

**Changing Literacy Instruction in Schools: Consequences of CSR
Program Participation on Teachers' Classroom Practice***

Richard Correnti and Brian Rowan

*School of Education
University of Michigan*

April, 2006

* The research reported here was conducted by the Consortium for Policy Research in Education through grants from the Atlantic Philanthropies, the William and Flora Hewlett Foundation, the National Science Foundation (Grants REC-9979863 and REC-0129421), and the U.S. Department of Education (grants OERI-R308A60003 and OERI-R308B70003). The opinions expressed in the paper are those of the authors, not the sponsors. We thank Stephen W. Raudenbush, Cecil G. Miskel, and Fred Morrison for insights about the data analyses reported here. The authors remain responsible for any errors in the work.

ABSTRACT

This study examined patterns of literacy instruction in schools adopting three of America's most widely disseminated CSR programs (the Accelerated Schools Project, America's Choice, and Success for All). A major goal of the study was to compare the average patterns of literacy instruction found in these schools to average patterns of literacy instruction found in a closely matched sample of comparison schools. Contrary to the common view that educational innovations seldom affect teaching practices in schools, the study found large differences in literacy instruction between teachers in America's Choice schools and comparison schools and between teachers in Success for All schools and comparison schools. By contrast, no differences in literacy teaching practices were found between teachers in Accelerated Schools Project schools and comparison schools. On the basis of these findings and our knowledge of the implementation support strategies pursued by the CSR programs under study, we conclude that well-defined and well-specified instructional improvement programs that are strongly supported by on-site facilitators and local leaders who demand fidelity to program designs can produce large changes in teachers' instructional practices.

**Opening Up the Black Box: Literacy Instruction in Schools
Participating in Three Comprehensive School Reform Programs**

One of the most dynamic trends in American education over the past decade has been the widespread adoption by elementary schools of what have come to be known as “comprehensive school reform” (CSR) programs. Gaining initial prominence through the efforts of the New American Schools Development Corporation, and later supported by the federal government’s Comprehensive School Reform Program, CSR programs promised to “break the mold” of American schooling by producing new and more effective patterns of instruction that would markedly improve student achievement in America’s schools (Berends, Bodilly, and Kirby, 2002). During the past decade, somewhere between 10 and 20 percent of all elementary schools in the United States adopted one of these innovative programs, making CSR one of the most widely disseminated education reforms of the past decade (for a discussion of factors leading to the emergence of CSR programs, see Rowan, Camburn, and Barnes, 2004).

Not surprisingly, the widespread adoption of CSR programs by local education agencies has given rise to a lively literature in the field of educational evaluation. Several consumer guides to CSR programs have been developed to inform potential adopters about the unique design features of specific programs (e.g., Herman et al., 1999; Northwest Regional Educational Laboratory, 2005). In addition, numerous studies of the implementation of CSR and other whole school reform programs have been conducted in schools and school districts across the country (e.g., Bodilly, 1996; Berends, Bodilly, and Kirby, 2002; Desimone, 2002; Mirel, 1994). Finally, an extensive meta-analysis has been conducted to summarize the effects of CSR programs on student achievement (Borman et. al., 2003).

In many ways, existing research on the adoption, implementation, and instructional effectiveness of CSR programs echoes a familiar theme in the literature on educational innovation in the United States. The CSR story begins when an influential and dedicated group of reformers (in this case business and government leaders) succeed in promoting (and, through legislation, institutionalizing) a new template for school reform (Rowan, Camburn, and Barnes, 2004). This template then diffuses widely and quickly through the education system, as several thousand schools adopt one or another CSR program. But, while adoption is seemingly quick and easy, implementation at local sites turns out to be difficult (Bodilly, 1996; Berends, Bodilly, and Kirby, 2002; Desimone, 2002; Mirel, 1994), and in addition, program evaluations gradually uncover a pattern of weak effects on the reform's intended goal—to improve the academic achievement of students (Borman et. al., 2003). As a result, enthusiasm for the new reform strategy wanes, and American educational policy veers away from what was once considered a promising approach to school reform in order to find a new magic bullet for school improvement.

In the case of CSR programs, this familiar story has a variety of analytic shortcomings. For one, a close inspection of the Borman et al. (2003) meta-analysis reveals that while CSR program effects on student achievement have been quite small on average (Cohen's $d_{sd} = .12$ in comparison group studies), there has been a great deal of program-to-program variability in effect sizes (e.g., Cohen's d_{sd} varies from $-.13$ to $+.92$ in comparison group studies). Thus, some CSR programs apparently improve student achievement outcomes more than others. This finding, in fact, has been common in research on innovative education programs in the United States, dating from the earliest evaluation of Follow Through (see, for example, House et al., 1978; Gersten, 1984). So, while existing research suggests that the

average effect of CSR programs on student achievement is small, variability in effectiveness from CSR program to CSR program is substantial.

Additionally, current research does not explain the variability in CSR program effectiveness very well. Borman et al. (2003), for example, sought to explain variation in program effects by looking at a variety of variables, including characteristics of the evaluation studies, variations in program features, and variation in study populations; but almost none of these factors explained why some programs did better than others. The paucity of findings in this analysis is understandable, especially in light of the weak measures of program features used. For example, in the Borman et al. (2003) meta-analysis, CSR programs were described as having (or not having) highly prescribed curricula and instructional practices, but this abstract indicator almost certainly glossed over important differences in the curricula and/or instructional practices that different CSR programs managed to implement in schools.¹

In general, the lack of descriptive data about the curricula and instructional practices implemented in CSR schools would seem to be a major stumbling block in explaining variability in achievement outcomes across programs. For example, there are two very plausible—yet unexplored—reasons why CSR programs might have varying effects on student achievement. The first would be that different CSR programs have been built around curricular and instructional practices that differ in their actual effects on student learning. In this scenario, if all CSR programs had equal rates of faithful implementation, we could assume that differences in curricular and instructional design were producing observed variability in program effects. To date, however, little attempt has been made to examine

¹ In this regard, Borman and colleagues (2003) were simply following a pattern laid down by other CSR researchers, who also characterized CSR programs in these abstract terms (e.g., Herman et al., 1999; Bodilly, 1996).

this hypothesis, largely because researchers have not collected the kinds of data on instruction and curriculum in CSR schools on a large enough scale to test this hypothesis.

Alternatively, we could assume that all CSR programs are built around curricular and instructional designs that are more effective than the norm in American education. And, if that were the case, then a reasonable explanation for program-to-program differences in effects on achievement might be the difficulties particular programs had in getting their preferred instructional reforms implemented in schools. As we shall see, previous research sheds some light on the extent to which some CSR programs are better than others at resolving problems of instructional implementation.

The Problem

Given these problems, this paper presents data on the curricular and instructional practices occurring in samples of schools implementing three of America's most widely disseminated CSR programs: the Accelerated Schools Project, America's Choice, and Success for All. At the time the study reported here was designed, the Accelerated Schools Project was operating in 561 schools across the United States, America's Choice was operating in 145 schools, and Success for All was in 1,103 schools. As we discuss below, schools were chosen for the current study from the groups of schools implementing these CSR programs in such a way as to match schools in the sample in terms of time implementing the program, neighborhood demographics, and geographic location, as well as to produce representative samples of schools implementing each of the CSR programs under study.

Within each of the schools under study, we administered instructional logs (or time diaries) to teachers (for a discussion of the methods used here, see Camburn and Barnes, 2004; Rowan, Camburn, and Correnti, 2004). This procedure allowed us to gain detailed,

reliable, and valid data on curricular and instructional practices in the schools under study. In this paper, we focus on literacy instruction in these schools, since this was the major reform target of the three CSR programs under study.

The purpose of this paper is to use these log data on literacy instruction to compare the average curricular and instructional practices occurring in the CSR schools in the sample to average literacy practices occurring in a matched sample of comparison schools not implementing these programs. The analyses presented below address two questions: (1) On average, did the CSR programs under study manage to get their prescribed literacy designs implemented? And, (2) on average, did the curricular and instructional practices occurring in the CSR schools under study differ from the literacy practices found in comparison schools?

Background

Our framing of the research problem suggests that CSR program developers need to address two issues in order to improve student achievement in the schools where they work. First, they need to create designs for instruction that, if implemented, are more effective than prevailing instructional practices in American schools, and second, they need to devise organizational change strategies that assure that these practices are in fact implemented in the schools where they work.

In this paper, we discuss the latter problem first, that is, we begin by discussing what is known about factors promoting successful implementation of new instructional practices in schools. We then turn to an understanding of how the CSR programs under study differed, both in terms of how they supported implementation, and in terms of the actual curricular and instructional practices they sought to implement in schools. Using this knowledge, we then formulate a series of hypotheses about the kinds of curricular and instructional

practices that we expect will differ significantly across CSR and comparison schools. The final step in the paper is to test these hypotheses using data from teacher logs.

The Paradox in Implementation Research

Our interest in CSR implementation arises from a curious paradox in research on innovative programs in American education. On one hand, conventional wisdom suggests that making change in American schools—and especially making changes in the instructional core—is extremely difficult. The origins of this view are several. For example, difficulties in getting new instructional programs and practices implemented in schools have been observed in studies of progressive education reforms (Cuban, 1990), NSF curricular reforms (Darling-Hammond and Snyder, 1992), the Head Start and Follow Through programs (Rivlin and Timpane, 1975), ESEA Title III and other federal programs supporting education change (Berman and McLaughlin, 1975), recent efforts at school restructuring (Elmore and McLaughlin, 1998; Fullan, 1991), and of course, CSR programs (Bodilly, 1996; Desimone, 2002).

What is interesting, however, is that alongside this body of research is a lesser-known and much less cited set of studies suggesting that faithful implementation of externally-designed instructional innovations is in fact quite possible (Firestone and Corbett, 1988). In the Follow Through evaluation, for example, the Direct Instruction model appeared to be far more faithfully implemented than were other models (Meyer, Gersten, Gutkin, 1983; Gersten, 1984; Gersten et al., 1986). Moreover, the study of Dissemination Strategies Supporting School Improvement (Crandall et al., 1982) likewise found a number of programs in which educational innovations—including instructional innovations—were transferred to local school sites with reasonable fidelity.

Resolving the Paradox

Fortunately, education researchers have worked for at least two decades to resolve this paradox in implementation findings. The key to this line of work has been to examine variation in the procedures used by external program developers to support innovation in schools and then to systematically study implementation rates. In this literature, so-called problems of implementation are often reconceived as problems of professional learning. The implication of this shift is to suggest that teachers are the key delivery mechanism in instructional innovations and that program developers wanting to implement new instructional practices in schools therefore need to devise successful strategies for helping teachers learn how to use new instructional practices in their specific work settings.

Over the years, this literature has suggested a number of factors that promote professional learning and increase rates of faithful implementation of instructional change efforts in schools. These factors include: (a) the innovative program is focused on changing specific, curriculum-embedded elements of instructional practice as opposed to more diffuse elements of instruction that cut across curricular areas or represent generic forms of teaching (Cohen and Hill, 2001; Desimone, Porter, Garet, Yoon, and Birman, 2002; Fennema et al., 1996); (b) within the particular curriculum area being addressed, the program has clearly defined goals for change, that is, a clear specification of what features of curriculum and instruction will be changed, and of the steps to be taken to achieve these changes (McLaughlin and Marsh, 1978; Elmore and Burney, 1997; Nunnery, 1998); (c) these goals are further clarified by the presence of extensive written materials and other documentary supports for teaching the new design to teachers (Peterson and Emrick, 1983); (d) the new practices to be implemented are ambitious and represent a marked change in existing practices (Datnow and Castellano, 2000; Gersten et al., 1986; Huberman and Miles, 1984);

(e) the program provides a knowledgeable, external facilitator whose job is to work closely with teachers in implementing these ambitious, new practices (Cox and Havelock, 1982; Crandall, Eiseman, and Lewis, 1986; Borko et al., 2003; McLaughlin and Marsh, 1978; Peterson and Emrick, 1983); and (f) external program designers, local program facilitators, and local administrative leaders all demand fidelity to these planned changes in instructional practice (Loucks, Cox, Miles, and Huberman, 1982; Huberman and Miles, 1984).

Hypotheses about the CSR Programs

Through analysis of program documents, as well as field research, we have found that the CSR programs participating in the Study of Instructional Improvement differed in important ways along the dimensions of implementation support just described. As a result, we hypothesize that these programs will also vary in the extent to which they produce distinctive patterns of instruction in schools. This section lays out our ideas about how these programs supported instructional change at the time we studied them and presents a set of hypotheses about the kinds of instructional differences that we expect to occur across CSR and comparison schools in the study.

Our hypotheses build on previous work on CSR implementation, in which we argued that the three CSR programs under study used three very different models of organizational control to stimulate faithful program implementation (Rowan, Camburn, and Barnes, 2004). In this work, we argued that the Accelerated Schools Project (ASP) used a process of “cultural control” to guide instructional change efforts in schools, that the America’s Choice program (AC) used a system of “professional control” to guide implementation efforts, and that Success for All (SFA) used a system of procedural controls to guide the process of instructional change. In the following sections, we describe how these different systems of

control corresponded (in differing degrees) to the factors that prior research has shown produce intended implementation outcomes.

The Accelerated Schools Project

ASP's strategy for bringing about instructional change can be likened to a system of "cultural control." That is, the program's approach to providing schools with implementation support revolves around promoting a normative commitment among school leaders and faculty to the program's abstract vision or ideal of "powerful learning" for all students. From the outset, ASP facilitators use the staff development process to emphasize the program's commitment to this abstract construct, and to define powerful learning as constructivist in nature, with an emphasis on authentic, learner-centered, and interactive forms of instruction. However, ASP is not prescriptive in nature. The program does not target particular school subjects for improvement, nor does it provide teachers with a great deal of explicit guidance about curriculum objectives or teaching strategies. Instead, ASP facilitators help schools use a systematic process of organizational development to design a unique path toward powerful learning and to adopt locally-appropriate forms of instructional practice consistent with this approach.

This description suggests that (at the time of our study) ASP's approach to producing instructional change lacked many of the features previous research identified as promoting implementation success. For one, the program's goals for change were generic in form—aiming at broad changes across the board rather than targeting specific areas of the curriculum for change. Moreover, the kinds of changes teachers were supposed to make were nowhere highly specified, and instead, each school (and each teacher within a school) was asked to "discover" the most appropriate means to producing powerful learning within their own particular context. Given this, schools and teachers were given a great deal of

autonomy in the ASP system, and there was, as a result, little real focus on implementation fidelity, either from external program facilitators, or from internal leaders. In fact, in previous research, we have found that ASP schools had the lowest levels of instructional leadership of all the schools in our study sample (Camburn, Rowan, and Taylor, 2003).

In this respect, ASP's approach to producing instructional change is much like the approach used by many of the federal programs supporting educational change studied by Berman and McLaughlin (1975). This study found very low levels of program implementation. But, as Loucks (1983: 11) noted, the federal programs studied by Berman and McLaughlin "stressed local initiative ...[and]...development and management, rather than changing specific classroom practices. As a result, not only were practices not defined enough so an outsider could see them ... in place, but they also changed continuously by design." Thus, researchers have argued that low levels of implementation were found because the designs largely stressed local problem solving (Datta, 1980). Because this situation also seems to characterize ASP's situation at the time of our research, we formulate the broad hypothesis that:

H₁: There will be no mean differences in literacy instructional practices across ASP and comparison schools in our sample.

America's Choice

The America's Choice program took a contrasting approach to instructional change at the time of our study, using what we call "professional controls" to stimulate instructional improvement. The AC program had its origins in the standards-based reform movement, and as a result, the program was built around some definite ideas about the curricular content that should be taught in schools and about methods of teaching inside classrooms, especially in the area of language arts. At the time of our study, for example, AC typically

began its work in local schools by focusing on the school's *writing* program (moving only later to changes in reading and mathematics programs). Moreover, AC typically provided teachers with a great deal of instructional guidance. For example, teachers in AC schools received a curriculum guide, were taught a set of recommended instructional routines for teaching writing (called "writers' workshop"), and worked with locally-appointed AC coaches and facilitators to develop "core writing assignments" and clear scoring "rubrics" for judging students' written work. Thus, in the area of writing instruction at least, AC was trying to implement a well-specified, standards-based curriculum grounded in professional consensus about what constitutes a desirable instructional program. AC also expected schools that adopted the program to create two new leadership positions—a design coach and a literacy coordinator. Design coaches were expected to help principals implement the program, while AC literacy coordinators were expected to work with classroom teachers. Previous research showed that levels of instructional leadership were highest in the AC schools in our study sample (Camburn, Rowan, and Taylor, 2003).

In our view, AC organizes the instructional improvement process in such a way as to produce high levels of faithful implementation—especially in the area of curriculum that it emphasized first: writing instruction. Writing instruction was central to the AC design at the time we studied the program. AC's work with schools began with implementation of writer's workshop in classrooms, and as part of this implementation focus, AC leaders expected teachers to spend more minutes per day of literacy instruction engaging students in writing than would be the norm in American elementary schools. In addition, AC sought to change the way writing was taught. For example, program leaders asked teachers to move beyond simply teaching writing mechanics (e.g., grammar, punctuation) in order to encourage the actual production of written text in essay assignments carried out within a

writing process model. Finally, AC's design also encouraged a closer integration of reading comprehension instruction with writing instruction. Implementation of reader's workshop was intended to follow implementation of writer's workshop, and teachers were expected to make explicit connections between the two so students would appreciate comprehension and writing as reciprocal processes.

Given AC's explicit emphasis on writing instruction, and the fact that its strategy of professional control included most of the features of implementation support that prior research has found will increase levels of implementation fidelity, we formulate the broad hypothesis that:

H₂: Teachers in AC schools will be more likely than teachers in comparison schools to integrate reading comprehension and writing instruction, thereby focusing more on writing instruction and placing more emphasis on students' production of extended, written text.

Success for All

Success for All provides yet a third model for promoting instructional change in schools: what we call "procedural controls". Of the three programs under study, SFA gave schools the clearest and most highly-specified plan for instructional improvement by producing a set of highly-specified instructional routines for the teaching of reading. In particular, the SFA program was built around a clear and well-defined reading curriculum that provided teachers with a weekly lesson sequence, and each lesson in this sequence was designed around a "script" intended to guide teaching activities through a 90-minute reading period. In grades K-2, moreover, these scripts were accompanied by program-provided curricular materials for use throughout the school.

SFA schools also were more centrally managed than other schools in our study. For example, schools implementing SFA were expected to appoint a full-time literacy coordinator, and this staff member was given substantial responsibility for school-wide coordination of the reading program, including the task of constituting reading groups and making teaching assignments to these groups on a school-wide basis every eight weeks. In addition, instructional leaders in SFA schools and SFA linking agents were asked to supervise implementation of SFA instructional routines. In prior research, levels of instructional leadership were found to be as high in SFA schools as in AC schools, and much higher than levels of instructional leadership found in ASP schools (Camburn, Rowan, and Taylor, 2003).

Clearly, SFA's approach to promoting instructional change encompasses many of the features prior research has shown produce faithful implementation. For one, SFA clearly focuses on the improvement of reading instruction, giving teachers an extraordinarily high degree of instructional guidance ranging from clear curricular guidelines to scripted lesson plans. In addition, leaders emphasize faithful implementation of the SFA reading program in schools, and closely monitor implementation progress.

Previous case study research (Datnow and Castellano, 2000), as well as a careful review of program materials, suggest several areas where reading instruction in SFA schools can be expected to differ from normative patterns of reading instruction in American schools. First, SFA is probably best characterized as a "skill-based" reading program calling for high levels of fast-paced, direct instruction in many different reading comprehension strategies. For example, the 5-day reading cycle described in SFA program documents calls for teachers to consistently teach a variety of reading comprehension strategies during a given lesson and to do so across all days of instruction. The strategies to be taught include: activating prior

knowledge, previewing and surveying text, self-monitoring for meaning, identifying story structure, sequencing and summarizing text, and so on. SFA lesson routines also call for teachers to employ a variety of instructional formats during lessons (e.g., use of explicit teaching, use of cooperative groups), and to have students engage in specific kinds of reading comprehension assignments during lessons (e.g., answering brief oral questions, answering multiple choice and/or fill in the blanks comprehension questions, writing brief answers to comprehension questions, discussing text with peers).

Given SFA's instructional design, and its strategy for promoting implementation, we propose the following hypothesis:

H₅: Teachers in SFA schools will be more likely than teachers in comparison schools to engage in direct/explicit teaching of reading comprehension, emphasizing low-level reading skills like literal comprehension and having students demonstrate comprehension through simple, direct responses to oral questions and/or short written work.

Data

Data to test these hypotheses were collected during the period AY 1999-2000 to AY 2003-2004 as part of the *Study of Instructional Improvement* (SII). SII was a large-scale, quasi-experiment that examined the design, implementation, and instructional effectiveness of three CSR programs – the Accelerated Schools Project (ASP), America's Choice (AC), and Success for All (SFA). In each school participating in the SII sample, two cohorts of students were studied, one group passing from grades K-2, the other from grades 3-5. Extensive data on the instruction received by these students was collected through frequently administered instructional logs, using procedures described by Rowan, Camburn, and Correnti (2004). In addition, students' achievement was assessed twice annually using

CTB McGraw Hill's Terra Nova. Finally, questionnaires were administered annually to teachers and school leaders and additional information about students' family and social backgrounds were collected through a parent interview upon each child's entry into the study.

Schools in SII

Schools in this study were chosen from a list of eligible schools using procedures outlined in Benson (2002). Overall, 31 AC schools, 30 SFA schools, 28 ASP schools, and 26 Comparison schools participated. These schools were located in 17 different states in the Northeast, Southeast, Midwest, Southwest and Northwest. Schools were chosen in order to balance the sample, as much as possible, in terms of geographic location and school demographic characteristics, and to achieve a representative sample of schools participating in each CSR program. By design, however, the final sample over-represented schools in the highest quartile of socio-economically disadvantaged schools in order to study instructional improvement in high-poverty settings.

Table 1 provides descriptive statistics on this sample of schools, broken down by CSR program participation. It shows that schools in the AC and SFA samples had higher minority concentrations and also served students with lower entering achievement. As we discuss later, we will control for these demographic differences using propensity score stratification methods prior to estimating differences in instructional practices across CSR program schools and comparison schools.

Insert Table 1 About Here

Data Collection within Schools

We collected data on literacy instruction by following two cohorts of students as they passed through these schools. In each school, samples of 8 students from each kindergarten

and 3rd grade classroom were randomly selected from the roster of students assigned to that classroom and followed over the course of the study. Since student mobility was high, however, student samples were “refreshed” annually by replacing students who left the school with a random sample of new students moving into the school. This strategy worked to combat the loss of student cases from the data set and also maintained the representativeness of student samples for each year in each school.

Instructional Data

Data on the literacy instruction received by these students were gathered from a language arts log administered to all teachers of cohort students.² Each log was a survey instrument containing roughly 100 items used to record information about a single day of instruction for a single student. The opening section of the log asked teachers to report on the amount of time spent by the focal student on reading/language arts instruction on the reporting day, as well as the amount of emphasis given in the focal students’ instruction to each of the following topics: word analysis, concepts of print, oral or reading comprehension, vocabulary, writing, grammar, spelling, and research strategies. Then, if teachers checked that word analysis, comprehension, or writing was an emphasis for a student on a given day, teachers completed additional items about the specific content that was taught in any of these focal domains, the methods used to teach that content, and the tasks and materials the focal student used that day.

To assure that log reports were representative of days of the school year and students in a classroom, every teacher of cohort students participated in three extended logging periods spaced evenly over the academic year, during which time they rotated daily log reports across

² Copies of the teacher log used in the study can be found at:
<http://www.sii.soe.umich.edu/instruments.html>

the sample of cohort students in their class. Overall, 89% of teachers who were asked to log did so, and they completed 90% of the logs they were administered. If students changed teachers during the course of the year (as did many SFA students), their new teachers completed logs.

To assure accuracy in teachers' log reports, SII researchers conducted a 1 day training for teachers, gave teachers a glossary defining and illustrating the terms used in the log, and encouraged teachers to consult a toll-free phone number with logging questions. An analysis of the correspondence between trained observers' log reports and teachers' log reports of the same lesson conducted during the pre-test phase of the research found that teacher-observer match rates were rarely more than a few percentage points different from observer-observer match rates for the same lessons, especially for the most commonly covered curriculum topics and most frequently used teaching practices (Camburn and Barnes, 2004).

Other analyses have demonstrated that instructional measures based on log data have adequate reliability and predictive validity. For example, using 3rd grade log data, Rowan, Camburn, and Correnti (2004) demonstrated that IRT measures of reading instruction had acceptable reliability when days of instruction were the object of measurement, and that when these measures were aggregated to form teacher-level measures, the measures reliably discriminated among teachers (with reliabilities of .75 and above). In two other studies (Correnti, Rowan, and Camburn, 2003; Rowan, Raudenbush, Correnti, and Johnson, 2005), SII researchers have shown that log-based measures of instruction had statistically significant and substantively meaningful effects on 1st and 3rd grade students' reading achievement, as assessed by *Terra Nova*.

Log Sample

In this study, 75,689 daily logs were collected in grades 1 through 5. Table 2 shows that across these 75,689 logs, there were 16,890 logs where a teacher reported teaching word analysis as a lesson focus, 38,635 where a teacher reported teaching reading comprehension as a lesson focus, and 32,660 where a teacher reported teaching writing as a lesson focus. Table 2 shows that in every grade, comprehension was taught around 50% of all days, writing was taught slightly more frequently in the lower grades than in the upper grades, and a focus on word analysis was highly concentrated in grades 1 and 2.

Insert Table 2 About Here

Outcome Measures

Since the different CSR programs under study sought to change different aspects of literacy instruction, the analyses we developed examined log data at a very explicit level of detail.

Frequency of Topic Coverage. In one part of the analysis, we looked at how frequently teachers in the CSR and comparison schools taught seven broad topics in the literacy curriculum, where these topics were: (1) reading comprehension; (2) writing; (3) word analysis; (4) reading fluency; (5) vocabulary; (6) grammar; and (7) spelling. Two additional topics—concepts of print and research strategies—occurred with such low frequency that they were dropped from these analyses. In this part of the analysis, we used all days of instruction as our lesson sample, that is, we calculated the percentages of days when one of the 7 topics just listed was taught using all 75,689 lessons in the data base.

Instructional Practice Measures. A second set of analyses examined how particular topics were taught on days when they were taught. Here, the samples included the 16,890 lessons when word analysis was taught, the 38,635 lessons when reading comprehension was taught, and the 32,660 lessons when writing was taught. The purpose of these analyses was to gain

greater insight into the nature of instruction across CSR and comparison schools, controlling for the sheer frequency of instruction in these topics.

Data Reduction

In both sets of analyses, we made a number of coding decisions to reduce the complexity of the log data. In the first analysis, where the focus was on the frequency with which certain large topics in the curriculum were taught, we coded a topic as being taught if a teacher reported that that topic was a major or minor focus of instruction, and untaught if he or she reported touching on it or not teaching it at all.

In the second analysis, where we examined how reading comprehension and writing were taught (when they were taught), we were using a large number of log items that would be difficult to analyze on an item-by-item basis. As a result, we developed a measurement strategy that reduced the item-level data by creating item groupings to indicate the presence or absence of underlying dimensions or characteristics of teaching practice, where lessons were coded as 1 = characteristic present or 0 = not present if a teacher marked any one of the constituent items thought to indicate the overarching construct as occurring on a given day. The item groupings used in these analyses are shown in Tables 3, 4, and 5. Each of these tables shows how we mapped specific items from the word analysis portion of the log, the reading comprehension portion of the log, and the writing portion of the log into larger measures of an instructional variable.

Insert Table 3 About Here

Insert Table 4 About Here

Insert Table 5 About Here

It is important to note that the item groupings shown in these tables have been empirically derived, but that in the analyses presented below, items were in fact grouped

based on prior literacy research and existing theory.³ As an example, in Table 4, the reader will note that we grouped two reading comprehension items (A1a, activating prior knowledge or making personal connections to text, and A1b, making predictions, previewing or surveying) together to form a measure of a single variable we called activate knowledge. In our view, items A1a and A1b are slightly different indicators of what is essentially the same reading activity – having students prepare to read text as a means of improving comprehension as they read.

Data on Additional Classroom Characteristics

In examining program differences in curriculum coverage and instructional practice, we also controlled for a variety of classroom level variables as an additional means of assuring that our samples of classrooms were equivalent across schools (for a complete list of these variables see Table 6). Among these variables are demographic characteristics of the teachers who headed each classroom, a variety of aggregate characteristics of classrooms such as students' prior achievement and socioeconomic status, and teachers' reports of the problem behaviors of students in a classroom. In addition, we included the grade level of each classroom in order to directly examine how instruction unfolded across grades.

Missing Data

Inevitably, data on teacher and classroom characteristics were missing in the data. To combat this problem, we used the SAS multiple imputation (MI) procedure to impute missing values for classroom cases in our data set. Peugh and Enders (2004) advocate for this approach to missing data, since listwise deletion is only robust under the assumption that data are missing completely at random (MCAR). By contrast, the MI procedure used in

³ In an analysis not shown here, for example, we used the statistical software program ORDFAC to conduct factor analyses to estimate the co-occurrence of items. These analyses confirmed that the item groupings derived theoretically, and shown in Tables 3-5, were nearly identical to the factors arising empirically in the data.

this study makes the far less severe assumption that data are missing at random (MAR). The MI procedure also assumes that data are multivariate normal, but Peugh and Enders (2004) report that MI is often robust to failures of this latter assumption.

In the MI procedure used here, more than 80 variables were used in the imputation phase. The wealth of available data increases the robustness of inferences to violations of the MAR assumption. The MI procedure creates several different data sets (in our case, five) each of which contain different plausible values of the missing data given the observed values on all variables and the underlying covariance matrix (for further discussion see Peugh and Enders, 2004). Table 6 provides descriptive statistics on the raw and imputed data for all of the classroom level variables imputed in our analyses.

Insert Table 6 About Here

It is important to note that the statistical software package HLM 6.0 (Raudenbush, Bryk and Congdon, 2000) used in the analyses reported here automatically calculates the average estimates of effects of independent variables on dependent variables across the multiple data sets and then produces standard errors of these estimates that account for the uncertainty in parameter estimates due to multiple imputation.

Statistical Models

HLM Logistic Regression Models

The outcome variables in this study were dichotomous variables measuring whether or not a particular curriculum topic was taught on a given day and whether or not a particular instructional practice or activity occurred. Outcomes on days, however, were nested within teachers, who in turn were nested within schools. To take account of this nesting, we used a three-level, hierarchical logistic regression model to test the hypotheses under study (Raudenbush and Bryk, 2002: Chapter 10). In these analyses, the level 1 sampling model for

the dichotomous outcome variables was a Bernoulli distribution, where the outcome being predicted was the log odds that the dependent variable would take on the value 1 = present on a given day of observation.

At level 1 of these HLM models, the log odds of an instructional outcome occurring on a given day was modeled as varying randomly around the mean response of a given teacher within a school, and as a function of the characteristics of the days on which a given log response was recorded, for example, the day of the week, day of the year (testing for both a linear and quadratic relationship for time), and whether or not the day was a holiday or adjacent to a holiday weekend. In the models, the effects of these lesson characteristics on outcomes were treated as fixed effects. Thus, the general form of the level one regression equations was:

$$[1] \eta_{ijk} = \log [\varphi_{ijk} / 1 - \varphi_{ijk}] = \pi_{0jk} + \sum_{p=1}^P \pi_{pik} a_{pijk},$$

where η_{ijk} is the log odds that an outcome will occur on day i for teacher j in school k , φ_{ijk} is the probability that the outcome occurred on day i for teacher j in school k ; π_{0jk} is the mean for the outcome for teacher j in school k , a_{pijk} are the independent variables (e.g., day of the week) that predict instruction, and π_{pik} are the corresponding level-1 regression coefficients that indicate the strength and direction of association between each characteristic a_p , and instruction for each teacher jk .

At level two of the HLM logistic regression model, we hypothesize that instructional outcomes among teachers within the same school vary randomly around school means for that outcome and are a function of several teacher and classroom characteristics that are treated as fixed effects in the model. Thus, the level two HLM equation in each analysis was:

$$[2] \pi_{0jk} = \beta_{00k} + \sum_{q=1}^{Q_p} \beta_{qpj} X_{qjk} + r_{pj}$$

where β_{00k} is the log odds that an instructional outcome will occur in school k , X_{qjk} are the teacher/classroom characteristics described earlier (e.g., teacher and student demographic characteristics, grade level) β_{qpj} are the corresponding level 2 coefficients that represent the strength and association between each teacher/classroom characteristic and the intercept for teacher j in school k , and r_{pj} is the random effect of teacher j in school k (assumed to be normally distributed with a mean of 0 and a variance τ_r).

At level three of the HLM models, we turn to modeling variation among schools in instructional outcomes. Here, the main question of interest is whether the log odds of an instructional outcome differ across CSR versus comparison schools. It should be pointed out that in the analyses discussed below the HLM models are estimated three different times, once each for an analysis of instructional outcomes in ASP vs. comparison schools, AC versus comparison schools, and for SFA versus comparison schools. Equation [3a] shows the level three HLM logistic regression model for each of these analyses:

$$[3a] \beta_{00k} = \gamma_{000} + \gamma_{001} (\text{CSR}) + u_{00k}$$

where γ_{000} is the log odds of instruction occurring in the sample of comparison schools, CSR is an indicator variable taking on a value of 0 if a school was in the comparison group and 1 if the school was in the focal CSR program being analyzed, γ_{001} is the corresponding school level coefficient representing the strength and direction of the association between CSR program participation by a school and the instructional outcome of interest, and u_{00k} is the random effect on the outcome for school k .

An issue in this analysis is that instructional outcomes often vary in systematic ways across grade levels—as our level 2 HLM model suggests. In particular, schools can vary how much word analysis, reading comprehension, or writing instruction they offer at particular grades (e.g., in the schools in our sample, word analysis instruction generally declines across grade levels). Schools also can vary teaching strategies, student work assignments, and so on across grades (e.g., in our study sample, the nature of texts being read, or the complexity and length of written assignments typically increases across grade levels). Thus, in the analyses presented below, we also examine the extent to which the effect of grade level (included at level 2 of the HLM model) also varies across CSR vs. comparison schools. Thus, an additional equation at level 3 of our HLM model is:

$$[3b] \quad \beta_{01k} = \gamma_{010} + \gamma_{011} (\text{CSR})$$

where β_{01k} is the effect of grade level on the instructional outcome of interest in school k , and γ_{010} is the grand mean for the grade level effect on the instructional outcome across all schools in the sample, CSR is an indicator variable taking on a value of 0 if a school was in the comparison group and 1 if the school was in the focal CSR program being analyzed, and γ_{011} is the corresponding school level coefficient representing the strength and direction of the association between CSR and the grade level slope in school k .

Propensity Score Stratification

The reader will note that the HLM models just discussed control for possible differences in instructional outcomes arising from differences in the days of the week or time of year when teachers completed logs, as well as the possible influences on instructional outcomes resulting from differences among teachers in professional background and classroom composition. However, schools in the SII sample were not randomly assigned to CSR programs, and differences in school characteristics (like those shown in Table 1) existed

among schools in each CSR sample as compared to schools in the comparison sample. To contend with this problem and strengthen the matching among CSR and comparison group schools in our analyses, we implemented the strategy of propensity score stratification discussed by Rosenbaum and Rubin (1983). A detailed discussion of the specific approach to propensity stratification used here is beyond the scope of this paper, although the interested reader can consult Appendix A for a detailed discussion. The important point is that our approach produced 4 propensity strata for the ASP versus comparison school statistical models, 5 strata for the AC versus comparison school statistical models, and 4 strata for the SFA versus comparison school statistical models. In each case, schools from both the CSR and comparison groups were included in each propensity strata, and within strata, schools were balanced on 34 school-level covariates, including all of those shown in Table 1. Thus, propensity stratification allows us to control for differences across samples in many school-level characteristics simply by adding a series of dummy variables into each HLM regression analysis.

Sensitivity Analyses

Stratifying schools based on their propensity to have been in the treatment allows for an estimation of a treatment effect purged of observed differences between treated and untreated schools. An additional concern for causal inference is whether or not there are omitted variables that could explain the treatment effects. Such an omitted variable would have to have a relationship (of a particular magnitude) with the treatment (CSR program affiliation in our case), and simultaneously, have a relationship (of a particular magnitude) with the outcome (literacy instruction in our case) in order to cause a spurious correlation between the treatment and outcome. Sensitivity analyses attempt to describe the magnitude of the relationship(s) an omitted variable would need to have in order to reduce the

treatment effect enough to accept the null hypothesis. We chose a rather conservative test of omitted variable bias for some of our most important findings. This conservative test examined whether the findings were sensitive to omitted variable bias assuming the rather implausible scenario that the omitted variable had a relationship with CSR program assignment equal to the maximum value of any observed covariate in our data set, *and*, simultaneously, also had a relationship with literacy instruction equal in magnitude to the maximum value of any observed covariate. Such a scenario is implausible for several reasons. First, we have compiled a rather large data set of observed covariates. This reduces the chances we have omitted an additional variable that could have caused the treatment effect we observe. Second, the covariates we have measured represent those currently thought to have meaningful relationships to teaching and learning in schools, including prior achievement and socioeconomic status. It is difficult to imagine an omitted variable with a more robust relationship to instruction than those covariates already measured. Third, and most important, it was extremely rare that the same observed covariate had the strongest association with both the selection into a CSR program and simultaneously with the outcome - literacy instruction. Therefore, it is extremely difficult to conceive of an omitted variable that exists that would exceed (in magnitude) this conservative test.

Results

The data analyses conducted here were voluminous. For example, for each comparison of instructional outcomes in CSR vs. control group schools, we estimated 40 different HLM logistic regressions, one for each of the instructional outcomes under study. Rather than present all 120 regressions in tabular form, we instead present the key results in graphical form, focusing solely on two effects drawn from these 120 statistical models: (a) the odds ratios of an instructional outcome occurring in schools in the focal CSR program vs. schools

in the comparison group, and differences in the log odds of an instructional outcome occurring at different grades due to a school's participation in a CSR program.⁴

Literacy Instruction in ASP Schools

Figure 1 graphically depicts these key results. The left hand side of the figure depicts an odds ratio that compares the odds of a literacy topic being taught in the average ASP school vs. the odds that it was taught in the average comparison school in the sample. In addition, the figure also presents the 95% confidence interval for these odds ratios.⁵ To interpret Figure 1, it is useful to recall that odds ratios of 1 for any outcome indicate that teachers in ASP and comparison schools were equally likely to have focused on that outcome across all lessons in the study; odds ratios greater than one indicate that ASP teachers were more likely to focus on a literacy topic; and odds ratios less than one indicate that ASP teachers were less likely than teachers in comparison schools to focus on a literacy topic. By placing a confidence interval around these odds ratios, instruction in ASP schools can be said to be statistically different from instruction in comparison schools when the line representing the 95% confidence interval for the ASP estimate does not cross the line representing an odds ratio of one.

Insert Figure 1 About Here

To see how this works, note that Figure 1 displays the estimated odds ratios for ASP vs. comparison schools as black squares and the confidence intervals around these estimates by the black lines running through the squares. This is done for each of the seven literacy

⁴ Readers interested in seeing all 120 regression tables can consult Correnti (2005).

⁵ The odds of teaching a particular topic can be calculated as the number of lessons when a topic was taught divided by the number of lessons when the topic was not taught. An odds ratio is simply the odds of teaching a topic for ASP teachers divided by the odds for comparison school teachers. The odds ratio can be considered a useful effect size metric for dichotomous outcomes, and is therefore valuable for assessing the magnitude of effects across all of the various outcomes reported here.

topics at issue. As Figure 1 shows, we found no significant differences in the likelihood that ASP and comparison teachers focused on writing, grammar, spelling, comprehension, word analysis, vocabulary, or reading fluency across all lessons in the study.

In addition, the far right column of Figure 1 shows whether or not differences between ASP and comparison schools increased or decreased as grade level increased. In the left panel of the table, for example, a — indicates that there was no difference in grade-to-grade coverage of topics, a ▲ indicates that differences between ASP and comparison schools were larger as grade increased, while a ▼ indicates that differences between ASP and comparison schools decreased as grade increased. As Figure 1 shows, there were no differences between ASP and comparison schools in the rates at which topic coverage either increased or decreased across grade levels.

Figure 2 graphically depicts ASP effects on 33 additional instructional outcomes representing the frequency with which teachers used varied teaching practices and/or covered various curricular topics during lessons when they taught word analysis, comprehension, and writing. Here too, we found very few differences in literacy instruction across ASP and comparison schools. As Figure 2 shows, ASP teachers were more likely than comparison teachers to have students discuss text when reading comprehension was taught, and they were less likely to have students provide brief answers to comprehension questions when they taught comprehension. But, in nearly all aspects of literacy instruction, there was in fact no mean difference in literacy instruction across ASP and comparison schools. Figure 2 also shows that there were very few differences between ASP and comparison schools on how teaching practices unfolded across grade levels. In sum, across the 40 dichotomous literacy outcomes analyzed in Figures 1 and 2, our analysis found only two significant differences between ASP and comparison schools – exactly the number of

differences that would be predicted to be found through chance alone using a 95% confidence interval. In this sense, the evidence strongly suggests that participation in the ASP program had virtually no effect on teaching practices in schools.

Insert Figure 2 About Here

Literacy Instruction in AC Schools

In Figures 3 and 4, we turn to differences in literacy instruction across AC and comparison schools. In contrast to ASP, where only two significant differences were found among the 40 statistical contrasts, Figures 3 and 4 show that half of all contrasts estimated here (20 out of 40) were statistically significant (at $p < .05$). Moreover, as predicted, most of the differences found between AC and comparison schools were in the rate at which AC teachers taught writing and in how writing was taught when it was a focus of a days' lesson.

For example, Figure 3 shows the odds ratios for AC vs. comparison schools in the frequency with which writing was taught across all days in the year. The odds ratio of 1.95 tells us that the odds an average teacher in an AC school taught writing was 1.95 times the odds that an average teacher in a comparison school taught writing. To better understand the implications of this finding, it is useful to translate this odds ratio into a difference in probabilities. Controlling for lesson, teacher, and school characteristics, estimates from our HLM models show that AC teachers focused on writing in 54% of all lessons, whereas comparison teachers focused on writing in just 38% of all lessons⁶. Furthermore, the results

⁶ These probabilities were calculated directly from the HLM model estimates as follows. Since the AC estimate was uncentered, the probability for teachers in the comparison schools (adjusting for lesson, teacher and school covariates) was simply a function of the model intercept. Specifically, the probability for comparison school teachers was calculated by the following formula: $1/(1+\exp(-\text{intercept}))$. The probability for the average AC teacher was determined by the formula: $1/(1+\exp(-(\text{intercept}+\text{AC estimate})))$. Probabilities can be converted to odds by the simple formula $(\text{prob.}/1-\text{prob.})$. Odds ratios can be calculated after determining the odds for a teacher in an AC school (odds_{AC}) and the odds for a teacher in a comparison school ($\text{odds}_{\text{comp}}$). Again, the odds ratio is simply $\text{odds}_{\text{AC}}/\text{odds}_{\text{comp}}$.

of our sensitivity analysis for this finding showed that the AC treatment effect is not sensitive to our conservative test for omitted variable bias.

While AC teachers were more likely to conduct lessons focused on writing, they were less likely to conduct lessons focused on spelling, reading fluency and vocabulary. Moreover, differences between AC and comparison school teachers on the frequency of teaching these latter topics increased with grade level (see the far right hand column of Figure 4), suggesting that AC teachers were even less likely than comparison teachers to cover these topics at higher grade levels than at lower grade levels.

Insert Figure 3 Here

Figure 4 shows that AC teachers' also differed in the instructional practices and curricular content they covered when they taught word analysis, reading comprehension, and writing. Consistent with AC's emphasis on implementing writer's workshop within schools' literacy programs, the largest instructional differences across AC and comparison schools were found for the frequency with which writing was taught on the same day as other literacy topics. For example, on days when AC teachers focused on word analysis, they were much more likely also to focus instruction on writing (mean OR = 4.48). Likewise, when instruction focused on comprehension, AC teachers were much more likely to have a general lesson focus on writing (mean OR = 4.91) and to directly integrate writing instruction with their work in reading comprehension (mean OR=3.00). Similarly, when AC teachers taught writing, they were more likely also to have taught comprehension (mean OR = 2.09) and more likely to directly integrate comprehension instruction into their work on writing instruction (mean OR=1.54). Indeed, these are the largest effect sizes in Figure 3 and indicate a clear and consistent pattern of writing having been much more likely to be integrated with other literacy content in AC schools.

Insert Figure 4 Here

In addition, Figure 4 shows that on days when writing was taught, AC teachers were more likely than comparison teachers to have engaged in 6 of the 10 writing-related instructional practices measured in this study. For example, AC teachers were more likely than comparison teachers to explicitly teach the writing process, and as the right hand side of Figure 4 shows, these differences were largest in the lower as opposed to higher grades, showing the special emphasis AC placed on teaching writing in the lower grades. Figure 4 also shows that AC teachers were more likely than comparison teachers to provide instruction on literary techniques and on different writing genres, and to have students share their writing and do substantive revisions to their writing. Finally, Figure 4 shows that AC teachers were more likely than comparison teachers to have their students write multiple connected paragraphs as they taught writing. Again, as the right hand part of the figure shows, this difference was largest in the lower elementary grades. Finally, Figure 4 shows many instances where AC teachers were less likely to focus on a variety of instructional practices or content areas in word analysis and comprehension. We reserve comments on these findings for the discussion section.

Sensitivity Analyses

We conducted sensitivity analyses for all of the outcomes in Figure 4 showing an AC treatment effect in writing or demonstrating the integration of writing with word analysis or comprehension instruction. Of the ten sensitivity analyses we conducted, we found that eight of the treatment effects reported above were not sensitive to omitted variable bias using a significance level $p < .05$. Thus, only two of our findings did not survive our most conservative test of omitted variable bias. They were, the extent to which comprehension was directly integrated into writing instruction, and the frequency teachers had students write

multiple connected paragraphs. The latter finding was sensitive to our most conservative test of omitted variable bias but was not sensitive to a different, less conservative test which assumed an omitted variable had a relationship with both the treatment and the outcome equal to the largest observed covariate in our data set.

Literacy Instruction in SFA Schools

Figures 5 and 6 show differences in literacy instruction across SFA and comparison schools. Across both figures, we found that literacy instruction outcomes in SFA schools were different from literacy instruction outcomes in the comparison schools for 22 of the 40 contrasts estimated, far exceeding what would be expected by chance. Specifically, in SFA schools, teachers were more likely to teach comprehension on a daily basis and also to teach comprehension differently from comparison teachers when they taught this subject. Also noteworthy is the magnitude of the differences, indicating clear preferences within SFA schools for and against certain instructional practices.

Insert Figure 5 Here

In particular, Figure 5 shows that teachers in SFA schools were more likely than teachers in comparison schools to teach reading comprehension (mean OR=1.82). Converting model estimates into probabilities (as discussed in footnote 6) showed that the average SFA teacher taught reading comprehension in 65% of all lessons, while the average comparison school teacher taught comprehension in 50% of all lessons. Furthermore, we conducted a sensitivity analysis on this finding and found that this SFA treatment effect is not sensitive to omitted variable bias given our most conservative test.

In addition, SFA teachers were more likely to teach word analysis, and much less likely to have taught grammar or spelling – neither of which are an explicit focus of the SFA 90-minute reading block. Finally, the findings on the grade level coverage of curricular

topics imply a sequenced progression of instruction in SFA schools. Relative to teachers in the comparison schools, teachers in SFA schools became more likely to focus on writing as grade level increased, and less likely to focus on spelling.

Figure 6 shows differences between teachers in SFA and comparison schools in how they taught comprehension when this topic was taught. Here, we see that differences between SFA and comparison schools exist for 4 of the 10 comparisons. While small in number, these differences are consistent with program features and revealing about the nature of instruction in SFA schools. For example, consistent with SFA guidelines for the 90 minute reading period, and as shown in prior qualitative research on SFA (Datnow and Castellano, 2000), teachers in SFA schools were more likely than comparison teachers to use teacher directed instruction in comprehension lessons, to focus on literal comprehension strategies, to check students' comprehension by eliciting brief answers from students, and (due to extensive use of cooperative grouping arrangements) to have students discuss text with one another. It is noteworthy that teachers in SFA schools did not compromise any other aspect of comprehension instruction in order to obtain these significant differences⁷. That is, in lessons where comprehension was taught, teachers in SFA schools were no less likely than comparison school teachers to focus on more advanced reading strategies or write extended text about what they read. But they did focus more during these lessons on direct instruction in literal comprehension and more frequently elicited brief answers from students.

Insert Figure 6 Here

⁷ Indeed, in analyses not shown here, when taking into account all lessons, teachers in SFA schools are more likely than teachers in comparison schools to have taught all of the instructional practices in comprehension.

Figure 6 also suggests that SFA teachers concentrated on some instructional practices more than others when word analysis and writing were taught. For example, when word analysis was taught, SFA teachers were far more likely than comparison teachers to teach sight words (mean OR=2.61) and far less likely to have students analyze the structure of words (mean OR=.48) or learn letter-sound relationships (mean OR=.33). When writing was taught, SFA teachers were far more likely also to have focused on comprehension (mean OR=3.17), far more likely to have students write sentences (mean OR=2.20), and far less likely to have taught literary techniques or genre study (mean OR=.47).

Figure 6 also suggests a sequence of instructional events in SFA schools that differs from that in comparison schools. Within word analysis, for example, SFA teachers were quicker to stop 5 of the 9 instructional practices as grade levels increased, as evidenced by the negative SFA effect on the grade level slope shown in the far right hand column of Figure 6. Moreover, in both comprehension and writing, relative to teachers in the comparison schools, teachers in SFA schools were more likely to concomitantly focus on comprehension and writing on the same day as grade level increased and they more likely to actively integrate work in both subject areas. Finally, writing at the paragraph level shows a greater increase in SFA schools as grade level increases. The resulting pattern in SFA schools suggests that the SFA design promotes a sequence of instructional events different from the one naturally occurring in American classrooms.

Sensitivity Analyses

We conducted sensitivity analyses for all of the outcomes in Figure 6 showing an SFA treatment effect in comprehension or an SFA treatment effect in the integration of comprehension with word analysis or writing. Of the seven sensitivity analyses we conducted, we found that six of the treatment effects reported above were not sensitive to

omitted variable bias using a significance level $p < .05$. The seventh finding – the extent to which the teacher used literal comprehension strategies in their comprehension instruction – was not sensitive to omitted variable bias using a significance level $p < .10$.

Discussion

At the beginning of this paper, we discussed the prevailing view among educational researchers that “most educational reforms never reach, much less influence, long standing patterns of teaching practice, and are, therefore, largely pointless if their intention is to improve student learning” (Elmore, 1996). While the results of this study have little to say about the impact of “most” educational reforms on instructional practices, they are wholly inconsistent with a view that educational reforms never have an impact on instructional practice. In fact, the results presented here show quite clearly that instructional changes can be produced by design in schools, but only when innovative programs are clearly targeted at particular curricular areas, when the interventions are well-designed and well-specified, and when local leaders promote implementation fidelity.

Our results thus confirm existing theories about what it takes to get new instructional practices faithfully implemented in schools. Of the three CSR program studied here, ASP relied heavily on locally developed and weakly specified designs for change, much like the designs in the now famous RAND change agent study (Berman and McLaughlin, 1975). Unsurprisingly then, schools’ participation in the ASP program led to very little change in instruction, as evidenced by the lack of differences in instructional practices across ASP and comparison schools. By contrast, the AC and SFA programs had well-specified instructional designs that were strongly supported by on-site facilitators and local leaders who demanded fidelity to program designs. The evidence presented here shows that these programs produced distinct instructional regimes in the schools where they worked. In particular,

teachers in AC and SFA schools were shown to do more and different instruction than comparison school teachers in the areas specifically targeted by these CSR designs, while instruction in ASP schools looked almost exactly like instruction in the comparison schools.

These findings occurred despite the fact that AC and SFA had different implementation strategies. As we argued earlier in this paper, the procedural controls used by SFA to encourage implementation fidelity differed in important respects from the professional controls used by AC (Rowan, Camburn, Barnes, 2004). The findings presented here, however, suggest that both procedural and professional control strategies can be successful for changing instruction in schools. Indeed, in the present research, each strategy for supporting implementation produced large mean differences in targeted areas of literacy instruction – for SFA, in reading comprehension, and for AC, in writing.

Moreover, our findings demonstrate that changes in teachers' instructional practice need not be confined to particular literacy content areas, or to particular teaching approaches. For example, AC and SFA sought to implement very different instructional regimes in the schools where they worked. SFA worked to implement what might be called a "skill-based" teaching regime that focused on direct teaching of basic comprehension skills using fast-paced lessons that increased students' opportunities to briefly demonstrate basic comprehension of text. By contrast, AC worked to implement what might be called a "literature-based" teaching regime that used the writing process to improve students' reading comprehension. Here, teachers were more likely to directly teach writing strategies, have students write extended passages of text, and to integrate writing instruction with their reading comprehension instruction, and vice-versa.

Despite these differences, both CSR programs managed to get their intended instructional regimes implemented. And, this, we argue, resulted not from differences across

the two CSR programs, but rather because of similarities—especially similarities in the implementation strategies used by these programs. Both AC and SFA focused their change efforts on a specific content area in literacy, and both programs challenged teachers to make substantial changes in their literacy instruction. Moreover, both programs provided teachers with written materials to be referenced as needed, especially when questions arose in practice. Finally, a crucial element in both programs was the continuous presence and support provided to teachers by well-trained, on site facilitators, and the press for implementation fidelity by school leaders.

Directions for Future Research

The question for future research is whether or not the differences in instructional practices observed across CSR and comparison schools in this study might explain the possible effects of these CSR programs on student achievement. At the beginning of this paper, we noted how a previous meta-analysis of this issue (Borman et al., 2003) found considerable program-to-program variation in CSR program effects on achievement but how, absent detailed data on the implementation of particular instructional practices in CSR schools, that meta-analysis was in a weak position to explain these differences. In this paper, we presented empirical data on the differential implementation of particular instructional practices across three different CSR programs, and that, we argue, is an important—even essential—first step toward understanding issues of instructional effectiveness related to these particular programs. As a result, in the final section of this paper, we want to speculate a bit about how our findings on instructional practices in AC and SFA might be used to formulate new hypotheses about the effects of these CSR programs on student achievement.

As we have seen, the data presented here show that SFA and AC implemented two, very different instructional regimes for teaching reading and writing in the schools where

they worked. SFA, for example, produced what we called a “skill-based” literacy teaching regime in the schools where it worked. Looking closely at the content covered by SFA teachers and at the kinds of academic tasks they focused reading lessons on, it makes sense to assume that this teaching regime would be more effective than normative approaches to literacy instruction in teaching students early reading skills. After all, as our data showed, SFA teachers spent more time teaching word analysis, and were more likely than comparison teachers to emphasize direct and explicit instruction on basic reading comprehension skills. Thus, it is not surprising that many studies have shown SFA’s advantage over comparison schools in producing early reading gains in the areas of word attack, word reading, and oral reading (e.g., Borman et al., 2005; Madden et al., 1993; Slavin et al., 1996).

On the other hand, we might hypothesize that SFA’s approach to literacy teaching will produce diminishing effects at higher grade levels, where students must engage in more difficult and cognitively demanding comprehension tasks such as comparing and contrasting sections of texts, evaluating conclusions, and so on. Indeed, this hypothesis is consistent with at least some existing evidence. For example, a few studies have found that SFA effects on reading achievement are smaller at higher grades and/or for achievement tests that assess other than word attack, word reading, and oral reading (Jones, Gottfredson, and Gottfredson, 1997; Ross and Smith, 1994; Smith, Ross, and Casey, 1996; Vensky, 1994). This conclusion has been disputed, however, by the developers, and contrary evidence can be found in Borman and Hewes (2002), as well as in Ross, Sanders, and Wright (1998) study examining the effects on students’ reading achievement of the closely related Roots and Wings upper grades reading program.

The point of our discussion, however, is not to resolve existing ambiguities about SFA program effects on reading achievement, but rather to suggest that a more detailed look at

the instructional practices implemented in that (or any) program can be used to generate additional hypotheses about the program's effects on reading achievement that can be empirically tested. Indeed, a similar point can be made about research on the effects of the AC program on students' reading achievement. As we have seen, AC seeks to implement we called a "literature-based" reading regime in schools, one that placed strong emphasis on writing about reading and developing reading comprehension through extended, written essays. But as we saw, that program (as implemented at the time of our study) placed about equal emphasis as comparison schools on word analysis and basic reading skills. Since the emphasis in this program was thus on what might be thought of as "higher order" or more advanced understanding of text, we might hypothesize that the advantages of the AC program over comparison schools in promoting reading achievement would emerge at later as opposed to earlier grades. And again, there is at least some evidence to support this hypothesis. For example, the best studies of AC program effects on reading achievement have been reported in Supovitz et al. (2002) and May, Supovitz, and Dada (2004). Both studies show no or extremely small positive effects of AC program participation on reading achievement at early grades (1-3), but larger effects at later grades (4-8). Again, this is consistent with the hypothesis we derived from a detailed look at instructional practices within schools implementing this program.

The larger point is that it is time for educational researchers interested in studying innovative programs in education to "open up the black box" of schooling and look more closely at the kinds of instruction occurring in schools that adopt innovative programs. By doing so, we have shown that much more can be learned about the conditions under which instructional interventions actually succeed in producing distinctive forms of instructional practice in schools, and about why such programs do or do not produce the effects on

achievement that are found in “black box” analyses that fail to measure instructional practices in program schools.

References

- Benson, G. (2002). Study of Instructional Improvement School Sampling Design. University of Michigan, Ann Arbor: Institute for Social Research.
- Berends, M., & Bodilly, S., & Kirby, S. (Ed.s) (2002). Facing the Challenges of Whole School Reform: New American Schools After a Decade. Santa Monica, CA: RAND.
- Berman, P., & McLaughlin, M. (1975). Federal Programs Supporting Educational Change, V. 4: The Findings in Review. Santa Monica, CA: RAND.
- Bodilly, S. (1996). Lessons from New American Schools Development Corporation's demonstration phase. Santa Monica: RAND.
- Borman, G.D., Hewes, G.M., Overman, L.T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research*, 73, 125-230.
- Borman, G.D., Slavin, R.E., Cheung, A., Chamberlain, A.M., Madden, N.A., & Chambers, B. (2005). Success for All: First-year results from the national randomized field trial. Educational Evaluation and Policy Analysis, 27, 1-22.
- Camburn, E. & Barnes, C.L. (2004). Assessing the validity of a language arts instruction log through triangulation. Elementary School Journal, 105, 49-76.
- Camburn, E., B. Rowan, and J. Taylor. (2003). Distributed leadership in schools: The case of elementary schools adopting comprehensive school reform models. Educational Evaluation and Policy Analysis, 25(4), 347-374.
- Cohen, D. & Hill, H. (2001). Learning Policy: When State Education reform Works. New Haven: Yale University Press.
- Correnti, R., Rowan, B. & Camburn, E. (2003). School Reform Programs, Literacy Practices in 3rd Grade Classrooms, and Instructional Effects on Student Achievement. Paper read at the Annual Meeting of the American Educational Research Association, Chicago, IL.
- Correnti, R. (2005). Literacy Instruction in CSR Schools: Consequences of Design Specification on Teacher Practice. Dissertation Abstracts International, A66(08). (UMI No. AAT 3186604)
- Cox, P., & Havelock, R. (1982). External facilitators and their role in the improvement of practice. A study of dissemination efforts supporting school improvement. Paper presented at the annual meeting of the American Educational Research Association, March 19-23, New York, NY.
- Crandall, D., & Bauchner, J., & Loucks, S., & Schmidt, W. (1982). Models of the school improvement process. A study of dissemination efforts supporting school improvement. Paper presented at the annual meeting of the American Educational Research Association, March 19-23, New York, NY.

Crandall, D., & Eiseman, J., & Louis, K. (1986). Strategic planning issues that bear on the success of school improvement efforts. Educational Administration Quarterly, v. 22(3), 21-53.

Cuban, L. (1993). How Teachers Taught: Constancy and Change in American Classrooms 1880-1990 (2nd ed.). New York: Teachers College Press.

Darling Hammond, L. and Snyder, J. (1992). Curriculum studies and traditions of inquiry: The scientific tradition. In P.W. Jackson (Ed.), Handbook of research on curriculum (pp. 41-77). New York: MacMillan.

Desimone, L. (2002). How can comprehensive school reform models be implemented? Review of Educational Research, 72(3), 433-480.

Datnow, A., & Castellano, M. (2000). Teachers' responses to Success For All: How beliefs, experiences and adaptations shape implementation. American Educational Research Journal, v. 37(3), 775-799.

Datta, L. (1980). Changing times: The study of federal programs supporting educational change and the case for local problem solving. Teachers College Record, 82(1), 101-116.

Desimone, L., & Porter, A., & Garet, M., & Yoon, K., & Birman, B. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. Educational Evaluation and Policy Analysis, 24(2), 81-112.

Elmore, R. (1996). Getting to scale with good educational practice. Harvard Educational Review, v. 66(1), 1-26.

Elmore, R., & Burney, D. (1997). Investing in teacher learning: Staff development and instructional improvement in school district #2, New York City. Philadelphia, Consortium for Policy Research in Education and the National Commission on Teaching and America's Future.

Elmore, R.F. and McLaughlin, M.W. (1988). Steady Work: Policy, Practice, and the Reform of American Education. Santa Monica, CA: RAND

Fennema, E., & Carpenter, T., & Franke, M., & Levi, L., & Jacobs, V., & Empson, S. (1996). A longitudinal study of learning to use children's thinking in mathematics instruction. Journal for research in Mathematics Education, v. 27, 403-434.

Firestone, W.A. and Corbett, H.D. (1988). Planned educational change. In N.J. Boyan (Ed.), Handbook of Research on Educational Administration (pp. 321-340). New York: Longman.

Fullan, M.G. (1991). The new meaning of educational change. New York: Teachers College Press.

- Gersten, R. (1984). Follow Through revisited: Reflections on the site variability issue. Educational Evaluation and Policy Analysis, v. 6(4), 411-423.
- Gersten, R., & Carnine, D., & Zoref, L., & Cronin, D. (1986). A multi-faceted study of change in seven inner-city schools. The Elementary School Journal, v. 86(3), 257-276.
- Herman, R., & Aladjem, D., & McMahon, P., & Masem, E., & Mulligan, I., & O'Malley, A., & Quinones, S., & Reeve, A., & Woodruff, D. (1999). An educator's guide to schoolwide reform. Washington, D.C.: American Institutes for Research.
- House, E., & Glass, G., & McLean, L., & Walker, D. (1978). No simple answer: Critique of the follow-through evaluation. Harvard Educational Review, v. 48 (2), 128-160.
- Huberman, A. & Miles, M. (1984). Innovation Up Close: How School Improvement Works. New York: Plenum Press.
- Loucks, S. (1983). Defining fidelity: A cross-study analysis. Paper presented at the annual meeting of the American Educational Research Association, April 11-15, Montreal, Quebec.
- Loucks, S., & Cox, P., & Miles, M., & Huberman, M. (1982). Portraits of the changes, the players and the contexts. A study of the dissemination efforts supporting school improvement. People, policies and practices: Examining the chain of school improvement, Vol. II. Network of Innovative Schools, Andover, MA. ERIC no. ED240714.
- Madden, Nancy A., Robert E. Slavin, Nancy L. Karweit, Lawrence Dolan and Barbara A. Wasik (1993). Success for All: Longitudinal Effects of a Schoolwide Elementary Restructuring Program. American Educational Research Journal, 30, 123-148.
- May, H. & Supovitz, J., & Lesnick, J. (2004). The impact of America's Choice on writing performance in Georgia: First Year Results. Philadelphia, Consortium for Policy Research in Education Research Brief.
- McLaughlin, M., and Marsh, D. (1978). Staff development and school change. Teachers College Record, v. 80(1), 69-94.
- Meyer, L., & Gersten, R., & Gutkin, J. (1983). Direct Instruction: A project Follow Through success story in an inner-city school. The Elementary School Journal, v. 84(2), 241-252.
- Mirel, J. (1994). School reform unplugged: The Bensenville New American School Project 1991-1993. American Educational research Journal, 31, 481-518.
- Nunnery, J. (1998). Reform ideology and the locus of development problem in educational restructuring. Education and Urban Society, v. 30(3), 277-295.
- Peugh, J.L. and Enders, C.K. (2004) Missing data in educational research: A review of reporting practices and suggestions for improvement. Review of Educational Research, 74(4), 525-556.

Peterson, S. & Emrick, J. (1983). Advances in Practice. In Paisley, W. & Butler, M. (Ed.s) Knowledge Utilization Systems in Education. (pp. 219-250). Beverly Hills: Sage Publications.

Raudenbush, S.W. and Bryk, A.S. (2002). Hierarchical Linear Models (Second Edition). Thousand Oaks: Sage Publications.

Raudenbush, S.W., Bryk, A.S., and Congdon, R. (2004). HLM 6.0. Lincolnwood, IL: Scientific Software International.

Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 17, 41-55.

Rowan, B., Camburn, E., & Barnes, C. (2004). Benefiting from comprehensive school reform: A review of research on CSR implementation. In C. Cross (Ed.), Putting the pieces together: Lessons from comprehensive school reform research (pp. 1-52). Washington, D.C.: National Clearinghouse for Comprehensive School Reform.

Rowan, B., E. Camburn, and R. Correnti. (2004). Using teacher logs to measure the enacted curriculum in large-scale surveys: Insights from the Study of Instructional Improvement. Elementary School Journal, 105, 75-102.

Rowan, B., Raudenbush, S.W., Correnti, R., Schilling, S.G., and Johnson, C. (2005). Studying “balance” in balanced literacy instruction: How different mixes of word analysis and text comprehension instruction affect first grade students reading achievement. Paper prepared for research seminar on learning from longitudinal data, National Center for Education Statistics, Washington, DC: May, 2005.

Slavin, R. E., Madden, N. A., Dolan, L. J., Wasik, B. A., Ross, S., Smith, L., & Dianda, M. Success for All: A summary of research. Journal of Education for Students Placed at Risk, 1996, 1, 41-76.

Stringfield, S., & Datnow, A. (1998). Scaling up school restructuring designs in urban schools. Education and Urban Society, v. 30(3), 269-276.

Supovitz, J., & Taylor, B., & May, H. (2002). The impact of America’s Choice on student performance in Duval County, Florida. Philadelphia, Consortium for Policy Research in Education Research Brief.

Table 1: Demographic Characteristics of Schools by CSR Program

| | ASP (N=28) | AC (N=31) | SFA (N=30) | Comp. (N=26) |
|--|---------------|--------------|---------------|-----------------|
| School Size | | | | |
| Number of Students in School | 485 | 563 | 465 | 498 |
| Elementary Students in State | 535,798 | 719,948 | 690,486 | 746,829 |
| Community Measures | | | | |
| Community Disadvantage Index | .26 | .64 | 1.06 | .79 |
| Proportion Households in Poverty | .14 | .19 | .23 | .22 |
| Proportion Unemployed in Community | .09 | .09 | .12 | .11 |
| Proportion Households Receiving Assistance | .09 | .14 | .19 | .15 |
| Student/Family Background-Proportion Students: | | | | |
| White | .36 | .12 | .19 | .29 |
| Black | .42 | .69 | .52 | .39 |
| Hispanic | .19 | .11 | .20 | .24 |
| Asian | .03 | .08 | .09 | .08 |
| Native American | .00 | .01 | .01 | .01 |
| Receiving Free/Reduced Lunch | .62 | .75 | .74 | .64 |
| From Single Parent Homes | .37 | .49 | .46 | .38 |
| Born to Teen Mother | .22 | .22 | .20 | .18 |
| Family Receiving AFDC | .08 | .14 | .15 | .13 |
| Pre-Treatment Aggregate Achievement | | | | |
| Woodcock-Johnson Language Arts – Entering Kindergartners | 97.68 | 102.32 | 94.15 | 103.31 |
| Woodcock-Johnson Mathematics – Entering Kindergartners | 99.32 | 94.22 | 97.25 | 103.62 |
| Percent Meeting State Proficiency Standards LA – Year prior to Treatment | 31.00 | 29.83 | 30.41 | 36.49 |
| Percent Meeting State Proficiency Standards Math – Year prior to Treatment | 32.21 | 24.40 | 29.52 | 31.63 |

Table 2: Number of lessons (percent within grade) with a topic focus on word analysis, comprehension, or writing across grades.

| | <u>Word Analysis</u> | <u>Comprehension</u> | <u>Writing</u> |
|-----------------------|----------------------|----------------------|----------------|
| 1 st Grade | 6,192 (40.6) | 7,567 (49.6) | 7,283 (47.8) |
| 2 nd Grade | 4,165 (27.8) | 8,121 (54.1) | 6,940 (46.3) |
| 3 rd Grade | 2,444 (15.5) | 8,105 (51.4) | 6,536 (41.4) |
| 4 th Grade | 2,259 (14.5) | 7,848 (50.3) | 6,412 (41.1) |
| 5 th Grade | 1,830 (13.0) | 6,994 (49.8) | 5,489 (39.1) |
| Total | 16,890 (22.3) | 38,635 (51.0) | 32,660 (43.2) |

Table 3: Word Analysis Measures

| <u>Measures</u> | <u>Components</u> | <u>Log Item</u> |
|------------------------------|---|--------------------------------|
| Letter–sound relationships | -letter-sound relationships | C1a |
| | -Counting the number of sounds in a word | C1b |
| | -sound spelling/invented spelling/developmental spelling | C1c |
| | -segmenting a part of the word | C1d |
| | -other segmenting tasks | C1e |
| | -blending initial sounds with a rhyming word (onset-rime) | C1f |
| | -blending individual phonemes into real words | C1g |
| | -blending phonemes into nonsense words | C1h |
| | -blending syllables | C1i |
| | -other blending tasks | C1j |
| | Sight words | -word recognition, sight words |
| Use picture/context cues | -use of context, picture, and/or sentence meaning and structure to read words | C1m |
| Use phonics cues | -use of phonics-based or letter-sound relationships to read words in sentences or stories | C1n |
| Structural analysis | -structural analysis, examining word families, prefixes, suffixes, contractions, etc. | C1l |
| Assess Student Ability | -I listened to the target student read | C4a |
| | -I took running records or conducted a miscue analysis | C4b |
| | -I administered a word analysis test | C4c |
| Teacher Directed Instruction | -I corrected the student’s errors or modeled the correct answer | C3a C3c |
| | -I prompted the student to use the context (other words in sentence, pictures, what they already know) to read the word | C3d |
| | -I gave oral cues – sounding out parts of the word for them | |
| Focus on Comprehension | Percent of word analysis lessons reading comprehension was also a focus | 4a |
| Focus on Writing | Percent of word analysis lessons writing was also a focus | 4b |

Table 4: Reading Comprehension Measures

| <u>Measures</u> | <u>Components</u> | <u>Log Item</u> |
|------------------------------|--|-----------------|
| Activate knowledge | -Activating prior knowledge or making personal connections to text | A1a |
| | -Making predictions, previewing or Surveying | A1b |
| Literal comprehension | -Answering questions that have answers directly stated in the text | A1j |
| | -Answering questions that require inferences | A1k |
| | -Explaining how to find answers or information | A1l |
| Story structure | -Using concept maps, story maps or text structure frames | A1i |
| | -Sequencing information or events | A1m |
| | -Identifying story structure | A1n |
| | -Summarizing important details | A1q |
| Analyze/synthesize | -Comparing and/or contrasting information or texts | A1p |
| | -Analyzing and evaluating text | A1r |
| Brief answers | -Answered brief oral questions | A3a |
| | -Answered multiple choice questions | A3e |
| | -Completed sentences filling in blanks | A3f |
| | -Wrote brief answers to questions | A3h |
| Students discuss text | -Discussed text with peers | A3b |
| | -Did a think-aloud or explained how they applied a skill or strategy | A3c |
| | -Generated questions about text | A3d |
| Extended answers | -Wrote extensive answers to questions | A3i |
| | -Worked on a literature extension project | A3j |
| Teacher directed instruction | -Teacher demonstrated or explained a skill | A4a |
| | -Teacher demonstrated or explained how to use a reading strategy | A4b |
| | -Teacher explained why or when to use a reading strategy | A4c |
| Integrate writing | -examining literary techniques or author's style | A1s |
| | -written literature extension project | A1t |
| | -examined literary techniques or author's style in writing | B1c |
| | -teacher explained how to write, organize ideas, revise or edit using a published author's writing | B3c |
| Focus on Writing | -Percent of reading comprehension lessons where writing was also a focus | 4b |

Table 5: Writing Measures

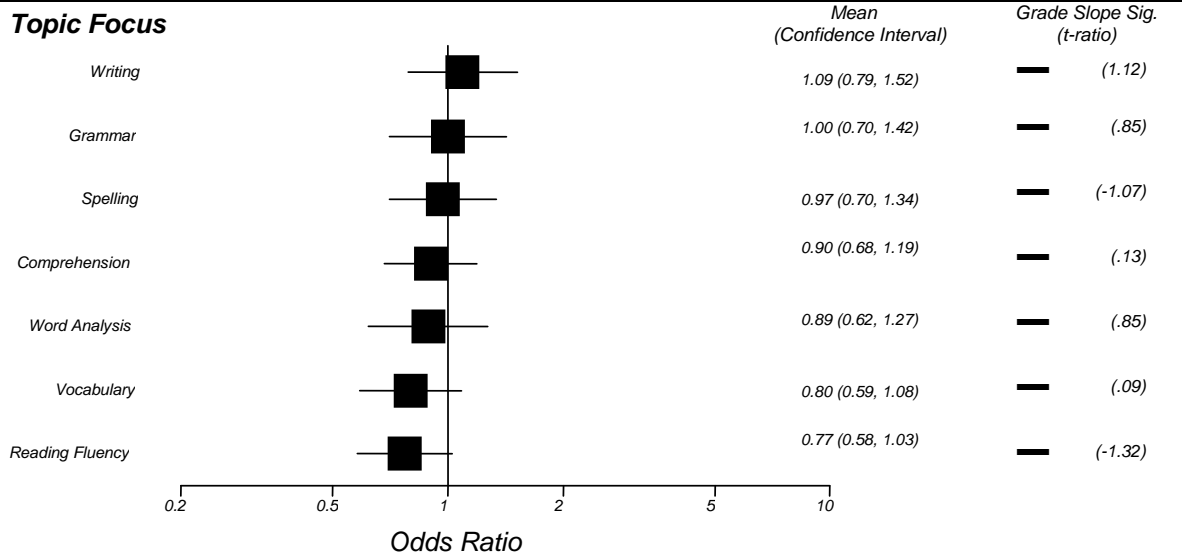
| <u>Measures</u> | <u>Components</u> | <u>Log Item</u> |
|----------------------------------|--|-----------------|
| Pre-Writing | -generating ideas for writing | B1a |
| | -organizing ideas for writing | B1b |
| Writing Practice | -writing practice | B1e |
| Revise writing | -revision of writing-elaboration | B1f |
| | -revision of writing-refining or re-organizing | B1g |
| Edit writing | -editing capitals, punctuation, or spelling | B1h |
| | -editing word use, grammar or syntax | B1i |
| Share writing | -sharing writing with others | B1j |
| Literary techniques /genre study | -literary techniques or author's style | B1c |
| | -writing forms or genres (e.g. letter, drama, editorial, Haiku) | B1d |
| Teacher comments on writing | -I commented on what the student wrote not how | B3f |
| | -I described what the student did well in his/her writing | B3g |
| Teacher directed instruction | -I demonstrated or did a think-aloud using my own writing | B3a |
| | -I explained how to write, organize ideas, revise or edit using st. writing | B3b |
| | -I explained how to write, organize ideas, revise or edit using a published author's writing | B3c |
| | -I led the student and his/her peers in a group composition | B3e |
| | -I commented on how the student could improve his/her writing | B3h |
| Focus on Comprehension | -Percent of writing lessons where reading comprehension was also a focus | 4a |
| Integrate Comprehension | -examining literary techniques or author's style | A1s |
| | -written literature extension project | A1t |
| | -wrote extensive answers to questions | A3i |
| | -worked on a literature extension project | A3j |
| Write words | -student's writing consisted of letter strings or words | B2a |
| Separate sentences | -student's writing consisted of separate sentences | B2b |
| Separate paragraph | -student's writing consisted of a single paragraph | B2c |
| Connected paragraphs | -student's writing consisted of connected paragraphs | B2d |

Table 6: Descriptive Statistics for Teacher Level Variables in HLM Analyses

| <i>Teacher Variables</i> | Raw Data | | | Imputed Data ^a | | |
|-----------------------------------|----------|----------|-----------|---------------------------|----------|-----------|
| | <u>N</u> | <u>M</u> | <u>SD</u> | <u>N</u> | <u>M</u> | <u>SD</u> |
| Grade | 1945 | 2.97 | 1.39 | 1945 | 2.97 | 1.39 |
| Male | 1817 | .10 | .30 | 1945 | .10 | .30 |
| White | 1792 | .58 | .49 | 1945 | .57 | .49 |
| Hispanic | 1792 | .11 | .31 | 1945 | .10 | .31 |
| Black | 1792 | .23 | .42 | 1945 | .23 | .42 |
| Asian | 1792 | .05 | .21 | 1945 | .05 | .21 |
| Other Race | 1792 | .03 | .18 | 1945 | .03 | .18 |
| ELA Specialist | 1833 | .07 | .25 | 1945 | .07 | .25 |
| Special Education Teacher | 1833 | .03 | .18 | 1945 | .03 | .18 |
| Has Masters Degree | 1833 | .63 | .48 | 1945 | .63 | .48 |
| Years Experience | 1818 | 12.45 | 9.85 | 1945 | 12.37 | 9.87 |
| Self-efficacy | 1816 | .06 | .98 | 1945 | .07 | .98 |
| Number of ELA Courses Taken | 1793 | 3.37 | 1.41 | 1945 | 3.36 | 1.41 |
| <i>Classroom Aggregates</i> | | | | | | |
| Average Student SES | 1927 | -.11 | .51 | 1945 | -.11 | .51 |
| Average Student Fall Achievement | 1656 | 579.94 | 49.04 | 1945 | 578.67 | 48.26 |
| Average Student Problem Behaviors | 1937 | 1.92 | .42 | 1945 | 1.92 | .42 |

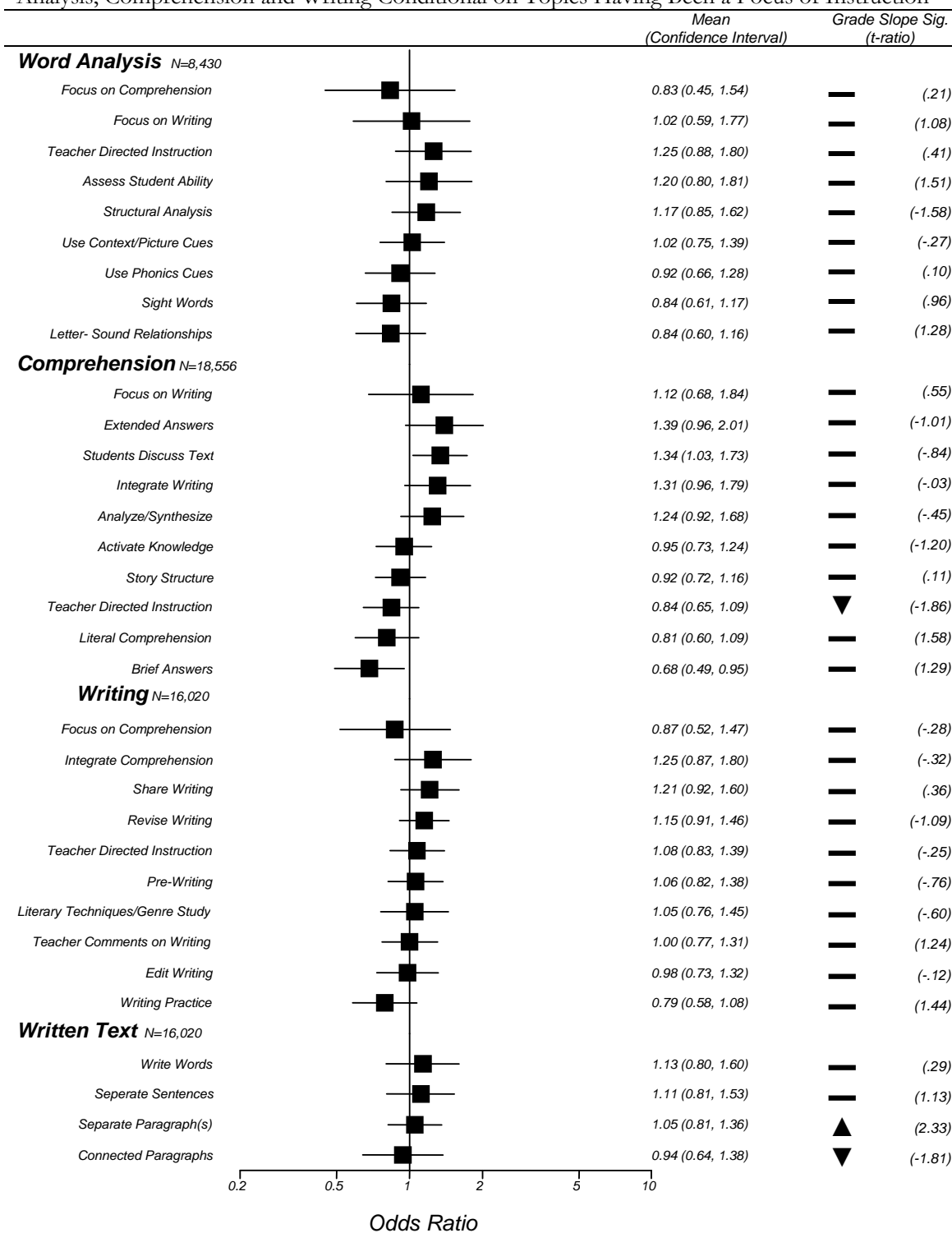
^a Means and SD's reported for the imputed data represent the average across all 5 data sets

Figure 1: Instructional Differences Between ASP and Comparison Schools in Literacy Topic Focus Across All Lessons (N=39,720)



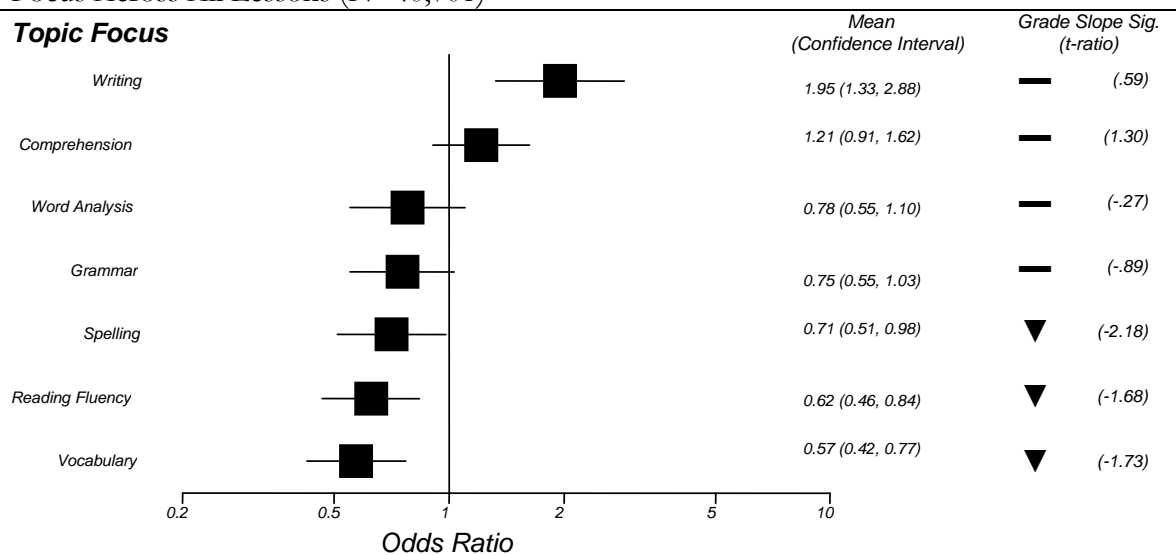
— Indicates ASP estimate was not significant on the grade level slope at $p < .10$

Figure 2: Differences Between ASP and Comparison Schools in Strategies Instruction in Word Analysis, Comprehension and Writing Conditional on Topics Having Been a Focus of Instruction



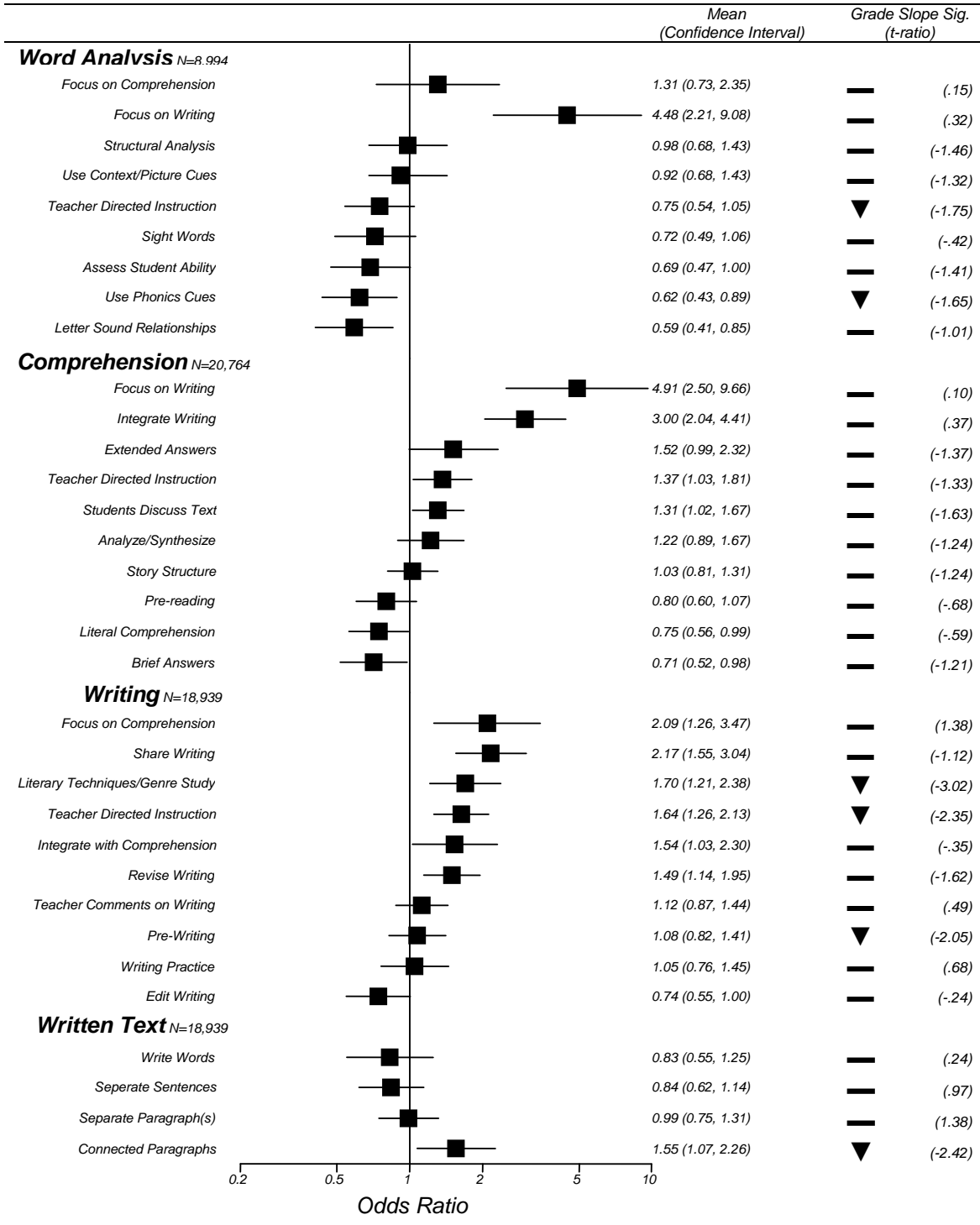
— Indicates ASP estimate on the grade level slope is not significant
 ▲ Indicates ASP estimate on the grade level slope is positive and significant at p<.10
 ▼ Indicates ASP estimate on the grade level slope is negative and significant at p<.10

Figure 3: Instructional Differences Between AC and Comparison Schools in Literacy Topic Focus Across All Lessons (N=40,701)



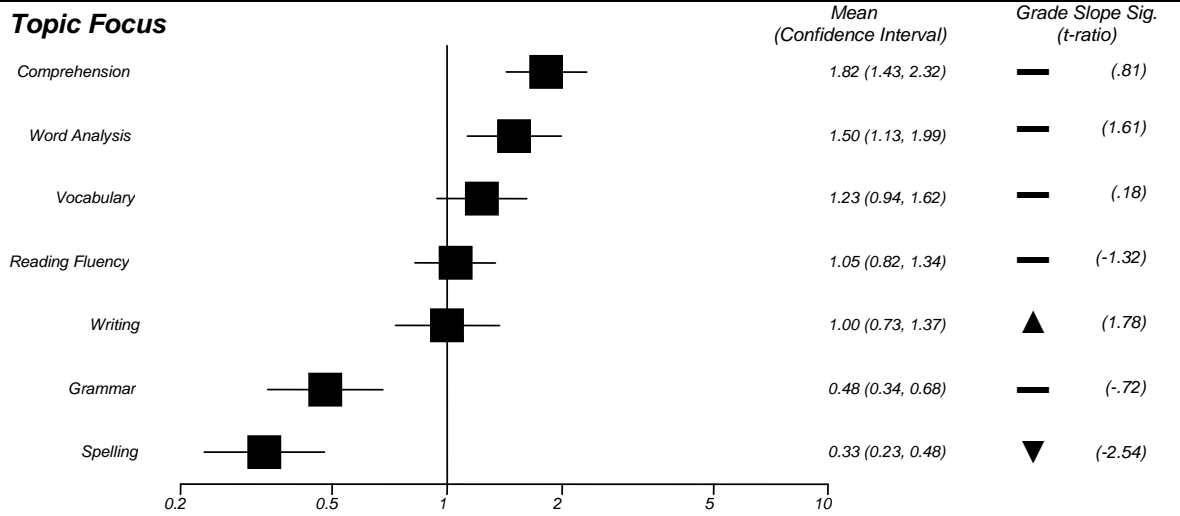
— Indicates AC estimate was not significant on the grade level slope at $p < .10$
 ▼ Indicates AC estimate on the grade level slope is negative and significant at $p < .10$

Figure 4: Differences Between AC and Comparison Schools in Strategies Instruction in Word Analysis, Comprehension and Writing Conditional on Topics Having Been a Focus of Instruction



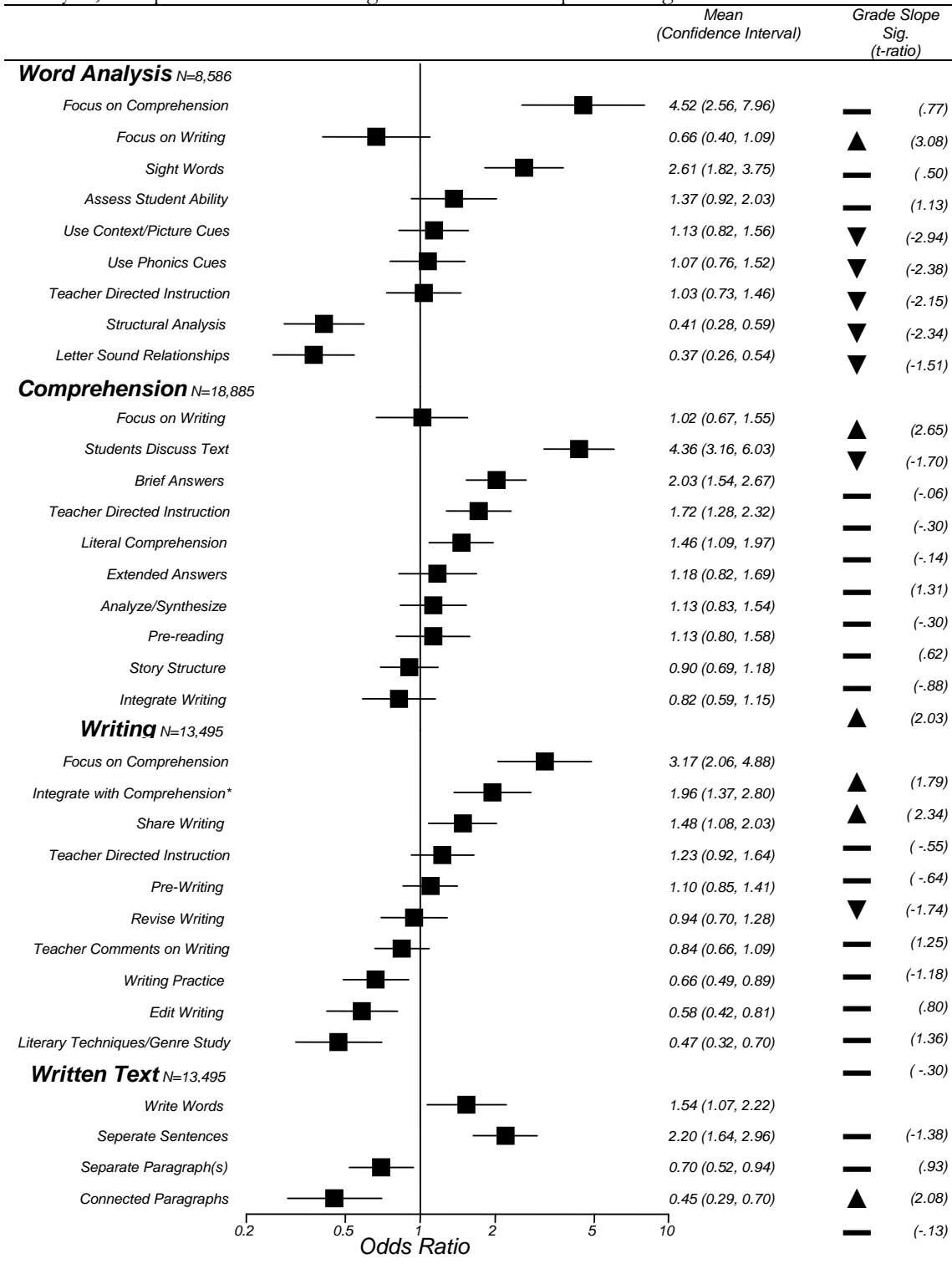
— Indicates AC estimate on the grade level slope is not significant
 ▲ Indicates AC estimate on the grade level slope is positive and significant at p<.10
 ▼ Indicates AC estimate on the grade level slope is negative and significant at p<.10

Figure 5: Instructional Differences Between SFA and Comparison Schools in Literacy Topic Focus Across All Lessons (N=34,182)



- Indicates SFA estimate was not significant on the grade level slope at $p < .10$
- ▲ Indicates SFA estimate on the grade level slope is positive and significant at $p < .10$
- ▼ Indicates SFA estimate on the grade level slope is negative and significant at $p < .10$

Figure 6: Differences Between SFA and Comparison Schools in Strategies Instruction in Word Analysis, Comprehension and Writing Conditional on Topics Having Been a Focus of Instruction



— Indicates SFA estimate on the grade level slope is not significant
 ▲ Indicates SFA estimate on the grade level slope is positive and significant at p<.10
 ▼ Indicates SFA estimate on the grade level slope is negative and significant at p<.10

Appendix A

The logic of propensity score stratification is as follows. Each unit, whether treated or not, has two potential outcomes Y_1 (if treated) and Y_0 (if control). The causal effect of the treatment is the difference between Y_1 and Y_0 for each unit. Since the unit either belongs to the control group or the treated group, it is impossible to observe both Y_1 and Y_0 for a given unit. However, we can estimate the average causal effect of a treatment in a population under the assumption that treatment assignment is independent of the potential outcomes. In that case, the average of the treated cases minus the untreated cases provides an unbiased estimate of $E(Y_1 - Y_0)$, which is the population average causal effect.

In the absence of random assignment, suppose it is possible to identify subsets of units (e.g., schools) which have the same distribution on all observed covariates but who differ in treatment assignment. Then, for this subset, treatment assignment is effectively random if no unobserved covariates predict treatment assignment. This is exactly what propensity score stratification accomplishes. It statistically equates subsets of units, in my case schools, on all observed covariates. Thus, we can estimate the average causal effect by pooling estimates of the within-stratum causal effect under the assumption of strongly ignorable treatment assignment. That assumption states that unobserved covariates are unrelated to treatment assignment given the observed covariates.

These methods were applied in a multi-step process. First, it was necessary to identify an exhaustive list of observed pre-treatment and exogenous characteristics of schools that could have theoretically confounded the treatment. The strength of the causal argument, under strongly ignorable treatment assignment depends on the assumption that the observed covariates are more likely to confound treatment than any unobserved covariates. Table A1 displays the 34 covariates we used to create the propensity score, and the source of the variables within the SII data.

Next, we examined each covariate individually to determine if there was a difference in means between each set of CSR schools and the set of comparison schools. Once covariates with significant differences were identified, we entered them as independent predictors of the probability of CSR (treatment) assignment. We ran a stepwise logistic regression model (with entry into the model conditional on a p-value of .10 or less). The stepwise model insures a parsimonious model is fit to the data. During this step, we saved each school's predicted probability (propensity) of being a treatment school. Next, we separated the schools into five equal strata based on their propensity to have received treatment.

Once schools were stratified it was important to check that schools were balanced within stratum on their propensity to be in the treatment and on each of the 34 observed pre-treatment covariates. First, we checked the difference in means between the predicted probabilities of the treated and control groups within each stratum. This confirmed that the continuous probability measure was roughly the same for treated and untreated schools within each stratum. Next, for all 34 covariates, we checked for within-stratum mean differences between treated and control groups. Using significance testing at a p-value of .05 would result in 95% of the 175 covariate contrasts being statistically insignificant through chance alone. In our final propensity models, we found that the AC propensity stratification yielded 5 strata and resulted in 97% of the contrasts on the covariates being insignificant, while SFA propensity stratification yielded 4 strata and resulted in 98% of the contrasts on the covariates being insignificant and ASP propensity stratification also yielded 4 strata and resulted in 98% of the contrasts on the covariates being insignificant. The final step in the

process of propensity score stratification involved taking the dummy-coded stratum variables and entering them into the regression models analyzing the outcomes of the study.

Table A1: List of School-Level Covariates Used to Obtain Propensity Scores for CSR Program Assignment

| <u>Variable</u> | <u>Source</u> |
|---|-------------------|
| Community Disadvantage Index – School Census Tracts | |
| Community Disadvantage Index – Community Census Tracts | |
| Proportion households with assistance income | |
| Proportion households in poverty | 1990 Census |
| Proportion individuals without a high school diploma | (pre-treatment) |
| Proportion of single parent households | |
| Proportion of unemployed individuals | |
| Inverse Median Income | |
| Percent of students in school... receiving free lunch | |
| ...White | |
| ...Black | |
| ...Hispanic | Common Core |
| ...Asian | of Data (CCD) |
| ...Native American | for year prior to |
| ...Other race | treatment |
| Number of students in school | |
| Number of students in district | |
| Number of schools in district | |
| Average socioeconomic status of students | |
| Average level of education attained by students' mothers | |
| Average number of students whose mother dropped out of high school | |
| Average number of people in students' household | |
| Average number of students' siblings | School Level |
| Percent of students from single parent home | Aggregate of |
| Percent of students born to a teenage mother | Student |
| Percent of students coming from households where parents ran out of food in last 12 months | Information |
| Percent of students coming from households where parents did not have resources to buy kids' clothing in last 12 months | Obtained from |
| Percent of students coming from households where parents received Aid for Families with Dependent Children (AFDC) in last 12 months | SII Parent |
| Percent of students coming from households where parents received food stamps in last 12 months | Interview |
| Percent of students who repeated a grade | |
| Average Reading Score on Woodcock-Johnson for entering Kindergarten Students | Aggregate |
| Average Math Score on Woodcock-Johnson for entering Kindergarten students | SII Student |
| | Achievement |
| | Data |
| Percent of students meeting state proficiency standards in Reading | State and |
| Percent of students meeting state proficiency standards in Mathematics | District Web- |
| | sites |