sufficient information on the CIERA data and results.

in the accountable talk graph. In the next section, we report a few illustrative findings on the frequency of observed literacy practices in San Diego.

Instructional Focus

Reading was, by far, the most common focus of literacy instruction in our case study schools, with an average of 87.3% of 5-min segments focused on reading instruction. Composition/writing and other language (e.g., grammar) made up a much smaller portion of instructional time, with, on average, only 11.6% and 5.6% of segments, respectively. This bias toward reading instruction could have also been a result of the timing of our observations. In some cases, teachers reported teaching writing outside of the literacy block that we observed.

Emphasis on comprehension of connected text. Our observation data revealed a strong instructional focus on comprehension of text in San Diego. Consistent with the CIERA study, we examined the frequency of instruction (including discussion, questioning, or writing) related to both lower- and higher-level meanings of text. As mentioned earlier, we considered instruction of lower-level meaning of text to include instances in which teachers posed questions and students were engaged in conversation or writing about what was explicitly stated in the text. The purpose of these questions was to assist students in comprehending the literal meaning of a particular text. By contrast, higher-level questioning or discussion required the students to go beyond what was explicitly stated in order to interpret the text.

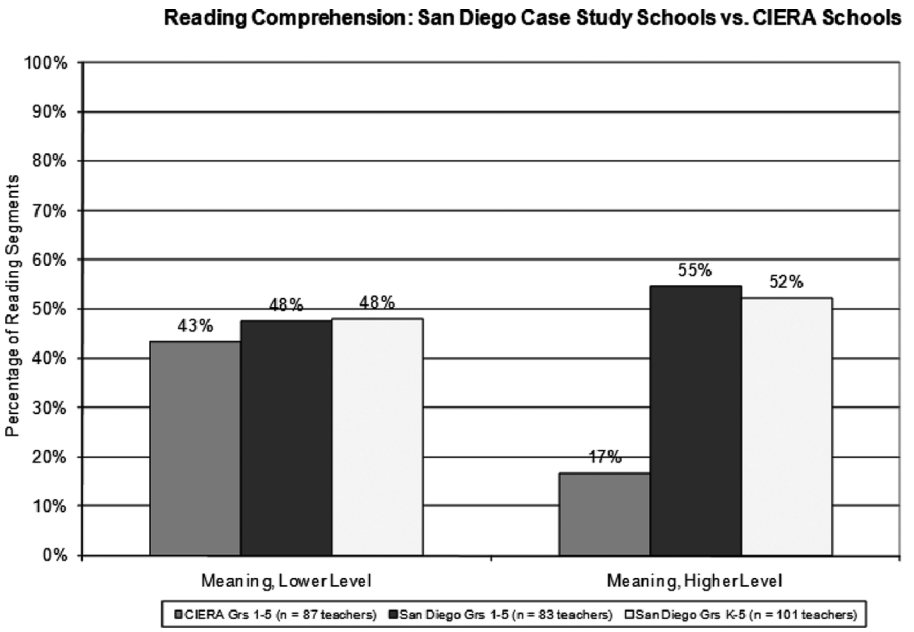

FIGURE 1 Reading instruction related to the meaning of text. *Note*. CIERA = Center for the Improvement of Early Reading Achievement.

Figure 1 depicts the frequency of instruction related to higher- and lower-level meaning of text in San Diego, as compared to patterns observed in the CIERA schools. The left-hand set of bars shows the average percentage of 5-min reading segments in which students talked or wrote about lower-level meaning of text in our study (48% in Grades 1–5), which was comparable to the observed percentage in the CIERA study (43%). The right-hand set of bars shows the average percentage of 5-min reading segments in which students were engaged in higher-level discussion about the meaning of text. We observed teachers utilizing higher-level questioning in 55% of their 5-min instructional segments, more than three times the 17% observed in the CIERA study. This large difference suggests that San Diego's focus on higher-level meaning-making has had an impact on classroom instruction.

Limited instruction on phonics and word study. Although we found an emphasis on reading comprehension in the classrooms we observed, our observation data revealed a limited focus on vocabulary, word study, and phonics in literacy instruction, lower than that found in the CIERA school change study. For example, we observed phonics on average in only 2.9% of segments. Vocabulary and word study (including word identification, sight words, and word recognition) were slightly more common, observed on average in 17.8% and 15.6% of segments, respectively. This pattern is not entirely unexpected, given the emphasis on comprehension in the San Diego reform. These results should be interpreted with caution, however, because some teachers reported that they taught word study later in the day, outside of the observed literacy block.

Teacher–Student Interactions

High levels of scaffolded instruction. The balanced literacy approach is designed to foster the gradual release of responsibility from teachers to students, moving from structured modeling (e.g., through read-alouds and shared reading) to scaffolded support (e.g., through guided reading) to independence of individual work.

Because the CIERA study found significant effects of coaching (positive) and telling (negative) on achievement, we analyzed the frequency of coaching and telling. Our assumption was that higher levels of coaching would indicate greater prevalence of instructional scaffolding. Recall that coaching occurs when teachers prompt or support students to elaborate on an answer or to deepen their understanding. The focus is on providing support to students that could transfer to other situations and thus foster independence. Telling, on the other hand, occurs when teachers simply give information to students.

As Figure 2 illustrates, the 2004–2005 data suggest that coaching strategies were more common in our case study classrooms than in the CIERA classrooms. Interestingly, the percentage of 5-min segments in which telling occurred also appeared to be higher in the San Diego case study classrooms. In other words, both coaching and telling took place with greater frequency in the San Diego case study schools than in the CIERA study schools. These frequencies could suggest that explicit instruction was taking place more frequently in San Diego than in the CIERA schools.[9]

[9]The higher frequency on both measures also may indicate that the San Diego coders were more sensitive in their coding. However, because the data collectors were trained by researchers from the CIERA study, who also helped to monitor reliability in our coding, we believe this explanation is less likely.

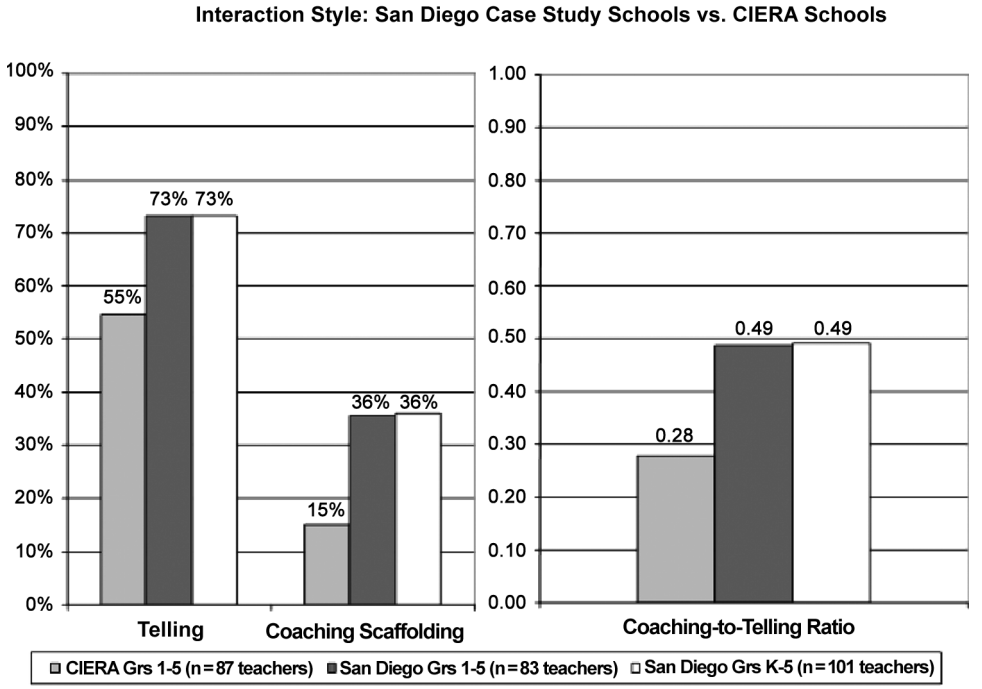

FIGURE 2 Interaction style as a percentage of instruction (all segments) and ratio of coaching to telling. *Note*. CIERA = Center for the Improvement of Early Reading Achievement.

Given the higher rate of both interaction styles, we calculated the ratio of coaching to telling in the two studies, to see if there were any differences in the balance between these styles. As shown in Figure 2, the coaching-to-telling ratio was almost twice as high in our case study schools as in the CIERA schools, suggesting that coaching and scaffolding may play a more significant role in instruction in San Diego, relative to schools participating in the CIERA study.

Accountable talk: Emphasis on ideas and evidence. Our analyses revealed variation in the prevalence of the eight dimensions of accountable talk measured for this study. As shown in Figure 3, the data indicated a high level of accuracy of student contributions and a high demand for evidence in the classroom. The extent to which conversations were focused on ideas was also relatively high, in comparison to other dimensions. These findings are consistent with the ways in which San Diego personnel described their goals for instruction and for students.

In contrast, the data indicate that teachers in the San Diego case study schools were relatively directive in topic selection and in their overall role, although they were slightly more apt to relinquish some control over conversational norms to the students. Specifically, the first three accountable talk measures shown in Figure 3 indicate that teachers generally control the conversation, select the topic, and play the role of discussion leader (rather than discussion participant), respectively. Other dimensions in which accountable talk was relatively low included the degree to which students responded to one another and elaborated on others' contributions. Students

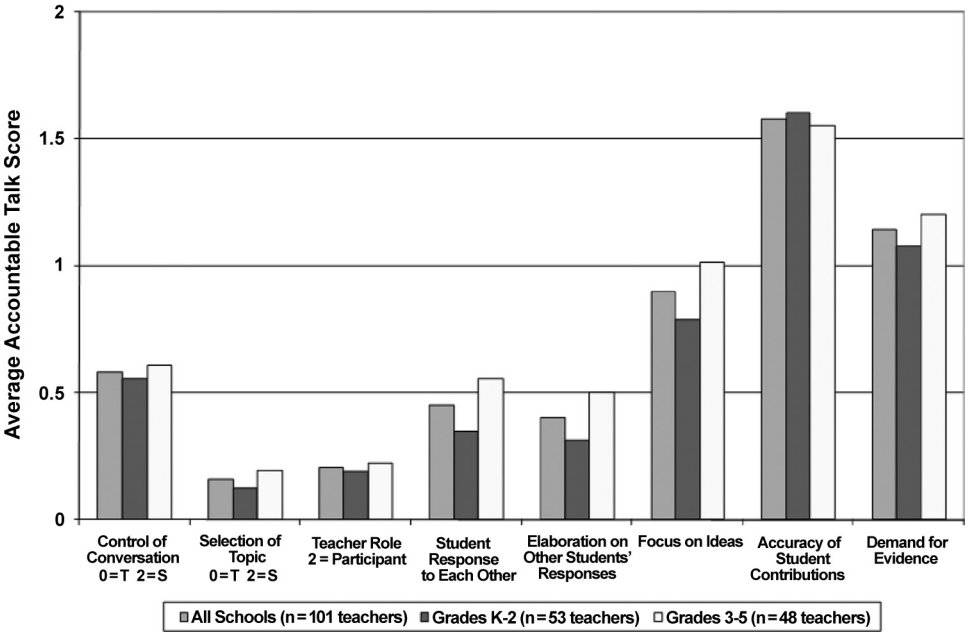

FIGURE 3 Average accountable talk scores for all case study schools and by grade span. *Note*. T = Teacher, S = Student.

tended to respond to the teacher rather than to each other, and students did not often systematically elaborate on what others (students or teachers) said.

Alignment With Schools' Instructional Focus

The distribution of literacy instructional activities previously described aligns well with the reported focus of literacy instruction, as described by principals and teachers in our case study schools. Specifically, school staff and administrators reported a focus on reading comprehension, with little (and in some cases insufficient) focus on other aspects of literacy such as word study, phonics, or grammar.

In several schools in our sample, for example, teachers' instructional goals were tied directly to the Units of Inquiry, a curricular map developed by the district and most often focused on higher-level comprehension skills such as identifying major themes in the text, understanding points of view within a text, or understanding character. This type of work would inevitably require students to engage in higher-level discussions of text, and teachers to ask higher-level questions about text. This emphasis was previously seen with the relatively high frequency of observed higher-level questions in the classroom. Even those schools that were not using the Units of Inquiry reported a focus on reading comprehension. For example, at one school that did not implement the Units of Inquiry, 9 of 10 teachers said that one of their main instructional goals was for students to understand the meaning, or main idea, of what they were reading.

Teachers at all schools also reported including word study and other literacy focuses, such as grammar and vocabulary, in the content of literacy instruction. However, they rarely discussed these aspects of literacy instruction when describing their primary instructional goals. Some schools increasingly focused, across the 2 years of the study, on these aspects of literacy instruction, particularly in the context of test preparation and English Language Development. However, reading comprehension and reading strategies continued to be cited more often as the overall focus of literacy instruction.

Which Literacy Instructional Practices Matter for Student Achievement?

The primary goal of these analyses was to explore the association between literacy practices aligned with San Diego's theory of instruction and student achievement in reading. In this section, we provide an overview of our analysis techniques and discuss the findings organized by student achievement measure.

Analysis Techniques

We used a two-level Hierarchical Linear Modeling (HLM) approach that nested students within classrooms to examine effects of classroom characteristics (including literacy practices) on student achievement (see the Appendix for model specifications). As previously mentioned, we conducted analyses for multiple outcome measures, including the DRP (the district-adopted standardized reading comprehension assessment) for Grades 2 and above, the CST for ELA, and the DRA (the district-adopted reading assessment for Grades K–3). Because these tests assess different sets of literacy skills and are administered at different grade levels, we expected that our analyses might yield different results depending on the outcome measure used. We focus here on the DRP and CST for the intermediate grades.[10]

Prior to fitting our analytic models, we ran two simple models without predictors for each of the outcome variables previously described; this enabled us to partition outcome variation occurring at the student and classroom levels respectively and to assess whether sufficient variation existed between classrooms to merit further analysis. In the case of the DRP, we found that 69% of the variance in test scores lay within classrooms, leaving a sizeable 31% to be explained by classroom differences. We then conducted a series of analyses on the predictors of this between-classroom variation, beginning with a replication of the models used in the CIERA study. We expanded and refined our analyses based on (a) the theory underlying San Diego's literacy instructional program, and (b) empirical findings from a set of iterative HLM analyses (i.e., retaining variables consistently showing a significant effect in the analyses).[11]

Each model included student-level controls for prior achievement, grade level, and demographic characteristics such as race/ethnicity, gender, and EL status. At the classroom level, we included a predictor variable for the presence of an experienced teacher (i.e., having at least

[10]Our findings from the DRA analyses were inconclusive. For more detail, see Bitter et al. (2007).

[11]We limited the students included in the HLM analysis to those who had been in the observed teacher's class for most of the year prior to DRP testing (i.e., since October 6, 2004). Students also needed to have values for all of the variables included in the analysis, including the prior achievement control.

two years of prior experience), as well as variables for relevant classroom instructional practices.[12]

Grades 3–5 Reading Comprehension (DRP)

Refined analysis results. Our primary measure of reading comprehension was the DRP assessment for Grades 3–5. The DRP score metric was the normal curve equivalent (NCE), a norm-referenced score that runs from 0 to 99.[13] In our refined HLM analyses of the DRP, three measures of literacy instruction demonstrated a consistently positive and statistically significant relationship to students' reading comprehension achievement. These were instruction focused on higher-level meaning of text, writing instruction, and the presence of accountable talk in classroom interactions (see Table 1).

Across all models, the strongest instructional predictor of increased reading comprehension was teachers' use of higher-level questioning and discussion about the meaning of text. Our final refined model indicated that for every standard-deviation increase in the coding of instruction related to higher-level meaning of text ($M = 51.98$, $SD = 18.47$), students' DRP scores increased by 1.61 NCE points, on average.

In the CIERA study, higher-level questioning was also the practice most consistently associated with student literacy growth (in both reading and writing). This link to student learning is supported by other research on literacy instruction. It is also relevant to reiterate here that higher-level comprehension instruction was emphasized in the San Diego reforms and was three times more prevalent in San Diego classrooms than those studied by the CIERA researchers.

Our refined analyses also suggested a link between writing and reading comprehension: Students in classrooms that included greater amounts of writing instruction improved their reading comprehension more than did students with less writing. Our coding scheme defined writing instruction to include both discussion about writing (e.g., about the content of the text being written or about writing strategies) and actual composition. For every standard-deviation increase in the coding of writing instruction ($M = 10.88$, $SD = 12.04$), students' DRP scores increased by 1.63 NCEs, on average.

Finally, in these models, we included a composite index for accountable talk as a measure of classroom interaction.[14] In the HLM DRP analysis, we found a positive and statistically significant relationship between the accountable talk measure and student achievement. This was a consistent finding across multiple models, some of which included accountable talk alone (with controls) and others of which included other literacy code variables. In our final refined model, we found that for a standard-deviation increase in the accountable talk index ($M = 0.74$, $SD = 0.27$), students' scores on the DRP increased by 1.04 NCEs, on average.

[12]Descriptive summary data for the sample of classrooms and students included in each analysis are available upon request.

[13]We lacked a prior achievement score on the DRP because the DRP was administered for the first time in the spring of 2005. The Spring 2004 CST score was considered to be the best available measure of prior achievement.

[14]To create this index for each teacher, we averaged the accountable talk score across all eight accountable talk variables. The Cronbach's alpha for this index was 0.87.

TABLE 1
Analysis Results for 2-Level HLM Analyses of DRP NCE Scores, Grades 3–5, 2004–2005

	Unconditional means model DRP NCE 0405[a]		*Unconditional Level 2 model DRP NCE 0405*		*Analysis model DRP NCE 0405*	
Fixed Effects	*Coefficient (b)*	*Standard Error (SE)*	*b*	*SE*	*b*	*SE*
CST ELA 0304[b]	—	—	11.802***	0.395	10.123***	0.434
English learner[c]	—	—	—	—	–4.240***	0.868
Gifted	—	—	—	—	3.190***	0.989
Female	—	—	—	—	0.547	0.669
White	—	—	—	—	4.196**	1.711
Black	—	—	—	—	–0.679	1.182
Other ethnicity[d]	—	—	—	—	3.027***	1.114
Grade 3	—	—	—	—	2.928**	1.201
Grade 4	—	—	—	—	–3.998***	1.128
Experienced teacher (2 or more years)	—	—	—	—	1.759	1.374
Literacy activity: Higher-level meaning of text[e]	—	—	—	—	1.606**	0.63
Literacy activity: Writing instruction[f]	—	—	—	—	1.633***	0.576
Accountable talk[g]	—	—	—	—	1.043*	0.581
Constant	36.727***	1.416	36.769***	0.726	36.538***	1.472
Number of students	843		843		843	
Number of classrooms	49		49		49	
Variance Components	Estimate (e)	SE	e	SE	e	SE
Classroom means	85.47	19.816	19.54	5.496	6.32	2.470
Student residual	190.65	9.567	97.53	4.910	91.10	4.577

	Partition of Total Variance	*% of Level 1 and 2 Variance Accounted for by Model*[i]
% of variance within classrooms (Level 1)	*69.0%*	*48.8%*	*52.2%*
% of variance between classrooms (Level 2)[h]	*31.0%*	—	*67.7%*

Note. *$p < 0.10$. **$p < 0.05$. ***$p < 0.01$.

[a]DRP NCE 0405 is the Normal Curve Equivalent Score for students in the spring of 2004–2005.

[b]CST ELA 0304 is the California Standards Test total scale score in English Language Arts for students in the spring of 2003–2004. This score has been standardized within grade for students in the sample.

[c]English learners are compared against students who are classified as Redesignated Fluent English Proficient, Initially Fluent English Proficient, and English Only.

[d]Students identified as White, Black, and Other ethnicity are compared against Hispanic students. Other ethnicity includes Asian, Pacific Islander, and American Indian students.

[e]Higher-level meaning of text measures the percentage of segments in which instruction focused on higher-level meaning of text was coded. This percentage has been standardized within the pool of teachers in the sample.

[f]Writing instruction measures the percentage of segments in which composition, discussions about the content of text being written, and discussions regarding writing strategies were coded. This percentage has been standardized within the pool of teachers in the sample.

[g]Accountable talk is an index calculated by averaging all eight measures of accountable talk across segments and observations. This index has been standardized within the pool of teachers in the sample.

[h]Also known as the intraclass correlation coefficient.

[i]Also known as Goodness of Fit or Pseudo R^2 statistics.

In sum, other things being equal, students in classrooms where teachers more frequently asked them to engage in higher-level interpretations and applications of text, to use evidence from the text to support their ideas and respond to one another, and to create their own meaningful text through writing showed higher levels of reading comprehension on average (controlling for prior achievement) than did students in classrooms where these practices were less common. Moreover, because these models assume that these influences were independent and additive, we can estimate a cumulative effect on achievement when all these practices are present: For a one standard-deviation increase in the coding of all three literacy practices together, students' DRP scores increased by 4.28 NCEs on average—approximately one-quarter of a standard deviation for our sample. This model explains two-thirds (67.7%) of the variance between classrooms on the DRP—a substantial portion.

Results of Taylor et al. (2003) replication analysis. In addition to our own elaborated and refined models, we also replicated the Taylor et al. (2003) analyses to determine the consistency of our results with the earlier study.[15] The resulting replication model for the DRP included only the variables that were found to be significantly associated with reading comprehension in the CIERA analyses: higher-level questioning (positive), comprehension skill instruction (negative), passive responding (negative), and percentage of students on task (positive).[16] Two of these variables were equivalent to variables we measured in San Diego: Higher-level meaning of text and percentage of students on task. For two other variables, however, we had to substitute similar measures from our study because we modified the original observation scheme. First, we substituted our accountable talk measure for the *passive responding* variable in the CIERA study. The *passive responding* measure included reading turn-taking, oral responding turn-taking, and listening. We hypothesized that including accountable talk in the model would better capture effects of active versus passive student response and talk. Second, we used our adapted measure of comprehension instruction, which was similar, but not identical, to the comprehension skill variable in the Taylor et al. (2003) study.[17]

Our analyses revealed effects similar to those identified in the CIERA analyses. First, instruction related to higher-level meaning of text was positively associated with reading achievement. In addition, we identified a positive association of accountable talk (a measure of active response and talk) with reading achievement. We also observed a negative association of comprehension activity instruction with reading achievement, although we hesitate to form conclusions about this finding because our measure was defined differently from that in the CIERA study and did not have a significant association with reading achievement in several other DRP models. Despite these similar results, our model did not demonstrate a significant positive relationship between the percentage of students on task and reading achievement.

[15]The results of this analysis were eliminated for space reasons and are available upon request.

[16]Taylor, Pearson, Peterson, and Rodriguez (2005) reported similar findings.

[17]*Comprehension skill* instruction in the CIERA study was coded when students were engaged in a comprehension activity at a lower level of thinking (e.g., traditional skill work such as identifying the main idea). Comprehension instruction in our study was coded when teachers and/or students were engaged in naming, defining, or pointing out a comprehension activity (such as identifying the main idea, or determining cause and effect) and/or reviewing how, when, or why one might engage in this activity.

Grades 3–5 ELA Performance (CST)

As mentioned earlier, we obtained individual student-level results, linked across years, for the CST, including both overall ELA scores and scores on individual strands (e.g., literary response, reading comprehension, etc.). One of the challenges in using the CST for our analyses is that the scores are not vertically equated across grade levels, which means that the scores of students at different grade levels are not directly comparable and cannot be used to draw conclusions about student growth. To address this issue, we standardized scores within grade level using the sample mean and standard deviation. Thus, an increase in standardized scores for a student from one grade to the next would indicate that the student had improved his/her performance relative to other students in the sample.

Overall CST score (ELA). Our HLM analyses of the overall CST scores in ELA resulted in fewer statistically significant relationships than we found using the DRP. Instruction focused on higher-level meaning of text and accountable talk did not have statistically significant effects in these CST models. However, statistically significant but minute relationships with two variables did emerge from a subset of these analyses: writing instruction (positive) and telling (negative). As seen in Table 2, for a standard-deviation increase in the coding of writing instruction ($M = 10.88$, $SD = 12.04$), students' scores on the CST ELA assessment increased by 0.075 standard deviations, on average. For a standard-deviation increase in the coding of telling ($M = 72.36$, $SD = 14.42$), students' CST scores decreased by 0.063 standard deviations.

As shown in Table 2, of the total variance in CST scores, 72.2% was within classrooms, and 27.8% was between classrooms. By adding in the predictors for the variables discussed above—higher-level meaning of text, writing, and telling—as well as the experienced teacher variable and the student controls, we can explain 29.3% of the variance between classrooms, much less than in the DRP models.

CST strands. One potential explanation of the inconsistency between the CST and DRP model results is that the CST measures a broad range of English language skills, some of which may be more or less influenced by specific instructional practice than is reading comprehension (as measured by the DRP). To explore this possibility further, we ran analyses using individual CST strands; that is, subsets of the CST items focused on each of the following literacy skills: literary response and analysis; reading comprehension; and word analysis, fluency, and systematic vocabulary development strands.[18] To form a more robust data set, we combined the scores for the literary response and analysis and reading comprehension strands together. The lack of scale scores for each measure makes the estimates from the strand analyses fairly rough. Nonetheless, some interesting patterns emerged.

Literary response and analysis and reading comprehension. The literary response strand addresses structural features of literature (e.g., genre) and narrative analysis of grade-level-appropriate texts, including comprehension of basic plots, comprehension of characters, and

[18]CST reports only raw scores (number correct) for each of the five strands (three reading and two writing strands). We standardized the results within the sample by grade to have a consistent measure across grades and used the 2003–2004 CST strand score to control for prior achievement.

TABLE 2
Analysis Results for 2-Level HLM Analyses of CST English Language Arts Scores, Grades 3–5, 2004–2005

	Unconditional Means Model CST ELA 0405[a]		*Unconditional Level 2 Model CST ELA 0405*		*Analysis Model CST ELA 0405*	
Fixed Effects	*Coefficient (b)*	*Standard Error (SE)*	*b*	*SE*	*b*	*SE*
CST ELA 0304[a]	—	—	0.796***	0.021	0.711***	0.023
English learner[b]	—	—	—	—	–0.228***	0.046
Gifted	—	—	—	—	0.182***	0.053
Female	—	—	—	—	0.133***	0.035
White	—	—	—	—	–0.199**	0.093
Black	—	—	—	—	–0.034	0.063
Other ethnicity[c]	—	—	—	—	0.154***	0.059
Grade 3	—	—	—	—	–0.002	0.078
Grade 4	—	—	—	—	–0.171**	0.069
Experienced teacher (2 or more years)	—	—	—	—	0.162*	0.092
Literacy activity: Higher-level meaning of text[d]	—	—	—	—	0.062	0.04
Literacy activity: Writing instruction[e]	—	—	—	—	0.075*	0.039
Teacher interaction style: Telling[f]	—	—	—	—	–0.063*	0.034
Constant	–0.016	0.081	–0.007	0.038	–0.072	0.094
Number of students	818		818		818	
Number of classrooms	49		49		49	
Variance Components	Estimate (e)	SE	e	SE	e	SE
Classroom means	0.27	0.064	0.05	0.015	0.04	0.011
Student residual	0.71	0.036	0.27	0.014	0.24	0.013
		Partition of Total Variance	*% of Level 1 and 2 Variance Accounted for by Model*[h]			
% of variance within classrooms (Level 1)		*72.2%*	*62.6%*		*65.5%*	
% of variance between classrooms (Level 2)[g]		*27.8%*	—		*29.3%*	

Note. $^{*}p < 0.10$. $^{**}p < 0.05$. $^{***}p < 0.01$.

[a]CST ELA 0405 and CST ELA 0304 are the California Standards Test English Language Arts total scale scores for students in the spring of 2004–2005 and 2003–2004. These scores have been standardized within grade for students in the sample.

[b]English learners are compared against students who are classified as Redesignated Fluent English Proficient, Initially Fluent English Proficient, and English Only.

[c]Students identified as White, Black, and Other ethnicity are compared against Hispanic students. Other ethnicity includes Asian, Pacific Islander, and American Indian students.

[d]Higher-level meaning of text measures the percentage of segments in which instruction focused on higher-level meaning of text was coded. This percentage has been standardized within the pool of teachers in the sample.

[e]Writing instruction measures the percentage of segments in which composition, discussions about the content of text being written, and discussions regarding writing strategies were coded. This percentage has been standardized within the pool of teachers in the sample.

[f]Telling measures the percentage of segments in which telling, or giving students information, was coded. This percentage has been standardized within the pool of teachers in the sample.

[g]Also known as the intraclass correlation coefficient.

[h]Also known as Goodness of Fit or Pseudo R^2 statistics.

identification of the underlying theme or author's message, among others. The reading comprehension strand of the CST addresses structural features of information materials (e.g., titles and table of contents) and comprehension and analysis of grade-level-appropriate texts, such as connecting information from the text with prior knowledge, recalling major points in the text, or distinguishing between the main idea and supporting details. Results of the HLM analyses using these two combined strands reaffirmed several of the patterns we observed in the DRP analyses.

In particular, we observed positive and statistically significant relationships between instruction related to higher-level meaning of text and writing instruction and strand achievement. We also observed a statistically significant negative relationship between telling and outcomes on the combined strands. These effects are shown in Table 3. Although the independent effect sizes for these variables appear quite small, recall that the measures are based on raw scores for relatively few test questions covering each domain. In addition, many of the statistically significant practices are intended to occur in combination in the San Diego approach.

As shown in Table 3, of the total variance in combined CST strand scores, 78.6% is within classrooms, and 21.4% is between classrooms. The aforementioned model explains 64.5% of the variance between classrooms, substantially more than the overall CST ELA model discussed in the previous section.

Word analysis, fluency, and vocabulary strand. This strand includes test items that address standards for decoding and word recognition, as well as vocabulary and concept development. Our findings were not as consistent and robust for this strand as they were for the strands previously discussed. However, in some models, we did observe statistically significant positive relationships between higher-level meaning of text instruction and writing instruction and the standardized strand scores, which provides further evidence to support our DRP analysis results. We also observed a positive effect of the proportion of students on task.

Literacy Activities in Practice

The quantitative modeling of instructional practices, as discussed in the previous section, can suggest which practices may positively influence reading achievement. In this case, we observed relationships between practices that were well aligned with the balanced literacy approach fostered by San Diego, such as higher-level discussions of text and accountable talk and reading comprehension. However, to gain a picture of what more effective literacy instruction looks like in practice, we turn to the qualitative scripting of the observed lessons. In the following, we draw on the scripted instructional data to provide an example of a classroom in which the teacher's instruction included relatively greater use of higher-level comprehension instruction and accountable talk.

In this classroom, the teacher combined several styles of instruction based on the balanced literacy approach. Across the classes we observed, she integrated shared reading, independent reading, individual student conferences, and read-aloud techniques. In shared reading segments, she incorporated higher-level questions to the group and fostered accountable talk. For example, in one observation, the students and teacher discussed an article that the teacher presented on an overhead projector about the inclusion of advertisements in students' textbooks. The teacher

TABLE 3
Analysis Results for 2-Level HLM Analyses of CST Reading Comprehension (RC) and Literary Analysis (LA) Strand Scores, Grades 3–5, 2004–2005

	Unconditional Means Model CST RC and LRA 0405[a]		*Unconditional Level 2 Model CST RC and LRA 0405*		*Analysis Model CST RC and LRA 0405*	
Fixed Effects	*Coefficient (b)*	*Standard Error (SE)*	*b*	*SE*	*b*	*SE*
CST RC & LRA 0304[a]	—	—	0.689***	0.025	0.588***	0.027
English learner[b]	—	—	—	—	–0.273***	0.057
Gifted	—	—	—	—	0.251***	0.063
Female	—	—	—	—	0.069	0.045
White	—	—	—	—	–0.017	0.117
Black	—	—	—	—	0.078	0.08
Other ethnicity[c]	—	—	—	—	0.144*	0.073
Grade 3	—	—	—	—	–0.016	0.07
Grade 4	—	—	—	—	–0.203***	0.072
Experienced teacher (2 or more years)	—	—	—	—	0.212***	0.08
Literacy activity: Higher-level meaning of text[d]	—	—	—	—	0.100***	0.035
Literacy activity: Writing instruction[e]	—	—	—	—	0.070**	0.034
Teacher interaction style: Telling[f]	—	—	—	—	–0.059**	0.03
Constant	–0.014	0.073	–0.005	0.038	–0.075	0.088
Number of students	818		818		818	
Number of classrooms	49		49		49	
Variance Components	Estimate (e)	SE	e	SE	e	SE
Classroom means	0.21	0.05	0.04	0.01	0.02	0.01
Student residual	0.77	0.04	0.43	0.02	0.40	0.02

	Partition of Total Variance	*% of Level 1 and 2 Variance Accounted for by Model*[h]	
% of variance within classrooms (Level 1)	*78.6%*	*44.5%*	*48.3%*
% of variance within classrooms (Level 2)[g]	*21.4%*	—	*64.5%*

Note. $^{*}p < 0.10$. $^{**}p < 0.05$. $^{***}p < 0.01$.

[a]CST RC & LRA 0405 and CST RC & LRA 0304 are the California Standards Test total combined raw scores for the Reading Comprehension and Literary Analysis and Response strands in the English Language Arts assessments for students in the spring of 2004–2005 and 2003–2004. The score has been standardized within grade for students in the sample.

[b]English learners are compared against students who are classified as Redesignated Fluent English Proficient, Initially Fluent English Proficient, and English Only.

[c]Students identified as White, Black, and Other ethnicity are compared against Hispanic students. Other ethnicity includes Asian, Pacific Islander, and American Indian students.

[d]Higher-level meaning of text measures the percentage of segments in which instruction focused on higher-level meaning of text was coded. This percentage has been standardized within the pool of teachers in the sample.

[e]Writing instruction measures the percentage of segments in which composition, discussions about the content of text being written, and discussions regarding writing strategies were coded. This percentage has been standardized within the pool of teachers in the sample.

[f]Telling measures the percentage of segments in which telling, or giving students information, was coded. This percentage has been standardized within the pool of teachers in the sample.

[g]Also known as the intraclass correlation coefficient.

[h]Also known as Goodness of Fit or Pseudo R^2 statistics.

stopped periodically to ask questions of students to discuss in pairs. Questions were focused on students' interpretation or judgment of the advertisements, including ''Do you think it is good or bad?'' or ''Who is for, against, or undecided?'' regarding the ads in textbooks. ''What did the author do to persuade you?'' ''How is the author getting his point across?'' ''How much information did he/she include that supported his/her own purpose?''

The teacher fostered accountable talk by encouraging students to respond to each other. For example, after one student responded about whether the article was for or against advertising in textbooks, another student followed with ''I agree with [Tom] because in the first paragraph it says [a reading from the text].'' Other students built on this response, asserting whether they agreed or disagreed with prior students' responses. This particular example also highlights students' use of evidence from the text, another element of accountable talk. The teacher frequently followed up with questions that pressed for clarification (e.g., ''I don't understand'' or ''Is that what you are saying?'').

In another lesson we observed in the following year, the teacher used similar techniques. She started by explaining to the students that she wanted them to ''pull out what is important'' to them in the text they were reading in a shared reading session. She also noted that she would be doing less modeling, as she trusted at this point that the students could do this work independently. In this remark, she demonstrated a release of responsibility to the students, a hallmark of the balanced literacy approach. As the teacher read aloud, she checked in with students periodically to ask, ''What are you thinking so far?'' Students responded by pointing out similarities of the book to their own lives. The teacher probed, asking for evidence from the text to support their ideas. The question then evolved into a discussion of why the author included a lot of description in the book. Students responded to each other, building on each other's comments and questions. For example, one exchange was as follows:

Student 1 (S1): Maybe [the author] wants us to imagine what is happening right now.
Student 2: I think the same as [S1], that the author wants us to imagine what it is. But he also wants us to feel it.
Teacher: Why feel it?
Student 3: Maybe the author is making you emotional.
Student 4: So we understand what he is trying to tell us.

The discussion grew deeper as the students progressed through the book, addressing possible themes and why the author used descriptive language.

DISCUSSION

As previously discussed, our analyses identified several literacy instructional practices that appear to be associated with student reading achievement. Our 2004–2005 findings both confirm those of prior research (specifically, Taylor et al., 2003, 2005) and identify additional important practices. Moreover, these findings suggest the benefit of a form of literacy instruction that aligns with San Diego's theory of instruction—that is, instruction that focuses on students' creating meaning from text and on supporting accountable talk in classroom interactions. Writing instruction was also found to benefit reading comprehension, suggesting that approaching comprehension from multiple perspectives could be beneficial.

The most consistent finding is that questioning and discussions related to higher-level meaning of text were positively associated with student reading achievement—this relationship emerged in analyses using the DRP and CST strands. This was also the most consistent finding in the studies conducted by Taylor et al. (2003, 2005). The fact that San Diego's instruction included a much greater prevalence of higher-level questioning, relative to the CIERA schools, suggests that the reform was successful in fostering this form of instruction. As mentioned earlier, during our study, SDCS was implementing a new initiative called the Units of Inquiry. This district curriculum focused on higher-level meaning of text, including author's message, author's bias, character analysis, and so forth. It may be that implementation of this standards-based curriculum also contributed to student reading growth.

In some models, we also found a positive effect of writing instruction on reading achievement, including both actual composition and discussions about writing. The focus of the writing instruction in San Diego was generally on the meaning of the text to be written. This typically took the form of the Writers' Workshop model. We should note that we did not observe a large amount of writing instruction (relative to other literacy practices) during this study, possibly because writing instruction had only recently reemerged as an instructional focus in the district. In addition, some teachers reported that they taught writing at another (unobserved) time in the day. Nonetheless, our findings suggest that continuing to include writing instruction, and perhaps integrating it with reading instruction, can contribute to students' reading skills.

Although some of our findings confirmed our expectations based on theory and past research, there were several additional relationships that we expected to observe but did not. In particular, based on the results from the CIERA study, we expected to see relationships between various interaction styles and reading achievement. For example, Taylor et al. (2003) found a positive relationship between teachers' modeling and coaching and measures of literacy achievement that our models did not show.[19] However, we did observe a positive effect of accountable talk. In San Diego, accountable talk may better capture the types of interaction styles fostered by the reform that influence student achievement. Our accountable talk measure included a range of interaction styles, including control of conversation and student response variables, as well as a focus on ideas, accuracy of response, and a press for evidence. In addition, our analyses of the overall ELA and literary response and analysis and reading comprehension strand CST scores demonstrated a negative relationship between telling and reading comprehension, further supporting the importance of interaction style to student growth in reading.

Limitations

The findings from these analyses of our 2004–2005 observations are well supported by prior and similar research. In addition, the models generally explained a large percentage (e.g., 67.7% for the refined DRP model for Grades 3–5) of the between-classroom variance, indicating they were fairly robust. However, to properly interpret these results, several limitations of this study are worth noting.

[19]We should note that coaching was significantly associated only with fluency, and modeling only with writing, in the CIERA study for Grades 2–5.

First and foremost, across both years of the study, we only captured snapshots of instruction. In the first year, we did so three times (for 1.5 hr each), and in the second year only twice. As we typically observed classrooms at a consistent time of the day, our ability to capture the wide range of literacy instruction incorporated into a full day of instruction was limited. For example, some teachers noted that they taught writing or word study primarily in the afternoon. We often observed these teachers regularly in the mornings and were, therefore, unable to capture these other instructional foci. We attempted to compensate for this limitation by including a relatively large number of teachers in the study across a range of schools.

Another limitation, which we noted previously, is that the observation scheme assumed quality was equal to quantity. Thus, we captured relationships among literacy practices and achievement by seeing which practices, when practiced more often, resulted in higher reading achievement. We did not examine whether the literacy practices, when used at the appropriate times, or when implemented in a higher quality way, resulted in achievement growth. Although this limitation did not allow us to explore the complexities associated with each literacy practice, it did allow us to maintain a high level of interrater reliability and to provide assurance to our study participants that our observations were nonevaluative.

In addition, these findings may be somewhat context-specific. We observed instruction only in San Diego, 5 years after the district's reform efforts began. Thus, instruction was relatively consistent among classrooms, well aligned with San Diego's theory of instruction, and also well aligned with many practices found to be beneficial in prior research, including the Taylor et al. (2003) study. This constrained variation in the frequency of classroom instructional activities may have limited our ability to detect significant effects of some of these activities on achievement. Indeed the effects we observed were smaller than those reported by Taylor et al. (2003). In addition, some of these practices may have a curvilinear effect. In other words, incorporating a certain level of the activity into instruction may be beneficial to raising achievement (when compared to not incorporating it at all or incorporating it to a minimal extent), but practicing the activities at a higher frequency may not be associated with an additional benefit. In some cases, the practices we measured in San Diego were implemented at very high frequencies overall, particularly in comparison to what was observed in the CIERA study.

Finally, when we replicated our analyses with the additional two observations from 2005–2006, despite similar distributions of literacy practices between the 2 years, we did not observe the same significant positive effects of higher-level comprehension instruction, writing, and accountable talk. These different results raised concerns about the generalizability of our findings, and we, therefore, conducted additional analyses to explore possible reasons for the discrepancies. We considered such factors as the timing of the visits (e.g., during preparation for the state tests), changes in teachers' scheduling of literacy activities during the day, and shifts in student demographics across the two years. Each of these changes may have had some influence, but it was the increase in the percentage of EL students in 2005–2006 that seemed to hold the greatest explanatory value. Our subsequent reanalysis of the data in this article, broken out by ELs and non-ELs (see O'Day, in this issue), indicated that some practices that were beneficial for fluent and native English speakers were less so for EL students, and vice versa. Hence, when the populations of students changed, so did our results. This finding has important implications for designing literacy instruction for the very diverse student populations in most urban districts and suggests that differential effectiveness of literacy practices for different groups of students is an important area for further research.

CONCLUSIONS

This study addressed two primary research questions: (a) whether instruction in San Diego classrooms was consistent with the approach to instruction hypothesized by the district to be effective, and (b) whether these instructional practices were associated with increased student reading achievement. The analyses relevant to the first question suggest that it is, indeed, possible to implement ambitious literacy instruction at scale in a diverse urban setting, as demonstrated by the widespread and frequent use of higher-level questioning and discussion about text, accountable talk among students, and scaffolding techniques. Analyses relevant to the second question lend support to the fundamental premise of SDCS reforms: Instruction focused on supporting students' active engagement with text can yield improvements in student learning outcomes, particularly in reading comprehension. Moreover, the consistency of findings between this study and that of Taylor et al. (2003, 2005) and across multiple outcome measures bolsters this latter conclusion.

At the same time, the generally smaller effects in this study, coupled with the findings of differential effects for EL students (see O'Day, in this issue) suggest areas for further investigation. Are the effects of certain practices, as measured by their frequency of use, actually curvilinear in nature? In what ways, and why, do effects differ for different types of students? How can we best assess the quality of particular practices (like coaching or asking higher-level questions about text) when frequency of use is insufficient? And how can effective practices be maintained in light of the frequent changes in leadership and reform strategies that plague American school systems? The latter question may prove to be the most important in the current reform context, for without some level of stability, even the most effective instructional strategies are unlikely to take hold and realize their potential benefit. That is a topic for another paper and future research.

ACKNOWLEDGMENTS

The research for this article was supported through generous grants from the William and Flora Hewlett Foundation, the Bill and Melinda Gates Foundation, and the Atlantic Philanthropies. We gratefully acknowledge the support from our funders, the time and participation of the SDCS district and school staffs, and the many contributions of the data collectors and analysts on the full research team, without whom this work would not have been possible. We also thank David Pearson for his invaluable suggestions and feedback during the research process, Gina Cervetti and Carolyn Jaynes for their work on the accountable talk codes and training in the CIERA instrument, and Marshall Smith and Gary Sykes for their thoughtful reviews of earlier versions of this article. The authors, however, bear sole responsibility for the findings and interpretations reported herein.

REFERENCES

Bitter, C. S., O'Day, J. A., Gubbins, P. M., Socias, M. J., & Holtzman, D. J. (2007, April). Measuring Instruction in San Diego City Schools: Tools and results. Paper presented at the 2007 Annual Meeting of the American Educational Research Association, Chicago, IL.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.) (2000). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

Campbell, J. R., Voelkl, K. E., & Donahue, P. L. (1997). *NAEP 1996 trends in academic progress*. Washington, DC: National Center for Education Statistics.

Chinn, C. A., Anderson, R. C., & Waggoner, M. A. (2001). Patterns of discourse in two kinds of literature discussion. *Reading Research Quarterly*, *36*, 378–411.

Cohen, D. K. (1996). Rewarding teachers for student performance. In S. H. Fuhrman & J. A. O'Day (Eds.), *Rewards and reform: Creating educational incentives that work* (pp. 60–112). San Francisco, CA: Jossey-Bass.

Cohen, D. K., & Ball, D. L. (1999). *Instruction, capacity, and improvement*, Philadelphia: University of Pennsylvania, Consortium for Policy Research in Education.

Cunningham, A. E., & Stanovich, K. E. (1998). What reading does for the mind. *American Evaluator*, *22*, 8–15.

Duffy, G. G., Roehler, L. R., Sivan, E., Rackliffe, G., Book, C., Meloth, M. S., et al. (1987). Effects of explaining the reasoning associated with using reading strategies. *Reading Research Quarterly*, *20*, 347–368.

Knapp, M. S., Adelman, N. E., Marder, C., McCollum, H., Needels, M. C., Padilla, C., et al. (1995). *Teaching for meaning in high-poverty classrooms*, New York: Teachers College Press.

No Child Left Behind Act of 2001, Pub. L. No. 107–110 (2001).

RAND Reading Study Group. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: RAND.

Resnick, L. B. (1987). *Education and learning to think*. Washington, DC: National Academy Press.

Roehler, L. R., & Duffy, G. G. (1984). Direct explanation of comprehension processes. In G. G. Duffy, L. R. Roehler, & J. Mason (Eds.), *Comprehension instruction: Perspectives and suggestions* (pp. 265–280). New York: Longman.

San Diego Unified School District Standards, Assessment, and Accountability Division. (2007). *Developmental Reading Assessment (DRA) Summaries: 2007–08*. Retrieved November 13, 2008, from http://studata.sandi.net/research/DRA/index.asp.

Schoenfeld, A. H. (1988). When good teaching leads to bad results: The disasters of ''well taught'' mathematics classes. *Educational Psychologist*, *23*, 145–166.

Schoenfeld, A. H. (1998). Making mathematics and making pasta: From cookbook procedures to really cooking. In J. G. Greeno & S. V. Goldman (Eds.), *Thinking practices in mathematics and science learning* (pp. 299–319). Mahwah, NJ: Lawrence Erlbaum Assocaites.

Taylor, B. M., Pearson, P. D., 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, P. D., Peterson, D. S., & Rodriguez, M. C. (2005). The CIERA School Change Framework: An evidence-based approach to professional development and school reading improvement. *Reading Research Quarterly*, *40*, 40–69.

APPENDIX: HLM ANALYSIS SUPPLEMENTAL INFORMATION

HLM Analysis Models

Prior to fitting models that included our substantive literacy code predictors, for each of our outcome measures, we ran an unconditional means model and an unconditional Level 2 model that enabled us to partition outcome variation, assess whether there was, indeed, systematic variation between classrooms that was worth exploring further, and gauge the extent to which the addition of predictors accounted for student- and classroom-level variation.

The general approach of these Hierarchical Linear Model (HLM) analyses is to use student-level data to understand differences across classrooms. At a first level, the dependent variable is the academic performance of a student in a given classroom. The intercept of this Level 1 model identifies the average student performance in each classroom, controlling for prior student academic achievement (see Unconditional Level 2 Model) and other relevant student characteristics, such as English language proficiency, gifted status, gender, ethnicity, and grade (see Generic

Analysis Model). The classroom average performance becomes the dependent variable in the second level of this HLM analysis. The relevant research question is then to try to understand how academic performance and literacy practices vary across classrooms (see Generic Analysis Model).

1. The Unconditional Means Model

The unconditional means model assesses whether there is potentially predictable outcome variation and, if so, where it resides (i.e., within or between classrooms). This analysis can help determine whether there is sufficient variation between classrooms to warrant further analyses. In the case of the DRP, we found that 69% of the variance in test scores lay within classrooms (Level 1) and 31% between classrooms (Level 2; see Table 1 in main text).

2. The Unconditional Level 2 Model

In the Unconditional Level 2 Model, the Level 2 variance component quantifies the amount of unpredicted variation in the Level 1 classroom parameters. Fitting this model allows us to assess whether interclassroom differences in the outcome measure are due to inter-classroom differences in intercepts.

Using this model with DRP scores as the outcome measure, we found that 49% of the within-classroom variation is explained by the 2003–2004 CST scale score in ELA.

3. The Generic Analysis Model

Level 1.

$$\begin{aligned} \mathrm{Yij} = &\pi 0\mathrm{j} + \pi 1\mathrm{j}(\mathrm{PRIORij}) + \pi 2\mathrm{j}(\mathrm{ELij}) + \pi 3\mathrm{j}(\mathrm{GIFTEDij}) + \pi 4\mathrm{j}(\mathrm{FEMALEij}) \\ &+ \pi 5\mathrm{j}(\mathrm{WHITEij}) + \pi 6\mathrm{j}(\mathrm{BLACKij}) + \pi 7\mathrm{j}(\mathrm{OTHETHij}) + \pi 8\mathrm{j}(\mathrm{GRADE3ij}) \\ &+ \pi 9\mathrm{j}(\mathrm{GRADE4ij}) + \xi\mathrm{ij} \end{aligned}$$

Yij = Outcome score for student i in classroom j
$\pi 0\mathrm{j}$ = Intercept for classroom j (Grand Mean – average baseline classroom performance level)
$\pi 1\mathrm{j}$ = Prior achievement score slope effect for classroom j
$\pi 2\mathrm{j}$ = Performance level adjustment for ELLs (relative to non ELL students)
$\pi 3\mathrm{j}$ = Performance level adjustment for gifted students
$\pi 4\mathrm{j}$ = Performance level adjustment for female students (relative to male students)
$\pi 5\mathrm{j}$ = Performance level adjustment of White students (relative to Hispanic students)
$\pi 6\mathrm{j}$ = Performance level adjustment for Black students (relative to Hispanic students)
$\pi 7\mathrm{j}$ = Performance level adjustment for students with other ethnicities (relative to Hispanic students)
$\pi 8\mathrm{j}$ = Performance level adjustment for Grade 3 students (relative to Grade 5 students)
$\pi 9\mathrm{j}$ = Performance level adjustment for Grade 4 students (relative to Grade 5 students)
$\xi\mathrm{ij}$ = Level 1 Student Residual

Level 2.

$$\pi 0j = \lambda 00 + \lambda 01(\text{EXP TEACHERj}) + \lambda 02(\text{LITCODESj}) + \eta 0j$$

$\lambda 00 =$ Population average of the Level 1 intercepts (classroom means)
$\lambda 01 =$ Performance level adjustment for classrooms with Experienced Teachers (relative to classrooms without experienced teachers)
$\lambda 02 =$ Marginal classroom performance adjustment for intensity of a given literacy activity observed in the classroom.
$\eta 0j =$ Level 2 Classroom Residual variance in true intercept

Level-2 Classroom Variance Explained

Comparing the residual variance of the unconditional Level 2 (UCL2) and any one of our analysis models (AM) allows us to assess the strength of the Level 2 classroom literacy codes (or any other classroom variable) as a predictor. In this comparison, R22 is calculated as:

$$R22 = [(\text{Classroom Means})\text{UCl2} - (\text{Classroom Means})\text{AM}]/(\text{Classroom Means})\text{UCL2}$$