AERA Open
July-September 2016, Vol. 2, No. 3, pp. 1–15
DOI: 10.1177/2332858416655782
© The Author(s) 2016. http://ero.sagepub.com

Supporting Common Core Instruction With Literacy Design Collaborative: A Tale of Two Studies

Joan Herman
Scott Epstein
Seth Leon

National Center for Research on Evaluation, Standards, and Student Testing

The article examines the results of two quasi-experimental studies of the implementation and impact of the Literacy Design Collaborative (LDC), an intervention designed to support secondary teachers' transition to Common Core State Standards in English language arts. The first study examines LDC implementation by eighth-grade social studies and science teachers in districts across Kentucky; the second study is set in sixth-grade advanced reading classes in a large urban district in Florida. Based on teacher surveys, logs, and analysis of classroom artifacts, the LDC was implemented with reasonable fidelity across both studies. Based on available assessment scores, results show statistically significant positive effects in Kentucky for reading but no corollary effect in Florida. There were no significant differences in writing scores in either site. The conclusion shares hypotheses that may explain the differences in results and reflects on implications for evidence-based practice.

Keywords: *evaluation, literacy, research utilization, quasi-experimental analysis, regression discontinuity, regression analyses, Literacy Design Collaborative, LDC, Common Core State Standards, evidence-based practice, artifact analysis*

Common Core State Standards (CCSS) have become a political punching bag in states and districts across the country. Controversy abounds over charges of federal intrusion eroding local control, pushback against new assessments being used to evaluate teachers, and assessments which are viewed as too time consuming. Yet remarkably, despite this political storm, institutional support for the content embedded in the standards remains widespread, particularly in English language arts (ELA). Only a handful of states did not initially adopt the Common Core ELA standards, and although a number of states have reviewed the standards in response to political pressure, only a small number have fully repealed them, with other states making smaller adjustments. Furthermore, new standards adopted in two of the repealing states, Indiana and South Carolina, were found to closely match the content of the CCSS (Heiten, 2015).

What this means is that the majority of students are being engaged with CCSS or CCSS-like ELA standards, which institutionalize key shifts in expectations for students and teachers. Among the key shifts are student engagement with complex texts and associated academic language in multiple disciplines, reading and writing grounded in evidence, and knowledge building through content-rich, informational texts (CCSS Initiative, 2015). The CCSS extends to content area teachers' new responsibility for students' literacy development and call on ELA teachers to teach a balance of literary and informational texts (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010).

The value of integrating literacy and content development in increasing student learning is well established. For example, randomized experimental evaluations of several interdisciplinary literacy interventions—such as the Seeds of Science/Roots of Reading program (Duesbery, Werblow, & Twyman, 2011; Goldschmidt, 2010; Wang & Herman, 2005) and the Great Exploration in Math and Science program (Pompea & Gek, 2002)—found statistically significant impacts on student content knowledge, vocabulary, and writing. Similarly, the Reading Like a Historian curriculum (Wineburg, Martin, & Monte-Sano, 2013), which develops students' disciplinary reading ability in history, was found to have a significant positive effect on students' historical thinking, ability to transfer new thinking strategies to their everyday lives, mastery of history content knowledge, and reading comprehension growth (Reisman, 2012).

Yet CCSS shifts are challenging for not only ELA teachers, who have traditionally focused instruction on literary rather than informational text, but also teachers in other content areas, who may lack pedagogical content knowledge to integrate literacy development into their teaching. The Literacy Design Collaborative (LDC) initiative offers one strategy to support teachers in shifting instructional practice

Creative Commons CC-BY: This article is distributed under the terms of the Creative Commons Attribution 3.0 License (http://www.creativecommons.org/licenses/by/3.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

and aligning classroom instruction to new expectations. The LDC is a template-based approach to designing instructional units that culminate in a content-based expository or argumentative writing-from-reading task. Featuring backward design processes (Tyler, 1949; Wiggins & McTighe, 1998), LDC tools provide scaffolding to ensure alignment with standards but also give teachers and instructional designers the freedom to build curricula that work for their own classrooms and schools. In this way, the LDC reflects long-standing theory and empirical research establishing the importance of what Berman and McLaughlin (1978) dubbed *mutual adaptation*, with schools and teachers changing their practice to meet the demands of new programs while adapting programs to meet their needs and changing circumstances.

In this article, we share the results of two studies of the LDC's early implementation and impact on student learning in two very different contexts. The first study examines the LDC in eighth-grade social studies and science classrooms in five small districts in Kentucky, while the second study examines the LDC implemented districtwide in sixth-grade Advanced Reading classrooms in a large diverse urban district in Florida. The studies include the first rigorous evaluation of the effectiveness of the LDC using quasi-experimental design (QED) analyses. At the same time, the studies provide a reminder of the importance of local context. Although mutual adaptation is essential to the success of educational interventions, variation in implementation can make it challenging to compare results across sites. In our evaluation, one study showed a positive impact on student achievement, while the other showed no impact. In our conclusion, we share hypotheses that may explain the differences in results, but further evaluation work is needed to build greater knowledge on the implementation factors and conditions essential to LDC success.

Research questions that we explore in this article include the following:

- How did teachers implement the LDC in the different sites?[1]
- What was the impact of the LDC on student learning in the different sites?
- How consistent are findings across sites?

Background on the LDC

With funding from the Bill and Melinda Gates Foundation, development of the LDC began in 2009 with literacy experts building a framework to help teachers incorporate literacy instruction into core subject instruction. The framework was piloted in 2010–2011 and refined with teacher feedback. Subsequently, the LDC's use has spread rapidly across the country, including statewide adoptions in Kentucky, Colorado, Louisiana, and Georgia.

At its heart, the LDC is a platform for instructional design that helps teachers incorporate literacy instruction into content-specific curricular units called *modules*. LDC modules, which typically involve 3 to 4 weeks of instruction, are built around a culminating writing assignment called the *LDC teaching task*. Teachers choose from a menu of templates to design a task that is relevant to their curriculum. Students base their writing, which can be explanatory or persuasive, on texts that they read during the LDC module. Here is an example of a filled-in template task:

> [Should industry continue to produce genetically modified crops for human use by utilizing genetic engineering?]
>
> Argumentative/Analysis at 3 levels: After reading [several current articles and scientific sources], write an [essay] that argues your position pro or con. Support your position with evidence from your research. L2. Be sure to acknowledge competing views. L3. Give examples from past or current events or issues to illustrate or clarify your position.

After deciding on the end-of-module writing task, teachers then use an LDC-specified framework—namely, the instructional ladder—to design activities to support students in developing the requisite literacy skills and content knowledge to successfully complete the culminating task. The steps of the ladder include core literacy activities that scaffold student learning and provide ongoing opportunities for formative assessment—such as note taking, identifying evidence to support claims, and evaluating contrasting positions. The instructional design platform also prompts teachers to align modules with literacy and content standards.

The LDC has a strong basis in theory and research. By aligning instructional design to college- and career-ready standards, the LDC aims to increase the rigor of classroom assignments and assessments, which itself is associated with greater student learning (Mitchell et al., 2005; Newmann, Bryk, & Nagaoka, 2001). By engaging teachers in the design process and analysis of student work, the LDC increases the capacity of teachers, and by exposing students to rigorous content and extended tasks, the LDC challenges them to engage with ELA standards and take a more active role in the learning process. The LDC therefore incorporates all three elements of Elmore's instructional core (2008), which he theorizes are the only ways to directly improve student learning at scale: raising the level of content, increasing the skill and knowledge of teachers, and increasing the level of active student learning.

Also, as noted earlier, mutual adaptation is a key feature of LDC implementation, which recent studies have linked to positive impacts on student achievement in a variety of contexts, including professional development for afterschool facilitators (Hirsch, Deutch, & DuBois, 2011), teachers as instructional leaders (Goldring et al., 2015), and teachers implementing the CCSS (Supovitz & Spillane, 2015). Similarly, other studies have found that some professional learning endeavors failed largely because of their inability to adapt to changing circumstances, whether the program was professional development in math (Borko, Koellner, & Jacobs, 2014; Koellner & Jacobs, 2015),

science (Trauth-Nare, 2016), or cross-disciplinary elementary classrooms (Hudson, 2015).

Study Contexts

We report the results of two parallel studies of the LDC in two implementation contexts (for full technical reports, see Herman et al., 2015a, 2015b). The first study examined LDC implementation in eighth-grade history/social studies and science classrooms in five districts in Kentucky. The districts were small- to medium-sized countywide rural and suburban districts across the state, with mostly White students and a substantial proportion of students eligible for free or reduced-price lunches. Depending on the district, teachers were required to participate in the LDC or volunteered to do so.[2] The implementation strategy was similar across districts: LDC teachers participated in summer professional development to orient them to the LDC and to develop initial modules, and teachers met during the school year with project leads to refine modules and analyze student work. Teachers typically worked in pairs to develop modules. Each teacher was required to implement at least one module during the fall and one during the spring, although the specific timing for module implementation was at the teacher's discretion.

The second study took place in a large county district in Florida serving a student population with diverse ethnicity and socioeconomic status. The study focused on LDC implementation in sixth-grade Advanced Reading classrooms. Relative to the state, the study district was more diverse in ethnicity and proportion of English language learners and slightly lower in socioeconomic status and student achievement, but students placed in Advanced Reading were on average higher performing than the mean student in the state, according to prior-year reading scores. Although labeled Advanced Reading, the class enrolled students in the middle- to high-achieving range.

The Florida district took a unique centralized approach to LDC implementation. District literacy leaders, on-site reading coaches, and teachers from 10 pilot schools initially developed four LDC modules to make up the core of the sixth-grade Advanced Reading curriculum. The modules were piloted in 2010–2011 at the 10 schools and revised with teachers during subsequent summer professional development sessions. Teachers were expected to implement the modules according to a districtwide pacing schedule and were provided detailed plans for the instructional activities and culminating performance tasks. Modules focused primarily on informational as opposed to literary texts.

Overall Study Design, Data, and Instrumentation

Study designs for the two states were very similar and examined the implementation and impact of the LDC in the 2012–2013 school year. Analyses focused on teachers with at least 1 year of LDC experience prior to 2012–2013 (in 2010–2011 or 2011–2012). Implementation measures included a twice-weekly teacher log, an end-of-year teacher survey, and a rubric-based analysis of the quality of module artifacts. Impacts on student learning were estimated via QED analyses. In both Kentucky and Florida, coarsened exact matching (CEM) was employed to identify matched samples of control students from outside the treatment districts who (a) were similar in demographics and prior achievement, (b) attended schools that were similarly effective prior to the LDC, and (c), in the case of the Kentucky study, were in classrooms led by teachers with similar prior effectiveness based on their students' state assessment scores in years prior to LDC implementation. Hierarchical linear model (HLM) regressions were then used to estimate the impact of the LDC on student learning, as measured by state assessment scores and, in Florida, local assessments. The Florida study also included regression discontinuity design (RDD) analyses, which compared the achievement of students within the district just below and above the threshold for entry into Advanced Reading. More detailed information on the methodological approach for the matched control group and regression discontinuity analyses can be found later preceding each set of results.

Sample

The teacher sample was defined as all teachers in the study districts charged with implementing the LDC in their classrooms during the 2012–2013 school year and who had at least one prior year of experience implementing the LDC. All teachers were included in the outcome analyses, but implementation data are limited to only those teachers who voluntarily agreed to complete those measures. Table 1 displays the overall teacher samples by state used for the QED analyses, as well as the completion rates for the different implementation measures. Roughly half of teachers in each site participated in the log and teacher artifact portions of the study, which were more time-consuming than the survey.

Table 2 displays the demographic characteristics of students in Kentucky and Florida who were taught by study teachers. The sample in Florida is considerably larger. Students in the Florida sample also were more diverse ethnically and were more likely to be eligible for free or reduced-price lunches or to be English language learners. Table 3 shows the distribution of Kentucky students across social studies and/or science classes. Over half the students in the Kentucky study sample received LDC instruction in both social studies and science classes; about one third were exposed in social studies but not science; and about 10% were exposed in science alone.

TABLE 1
Teacher Sample for Quasi-Experimental Design Analyses and Implementation Measures by State

	Kentucky		Florida	
Study Component	*n*	%[a]	*n*	%[a]
Total teacher sample	36	—	101	—
Teachers completing implementation measures				
Logs	18	50	52	51
Teacher artifacts	18	50	—[b]	—[b]
Survey	27	75	56	55

[a]Relative to all eligible teachers.
[b]All Florida teachers implemented the same modules, so artifacts were not submitted by individual teachers.

TABLE 2
Demographic Characteristics of Kentucky Eighth-Grade Students and Florida Sixth-Grade Advanced Reading Students Taught by Literacy Design Collaborative Study Teachers

Demographic Characteristic	Kentucky (*n* = 2,529)	Florida (*n* = 6,926)
Ethnicity		
Hispanic	3.0	26.1
White	90.0	47.8
Black	3.0	14.7
Other (Asian, American Indian, Alaskan Native)	4.0	11.3
Free/reduced-price lunch eligible	47.0	51.4
English-language learner	0.4	1.3
Gender: female	49.0	52.3
Special education	10.0	4.2

TABLE 3
*Distribution of Literacy Design Collaborative Students by Social Studies and/or Science (*n *= 2,529)*

Literacy Design Collaborative Exposure	*n*	%
In social studies and science	1,429	56.5
In social studies only	827	32.7
in science only	273	10.8

Implementation Measures

Our implementation measures draw on research on instruction and instructional change, given that the ultimate goal of the LDC intervention is to align teachers' instruction to college- and career-ready standards. Classroom practice is notoriously impervious to reform (Cuban, 1984; Lortie, 1975); however, an emerging body of research has documented the relationship between student achievement and specific instructional practices that create opportunities to learn (see Bryk, Sebring, Allenworth, Luppescu, & Easton, 2010; Rowan & Correnti, 2009; Winters & Herman, 2011). Our implementation measures thus focus on classroom instruction while recognizing that multiple factors influence and inhibit teacher innovation and instructional change.

Teachers in both states were asked to complete a web-based teacher log twice weekly during implementation for each of two LDC modules. In the Florida study, teachers were asked to log specifically on two of the four main district-required modules. In Kentucky, teachers were asked to log on one module during the fall and one in the spring. Typically, these two modules represented the entirety of Kentucky teachers' module instruction during the school year. The logs focused on (a) the degree to which instruction aligned with the structure of the LDC intervention, (b) the degree to which instruction explicitly specified and addressed the discrete literacy skills required to complete the summative task, and (c) the quality and extent of formative assessment practices incorporated into LDC instruction. The log was designed to capture descriptive data on classroom instruction on the particular days that the log was completed and focused on only one of the teacher's classes—the same class for all logs.

Our research team analyzed the quality of LDC modules in both states. In Florida, the district provided the four sixth-grade Advanced Reading modules with detailed daily lesson plans, texts, and activities. In Kentucky, we requested that participating teachers submit artifacts in conjunction with logging. The artifacts requested included (a) a completed teaching task (often printed from an online LDC design tool available to many teachers), (b) copies of all texts used in the module, (c) one sample of supplementary materials used during the reading component and one from the writing component, and (d) three samples of student work on the culminating writing task, marked *high, medium*, and *low*.

Our module quality rubric included nine dimensions of quality. Attending to both content and literacy demands, the dimensions address the quality of the central writing task and the texts that it draws on, the quality of the instructional ladder, and overall module coherence. Each dimension was scored on a scale of 1–5, where 1 indicated poor quality, 3 that the quality of the dimension was moderately realized, and 5 that the quality was fully realized (see online Appendix A for a copy of the rubric).

For the Kentucky analyses, social studies and science teachers were recruited as scorers and received special training to ensure that they could consistently apply the rubric. Scorers established their consistency before scoring, and consistency was checked throughout the scoring process. The measurement quality of the resulting scores was established through generalizability, factor analysis, and decision study methodologies.[3] Both the social studies and science analyses revealed low rater variance across scoring

dimensions (between 0% and 14% of total variation depending on the dimension and subject) and high teacher and/or teacher by module variation (between 28% and 72% depending on the dimension and subject), suggesting that the scores were capturing real differences in module implementation across teachers. Moreover, based on factor analyses, all nine dimensions loaded on a single factor for both subjects, making the case that our module quality rubric effectively measures a coherent trait that might be understood to be LDC implementation or, perhaps more generally, instructional quality in the integration of literacy and content.

For the Florida analysis, which involved a much smaller number of modules, a simpler rating process was employed with the lead rubric developer and a middle school social studies teacher trained for the Kentucky analyses scoring each of the four modules. Any score discrepancies were addressed through consultation and consensus.

The teacher survey—developed with a research partner studying broader implementation and scale-up issues in a larger sample of LDC teachers—was administered online at the end of the school year. Survey questions asked teachers to reflect on LDC implementation during the school year and included a section on module implementation with items design to mirror the intent of the teacher log, as well as sections on experience using the LDC, attitudes regarding literacy instruction, extent of professional development, leadership support, and collaboration. Like the logs, the survey intent was to provide descriptive data on LDC implementation.

Outcome Measures

QED analyses rely primarily on state assessment scores to measure the impact on student learning. The Kentucky study used data from the Kentucky Performance Rating for Educational Progress (K-PREP) in reading, writing, and social studies. K-PREP contained multiple-choice and short constructed-response items in a blended model of criterion- and norm-referenced testing. According to test specifications, the reading items are evenly distributed across the domains of key ideas, craft and structure, integration of ideas, and vocabulary acquisition, reflecting the major domains of the CCSS. Passage genres are weighted toward informational text (55% vs. 45%, respectively; Pearson, 2013a). The writing assessment features on-demand writing samples. At Grade 8, students respond to a passage-based argumentative prompt and have a choice of stand-alone narrative or explanatory prompts. The social studies assessment uses multiple-choice items supplemented with three extended-response items to address the domains of government and civics, cultures and societies, economics, geography, and historical perspective. Reported reliability for the eighth-grade reading and social studies tests are .87 and .90, respectively (Pearson, 2013b).

In Florida, our primary outcome measure was developmental scale scores from the 2012–2013 Florida Comprehensive Achievement Test (FCAT 2.0) in reading. The FCAT 2.0 reading development scale scores are vertically aligned to track student longitudinal progress from year to year, from Grades 3 to 10. The sixth-grade FCAT 2.0 reading test is composed of 50 to 55 multiple-choice items. According to test specifications (Florida Department of Education, 2012), the items on the sixth-grade reading test are allocated proportionally into the following categories: vocabulary (20%); reading application (30%); literary analysis, fiction/nonfiction (30%); and informational text (20%).

We additionally received student scores on the district's writing assessment. The district test mirrored the Florida Writes assessment (which was part of the state assessment system for Grades 4, 8, and 10) for grades not assessed by the state and was administered at the beginning and end of the year. In contrast to the writing-from-reading emphasis of the LDC, district and state writing assessments used prompts that could be answered solely with students' prior knowledge. The prompts called for narrative, expository, or persuasive writing.

Implementation Results

We now present abbreviated summaries of implementation data to provide context for outcome results.

Module Quality

Table 4 presents descriptive results for the artifact analysis, including mean scores and standard deviations on each dimension for each state and subject. Kentucky social studies and science modules were analyzed and are reported separately because raters exclusively scored modules in their subject area. The majority of average dimension scores are ≥3.0, and nearly all are ≥2.7. However, as evidenced by the reported standard deviations, there is considerable variation in quality within each of the three groups of modules, particularly in Kentucky.

The artifact analysis found that module designers were generally successful in building effective writing tasks and selecting texts. Developing quality instructional strategies to support student learning and ensuring that the modules were clear and coherent were more challenging tasks for some of the teachers in Kentucky.

Teacher Logs

Teachers in Kentucky spent on average 3 to 4 weeks on each of two modules that they implemented in the school year. In Florida, module instruction made up a majority of the instructional year, with about 6 weeks spent on each module. Teacher logs in both studies revealed that teachers spent module

TABLE 4
Descriptive Statistics for Module Quality Rubric Domain Ratings by State and Subject

	KY Social Studies (n = 22)		KY Science (n = 15)		FL Advanced Reading (n = 4)	
Dimension	*M*	*SD*	*M*	*SD*	*M*	*SD*
Effective writing task	3.4	1.3	3.3	1.1	3.8	0.8
Alignment to literacy and content standards	2.1	1.2	3.2	1.2	3.1	0.5
Text alignment	3.5	1.3	3.6	1.3	3.9	0.5
Text appropriateness	3.4	1.0	3.7	1.0	3.4	0.6
Text rigor	3.6	1.3	3.3	1.4	3.3	0.8
Fidelity to LDC module instruction	2.7	1.3	4.0	1.0	4.0	0.0
Quality instructional strategies	2.9	1.1	3.2	1.2	3.8	0.8
Coherence and clarity of module	2.8	1.3	3.5	1.2	3.8	0.8
Overall impression	2.7	1.0	3.3	1.2	3.3	0.4

Note. Ratings are on a 1- to 5-point scale, where a score of 1 indicates that a dimension is not in evidence, 3 indicates that quality was moderately realized, and 5 indicates that quality is fully realized. LDC = Literacy Design Collaborative.

instructional time in a variety of ways—including lecturing on subject matter content, giving mini-lessons, supporting student skill development via explicit strategy instruction, leading whole-class discussion, facilitating small group work, and allowing time for independent reading and writing. In each state and subject, teachers on average reported spending a larger proportion of time on independent reading and writing than on any other classroom activity. The log had branch-out questions for the four LDC module stages: introducing the module, reading process, transitioning to writing, and writing process. Patterns of log responses demonstrated that in both studies, teachers followed this basic LDC structure. Within each stage of module implementation, teachers focused on a variety of student skills, although there was considerable variation across teachers. During the reading process stage of module instruction, there was a strong focus on basic skills, such as independent reading and research, note taking, and summarizing important points by all groups of teachers. There is also evidence that many teachers focused on more critical reading skills, such as drawing conclusions from text, citing textual evidence to support claims, and evaluating the strength of evidence. Even more advanced skills were less of a focus, such as comparing arguments in two or more texts and examining authors' biases. In the writing process stage, examining text structure, writing different types of paragraphs, and incorporating evidence were focuses of instruction across both states. Teachers also reported using a variety of strategies for assessing student understanding and providing feedback when formative assessment revealed misunderstandings.

Teacher Surveys

Survey findings on classroom implementation of the LDC generally echo the log findings. Teachers in all settings reported spending classroom time on each of the four module stages, focusing at least some attention on a broad range of reading and writing skills, and using a variety of formative assessment and feedback strategies.

Teachers in both states reported that the LDC initiative was supported by district and school administrators, although some teachers felt that administrators did not have a firm understanding of the LDC. Although school administrators generally encouraged teachers to participate in the LDC, they were less likely to be actively engaged in the LDC by giving teachers feedback on LDC instruction.

A majority of teachers reported participating in professional development, viewed their LDC colleagues as collaborative, and reported that collaboration helped them implement the LDC in a variety of ways. Perhaps as a result of collaboration and other support, respondents seemed confident in their ability to implement modules. Some teachers, however, did report barriers to successful implementation, such as difficulty locating content-rich reading materials and finding time to prepare for module instruction and provide feedback to students.

Overall, teachers reported that they found the LDC to be a useful tool in meeting a range of instructional goals, including implementing standards, integrating literacy into content area instruction, and increasing the rigor of writing assignments. Despite this general support, opinions were mixed about student engagement relative to regular instruction, with many teachers reporting that substantial proportions of students struggled with LDC demands. Nevertheless, most teachers agreed that the LDC resulted in higher-quality student writing and supported students' college readiness.

Kentucky Outcomes Analysis

Method

Our Kentucky LDC teacher sample for the outcomes analysis includes 37 eighth-grade social studies and science teachers in the five target Kentucky school districts who

began teaching the LDC in either 2010–2011 or 2011–2012 and continued implementing the LDC in 2012–2013. The eligible student sample for the analysis includes 2,529 students (a) who were enrolled in an eighth-grade social studies or science class taught by one of the 37 teachers and (b) for whom prior achievement scores were available.

We used CEM to identify comparison students from other districts across Kentucky. CEM is a flexible approach that allows the researcher to specify the precise conditions under which a comparison student may be matched with an intervention student. This process was applied for each of the three outcome measures (K-PREP reading, writing, and social studies), resulting in three matched data sets. Creating separate matched data sets for each outcome maximized the sample size for each outcome analysis, as patterns of missing data varied across outcomes. Although we included indicators for students, teachers, and schools, all matching was at the student level. Student characteristics in the model included demographic variables (race/ethnicity categories, gender, free and reduced-price lunch eligibility, etc.) as well as prior achievement on reading and science state assessments (one prior achievement variable was based on the combined predictive capacities of the two assessments). In addition, our matching methodology selected comparison students whose teachers had similar prior effectiveness. Prior effectiveness was produced by calculating a teacher's value added on student learning in 2009–2010. The assessments used for this variable depended on the outcome measure that we were testing; the matching model used to test the impact of the LDC on writing used writing scores for 2008–2009 and 2009–2010 to calculate prior teacher effectiveness and likewise for reading and social studies. Students under teachers without prior effectiveness data were matched to comparison students under teachers with missing data as well (most were likely new to the profession). Finally, a school prior effectiveness variable was calculated with prior seventh-grade science, math, and reading assessment data. Seventh-grade data were used to ensure that the school effectiveness variable was independent of the teacher effectiveness variable in the matching model. Valued-added models used to produce the teacher and school effectiveness variables can be found in online Appendix B. These variables were also used in the HLM regressions estimating the impact of the LDC on student achievement, as described later.

The CEM process found similar matches for a majority of the eligible 2,529 LDC students. Ninety-one percent of treatment students were retained in the sample for the writing outcome analysis after matching, as were 88% of treatment students for the reading analysis and 90% for the social studies analysis. See Table 5 for a summary of the number of treatment and control students before and after matching. The matching models were effective in achieving close balance with regard to prior student scores and demographics, as well as for teacher and school effectiveness indicators.

TABLE 5
Summary of Kentucky Treatment and Comparison Samples by Outcome

Sample, *n*	Treatment	Comparison
Eligible for matching	2,529	43,333
Matched sample		
For writing	2,300	12,208
For reading	2,232	13,174
For social studies	2,284	18,265

For each of the three outcome measures, two separate two-level HLMs were employed. Each HLM modeled students' dosage under treated and nontreated teachers in eighth-grade science and social studies courses. Results are robust across model specification, and we therefore present results from only one set of models. In the models presented, each observation at Level 1 represents one student. Level 2 observations represent the combination of a social studies and science teacher. To simplify the design, students with more than one science or social studies teacher were randomly assigned to one of those multiple teachers. Therefore, each Level 1 observation is associated with one Level 2 observation. This does not present a substantial problem, as a majority of students in 2012–2013 were associated with only one science and one social studies teacher. Prior teacher and school effectiveness indicators were aggregated as cumulative sums for the teacher combination at Level 2. The two-level HLM equation employed to analyze student effects for each outcome can be seen in online Appendix B.

The measures of teacher and school effectiveness described earlier, as well as student-level demographic and prior achievement variables, were also used as value-added controls at the HLM regression stage (see online Appendix B for model equations). Our analysis therefore controlled for observables in two ways, at the matching and modeling stages. The models also examined potential interactions between the LDC treatment and prior school and teacher effectiveness as well as student characteristics. These interaction variables were intended to test whether the LDC had differential effects on student learning depending on the school, teacher, and/or individual student's standing on the given variable. Given the relatively small teacher sample size overall and the fact that the interaction analyses cut the data into yet smaller slices, we consider the interaction results somewhat exploratory. Nevertheless, we think that they are worth noting, particularly because of the similarity in some of the interaction findings across the two states.

Results

HLM results for each of the three primary outcomes are displayed in Tables 6–8. The tables include a number of

TABLE 6

LDC Student Effect Estimates on K-PREP Reading: 2012–2013—Including Interactions With Prior Teacher Effectiveness and Student Characteristics

Level 2 Variables	Model Coefficient (*SE*)
LDC treatment	0.058 (0.023)*
LDC treatment by teacher effectiveness	–0.181 (0.202)
LDC treatment by student characteristic interactions	
Female	–0.004 (0.017)
Special education	–0.110 (0.034)*
Free/reduced-price lunch eligible	0.053 (0.017)*
Prior achievement	0.034 (0.011)*

Note. Fixed effects for demographic predictors and for prior school and teacher effectiveness not shown. K-PREP = Kentucky Performance Rating for Educational Progress; LDC = Literacy Design Collaborative.
*$p = .05$.

TABLE 7

LDC Student Effect Estimates on K-PREP Social Studies: 2012–2013—Including Interactions With Prior Teacher Effectiveness and Student Characteristics

Level 2 Variable	Model Coefficient (*SE*)
LDC treatment	–0.026 (0.023)
LDC treatment by teacher effectiveness	–0.288 (0.082)*
LDC treatment by student characteristics interactions	
Female	0.013 (0.016)
Special education	–0.007 (0.037)
Free/reduced-price lunch eligible	0.039 (0.019)*
Prior achievement	0.050 (0.017)*

Note. Fixed effects for demographic predictors and for prior school and teacher effectiveness not shown. K-PREP = Kentucky Performance Rating for Educational Progress; LDC = Literacy Design Collaborative.
*$p = .05$.

TABLE 8

LDC Student Effect Estimates on K-PREP Writing: 2012–2013—Including Interactions With Prior Teacher Effectiveness and Student Characteristics

Level 2 Variable	Model Coefficient (*SE*)
LDC treatment	0.030 (0.042)
LDC treatment by teacher effectiveness	0.004 (0.120)
LDC treatment by student characteristics interactions	
Female	–0.032 (0.031)
Special education	0.031 (0.047)
Free/reduced-price lunch eligible	–0.002 (0.027)
Prior achievement	0.016 (0.016)

Note. Fixed effects for demographic predictors and for prior school and teacher effectiveness not shown. K-PREP = Kentucky Performance Rating for Educational Progress; LDC = Literacy Design Collaborative.

interactions between treatment status and student characteristic variables, as well as the interaction between treatment status and the prior effectiveness of the teacher. Each teacher combination observation at Level 2 received a value of 0 if neither teacher was treated, 1 if one of the two was treated, and 2 if both teachers were treated. Thus, the treatment effect coefficients for each model represent the effect of one treated teacher. While the models controlled for all student, teacher, and school indicators previously discussed, we limit our presentation and discussion in the body of this article to the intervention effects of interest; full results can be found in online Appendix B. Table 6 shows HLM results for the K-PREP reading scores. The data indicate that the LDC had a small, statistically significant positive effect on students' reading performance. LDC students scored higher in reading than did the comparison group, demonstrating that the LDC had a measurable effect on students' literacy learning.

To provide a benchmark for interpreting this effect, we used a relatively new methodology to convert the effect size into a gross indicator of the number of months of learning that it represents (see Hill, Bloom, Black, & Lipsey, 2007). Following this approach, we used available data to estimate the growth in K-PREP reading scores from eighth to ninth grade. We then determined the proportion of typical growth represented by the observed LDC effect size—that is, the LDC effect size divided by the effect size expected from Grade 7 to Grade 8. We then used this proportion to calculate the number of additional months of growth that can be associated with LDC relative to a 9-month academic year. Relative to typical growth in reading from eighth to ninth grade, the calculation found that the effect size for the LDC represents 2.2 months of schooling. Given that a typical Kentucky teacher spent 4 to 8 weeks teaching the LDC, the effects of the LDC appear substantial.

The data also show interactions between LDC effects and student characteristics. Students' prior achievement, based on their prior-year K-PREP scores, and students' socioeconomic status, as revealed by their free or reduced-price lunch status, both show positive interactions with the treatment. That is, LDC students who were relatively higher achieving prior to their LDC experience showed relatively greater benefit than did those who started relatively lower achieving, although the observed effect is very small. Interestingly, LDC students eligible for free or reduced-price lunch also appear to have benefited more from the LDC, after controlling for other variables. Although the observed effect was very small, we speculate that because the LDC intervention provided a common approach for teachers across subjects and grades, it may have encouraged districts and schools to utilize Title I and other resources for disadvantaged students in a more strategic, coordinated way in LDC schools—but these interaction results require replication and, if so, further inquiry. We did not find evidence of differential effects of the LDC by gender. Controlling for other factors, special

education students appear to have done less well under the LDC; however, the share of students falling into this category was small.

The results for K-PREP social studies are shown in Table 7. The coefficient for the main effect for the LDC is small and not statistically significant, which indicates that the LDC's addition of literacy to course requirements did not diminish students' content performance. Table 7 also reveals a significant interaction between prior teacher effectiveness and the LDC. LDC students taught by teachers who were relatively less effective prior to the LDC benefited more than did students of relatively more effective teachers. However, this interaction is difficult to interpret and should be treated cautiously given that all teachers', including science teachers', prior effectiveness scores were based on their students' eighth-grade social studies performance for the study's baseline year (Kentucky does not assess science in eighth grade). Students' prior-year performance on the K-PREP and their free or reduced-price lunch status show the same, small positive interaction with LDC treatment status as in the reading outcome model. LDC students who started the year performing at a relatively higher level experienced more benefit from the LDC in their social studies performance, as did students who were from a relatively lower socioeconomic status. We did not find differential treatment effects of the LDC by gender or special education status.

K-PREP writing results, as shown in Table 8, show neither main nor interaction effects for the LDC. There is no evidence of any impact of the LDC intervention on this particular writing assessment.

Florida Outcomes Analysis

Method

Two types of QEDs were utilized in the Florida study to estimate the effect of the LDC on student achievement: a matched control group design similar to that applied in Kentucky and a RDD. The matched control group analyses compare the full population of LDC students in sixth-grade Advanced Reading courses in the study district with selected matched students from across Florida in Advanced Reading or language arts classes. As in Kentucky, CEM was used to select matched students. The RDD takes advantage of a natural experiment created by the study district's selection process for entry into the Advanced Reading course. The design focuses on students near the cut point for entry into Advanced Reading and compares the performance of students that just made it into the course with students who just missed being assigned to the course.

Each of the two designs has advantages and disadvantages. The matched control group design includes a much more complete set of students receiving LDC instruction in sixth-grade Advanced Reading. However, because the LDC was implemented districtwide and matched control students were selected from outside the district, it is difficult to tease out the impact of the LDC from the effect of other district programs and conditions. That is, we cannot rule out the possibility that differences between treated and matched control samples are due to other district-specific effects. With the RDD, these district effects were controlled because both treatment and control group students came from the district. However, the sample of students came from a specific portion of the prior achievement distribution adjacent to the cut point; therefore, the estimates may not be generalizable to the full population of LDC students. Another advantage of the RDD is that it allows us to test the effect of the LDC on the local district writing measure, which is not possible in the matched control group analysis.

Two types of HLMs were employed in our matched control group analyses: (a) a three-level model with student at Level 1, school by year at Level 2, and school at Level 3 and (b) a two-level model with student at Level 1 and school at Level 2. Equations for both models can be found in online Appendix C. The three-level model used 2009–2010 as the baseline year and estimated effects in each subsequent year (2010–2011, 2011–2012, and 2012–2013) as compared with baseline. This model provides information on achievement trends for study district Advanced Reading students relative to matched controls across all of the years of implementation. The analysis shows a large dip in performance in the study district from 2009–2010 to 2010–2011 relative to statewide control students for both cohorts of schools: the Phase 1 pilot schools that began implementing the LDC in 2010–2011 and Phase 2 schools that had not begun LDC implementation yet. Given that this negative effect is seen for both Phase 1 and Phase 2 students, it is unlikely to have been the result of LDC implementation but rather the result of other unexplained district factors or conditions. One possible explanation for the negative effect in 2010–2011 is the introduction of the new state reading test: FCAT 2.0. Decreases in performance are common with the introduction of new assessments, and the study district may have been affected by this shift in a differential way from the state at large due to unexplained district conditions.

Given the overall drop in performance in the study district in 2010–2011 and the fact that the state assessment changed in the same year, we rely and report on analyses that use data only from the new test (starting with 2010–2011). A two-level HLM was implemented estimating LDC impact in 2012–2013 while controlling for student prior achievement in 2011–2012 and the prior effectiveness of schools in 2010–2011. We excluded the 10 pilot schools from this analysis, as they had already started implementing the LDC in the baseline year.

As in Kentucky, CEM was used to identify matched students from outside the target district. In Florida, the process did not include teacher prior effectiveness, as those data were not available. Matching incorporated both student and

TABLE 9
Summary of Florida Treatment and Comparison Samples

Sample, *n*	Treatment	Comparison
Eligible for matching	5,548	14,523
Matched sample	5,338	9,241

school characteristics, but all matching was at the student level. Prior school effectiveness was obtained by running a two-level HLM that controlled for student characteristics at Level 1 and by saving out empirical Bayes estimates of school value added at Level 2. The equation as well as the results for this value-added model can also be found in online Appendix C.

The process found matches for a majority of the eligible 5,548 LDC students, as seen in Table 9. The matching models were effective in retaining a large percentage of treatment observations (96.2) and in achieving close balance on prior student scores and demographics as well as school effectiveness.

As with the Kentucky analyses, student and school covariates were included as controls in the HLM regression model as well as the matching protocol, and our estimates are therefore double robust. Interaction analyses were also included.

We turn now to the methodology for the RDD analyses. We implemented RDD analyses for four school years: 2009–2010, 2010–2011, 2011–2012, and 2012–2013. As noted earlier in the report, the LDC intervention began with 10 Phase 1 pilot schools in 2010–2011, followed by expansion to all middle schools in the district in 2011–2012 and 2012–2013. For the 2010–2011 analysis, we removed students in the 10 Phase 1 pilot schools in 2010–2011. This methodological decision allowed us to analyze two cohorts of students prior to LDC intervention and two cohorts of students in the postintervention period.

We implemented several standard tests to determine whether our setting was amenable to the RDD, including tests to determine (a) if the assignment measure (FCAT prescore) was used to determine assignment to Advanced Reading, with an apparent threshold at a cut score; (b) what the threshold on the assignment measure was; (c) whether the density of the assignment measure was continuous through the threshold indicating potential manipulation/gaming of the assignment measure; and (d) whether the observable characteristics of students were continuous through the threshold.

The primary measure used for determination of assignment in Advanced Reading was the prior-year FCAT reading score. From graphical and frequency-based analyses, it is clear that while there was no single cut point that correctly classified every student into Advanced Reading classes, there were cut points that fairly sharply assigned the students. For example, in 2012–2013 only 2.1% of students with a prior 2011–2012 FCAT score of 214 or 215 were assigned into Advanced Reading, in contrast to 96.0% of students with a score of 216 or 217. Thus, we use a cut point of 216 for the 2012–2013 RDD analysis. Graphical analysis also verified the discontinuity in the probability of Advanced Reading assignment on either side of the cut point, suggesting that a "fuzzy" RDD would be appropriate. Figure 1—graphing the relationship between FCAT prior year scores and assignment into Advanced Reading in 2012–2013—clearly demonstrates this discontinuity. Similar graphs for 2011–2012, 2010–2011, and 2009–2010 can be seen in online Appendix D (Figures D1–D3). In addition, we performed a McCrary test and visually examined density graphs in each year to test for evidence of manipulation around the cut points; we found none.

Having taken the proper predesign steps, we then proceeded with RDD outcome analyses. Analytic steps in the outcome analyses included identifying a bandwidth of students around the assignment cut point for inclusion in each model and producing a local average treatment effect estimate and standard error. The two-stage least squares approach is outlined in online Appendix D. In addition, a graph plotting the assignment measure and outcome estimate was used for visual inspection of potential discontinuity.

Results

Results for the matched control group analysis (two-level HLM) presented in Table 10 show no evidence of LDC impact on FCAT reading performance in either direction. The interaction of treatment status with prior student achievement is positive and statistically significant at the 5% level, suggesting that the LDC may be more effective for students with higher levels of prior achievement, consistent with Kentucky findings. We did not find statistically significant interaction effects for other demographic variables. The model reported here excludes schools without prior effectiveness data in 2010–2011 (i.e., new schools). An alternative specification including these schools found similar results. Full results for the model reported in Table 10 can be found in Table C2. The online Appendix also includes results from the three-level model described earlier.

Regression discontinuity results for each year are presented in Table 11. The assignment variable was standardized around a cut point score of 0 to ease comparison across years, as the scale of the FCAT score changed over time. The bandwidths shown in Table 11, which determine the observations included in the analysis, were obtained with the Imbens and Kalyanaraman optimal bandwidth algorithm. The results indicate no evidence of a discontinuity in outcomes around the threshold for assignment to Advanced Reading in either the years before or the years after the LDC intervention began. Figure 2 shows this graphically for

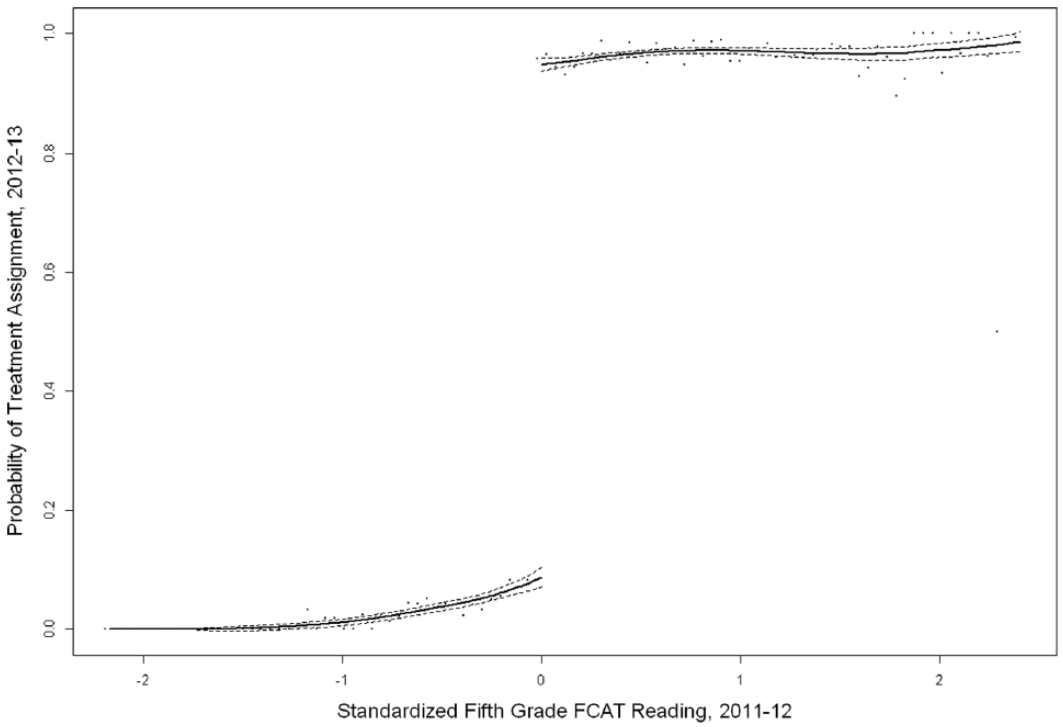

FIGURE 1. *Relationship between assignment variable (fifth-grade reading in 2011–2012) and probability of assignment into advanced reading in 2012–2013. FCAT = Florida Comprehensive Achievement Test.*

TABLE 10
LDC Student Effect Estimates on FCAT Reading: 2012–2013—Including Interactions With Student Characteristics

Level 2 Variables	Model Coefficient (*SE*)
LDC treatment	–0.051 (0.048)
LDC treatment by student characteristics interactions	
Gender	0.073 (0.038)
Free/reduced price lunch eligible	–0.049 (0.038)
Prior achievement	0.056 (0.025)*

Note. Fixed effects for demographic predictors and for prior school effectiveness not shown. FCAT = Florida Comprehensive Achievement Test; LDC = Literacy Design Collaborative.
*$p = .05$.

TABLE 11
LDC Student Regression Discontinuity Design Local Average Treatment Effect Estimates on Sixth-Grade FCAT Reading Scores: 2009–2010 to 2012–2013

Year	Bandwidth	Observations	Estimate (*SE*)
2009–2010	1.032	8,324	0.031 (0.035)
2010–2011	0.749	5,136	–0.025 (0.057)
2011–2012	0.692	6,112	–0.015 (0.029)
2012–2013	0.597	6,313	–0.002 (0.022)

Note. FCAT = Florida Comprehensive Achievement Test; LDC = Literacy Design Collaborative.

2012–2013, the key postintervention year. There is no evidence of discontinuity at the standardized cut point of 0 on the prior FCAT reading score (the assignment variable). Similar graphs for 2011–2012, 2010–2011, and 2009–2010 can be seen in online Appendix D (Figures D4–D6).

Regression discontinuity results for the share of students achieving a basic level of proficiency on the district writing assessment are presented in Table 12. The district writing assessments were scored with a 6-point rubric in 2010–2011 and 2011–2012 based on the rubric for the state writing test. In 2012–2013, the study district began using the LDC 4-point rubric to score writing assessments. Basic performance constituted a score of 3 on the earlier rubric and a score of 2 on the later rubric. Otherwise, the methodology was similar to that used for the analysis of FCAT reading outcomes. As with reading, we see no evidence of a discontinuity in the share of students who have at least a basic level of proficiency according to the writing assessment at the threshold for Advanced Reading assignment in either the year before or the years after the LDC intervention began. Graphs found in online Appendix D confirm these findings (see Figures D7–D9).

Discussion and Conclusions

We return now to a review of our research questions and what they reveal about the LDC's support for the CCSS implementation and impact. In examining whether and how

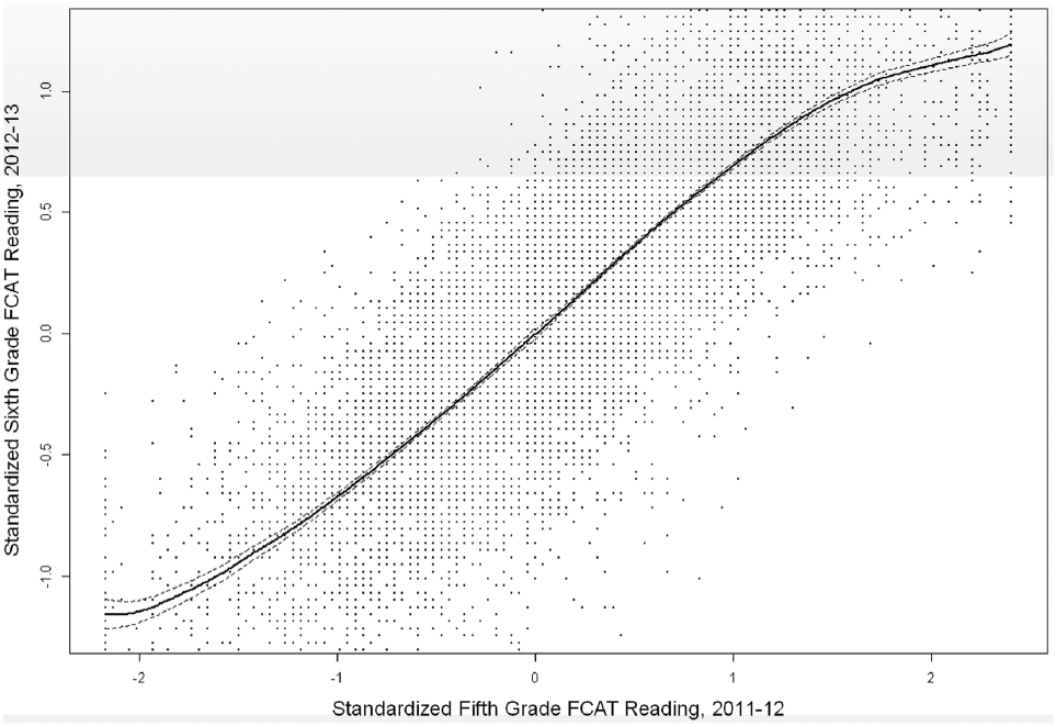

FIGURE 2. *Regression discontinuity graph for 2012–2013 showing relationship between assignment variable (fifth-grade FCAT reading in 2011–2012) and the outcome variable (sixth-grade FCAT reading in 2012–2013). FCAT = Florida Comprehensive Achievement Test.*

TABLE 12
LDC Student Regression Discontinuity Design Local Average Treatment Effect Estimates on Probability of Scoring at a Basic Level on the District Writing Measure

Year	Bandwidth	Observations	Estimate (*SE*)
2010–2011	0.968	5,378	0.043 (0.032)
2011–2012	1.006	6,986	0.001 (0.026)
2012–2013	0.574	5,299	0.028 (0.034)

Note. LDC = Literacy Design Collaborative.

teachers implemented the LDC across both studies, our data indicate that overall the LDC was implemented with a reasonable level of fidelity (although implementation quality did vary across teachers). Because the LDC model requires that teachers closely align their instructional design and teaching to the CCSS, our results thus suggest that the LDC was successful in the study districts in terms of engaging students and teachers with Common Core content. That the data also found positive attitudes among teachers regarding LDC usefulness is a reason for encouragement at this early stage of implementation.

Our outcome analyses reveal a less consistent picture. We found a positive effect on student reading scores in Kentucky but no corollary effect in Florida. Impacts on social studies and writing scores were not found. Importantly, however, both studies found evidence that the LDC has a differential impact on students based on their prior achievement. Interaction effects suggested that that lower-ability students were struggling to meet the LDC's CCSS-aligned demands, a finding mirrored in teacher results. This latter finding is worth underscoring as one that demands attention in other studies of the CCSS and one that the LDC is already working to remedy.

That student outcome findings diverged could have perhaps been foreseen given differences in study contexts. We can speculate on some of the implementation and technical factors that may have influenced the relative effectiveness:

Difference in course context: Through implementation in content courses in Kentucky, the LDC functionally may have added literacy instruction beyond what students were receiving in their ELA classes alone, while in Florida, the LDC changed the ELA coursework. In fact, a majority of students in our Kentucky QED were exposed to the LDC in both their social studies and science classes, as seen in Table 3. Did the additional literacy time in Kentucky make a difference?

Bottom-up versus top-down implementation strategy: In Kentucky, teachers developed their own modules; in Florida, module development was centralized and

mandated as a nearly full-year curriculum. Did teachers feel greater ownership in Kentucky and thus more fully implement LDC modules than they did in Florida (see Berman & McLaughlin, 1978)? Even though modules in Florida were rated higher in quality than in Kentucky, might Florida teachers have been overwhelmed by the many and detailed literacy strategies the modules laid out?

Sensitivity of outcome measures: Kentucky's K-PREP reading assessment claims alignment with the CCSS and, consistent with the LDC, features a balance of informational and literary text passages, while Florida's FCAT at the time of the study was less aligned with Common Core and prioritized comprehension of literary text. Do the differences in study findings reflect the relative sensitivity of the two measures? Neither state's assessment closely mirrored the LDC requirement for writing that draws on multiple sources, which may help explain lack of impact in either state.

Incomplete implementation data: As is routine for protection of human subjects, teachers' completion of study implementation measures was voluntary. Only about 50% volunteered to participate; how and whether nonresponders implemented the LDC is unknown, as is how their implementation influenced student outcomes.

Unmeasured variables: Given the quasi-experimental nature of the studies' designs, any number of unmeasured variables may have influenced study results, in addition to those above. We return to a consideration of these in the next section.

These and other factors represent caveats to current study results as well as hypotheses about the reasons for observed differences. Future research will need to grapple with questions that these study data cannot answer.

These same caveats also cause us to reflect on current mandates for the use of evidence-based practices, most recently renewed in the Every Student Succeeds Act (Civic Impulse, 2015). The dictum seems so obvious yet is difficult to implement in practice. The pedagogical shifts required by the CCSS require teachers to change their practices and incorporate new strategies, but it takes time to accumulate evidence of the impact of such new strategies. In addition, constraints on research may limit the generalizability of that evidence, as current results demonstrate. That it takes time for teachers to gain comfort and expertise with new practices extends the time required to accumulate solid evidence even further. In fact, our study may have been premature, and in the rush to evaluate impact, we may have shortchanged the LDC. Even so, states, districts, and schools need to make more immediate decisions than what the timing of rigorous research may enable.

The reality of how districts and schools implement reforms is a second conundrum for evaluation. Our study focused on LDC implementation in classroom practice and its impact on student learning. Although we queried issues of professional development and leadership support, we could not adequately take into account a host of other, multilevel state, district, and school policies and practices involved in whether and how reform actually gets delivered (Fixsen, Naoom, Blasé, Friedman, & Wallace, 2005). For example, in addition to clear differences in implementation plans, Kentucky was the first state to adopt the CCSS, and it initiated a strong statewide architecture to promote focus on standards implementation, including the LDC (Holliday & Smith, 2012). In the Florida district that we studied, there were a number of competing initiatives, including a nationally visible districtwide initiative on teacher evaluation (Curtis, 2012). In hindsight, we suspect that these differences in foci may have had an influence, but they are just one of a number of interacting factors that likely influenced LDC implementation and effects, just as any number of community and local context variables no doubt make a difference in the messy arena of state and local use of research (Ansen, 2015).

A closely related issue is the nature of successful interventions, which may work against generalizability of findings. We speak here of the previously cited importance of local adaptation (Berman & McLaughlin, 1978), meaning that implementation can differ greatly across sites, which can be beneficial but also potentially produce differential effects, as observed in our two studies. With districts and schools adapting implementation structures to local contexts, how should fidelity of implementation be defined? How much consistency should be expected in observed effects?

Furthermore, the alignment of intervention goals with outcome measures continues to challenge validity in the identification of evidence-based practices. Because of the cost of developing and/or administering special study outcome measures, evaluators commonly draw on existing state or district assessments; in fact, much of our knowledge on intervention effects is built on these measures. Yet, if the constructs measured by the existing instruments are at odds with intervention goals, as in our Florida study, can we expect to see an intervention effect (Yarbrough, Shulha, Hopson, & Caruthers, 2011)? As states move away from PARCC (Partnership for Assessment of Readiness for College and Careers) and Smarter Balanced, alignment with Common Core's deeper learning expectations may be an issue (Yuan & Le, 2014). Furthermore, the possibility of common, available measures in cross-state studies is reduced.

These reflections are not meant to diminish the value of rigorous research but rather to point out the complexity of research use that goes beyond the simple dictums of "evidence-based practice" to help us understand whether, how, and why interventions work and provide evidence that can

inform policy decision making. We have established standards for conducting rigorous outcomes research (What Works Clearinghouse, 2014), but we think that one implication of our work, as others have noted (Goodson, 2015), is the need to move to more rigorous and useful implementation research as well.

Acknowledgments

We acknowledge and thank Dr. Sara Reber and Dr. Jia Wang for their valuable contributions the work reported here. We are also grateful to the teachers and school and district coordinators for their participation in our studies. Without their cooperation, our research would not be possible.

Funding

The work reported herein was supported by Grant 20114496 from the Bill and Melinda Gates Foundation through Research for Action Inc., with funding to the National Center for Research on Evaluation, Standards, and Student Testing. The findings and opinions expressed in this report are those of the authors and do not necessarily reflect the positions or policies of Research for Action Inc. or the Bill and Melinda Gates Foundation.

Notes

1. For space reasons, our coverage of implementation findings is brief and focuses on setting the context for outcome results. For more information on our extensive implementation analyses and results, see Herman et al. (2015a, 2015b).

2. The eighth-grade study also included six small districts in Pennsylvania, but unfortunately, data access did not allow for rigorous quasi-experimental analyses of outcomes. Implementation analyses largely showed similar results to those of Kentucky.

3. Generalizability, factor analysis, and decision studies were conducted within subjects across Kentucky and Pennsylvania teachers. See Reisman, Herman, Luskin, and Epstein (2013) for more information.

References

Ansen, R. (2015). *Democracy, deliberation, and education*. University Park, PA: Penn State Press.

Berman, P., & McLaughlin, M. W. (1978). *Federal programs supporting educational change: Vol. 8. Implementing and sustaining innovations*. Santa Monica, CA: RAND.

Borko, H., Koellner, K., & Jacobs, J. (2014). Examining novice teacher leaders' facilitation of mathematics professional development. *Journal of Mathematical Behavior*, *33*, 149–167.

Bryk, A. S., Sebring, P. B., Allenworth, E., Luppescu, S., & Easton, J. Q. (2010). *Organizing schools for success*. Chicago, IL: University of Chicago Press.

Civic Impulse. (2015). *S. 1177—114th Congress: Every Student Succeeds Act*. Retrieved from https://www.govtrack.us/congress/bills/114/s1177

Common Core State Standards Initiative. (2015). *Key shifts in English language arts*. Retrieved from http://www.corestandards.org/other-resources/key-shifts-in-english-language-arts/

Cuban, L. (1984). Transforming the frog into a prince: Effective schools research, policy and practice at the district level. *Harvard Educational Review*, *54*(2), 129–151.

Curtis, R. (2012). *Building it together: The design and implementation of Hillsborough County Public Schools' Teacher Evaluation System*. Washington DC: Aspen Institute.

Duesbery, L., Werblow, J., & Twyman, T. (2011). *The effect of the Seeds of Science/Roots of Reading curriculum (planets and moons unit) for developing literacy through science in fifth grade*. Berkeley, CA: University of California Berkeley, Graduate School of Education.

Elmore, R. (2008). *Improving the instructional core*. Retrieved from http://www.eastbaycharterconnect.org/uploads/7/1/7/6/7176220/improving_the_instructional_core_elmore_2008.pdf

Fixsen, D., Naoom, S., Blasé, K. A., Friedman, R., & Wallace, F. (2005). *Implementation research: A synthesis of the literature*. Retrieved from http://nirn.fpg.unc.edu/sites/nirn.fpg.unc.edu/files/resources/NIRN-MonographFull-01-2005.pdf

Florida Department of Education. (2012). *FCAT 2.0 reading test item specifications, Grades 6–8*. Retrieved from http://www.fldoe.org/core/fileparse.php/5682/urlt/0077907-fl10spisg68rwtr3gfinal.pdf

Goldring, E., Grissom, J. A., Neumerski, C. M., Murphy, J., Blissett, R., & Porter, A. (2015). *Making time for instructional leadership*. New York, NY: Wallace Foundation.

Goldschmidt, P. (2010). *Evaluation of Seeds of Science/Roots of Reading: Effective tools for developing literacy through science in the early grades*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Goodson, B. D. (2015). *Evidence at the crossroads: Part 5. Improving implementation research*. Retrieved from http://blog.wtgrantfoundation.org/post/134340131577/evidence-at-the-crossroads-pt-5-improving

Heiten, L. (2015). *South Carolina's new math standards depart little from Common Core*. Retrieved from http://blogs.edweek.org/edweek/curriculum/2015/06/south_carolinas_new_math_stand.html?_ga=1.238603846.732062923.1449770714

Herman, J. L., Epstein, S., Leon, S., Dai, Y., La Torre Matrundola, D., Reber, S., & Choi, K. (2015a). *The implementation and effects of the Literacy Design Collaborative (LDC): Early findings in eighth-grade history/social studies and science courses* (CRESST Report No. 848). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Herman, J. L., Epstein, S., Leon, S., Dai, Y., La Torre Matrundola, D., Reber, S., & Choi, K. (2015b). *The implementation and effects of the Literacy Design Collaborative (LDC): Early findings in sixth-grade Advanced Reading courses* (CRESST Report No. 846). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Hill, C., Bloom, H., Black, A. R., & Lipsey, M. W. (2007). *Empirical benchmarks for interpreting effect sizes in research*. New York, NY: MDRC.

Hirsch, B. J., Deutsch, N. L., & DuBois, D. L. (2011). *After-school centers and youth development: Case studies of success and failure*. Cambridge, UK: Cambridge University Press.

Holliday, T., & Smith, F. (2012, September–October). Leading Common Core implementation. *Principal*, pp. 12–15.

Hudson, Q. (2015). *The effectiveness of professional learning communities as perceived by elementary school teachers* (Unpublished doctoral dissertation). Walden University, Minneapolis, MN.

Koellner, K., & Jacobs, J. (2015). Distinguishing models of professional development: The case of an adaptive model's impact on teachers' knowledge, instruction, and student achievement. *Journal of Teacher Education*, *66*(1), 51–67.

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

Mitchell, K., Shkolnik, J., Song, M., Uekawa, K., Murphy, R., Garet, M., & Means, B. (2005). *Rigor relevance and results: The quality of teacher assignments and student work in new and conventional high schools*. Retrieved from http://smallhs.sri.com/documents/Rigor_Rpt_10_21_2005.pdf

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects*. Retrieved from www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf

Newmann, F., Bryk, A. S., & Nagaoka, J. K. (2001). *Authentic intellectual work and standardized tests: Conflict or coexistence?* Chicago, IL: Consortium on Chicago School Research.

Pearson. (2013a). *Kentucky Performance Rating for Educational Progress, 2012–13 technical manual*. Retrieved from http://education.ky.gov/AA/KTS/Documents/2012-13%20K-PREP%20Technical%20Manual%20v1%203.pdf

Pearson. (2013b). *Kentucky Performance Rating for Educational Progress, 2012–13 yearbook: Version 1.0*. Retrieved from http://education.ky.gov/aa/kts/documents/2012-13%20k-prep%20yearbook%20v1.pdf

Pompea, S. M., & Gek, T. K. (2002). Optics in the Great Exploration in Math and Science (GEMS) program: a summary of effective pedagogical approaches. In *Education and training in optics and photonics 2001* (pp. 103–109). Bellingham, WA: International Society for Optics and Photonics.

Reisman, A. (2012). Reading like a historian: A document-based history curriculum intervention in urban high schools. *Cognition and Instruction*, *30*(1), 86–112.

Reisman, A., Herman, J., Luskin, R., & Epstein, S. (2013). *Summary report: Developing an assignment measure to assess quality of LDC modules*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Rowan, B., & Correnti, R. (2009). Studying reading instruction with teacher logs: Lessons from a study of instructional improvement. *Educational Researcher*, *38*(2), 120–131.

Supovitz, J. A., & Spillane, J. (2015). *Challenging standards: Navigating conflict and building capacity in the era of the Common Core*. New York, NY: Rowman & Littlefield.

Trauth-Nare, A. (2016). Re-envisioning scientific literacy as relational, participatory thinking and doing. *Cultural Studies of Science Education*, *11*, 327–334.

Tyler, R. W. (1949). *Basic principles of curriculum and instruction*. Chicago, IL: The University of Chicago Press.

Wang, J., & Herman, J. (2005). *Evaluation of Seeds of Science/Roots of Reading project: Shoreline Science and Terrarium Investigations*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

What Works Clearinghouse. (2014). *Procedures and standards handbook: Version 3.0*. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_standards_handbook.pdf

Wiggins, G., & McTighe, J. (1998). *Understanding by design*. Upper Saddle River, NJ: Merrill Prentice Hall.

Wineburg, S., Martin, D., & Monte-Sano, C. (2013). *Reading like a historian: Teaching literacy in middle & high school classrooms*. New York, NY: Teachers College Press.

Winters, L., & Herman, J. L. (2011). *The turnaround toolkit: Managing rapid, sustainable school improvement*. Thousand Oaks, CA: Corwin Press.

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

Yuan, K., & Le, V. (2014). *Measuring deeper learning through cognitively demanding test items: Results from the analysis of six national and international exams*. Santa Monica, CA: RAND.

Authors

JOAN HERMAN, EdD, is Director Emeritus and currently senior research scientist at the National Center for Research on Evaluation, Standards and Student Testing (CRESST), Graduate School of Education and Information Sciences, University of California, Los Angeles, 300 Charles E. Young Dr. North, Los Angeles, CA 90095-1522; herman@cresst.org. Her primary research interests focus on the design and use of assessment to support deeper learning and the evaluation of innovative programs.

SCOTT EPSTEIN, MPP, is a research associate at the National Center for Research on Evaluation, Standards and Student Testing (CRESST), University of California, Los Angeles, 300 Charles E. Young Dr. North, Los Angeles, CA 90095-1522; epstein@cresst.org. His research interests include the evaluation of educational interventions, and educational accountability and reform.

SETH LEON is a senior statistician at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) at University of California, Los Angeles, 300 Charles E. Young Dr. North, Los Angeles, CA 90095-1522; leon@cresst.org. His primary research interests include matching based methodologies used to draw causal inferences in the evaluation of student achievement.