ASSESSING THE VALUE-ADDED EFFECTS OF LITERACY COLLABORATIVE PROFESSIONAL DEVELOPMENT ON STUDENT LEARNING

ABSTRACT

This article reports on a 4-year longitudinal study of the effects of Literacy Collaborative (LC), a school-wide reform model that relies primarily on the one-on-one coaching of teachers as a lever for improving student literacy learning. Kindergarten through second-grade students in 17 schools were assessed twice annually with DIBELS and Terra Nova. Scores from the study's first year, before coaching began, offered a baseline for assessing the value added to student learning over the following 3 years. A hierarchical, crossed-level, value-added-effects model compared student literacy learning over 3 years of LC program implementation against observed growth under baseline conditions. Results demonstrated increasing improvements in student literacy learning during LC implementation (standard effect sizes of .22, .37, and .43 in years 1, 2, and 3, respectively), and the benefits persisted through subsequent summers. Findings warrant a claim of substantial effects on student learning for the LC coaching model.

Gina Biancarosa
UNIVERSITY OF OREGON

Anthony S. Bryk
THE CARNEGIE FOUNDATION FOR THE ADVANCEMENT OF TEACHING

Emily R. Dexter
LESLEY UNIVERSITY

THE ELEMENTARY SCHOOL JOURNAL VOLUME 111, NUMBER 1
 0013-5984/2010/11101-0002 $10.00

THE use of school-based literacy coaches as a professional development (PD) strategy has become widespread in U.S. schools. Many school districts have made large investments in initiatives to train and support literacy coaches. While an extensive literature advocates for this approach, few empirical studies of coaching and its effects on teaching practice and student achievement exist. The current article examines the value-added effects of literacy coaching on kindergarten through second-grade students' literacy learning based on a 4-year longitudinal field trial of the effectiveness of PD in the Literacy Collaborative (LC) program, which relies heavily on one-to-one literacy coaching as a means of improving student literacy learning.

Background

There is a growing body of educational literature on coaching and on literacy coaching in particular. Most of the publications are descriptive and prescriptive in orientation. Allen (2006), Bean and Carroll (2006), Blachowicz, Obrochta, and Fogelberg (2005), Casey (2006), Toll (2007), and Walpole and McKenna (2004), for example, offer descriptions of the literacy coach's role as well as recommendations on how best to fulfill this role. The International Reading Association (2004, 2006) describes qualifications and ability standards for literacy coaches. However, as Neufeld and Roper (2003) note, "No one, as yet, has proven that coaching contributes significantly to increased student achievement. Indeed, there are scant studies of this form of PD and how it influences teachers' practice and students' learning" (p. 1).

Literacy Coaching's Effects on Student Learning

Most of the empirical literature on literacy coaching is in the form of program evaluations and is grounded in qualitative methods (Gibson, 2006; Neufeld & Roper, 2003). Neufeld and Roper (2003), for example, used qualitative data to describe the actual work of coaches in four urban districts. Poglinco et al. (2003) conducted a descriptive study of coaching in 27 schools that implemented a comprehensive school reform model. Although they used a four-point rubric to evaluate teachers' practices against program standards, they did not assess the effects of coaching on these practices and, most critically, they did not examine the effects of either coaching or teacher practices on student learning. In another 3-year study of Institute for Learning reform efforts, two of three districts placed full-time English language arts (ELA) coaches in all of their schools (Marsh et al., 2005). Based on changes in the schools' ELA proficiency percentages over multiple years, the study's authors concluded that one district showed "substantial" improvement in district scores, while the other showed "limited" improvement (Marsh et al., 2005). However, although teachers reported benefits from coaching, both student-level outcome data and causal analyses of coaching effects on these outcomes were not examined. Several statewide evaluations of literacy coaching programs have been conducted, including evaluations in Alabama (Norton,

2007), Alaska (Barton & Lavrakas, 2006), and Idaho (Reed & Rettig, 2006), but again the impact of coaching on student outcomes was not rigorously examined.

To date, only two studies have offered empirical evidence of the effects of coaching on student literacy growth: one study focused on literacy coaching in second grade (Garet et al., 2008) and the other focused on middle school (Marsh et al., 2008). These studies showed minimal (Marsh et al., 2008) or null (Garet et al., 2008) effects on students' learning. However, in both of these studies, the coaches were trained for a week or less before they began their coaching work in the schools. Moreover, the coaching models in the two studies were not well established; one was created for the experimental study (Garet et al., 2008) and the other was the product of a rapid, statewide scale-up of coaching over the course of a few years (Marsh et al., 2008). The coaching model studied here offers far more intensive PD for coaches prior to their coaching of teachers and is embedded in a well-established and comprehensive literacy framework.

Literacy Collaborative

Established in 1993, LC is a comprehensive school reform program designed to improve elementary children's reading, writing, and language skills primarily through school-based coaching. The program builds on 30 years of research and development grounded in the reading theories of Marie Clay (1979, 1991, 2004) and elaborated by Fountas and Pinnell (1996, 2006). LC is committed to the idea that teachers need both training in particular procedures and opportunities to analyze their teaching with a "more expert other" (i.e., the coach; Norlander-Case, 1999). Grounded in Bruner's theory of instruction as scaffolding (Bruner, 1986, 1996) and Vygotsky's theory of the zone of proximal development (Gallimore & Tharp, 1990; Vygotsky, 1978), such PD aims to support over time the development of the deep understandings that teachers need to continuously improve their practice.

LC trains and supports school-based literacy coaches, who are teachers selected by their schools to lead local instructional improvement efforts. Over the course of one year, the coaches attend an intensive, graduate-level training program while also teaching children. The rigorous training includes coverage of the theory and content of literacy learning, how to teach children within LC's instructional framework, and how to develop these understandings in other teachers through site-based PD and coaching. Regarding the final goal, LC coaches learn how to lead a PD course to introduce theories and instructional practices to teachers and how to use one-on-one coaching as a mechanism to support individual professional growth and development. After their training year, coaches reduce their teaching time[1] and spend approximately half of their time providing PD and coaching to their school colleagues. As part of their coaching role, they also participate with administrators and teacher leaders in a schoolwide leadership team that monitors student achievement and supports implementation of LC professional development and instructional practices.

Teachers' entry into the LC program begins with participation in a 40-hour course led by the coach, who introduces the basic elements of the comprehensive literacy instructional framework used by LC (Fountas & Pinnell, 1996, 2006; Lit-

eracy Collaborative, 2009). Ongoing courses after the initial year offer 10–12 hours of PD annually. Research has shown that while PD sessions of this sort can provide a common knowledge base and shared perspective among teachers, these meetings alone afford little guidance on what to do about particular problems of practice emerging in an individual teacher's classroom (Kohler, Crillery, Shearer, & Good, 1997; Lieberman, 1995; Schön, 1983). To address this need, LC relies on coaches working one-on-one with teachers in their classrooms: observing, modeling, and catalyzing teachers' development toward more expert practice. As such, the LC model relies on one-to-one coaching sessions as the coaches' high-leverage activity by which they are able to most effectively help teachers develop their instructional practices.

LC coaching is centered on a comprehensive approach to literacy instruction that focuses on engaging students at all ability levels in the reading and writing processes. Although the program was developed for students through grade 8, the K–2 program was the primary focus of this study. The LC program targets all components of reading, writing, and language development, including, but not limited to, direct and embedded instruction in phonics and phonological awareness, vocabulary and word structure, fluent reading, and literal, inferential, and critical thinking about texts.

Six core components form the LC comprehensive literacy framework for kindergarten through second grade: interactive read-aloud, shared reading, guided reading, interactive writing, writing workshop, and word study. The components vary in their use of student grouping and the level of scaffolding provided, as well as in their focus on reading, writing, or word-level skills and knowledge. For example, during interactive writing the teacher and children (either as a whole class or in small groups) collaboratively compose a text by writing it out word-by-word (generally on a large chart). At several carefully selected points, the teacher invites individual children to come up to the chart and make contributions by adding letters or words that have high instructional value in helping children learn about the construction of words (phonics) and the writing process (McCarrier, Pinnell, & Fountas, 2000). In contrast, during the writing workshop, the teacher provides a mini-lesson on some aspect of writing based on student needs; the students then write independently as individuals confer with the teacher before the group reconvenes to briefly share their progress. The teacher uses the mini-lesson and sharing period to reinforce a principle of writing and uses the conferences to offer individualized support and instruction to students. Together, the six components of K–2 reading and writing instruction constitute a repertoire of practices that teachers orchestrate based on their pedagogical knowledge and in response to their observation of children so that children are supported in acquiring new principles across multiple components and contexts (Scharer, Pinnell, & Bryk, 2008).[2]

Research Design

The current article reports findings from a longitudinal study of the LC program effects at the school and teacher levels. The study was designed to specifically

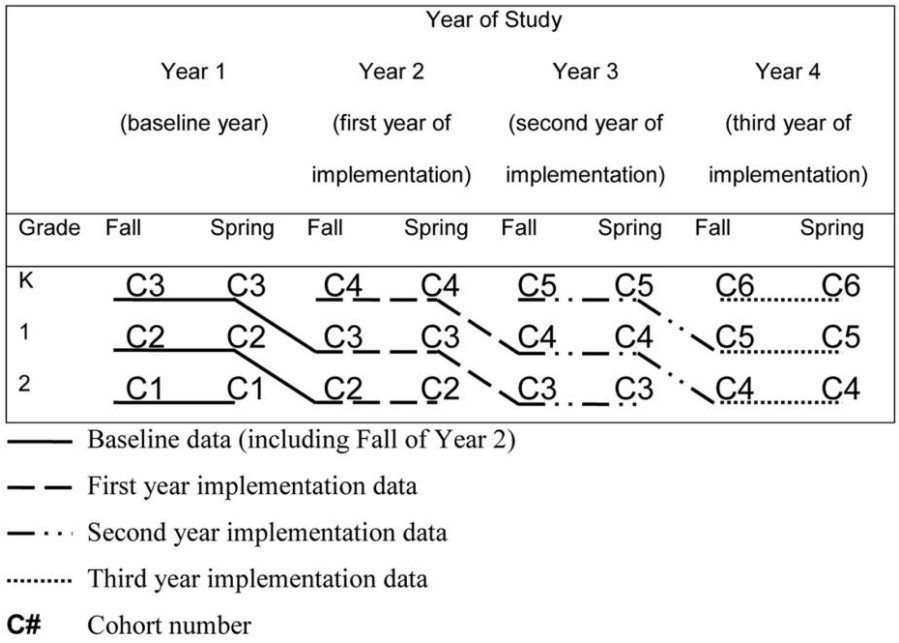

Figure 1. Accelerated multiple cohort design: six cohorts in four grades across 4 years with Literacy Collaborative implementation in grades K–2.

isolate the effects of the introduction of LC professional development in a diverse set of schools. During the first year of the study, literacy coaches trained for their new role, and therefore did not conduct literacy PD activities at their respective schools. Thus, the first year of this 4-year study represents a no-treatment period and affords baseline data on student achievement for each school and classroom prior to program initiation. Implementation of LC professional development in kindergarten through second-grade classrooms began during the study's second year. The analyses reported in this article focus on the effects on student literacy learning in these grades over the following 3 years of implementation compared to the baseline year.

An Accelerated Longitudinal Cohort Design

The current study utilized an accelerated multicohort, longitudinal, quasi-experimental design. We collected fall and spring student achievement data from multiple student cohorts at three grade levels (K–2) over the course of 4 years. The study involved children from six different cohorts who entered at different grades and in different years. Figure 1 depicts these cohorts and the timing of LC implementation for each. For example, Cohort 3 entered the study as kindergartners during the baseline year of the study, and attended first and second grade in the first and second years of implementation, respectively.

Student achievement data from the first three waves of data collection (fall and spring of the first year and fall of the second year) offer baseline information because they occurred prior to LC implementation. This is denoted with solid

black lines in Figure 1. Student achievement data at subsequent time points, during which LC was implemented, are compared to these baseline data. The first year of LC effects are denoted by dashed lines in Figure 1, the second year by alternating dashed-and-dotted lines, and the third year by dotted lines.

Logic of Value-Added Modeling

In general, the data collected on student learning during the baseline, or no-treatment, period in an accelerated multicohort, longitudinal, quasi-experimental design allow us to estimate the value-added effects of a subsequent intervention on student learning. The current application of value-added modeling is rooted in the idea that each child has an individual latent growth trajectory. This trajectory describes the expected achievement growth in grades K–2 for each child if exposed to the average instructional conditions prevalent in the school during the baseline period. We then compare the observed student growth trajectories in the 3 years of LC implementation to these expected (or latent) growth trajectories under baseline conditions. The value-added effects represent the difference between these observed and expected outcomes. In principle, each teacher and school may have a unique value-added effect during each time period. Our analyses focus on the value-added effects in the study's second, third, and fourth years because these years include potential effects associated with LC coaching above and beyond any teacher or school effects present in the baseline year.

Participants

The final study sample included 4 years of data amounting to 27,427 observations of 8,576 students in 17 schools located in eight states across the eastern United States.[3] During the course of the study, students attended 287 teachers' classrooms.

Students. During each year of the study, approximately 1,150 students were assessed in each grade level from kindergarten through second grade. This represents a student participation rate of 90% or higher at each testing occasion.

Overall, approximately 61% of the student sample has complete data. These students have test scores at every occasion for which their cohort was eligible to be assessed (see Table 1). Of the students with incomplete data, most either entered a study school after data collection began for their cohort or transferred out of a study school prior to second grade. Only 3% of children missed testing on one or more occasions for which they were eligible to be assessed (e.g., due to absences).

The total K–2 sample includes 8,829 children and 28,935 observations. A small number of students were excluded who had repeated a grade or were tested mistakenly with materials other than those appropriate for their grade level. This adjustment resulted in the loss of 230 children and reduced the analytic sample to 8,599 children and 27,839 observations. In addition, we excluded a small number of individual test scores with unusually large standard errors of measurement that raised questions about the reliability of these individual test administrations. This adjustment reduced the analytic sample to the final total of 8,576 children and

Table 1. Number of Observations per Child by Cohort for Analytic Sample (n = 8,576)

Cohort	Observations							Percent with Incomplete Data
	1	2	3	4	5	6	Total	
1 (max = 2)	183	1,027	0	0	0	0	1,210	15.1
2 (max = 4)	150	257	193	829	0	0	1,429	42.0
3 (max = 6)	200	299	115	212	138	681	1,645	58.6
4 (max = 6)	180	309	96	234	128	720	1,667	56.8
5 (max = 4)	157	275	126	891	0	0	1,449	38.5
6 (max = 2)	119	1,057	0	0	0	0	1,176	10.1
Total	989	3,224	530	2,166	266	1,401	8,576	39.3

27,427 observations. The sample counts presented in Table 1 are based on this final analytic sample.

Teachers. Not including the literacy coaches, 287 teachers taught in kindergarten through second-grade classrooms in the 17 study schools at some point during the study's 4 years. Of these 287 teachers, 111 teachers were present for all 4 years (38.7% of the teacher sample), 39 teachers (13.6%) were present for 3 years, 40 teachers (13.9%) were present for 2 years, and 97 teachers (33.8%) were present for only 1 year. Within this group, 259 kindergarten through second-grade teachers participated in one-on-one coaching at least once during the 3 years of LC implementation.

Schools. As Table 2 demonstrates, schools varied widely in their student composition. More than 90% of students were white in several schools, while in other schools, 30% or more of the students were African American or Latino. Similarly, the schools ranged in their socioeconomic composition, with the percentage of students receiving free or reduced-price lunches ranging from a low of 19% in one school to a high of 86% in another.

Measures

We used a mix of reading assessments in order to broadly assess students' literacy learning within the primary grades studied. Depending on a child's cohort

Table 2. Percentages of Key Demographics of the Student Sample in the Base Year

Characteristic	Overall	School																
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Low income	46.0	47	33	20	35	36	19	49	38	60	50	60	73	43	45	50	86	38
Race/ethnicity:																		
African American	15.5	3	31	30	1	7	20	1	1	41	3	1	50	0	35	0	36	4
Latino	5.8	31	8	12	2	20	4	0	3	2	1	0	6	0	4	0	4	2
Other	7.2	5	10	11	2	1	15	0	1	28	2	0	29	12	0	3	0	3
White	70.6	61	51	47	95	72	54	99	95	21	94	99	15	88	61	97	60	91

Note.—Racial ethnic group percentages may not total 100% due to rounding.

and the length of enrollment at a study school, this resulted in children being tested a maximum of six times (see Table 1).

Dynamic Indicators of Basic Early Literacy Skills (DIBELS). Participating students took a variety of subtests from DIBELS beginning in the fall of kindergarten through the spring of second grade. These subtests tap a range of early literacy skills, including letter recognition, phonological awareness, decoding, and oral reading fluency. The choice of subtests to administer at each grade level and semester (fall and spring) was based primarily on publisher recommendations (Good & Kaminski, 2002). However, in some instances we chose to include an additional, more difficult subtest. For example, we added the oral reading fluency subtest in the fall of first grade, in order to improve our assessments' capacity to discriminate effectively among students with higher levels of literacy learning. Table 3 provides a schedule of the specific subtests administered each semester in each grade and the reported reliability and validity of these subtests (Good, Wallin, Simmons, Kame'enui, & Kaminski, 2002). Concurrent validity statistics for our study sample are consistent with published validity information (see Table 3).

Terra Nova. Each spring, participating first- and second-grade students took the reading comprehension subtest from the Terra Nova Multiple Assessments of Reading, a group-administered, standardized, norm-referenced reading test. (See McGraw-Hill [2001] for information on the reliability and validity of this test.)

Rasch scaling. The DIBELS and Terra Nova results were scaled together using Rasch modeling (Wright & Master, 1982). The resultant vertical scale allowed us to fully exploit the longitudinal character of our student literacy learning data and resolve several difficulties with the use of DIBELS assessments in program effects studies. The final vertical scale also more closely approximates the principle of a single ruler or metric where a one-unit difference on the scale at any level of ability implies an equal difference on the trait measured (reading in this case), which is an assumption for parametric growth curve analyses (Raudenbush & Bryk, 2002). The full details of the Rasch analysis are reported elsewhere (Luppescu, Biancarosa, Kerbow, & Bryk, 2008).[4] Fewer than 2% of all items in the resulting scale exhibited signs of misfit, and the average infit (information-weighted mean square fit) was 1.00, which is the expected fit in a good scale. Moreover, student scale measures correlated well with raw scores on all constituent subtests of the DIBELS and Terra Nova, ranging from a low of .52 with DIBELS initial sound fluency to highs of .77 and .85 for the Terra Nova and DIBELS oral reading fluency subtest, respectively.

As is customary in Rasch scaling, the final measures are reported in a logit metric. Since logits are not intrinsically meaningful, we illustrate here the differences in literacy status one would likely find among students scoring at different values on our scale. For example, a child scoring at 1.0 logit (approximately an average child in the fall of kindergarten) typically can name about 30 letters in a minute, thus indicating good letter-name knowledge. That same child most likely discerns a few initial phonemes, but not many, and has very little chance of being able to segment words into phonemes. In contrast, a child scoring at 2.0 logits (approximately an average child in the spring of kindergarten or fall of first grade) is both accurate and fluent in letter-name knowledge and has almost mastered initial sound identification, but is still largely unable to segment words phonemi-

Table 3. DIBELS and Terra Nova Testing Schedule by Grade and Time of Year, Alternate Form Reliability, and Concurrent and Predictive Validity

	Kindergarten		Grade 1		Grade 2			Validity	
Measures	Fall	Spring	Fall	Spring	Fall	Spring	Reliability	Concurrent	Predictive
DIBELS:									
Initial sound fluency (ISF)	X	X					.72[a]	.48 PSF[a] .36 WJ readiness[a] .60 CTOPP phonological awareness[b] .41–.45 LNF[e] .41–.46 PSF[e] .39 NWF[e]	.45 TORF[a] .36 WJ total reading[a]
Letter name fluency (LNF)	X	X	X				.88[a]	.70 WJ readiness[a] .53 CTOPP phonological awareness[b] .58 CTOPP rapid naming[b] .41–.45 ISF[e] .43–.57 PSF[e] .58–.64 NWF[e] .52 ORF[e]	.65–.71 TORF[a]
Phonemic segmentation fluency (PSF)	X	X	X	X			.79–.88[a]	.54 WJ readiness[a] .53 CTOPP phonological awareness[b] .41–.46 ISF[e] .43–.57 LNF[e] .41–.46 NWF[e] .17–.24 ORF[e] .20 TN[e]	.62 NWF[a] .62 TORF[a] .68 WJ total reading[a]
Nonsense word fluency (NWF)		X	X	X			.83[a]	.36–.59 WJ readiness[a] .39 ISF[e] .58–.64 LNF[e] .41–.46 PSF[e] .70–.73 ORF[e] .45 TN[e]	.62–.82 TORF[a] .66 WJ total reading[a]
Oral reading fluency (ORF)			X	X	X	X	.94[a]	.92–.96 TORF[a] .74–.75 ITBS[c] .52 LNF[e] .17–.24 PSF[e] .70–.73 NWF[e] .68 TN 3	.68–.75[c]
Terra Nova (TN)				X		X	.67–.84[d]	.20 PSF[e] .45 NWF[e] .68 ORF[e]	.67–.82 PSSA[d]

Note.—CTOPP = Comprehensive Test of Phonological Processing; ITBS = Iowa Test of Basic Skills reading comprehension; PSSA = Pennsylvania State System of Assessment; TORF = Test of Oral Reading Fluency; WJ = Woodcock Johnson Psycho-Educational Battery.

[a] Good, Wallin, Simmons, Kame'enui, and Kaminski (2002).

[b] Hintze, Ryan, and Stoner (2003).

[c] Schilling, Carlisle, Scott, and Zeng (2007).

[d] Brown and Coughlin (2007).

[e] Current sample.

Table 4. Mean Scores and Standard Deviations in Logits for Analytic Sample of Kindergarten (K) through Second-Grade Students by Grade, Semester, and Year (n = 8,576)

	Year of Study							
	1 (Baseline Year)		2 (First Year of Implementation)		3 (Second Year of Implementation)		4 (Third Year of Implementation)	
Grade	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring
K	.87	1.76	.92	2.18	1.02	2.29	.69	2.25
	(1.34)	(1.17)	(1.38)	(1.21)	(1.44)	(1.52)	(1.36)	(1.66)
	(n = 1,050)	(n = 1,076)	(n = 1,145)	(n = 1,150)	(n = 1,175)	(n = 1,164)	(n = 1,130)	(n = 1,103)
1	1.68	2.92	1.65	3.01	1.92	3.16	1.96	3.22
	(1.12)	(.87)	(1.20)	(.89)	(1.33)	(1.02)	(1.28)	(1.05)
	(n = 1,127)	(n = 1,155)	(n = 1,173)	(n = 1,182)	(n = 1,201)	(n = 1,212)	(n = 1,185)	(n = 1,125)
2	3.45	4.23	3.12	4.21	3.34	4.41	3.25	4.29
	(1.96)	(1.24)	(2.00)	(1.20)	(2.01)	(1.18)	(2.06)	(1.25)
	(n = 1,074)	(n = 1,163)	(n = 1,096)	(n = 1,181)	(n = 1,120)	(n = 1,165)	(n = 1,155)	(n = 1,120)

cally. The child can read a handful of words in a minute when given a passage of continuous text, but has little success at reading nonsense words (an indicator of decoding skill out of context). A child scoring at 3.0 logits (approximately an average child in the spring of first grade or fall of second grade) has mastered letter names and initial sounds, can read 50–60 words per minute accurately, and may answer correctly about a third of the first-grade Terra Nova comprehension questions. This child also does well on all but the hardest phonemic segmentation and nonsense word-reading tasks, but may not be very fast at these tasks overall and is generally better at the former than the latter. Finally, a child scoring at about 4.0 logits (approximately an average child in the spring of second grade) has mastered component reading skills (e.g., letter name knowledge, phonemic segmentation, decoding), reads about 90 words correctly per minute, and does well on two-thirds of the first-grade Terra Nova comprehension questions and on about a third of the second-grade questions.

Analyses

We began our analyses by visually examining the observed mean outcomes separately for each cohort. Findings from this examination guided our approach to analyzing these data through hierarchical, crossed-level, value-added-effects modeling.

Empirical Student Literacy Learning Trajectories

We describe in this section the basic growth patterns found in the observed data. Table 4 reports the mean Rasch literacy development scores for K–2 students in the final analytic sample by grade, semester (i.e., fall or spring), and study year; Figure 2 depicts this same information and identifies the longitudinal data for

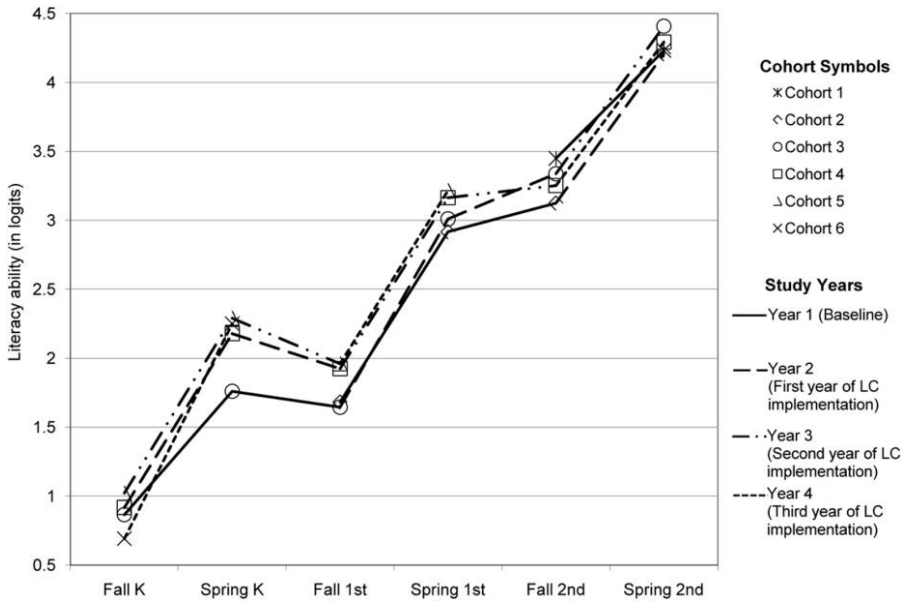

Figure 2. Means by cohort and year of Literacy Collaborative implementation.

each separate cohort by a distinct symbol. We first discuss the baseline trends that informed the model-building process before turning to implementation trends.

Baseline trend. Data collected in the fall and spring of the first year and fall of the second year (i.e., prior to the initiation of LC school-based PD) constitute the baseline trend for assessing subsequent program effects. Data from three different student cohorts constitute this baseline (see Fig. 2): Cohort 3 (represented by open circles) began the study in the fall of kindergarten, Cohort 2 (represented by open diamonds) began in the fall of first grade, and Cohort 2 (represented by asterisks) began in the fall of second grade.

Under an accelerated longitudinal cohort design, the results from the different baseline cohorts should connect smoothly as one overall growth trajectory. The resultant longitudinal trajectory is the baseline against which subsequent program effects are evaluated. Note that, as expected under an accelerated longitudinal cohort design, we found a near perfect overlap in mean achievement at the fall of first grade where Cohorts 2 and 3 join. However, a small gap of about .25 logits was found where Cohorts 1 and 2 join in the fall of second grade. This indicates that prior to program implementation, these two cohorts differed somewhat in their average literacy ability, at least at this one point in time. As a result, we have introduced a set of statistical adjustments for possible cohort differences in the hierarchical, crossed-random-effects models estimated below.

Implementation years. Subsequent to examining the baseline trends, we plotted the subsequent LC implementation years' data on top of the baseline trend to provide a first look at possible program effects. As noted above, Figure 2 illustrates the mean student outcomes at each testing occasion during the 3 years of LC implementation.

Again, the longitudinal data for each separate cohort are identified by a distinct symbol. For example, the trajectory for Cohort 3 is identified by open circles and

includes data from the baseline year and 2 years of implementation. The trajectory begins as a solid line in the fall of kindergarten and continues to the fall of first grade when implementation began. The trajectory continues, therefore, as a dashed line through the fall of second grade, which represents first-year LC implementation effects. Following the same cohort through the spring of second grade incorporates second-year LC implementation effects on this group, which is represented by a dashed-and-dotted line.

Of primary interest here is a comparison of the slopes representing student learning during the academic years and the change in these slopes over the course of the study. Specifically, the increasing steepness of the slopes from fall to spring within each grade (from solid line, to dashed line, to dashed-and-dotted line) suggests positive overall value-added effects associated with the LC program. These value-added effects are most apparent in kindergarten when students' fall entry status is almost identical for Cohorts 3, 4, 5, and 6, but there is increasing separation in achievement among these three groups by the following spring.

Key observations for value-added modeling. In addition to the possible cohort effects in the baseline results noted earlier, several other distinct features in these longitudinal data have important implications for subsequent value-added modeling. First, growth during academic years (from fall to spring) is markedly steeper than growth during the summer break periods (from spring to fall). This means that we must separately parameterize the rates of student learning in these two periods. Second, as noted earlier, the academic learning rates (slopes) appear to vary across grade levels, with larger gains observed in kindergarten and first grade than second grade. Thus, we also need to introduce a set of fixed effects in the model to capture these departures from strict linearity.

Finally, there is some evidence in Figure 2 that program effects may vary by year of implementation. Thus, in the analyses assessing teacher- and school-level value-added effects to student literacy learning that follow, we estimate separate effects for each year.

Hierarchical, Crossed-Level, Value-Added-Effects Modeling

The accelerated longitudinal cohort design used in the current study lends itself naturally to value-added modeling because our data consist of repeated measures on students who cross teachers within school over time. The hierarchical, crossed-level, value-added-effects model that we applied can be conceptualized as the joining of two separate hierarchical models, the first of which is a two-level model for individual growth in achievement over time, and the second of which is a two-level model of the value added that each teacher and school contributes to student learning in each particular year. In essence, the core evidence for LC effects consists of comparing learning gains in each teacher's classroom during each year of program implementation to the gains in that same teacher's classroom during the baseline year. The observed gains in each classroom, however, are now adjusted for any differences over time in the latent growth trends for students being educated in each classroom. In contrast to the simple descriptive statistics presented in Figure 2, a hierarchical, crossed-level, value-added-effects model allows us to take full advantage of the longitudinal character of the data on

each student and adjusts for any outcome differences associated with individual latent growth trends in estimated teacher-classroom and school value-added effects.

Basic individual growth model. We began building our model by specifying a level 1 model for the literacy score at time *i* for student *j*. We specified five level 1 predictors to capture key characteristics noted in the empirical growth trajectories in Figure 2. Specifically, we created three indicators for academic year learning: a base learning rate during kindergarten and two grade-specific deviation terms for first and second grade, respectively. Since summer learning rates appeared to vary between kindergarten and first grade versus between first and second grade, we created two additional indicators of growth: a K–grade 1 summer indicator and a grade 1–grade 2 summer indicator. A system was developed for coding the indicators such that the intercept represented the latent score for student *j* at entry into the data set, regardless of the specific time occasion when this may occur.[5]

Since we assume that each child has a unique latent growth trajectory, the intercept and base academic year learning rate were specified as randomly varying among individual students. These parameters capture the variation between children in their initial literacy status and their latent growth rate in literacy learning. In preliminary analyses we also considered modeling summer period effects as randomly varying. However, we were unable to reliably differentiate among children in this regard once random intercepts and random academic year learning effects were included in the model. Therefore, the summer period effects were treated as fixed at the individual child level.

Adjusting for possible time of entry effects. As noted earlier, a key feature of an accelerated multiple cohort design is that students entered the data set at different points in time by virtue of their cohort (i.e., Cohort 1 entered in second grade, Cohort 2 began in first grade, and Cohorts 3–6 entered in kindergarten). In addition, some students transferred into a school after the start of the study and were absorbed into their respective cohorts.

To account for the fact that many children did not enter the dataset in the fall of kindergarten, we added another set of dichotomous indicator variables (termed "Entry") as fixed effects in the child-level model. Each indicator variable corresponds to a specific grade and time of year and represents the first occasion at which a student appears in the dataset.

For example, if a student joined the study in the fall of first grade, that student's Entry would be a three, and the Entry_3 indicator variable would be scored as one and all other Entry variables would be zero. The reference category for this set of indicators is the fall of kindergarten, or Entry_1. As a result, the intercept in the model represents the predicted initial literacy status in the fall of kindergarten and the Entry variables for Entry locations two through six represent the mean differences in literacy status among students, depending on when they first entered the dataset.[6]

Adjusting for cohort differences. Because the earlier descriptive analysis indicated small average differences in initial literacy status among some baseline cohorts, we also included a set of fixed effects for each cohort to absorb these and any other potential residual differences between cohorts.[7] These dichotomous indicator variables were also included as fixed effects at the child level. We note that cohort and entry effects are not redundant because students could enter a cohort after data collection was initiated for their specific cohort.[8]

The value-added model. The value-added model estimates both the average value added by LC in each year of implementation and random value-added effects associated with each teacher and with each school year of implementation. We describe these effects below.

Average LC value-added effects. Separate fixed effects were estimated to assess the average value added of LC during each of the 3 years of program implementation. These estimates represent the increments to learning in comparison to the growth trends observed in the baseline period. These are captured through three dichotomous indicator variables (LC_Year_1, LC_Year_2, and LC_Year_3). The coefficients associated with each of these variables estimate average LC effects after the first, second, and third years of implementation, respectively. The LC_Year_1 indicator was set to one for observations collected in the spring of the second year of the study, LC_Year_2 was equal to one in the spring of the third year of the study, and LC_Year_3 was equal to one in the spring of the fourth year of the study; otherwise, all three were set to zero.

We also estimated the average effects after the summers of the first and second years of implementation, which are termed LC_Summer_1 and LC_Summer_2, respectively. LC_Summer_1 was equal to one at the fall of the third year of the study and LC_Summer_2 was equal to one at the fall of the fourth year of the study; otherwise, both were set to zero. Note that these LC summer effects allow us to estimate the extent to which the value added observed at the end of the previous year (i.e., the spring testing) was maintained through the summer (i.e., the subsequent fall testing). If, in fact, the LC academic effects are sustained, then the estimate for the summer LC effects should be similar in magnitude to the corresponding academic year LC effects.

All five of the LC implementation indicators are set at zero at the first occasion when the student appears in the data set. In this way, we preserve the meaning of the intercept.[9]

Random value-added effects. To estimate value-added effects for schools and teachers, students were linked to their school's identifier at all time points and to their teacher's identifier each spring. Each fall, students were linked to a school- and grade-specific identifier. This linking of students with their teachers in the spring allows us to estimate the value-added effect of each teacher on students' learning from fall to spring each year.

Based on these links, we estimated several random value-added effects for schools and teachers that captured their contributions to student learning during the baseline year and each of the 3 years of the LC program. We review first the random effects in baseline growth, followed by the 3 years of implementation.

We incorporated five random effects to capture possible teacher and school value-added effects to students' learning during the baseline period. First, to capture differences among schools in the student achievement levels at entry, the intercept was allowed to vary randomly among schools. This specification controls for possible selection effects associated with prior achievement. Second, to capture variation among schools in academic year learning rates during the baseline period, the base academic year growth indicator was allowed to vary randomly at the school level. Third, we added a random effect, "Base_Tch_VA," to represent the variation among teachers within schools in their value added to

student learning in the spring of the baseline year. The dichotomous indicator variable, Base_Tch_VA, takes on a value of one only at this time point and only for those children who were also present in the data set in the fall of the baseline year. The fixed effect for Base_Tch_VA was set to zero so as not to be redundant with the fixed effect for the base academic year growth indicators described earlier. Finally, we allowed the two summer indicators to vary randomly at the school level. These effects captured the extent to which schools varied in learning (or loss) during the summers. Since there is no assigned teacher during the summer periods, a random teacher effect for these occasions was not deemed sensible.

Finally, we also allowed LC value-added effects in all 3 years of implementation to vary randomly at both the school and teacher levels. This specification allowed us to estimate the amount of variation among schools and among teachers within schools in LC effects and generated empirical Bayes estimates of individual teacher and school effects.

Final model. Assembling all of these components together produced the following final model. The outcome, $Y_{ijk\ell}$, was defined as literacy status in logits at time i for student j in teacher k's class in school ℓ. Due to space limitations, we report here only the final mixed model, which was as follows:

$$
\begin{aligned}
Y_{ijk\ell} = {} & \delta_{000} + \delta_{0100} \times (\text{Entry\_2}_{j\ell}) + \delta_{0200} \times (\text{Entry\_3}_{j\ell}) + \delta_{0300} \times (\text{Entry\_4}_{j\ell}) \\
& + \delta_{0400} \times (\text{Entry\_5}_{j\ell}) + \delta_{0500} \times (\text{Entry\_6}_{j\ell}) + \delta_{0600} \times (\text{Cohort\_1}_{j\ell}) \\
& + \delta_{0700} \times (\text{Cohort\_2}_{j\ell}) + \delta_{0800} \times (\text{Cohort\_4}_{j\ell}) + \delta_{0900} \\
& \times (\text{Cohort\_5}_{j\ell}) + \delta_{01000} \times (\text{Cohort\_6}_{j\ell}) + b_{00j\ell} + d_{000\ell} \\
& + \delta_{1000} \times (\text{Academic\_year}_{ijk\ell}) + b_{10j\ell} \times (\text{Academic\_year}_{ijk\ell}) + d_{100\ell} \\
& \times (\text{Academic\_year}_{ijk\ell}) + c_{60k\ell} \times (\text{Base\_Tch\_VA}_{ijk\ell}) \\
& + \delta_{2000} \times (\text{First\_grade\_dev}_{ijk\ell}) \\
& + \delta_{3000} \times (\text{Second\_grade\_dev}_{ijk\ell}) \\
& + \delta_{4000} \times (\text{Summer\_K–1}_{ijk\ell}) + d_{400\ell} \times (\text{Summer\_K–1}_{ijk\ell}) \\
& + \delta_{5000} \times (\text{Summer\_1–2}_{ijk\ell}) + d_{500\ell} \times (\text{Summer\_1–2}_{ijk\ell}) \\
& + \delta_{7000} \times (\text{LC\_Year\_1}_{ijk\ell}) + c_{70k\ell} \times (\text{LC\_Year\_1}_{ijk\ell}) + d_{700\ell} \\
& \times (\text{LC\_Year\_1}_{ijk\ell}) \\
& + \delta_{8000} \times (\text{LC\_Year\_2}_{ijk\ell}) + c_{80k\ell} \times (\text{LC\_Year\_2}_{ijk\ell}) \\
& + d_{800\ell} \times (\text{LC\_Year\_2}_{ijk\ell}) \\
& + \delta_{9000} \times (\text{LC\_Year\_3}_{ijk\ell}) + c_{90k\ell} \times (\text{LC\_Year\_3}_{ijk\ell}) \\
& + d_{900\ell} \times (\text{LC\_Year\_3}_{ijk\ell}) \\
& + \delta_{10000} \times (\text{LC\_Summer\_1}_{ijk\ell}) \\
& + \delta_{11000} \times (\text{LC\_Summer\_2}_{ijk\ell}) \\
& + e_{ijk\ell}.
\end{aligned}
$$

Table 5 summarizes the interpretation for the parameters in this model that are of primary interest to the research questions.

Results

Estimates for the final model were derived using the HCM3 subprogram with HLM (version 7.0). HCM3 is a flexible program that allows for estimating a variety of four-level data models. In our application, the model consisted of repeated measures on students crossing teachers nested within schools. Formally, we represented this as repeated measures: (students × teachers): schools. All random effects at the teacher and school levels were treated as cumulative within HCM3. The full final fitted model is reported in Tables 6 and 7; however, only the most relevant results are discussed below.

As a check on model fit, we undertook a number of posterior predictive validity tests in which we used the results from the fitted model to predict the overall outcomes and school by school. Residuals between observed data and model-based outcomes were then examined. We found no evidence of any systematic variation in the posterior predictions as compared to the observed data.

Student-Level Variance in Growth

Results indicate that average entry literacy learning status (δ_{0000}) in kindergarten was .87 logits and the average literacy learning rate (δ_{1000}) was 1.02 during the academic year. Controlling for teacher and school effects, there was significant variance between children at entry into the dataset and in their growth rates (see Table 7). The student-level standard deviation at entry was 1.17, indicating a wide range of variability in students' incoming literacy learning. The student-level standard deviation for the academic year literacy learning rate was .25, indicating considerable variability in learning among children even after controlling for their specific classrooms and schools. Finally, these random effects are moderately and negatively correlated, meaning that children who entered with lower literacy tended to learn at a faster rate than those who enter with higher literacy.

Average LC Value-Added Effects

In terms of program effects, the average value added during the first year of implementation (δ_{7000}) was .16 logits. This represents a 16% increase in learning as compared to the average baseline growth rate of 1.02. During the second year of implementation, the estimated value added (δ_{8000}) was .28, which represents a 28% increase in productivity over baseline growth. The third year yielded an estimated value added (δ_{9000}) of .33 logits, which represents a 32% increase in productivity over baseline. These value-added effects convert into standard effect sizes of .22, .37, and .43, respectively, based on the residual level 1 variance ($e_{ijk\ell}$) estimated under the HCM3 model.

Table 5. Interpretation of Coefficients in the Final Model of Primary Interest to Answering the Research Questions

Variable	Variable Description
Intercept:	
δ_{0000}	Base initial literacy status for children in the fall of kindergarten
$b_{00j\ell}$	Variability in initial literacy status between children nested within schools
$d_{000\ell}$	Variability in initial literacy status between schools
Academic year growth parameter:	
Academic_year:	
δ_{1000}	Baseline literacy learning rate during the academic year, or the average growth in literacy status from fall to spring, after adjusting for cohort, entry, and value-added effects
$b_{10j\ell}$	Variability in latent academic year learning rates among students
$d_{100\ell}$	Variability in school value-added effects on student learning during the baseline period
Random effect for teacher baseline year effects:	
Base_Tch_VA:	
$c_{60k\ell}$	Variability in teacher value-added effects on student learning (within schools) during the baseline period (i.e., from fall to spring in the first year of the study)
Adjustments for grade-specific growth rates during academic year:	
First_grade_dev:	
δ_{2000}	Average adjustment to literacy learning rate (Academic_year) for first grade, controlling for cohort and entry
Second_grade_dev:	
δ_{3000}	Average adjustment to literacy learning rate (Academic_year) for second grade, controlling for cohort and entry
LC implementation effects:	
LC_Year_1:	
δ_{7000}	Average value-added effect for the first year of implementation averaged across schools and teachers
$c_{70k\ell}$	Variability in the value-added effects for the first year of implementation among teachers within schools
$d_{700\ell}$	Variability in the value-added effects during the first year of implementation among schools
LC_Year_2:	
δ_{8000}	Average value-added effect during the second year of implementation averaged across schools and teachers
$c_{80k\ell}$	Variability in the value-added effects for the second year of implementation among teachers within schools
$d_{800\ell}$	Variability in the value-added effects for the second year of implementation among schools
LC_Year_3:	
δ_{9000}	Average value-added effect during the second year of implementation averaged across schools and teachers
$c_{90k\ell}$	Variability in the value-added effects for the second year of implementation among teachers within schools
$d_{900\ell}$	Variability in the value-added effects for the second year of implementation among schools
LC_Summer_1:	
δ_{10000}	Sustaining effect of the first year value added through the summer, averaged across schools and teachers
LC_Summer_2:	
δ_{11000}	Sustaining effect of the second year value added through the summer, averaged across schools and teachers

Table 6. Final Hierarchical Crossed-Level Value-Added Model Fixed Effects

Fixed Effects	Coefficient	*SE*	*t*	*p*-value
Intercept, δ_{0000}	.866	.120	7.206	<.001
Entry_2, δ_{0100}	.948	.083	11.352	<.001
Entry_3, δ_{0200}	.775	.053	14.553	<.001
Entry_4, δ_{0300}	1.835	.090	20.485	<.001
Entry_5, δ_{0400}	2.164	.069	31.274	<.001
Entry_6, δ_{0500}	2.870	.104	27.564	<.001
Cohort_1, δ_{0600}	.397	.075	5.279	<.001
Cohort_2, δ_{0700}	.061	.058	1.054	.292
Cohort_4, δ_{0800}	.051	.045	1.142	.254
Cohort_5, δ_{0900}	.128	.049	2.613	.009
Cohort_6, δ_{01000}	−.145	.055	−2.651	.008
Academic_year, δ_{1000}	1.019	.079	12.952	<.001
First_grade_dev, δ_{2000}	.092	.043	2.127	.033
Second_grade_dev, δ_{3000}	−.193	.044	−4.350	<.001
Summer_K–1, δ_{4000}	−.184	.068	−2.727	.015
Summer_1–2, δ_{5000}	.362	.093	3.882	.001
LC_Year_1, δ_{7000}	.164	.038	4.271	<.001
LC_Year_2, δ_{8000}	.280	.049	5.669	<.001
LC_Year_3, δ_{9000}	.327	.068	4.824	<.001
LC_Summer_1, δ_{10000}	.205	.030	6.732	<.001
LC_Summer_2, δ_{11000}	.149	.039	3.852	<.001

Variation in LC Value-Added Effects

The school and teacher random effects in our model enabled the estimation of variation among schools and teachers in their value added (or extra contribution) to student learning each year (see Table 7). All of the variance components were statistically significant.

Variance in school-level effects. Figure 3 illustrates the distribution of value-added effects among schools. We display empirical Bayes estimates for the LC value-added effect in each of the 17 schools for each year of implementation. As reference points, the average first-, second-, and third-year effects are also displayed in Figure 3 as three dashed lines. Recall that a value-added effect of .0 signifies no improvement in academic year learning as compared to the baseline rate in that school. In almost every case, the estimated individual school effects are positive. The only exception was school 12 during the first year of LC implementation. Notice also that school value-added effects increased over time in most cases, with the most notable exception being school 16. Increasing variation over time in the magnitude of the school effects is also manifest in this display. For example, the largest individual school effect in the first year was about .30. In contrast, by the third year, several schools had effects of .35 or higher. This is equivalent to a 35% improvement during the final year of LC implementation in these schools as compared to the overall baseline academic learning rate of 1.02.

The variation among schools in academic growth rates during the baseline period was .082. Variance in LC effects across the 3 years of implementation increases from .013 to .036. There is a weak, positive relationship between the baseline academic year growth rate in a school and its first-year LC value-added

Table 7. Final Hierarchical Crossed-Level Value-Added Model Variance Components and Correlations among Them

Random Effects	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Residual:							
(1) σ^2	.584						
Student level:							
(1) Intercept, b_{00}	1.380						
(2) Baseline academic growth rate, b_{10}	−.428	.061					
Teacher level:							
(1) Baseline value-added, c_{60}	.056						
(2) Literacy Collaborative year 1 value added, c_{70}	.655	.068					
(3) Literacy Collaborative year 2 value added, c_{80}	.439	.613	.131				
(4) Literacy Collaborative year 3 value added, c_{90}	.378	.709	.498	.217			
School level:							
(1) Intercept, d_{000}	.221						
(2) Baseline value added, d_{100}	−.841	.082					
(3) Baseline K–1 summer, d_{400}	.399	−.684	.066				
(4) Baseline 1–2 summer, d_{500}	.883	−.878	.575	.136			
(5) Literacy Collaborative year 1 value added, d_{700}	−.372	.288	−.612	−.511	.013		
(6) Literacy Collaborative year 2 value added, d_{800}	−.045	−.030	−.477	−.104	.517	.020	
(7) Literacy Collaborative year 3 value added, d_{900}	.465	−.622	.230	.331	.226	.585	.036

Note.—Diagonal values are variance components and off-diagonal values are correlations among random effects.

effect, meaning that schools with higher growth at baseline accrued slightly larger value-added effects that year. In the second year, there is no relationship between the baseline academic year growth rate in a school and the LC value-added effects. By the third year, the correlation is strong and negative, indicating that larger LC value-added effects accrued in schools that had lower baseline academic year growth rates. Correlations between LC school value-added effects are moderately strong between the first and second years of implementation and the second and third years of implementation, but are weak between the first and third years of implementation.

Variance in teacher-level effects. Similarly, we found increasing variance between teachers within schools over time from .056 in the baseline year to .217 in the final year of implementation. The correlations among the teacher effects over time are moderate to strong, ranging from .38 to .71. These capture the consistency in the teacher effects from one year to the next after controlling for differences in latent growth trajectories of the children educated in each classroom.

Figure 4 illustrates the variability in teacher value-added effects over time. Each box plot represents the empirical Bayes estimates for the teacher effects during each year of LC implementation in each school. The increasing variability over time among teachers within schools is clearly visible in the increasing range of the box plots here. The vast majority of teachers in most of the participating schools

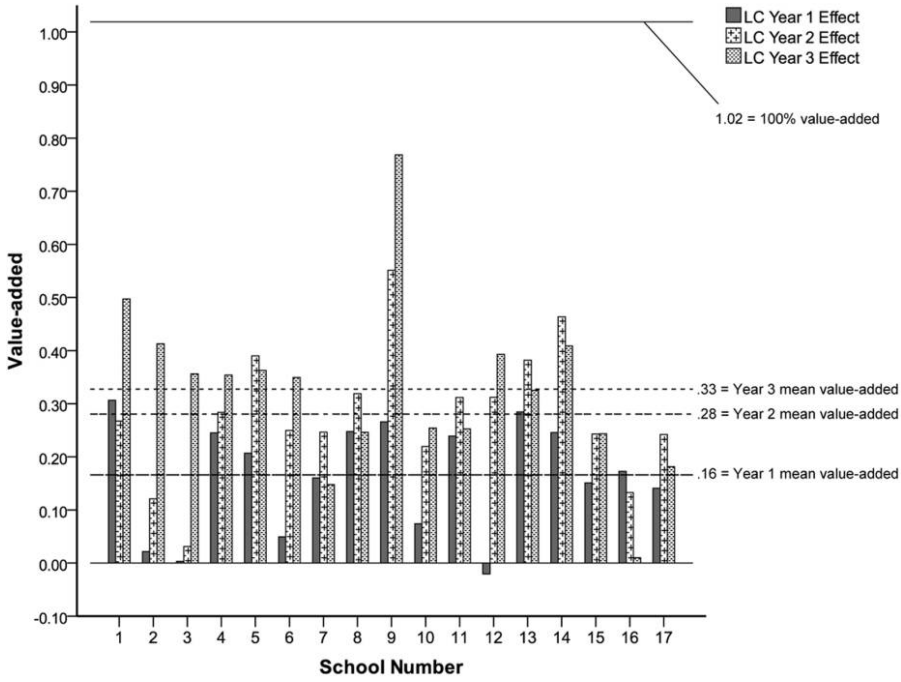

Figure 3. Variation in school value added over 3 years of LC implementation compared to the average value added across schools.

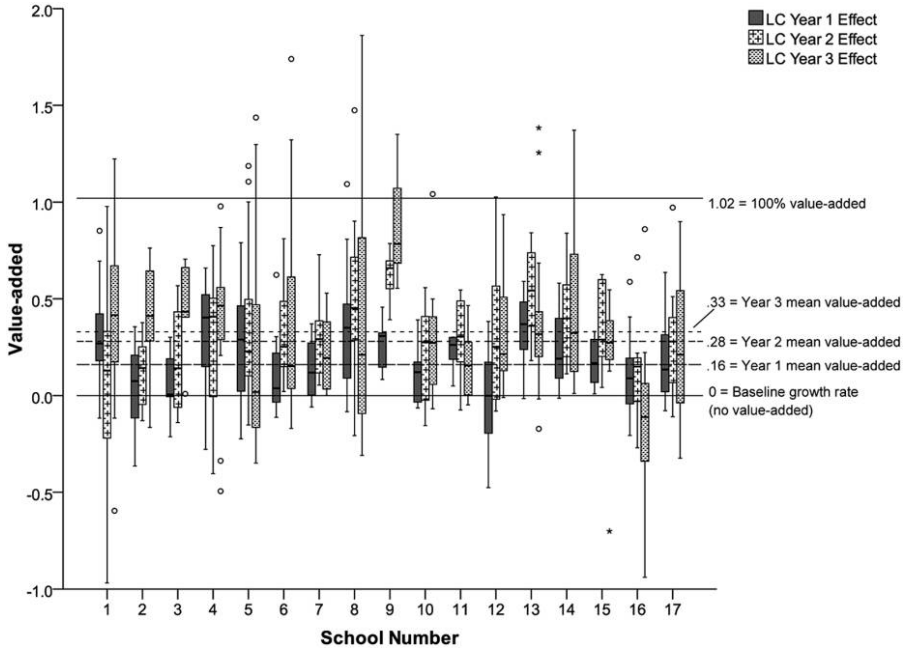

Figure 4. Variation in teacher value added over 3 years of LC implementation compared to school value added and average value added across schools.

showed substantial value-added effects by the end of the study. Moreover, a fair number of teachers show value-added effects of 1.0 or more by the third year.

Persistence of Effects over the Summer

The model also allowed examination of whether the value-added effects estimated at the end of the first and second years of implementation were maintained over the subsequent summers. The results indicate that program effects did persist through the subsequent fall testing. The average value-added effect after the first summer of LC implementation (δ_{10000}) was .21 logits, which is somewhat higher than the first-year implementation effect of .16 logits. The average value-added effect after the second summer of LC implementation (δ_{11000}) was .15 logits, which is about half the magnitude of the second-year implementation effect of .28 logits. An omnibus test comparing LC effects at the end of the summer to those estimated at the end of the preceding academic year yielded statistically indistinguishable results ($\chi^2 = 1.43$, $df = 1$, $p = .2303$). Improvements during the academic year of LC implementation appear to persist through the following fall assessments.

As a check on model fit, we conducted a number of tests of the posterior predictive validity of our model during which we used the results from the fitted model to predict outcomes school by school, as well as across schools. Specifically, observed data were compared to model-based residuals and we found no evidence of any systematic variation in the predictions as compared to the observed data.

Discussion

This article presents the results from a 4-year longitudinal study of the effects of the LC program on student learning in 17 schools. A rigorous quasi-experiment was designed to estimate separate program effects by year, school, and teacher. Inferences about these program effects are based on a value-added analysis that compares student learning gains during LC implementation to those achieved in the same classrooms and schools prior to the program intervention.

Results demonstrate significant gains in student literacy learning beginning in the first year of implementation and that the effect's magnitude grew larger during each subsequent year of implementation. On average, children in participating schools in the first year of implementation made 16% larger learning gains than observed during the baseline no-treatment period. In the second year, children learned 28% more compared to the baseline data, and by the third year they had learned 32% more. Our analyses also indicate that these results persisted across summer periods as verified through the follow-up of students in the fall of the subsequent academic year.

These results contrast with findings from two other recent studies, which reported that literacy coaching had little (Marsh et al., 2008) or no (Garet et al., 2008) impact on student learning. One reason for the novel findings may be that the coaching strategies evaluated in the two previous studies differed significantly from the LC model investigated in this article. As noted in the background to the current study, the LC model involves a full year of PD for coaches before they

begin working with teachers, whereas coaches in the aforementioned studies received only a few days of training prior to engaging in their new roles. In addition, coaching in the LC model is organized around a detailed and well-specified literacy instructional system that includes a repertoire of instructional practices. In contrast, the coaching models in the previous studies were designed more as supplements to extant and more varied curricula. In the Garet et al. (2008) study, coaching was literally an add-on to another PD curriculum, and in the Marsh et al. (2008) study, the coaching framework was set forth at a state level with an expectation of variation in content and implementation across local school districts. Either of these differences, as well as other possible differences among the coaching programs studied, could account for the significant differences in estimated effects found here and in previous research. Continued research on multiple models of coaching across multiple contexts is needed to resolve whether, and under what conditions, coaching can stimulate improvement in student learning. At a minimum, the current study does suggest that well-specified and well-supported coaching initiatives can effect positive changes in student learning.

The overall pattern of effects—increasing over time in both size and variability within and between schools—also merits comment. In each study school, coaching was a new professional undertaking for the individuals who took on this role. Both LC program documents and other more general clinical accounts of instructional coaching detail a complex professional role that may take several years to learn well (Gibson, 2005, 2006). Just as novice teachers improve during the first few years of learning to teach (Borko & Putnam, 1996; Darling-Hammond & Bransford, 2005; Kane, 1993; Snow, Griffin, & Burns, 2005), it is reasonable to hypothesize that coaches do as well. The increasing size of the estimated program effects from the first year to the third year is certainly consistent with the aforementioned hypothesis.

Thus, increasing coaching expertise over time is one plausible explanation for the temporal patterns observed in the current study. Unfortunately, reliable and valid measures of coaching expertise do not yet exist and therefore could not be employed in the current study. The current findings cohere with previous calls for research to develop coaching expertise measures and explore the potential role of a coach's developing expertise as an explanatory factor for differences in effects on both teachers and students (Neufeld & Roper, 2003).

The observed increasing effects over time may also have resulted from changes in the informal professional networks around literacy instruction within participating schools. If the professional context of a school changes as a result of coaching, teachers, especially new teachers, may benefit not only from the mentoring of a more experienced coach, but also through the social learning that can occur as they interact with increasingly more expert colleagues. This represents another plausible hypothesis for future research to investigate. Some preliminary descriptive evidence drawn from the current study that supports this latter account has been reported elsewhere (Atteberry & Bryk, 2009).

In terms of the increasing variability in effects over time, both within and between schools, this too seems quite plausible given the nature of a coach's work. In principle, the effects of coaching accrue over time through the one-on-one interactions that occur between a coach and an individual classroom teacher.

Stated somewhat differently, coaching is a relational practice whose efficacy presumably depends on the quality of the relationship that a coach is able to establish with each individual teacher (Bean & Carroll, 2006). Variations in the quality of these connections, as well as the duration of these interactions, might well account for the increasing variability in teacher value-added effects on student learning reported in this study. This pattern is consistent with the findings of Marsh et al. (2008), who reported that larger effects on student literacy learning were associated with coaches who had been coaching for a longer period of time. Individual variability in teacher effects may well be a structural characteristic of instructional coaching efforts; if so, this variability merits further consideration by researchers and coach trainers so that future coaching initiatives effect greater consistency in teachers' instructional improvement over time.

The overall pattern of increasing variability in teacher effects found in this study is consistent with a work organization where literacy coaches operate with considerable discretion in engaging individual teachers in extended instructional coaching. By design, coaching is an intervention from which we might reasonably expect variable effects to accrue depending on the quality of the coach, the school context in which the coach works, and the varying amounts of coaching that each individual teacher receives. Future studies should include larger samples of schools and coaches in order to examine more thoroughly these possible sources of variability in effects.

Limitations

This study was undertaken to assess the effects of the LC program on student literacy learning and found statistically and substantively significant impacts. Even though the study involved a rigorous quasi-experimental design, we need to consider the plausibility of the major competing hypotheses for the observed results reported here.

Foremost, we must consider whether the observed improvements in learning gains over time are in fact school improvement effects or might alternatively be attributable to student selectivity effects. Student selectivity is generally less problematic in an accelerated multicohort longitudinal design because value-added estimates are based on changes in student learning as compared to baseline results from the same students, classrooms, and schools. In other words, each student, classroom, and school serves as its own individual control. Only if the student composition in these schools were changing over time coterminous with the study would a plausible alternative explanation exist. However, we found no evidence of this. We collected basic student demographic data each year, and no substantial changes in student characteristics were observed over time. More significantly, the data presented in Figure 2 document almost identical mean achievement scores for students at entry into kindergarten over all 4 years of the study. In addition, our model included cohort effects, which captured differences in baseline abilities between each cohort (see Table 6). Therefore, the value-added effects are observed even after controlling for historical differences in cohorts. For all these reasons,

taken together, the current empirical results make a student selectivity hypothesis implausible.

Therefore, we conclude that the current results provide sound evidence that sample schools actually did improve performance during the course of the study. However, we still have to consider plausible competing hypothetical causes, other than the LC program, that might also account for these effects. That is, might there have been a concomitant alternative treatment effect, occurring during the same time period in these same schools, that could explain the observed results? On balance, school districts are complex organizations with multiple sources of programs and funding. In fact, the study was carried out during the period when NCLB and Reading First initiatives were being introduced. These federal initiatives focused districts on improving reading skills through a variety of mechanisms, including structured literacy block times, a focus on skill work in the early grades, and structured accountability procedures. In any given school, initiatives of this sort could in principle account for some or all of the effects observed in this study.

This alternative explanation for our findings strikes us as unlikely for several reasons. First, the LC program required a substantial fiscal commitment on the part of participating schools. While the study did offer districts a partial reduction in costs for the LC training of school coaches, districts still incurred substantial out-of-pocket expenses for residual tuition costs, travel costs to send novice coordinators for training to Boston, Massachusetts, or Columbus, Ohio, and the salary costs required to free a teacher half time to prepare for and assume the literacy coach role. In short, districts were asked to make a major commitment to LC as their literacy reform strategy and to use their available discretionary resources to fund it. That a second equally significant reform initiative might have coexisted in these same schools at the same point in time seems dubious.

Second, although base individual teacher effects are included as controls in the value-added model, changes in these effects over time could be influenced by other exogenous professional improvement opportunities simultaneously being afforded teachers. This theory seems unlikely given the scope of program effects documented above. Specifically, at the school building level, LC participation makes substantial time demands on teachers' time. As a result, the likelihood of teachers being engaged simultaneously in some other literacy PD seems improbable. Any exogenous teacher-improvement efforts would have to have been both deep and widespread to possibly account for the observed data.

Third, we conducted annual interviews with each literacy coach about the progress of the initiative in their respective buildings. We had no reports of concurrent competing initiatives that might account for the broad base of results observed.

Fourth, the fact that the effects were broadly based, accruing by the third year in almost all of the 17 schools, adds further doubt to the alternative treatment-effect hypothesis. Study schools were located in eight different states and nine different districts. For an alternative concurrent treatment to account for the observed results, most of these districts would have had to develop another competing and equally effective literacy-improvement program in the same schools at the same time this initiative took place. While it is possible that other effective

local initiatives may have been introduced in one or more schools during this period, achieving such coincidental effective change consistently across the diverse sites engaged in this study seems unlikely.

Finally, the overall magnitude of the effects documented in the current article militates against a plausible alternative treatment effect as a full explanation for the observed outcome. Even if we were to assume that some significant alternative program effects existed in some schools, and even if they accounted for half of the observed effects (all of which seems doubtful given the evidence and arguments already outlined), the residual effect sizes would still represent a meaningful contribution to improving student learning.

Conclusion

For all of the aforementioned reasons, we conclude that the study schools' participation in the LC program led to positive program effects on children's literacy. Given the prominent role of coaching in this program as a lever for enacting change in teachers' practice and consequently in students' learning, this study contributes important new evidence of the potential for literacy coaching to yield improvements in student literacy outcomes. Nonetheless, the evidence supporting coaching's effectiveness is still slim, and it is unclear whether the effects found here are specific to LC or can be observed in other coaching models. Studies of the variations within and between coaching models, particularly with a focus on coach preparation and developing expertise, would help to clarify the mechanisms by which coaching can be an effective lever for change in student achievement.

Notes

The work described in the current article was supported by a Teacher Quality Grant from the Institute for Educational Sciences (IES), R305M040086. We are appreciative of the support provided by IES. All errors of fact, omission, and/or interpretation are solely the authors' responsibility. The research team involved in this study included affiliates of the coaching program being investigated. This collaboration informed the study design and the development of tools to measure teacher practice. The analytical team worked independently from those affiliated with the program to ensure objectivity. Correspondence regarding this article should be addressed to Gina Biancarosa, College of Education, 5261, University of Oregon, 1655 Alder Street, Eugene, OR 97403. E-mail: ginab@uoregon.edu.

1. LC coaches may remain active teachers either by having a partner teacher who typically covers science and math instruction while coaching occurs, or by co-teaching in others' classrooms for extended periods of time, ranging anywhere from one month to an entire year.

2. For further details about the LC program, see http://www.literacycollaborative.org/.

3. The initial design involved 18 schools, but one school was subsequently lost because a coach for the school was never certified by LC due to the candidate coach's failure to complete the Literacy Coordinator training. Thus, the school never implemented coaching.

4. Technical information that could not be incorporated into the current article due to length constraints is available at The Carnegie Foundation for the Advancement of Teaching's web site (http://www.carnegiefoundation.org/elibrary) as well as directly from the first author.

5. Further details on the functioning of this set of indicators can be found on the Carnegie Foundation web site (http://www.carnegiefoundation.org/elibrary) as well as directly from the first author.

6. Further details on the functioning of this set of indicators can be found on the Carnegie Foundation web site (http://www.carnegiefoundation.org/elibrary) as well as directly from the first author.

7. Cohort 3, which entered kindergarten during the study's baseline year, served as a reference category.

8. Further details on the functioning of this set of indicators can be found on the Carnegie Foundation web site (http://www.carnegiefoundation.org/elibrary) as well as directly from the first author.

9. An illustration of the simultaneous operation of these five indicators can be found in the technical documentation on The Carnegie Foundation for the Advancement of Teaching's web site (http://www.carnegiefoundation.org/elibrary) as well as directly from the first author.

References

Allen, J. (2006). *Becoming a literacy leader: Supporting learning and change.* Portland, ME: Stenhouse.

Atteberry, A., & Bryk, A. (2009, April). *The role of schools' social networks in intervention diffusion.* Paper presented at the annual meeting of the American Education Research Association, San Diego.

Barton, R., & Lavrakas. (2006). Finding gold at the end of the rainbow: Anchorage's investment in literacy coaching pays big dividends. *Northwest Regional Education Laboratory,* **12**(1), 6–11.

Bean, R. M., & Carroll, K. E. (2006). The literacy coach as a catalyst for change. In C. Cummins (Ed.), *Understanding and implementing Reading First initiatives: The changing role of administrators* (pp. 139–152). Newark, DE: International Reading Association.

Blachowicz, C. L. Z., Obrochta, C., & Fogelberg, E. (2005). Literacy coaching for change. *Educational Leadership,* **62**(6), 55–58.

Borko, H., & Putnam, R. T. (1996). Learning to teach. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 673–708). New York: Macmillan.

Brown, R. S., & Coughlin, E. (2007). *The predictive validity of selected benchmark assessments used in the mid-Atlantic region* (Issues & Answers, REL 2007-No. 017). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic.

Bruner, J. (1986). *Actual minds, possible worlds.* Cambridge, MA: Harvard University Press.

Bruner, J. (1996). *The culture of education.* Cambridge, MA: Harvard University Press.

Casey, K. (2006). *Literacy coaching: The essentials.* Portsmouth, NH: Heinemann.

Clay, M. (1979). *Reading: The patterning of complex behavior* (2nd ed.). Portsmouth, NH: Heinemann.

Clay, M. M. (1991). *Becoming literate: The construction of inner control.* Portsmouth, NH: Heinemann.

Clay, M. (2004). *Change over time in literacy learning.* Auckland: Heinemann Educational.

Darling-Hammond, L., & Bransford, J. (2005). *Preparing teachers for a changing world: What teachers should learn and be able to do.* San Francisco: Jossey-Bass.

Fountas, I. C., & Pinnell, G. S. (1996). *Guided reading: Good first teaching for all children.* Portsmouth, NH: Heinemann.

Fountas, I. C., & Pinnell, G. S. (2006). *Teaching for comprehending and fluency: Thinking, talking, and writing about reading, K–8.* Portsmouth, NH: Heinemann.

Gallimore, R., & Tharp, R. (1990). Teaching mind in society: Teaching, schooling, and literate discourse. In L. C. Moll (Ed.), *Vygotsky and education: Instructional implications and applications of sociohistorical psychology* (pp. 175–205). Cambridge: Cambridge University Press.

Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., et al. (2008). *The impact of two professional development interventions on early reading instruction and achievement* (NCEE 2008-4030). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Gibson, S. (2005). Developing knowledge of coaching. *Issues in Teacher Education,* **14**, 63–74.

Gibson, S. (2006). Lesson observation and feedback: The practice of an expert reading coach. *Reading Research and Instruction,* **45**, 295–318.

Good, R. H., & Kaminski, R. A. (2002). *Dynamic indicators of basic early literacy skills* (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement.

Good, R. H., Wallin, J., Simmons, D. C., Kame'enui, E. J., & Kaminski, R. A. (2002). *System-wide percentile ranks for DIBELS benchmark assessment* (Technical Report 9). Eugene: University of Oregon.

Hintze, J. M., Ryan, A. L., & Stoner, G. (2003). Concurrent validity and diagnostic accuracy of the Dynamic Indicators of Basic Early Literacy Skills and the Comprehensive Test of Phonological Processing. *School Psychology Review,* **32**, 541–556.

International Reading Association. (2004). *The role and qualifications of the reading coach in the United States.* Newark, DE: Author.

International Reading Association. (2006). *Standards for middle and high school literacy coaches.* Newark, DE: Author.

Kane, R. G. (1993). *Knowledge in action: A longitudinal study of the propositional and procedural knowledge base of the beginning teacher.* Unpublished master's thesis, Griffith University, Brisbane, Australia.

Kohler, F., Crillery, K., Shearer, D., & Good, G. (1997). Effects of peer coaching on teacher and student outcomes. *Journal of Educational Research,* **90**, 240–250.

Lieberman, A. (1995). Practices that support teacher development. *Phi Delta Kappan,* **76**(8), 591–596.

Literacy Collaborative. (2009). *Implementation and training.* Retrieved October 22, 2009, from http://www.literacycollaborative.org/about/phases/

Luppescu, S., Biancarosa, G., Kerbow, D., & Bryk, A. S. (2008, July). *Testing the fluency-comprehension relationship through equating: A Rasch equating of DIBELS and Terra Nova in grades K–3.* Paper presented at the annual meeting of the Society for Scientific Studies of Reading, Asheville, NC.

Marsh, J., Kerr, K., Ikemoto, G., Darilek, G., Suttorp, H., Zimmer, R., et al. (2005). *The role of districts in fostering instructional improvement: Lessons from three urban districts partnered with the Institute for Learning.* Santa Monica, CA: RAND.

Marsh, J. A., Sloan McCombs, J., Lockwood, J. R., Martorell, F., Gershwin, D., Naftel, S., et al. (2008). *Supporting literacy across the sunshine state: A study of Florida middle school reading coaches.* Santa Monica, CA: RAND.

McCarrier, M. C., Pinnell, G. S., & Fountas, I. C. (2000). *Interactive writing: How language and literacy come together.* Portsmouth, NH: Heinemann.

McGraw-Hill. (2001). *Terra Nova technical quality: Reliable, useful results based on psychometric excellence.* Monterey, CA: CTB McGraw-Hill.

Neufeld, B., & Roper, D. (2004). *Coaching: A strategy for developing instructional capacity.* Cambridge, MA: Education Matters.

Norlander-Case, K. A. (1999). *The professional teacher: The preparation and nurturance of the reflective practitioner.* San Francisco: Jossey-Bass.

Norton, J. (2007). Adding layers of support: Alabama's program helps site-based coaches succeed. *National Staff Development Council,* **28**, 20–25.

Poglinco, S. M., Bach, A. J., Hovde, K., Rosenblum, S., Saunders, M., & Supovitz, J. A. (2003). *The heart of the matter: The coaching model in America's Choice schools.* Philadelphia: Consortium for Policy Research in Education.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Reed, B., & Rettig, R. (2006). Side by side: The Idaho Reading First program finds success by offering targeted professional development to its literacy coaches. *Northwest Education,* **12**(1), 18–21.

Scharer, P. L., Pinnell, G. S., & Bryk, A. S. (2008). Supporting learning through a multi-layered professional community: Making change work with Reading Recovery and Literacy Collaborative. *Journal of Reading Recovery,* **7**(2), 42–52.

Schilling, S. G., Carlisle, J. F., Scott, S. E., & Zeng, J. (2007). Are fluency measures accurate predictors of reading achievement? *Elementary School Journal,* **107**, 429–448.

Schön, D. A. (1983). *The reflective practitioner: How professionals think in action.* New York: Basic.

Snow, C. E., Griffin, P., & Burns, M. S. (2005). *Knowledge to support the teaching of reading: Preparing teachers for a changing world.* San Francisco: Jossey-Bass.

Toll, C. (2007). *Lenses on literacy coaching: Conceptualizations, functions, and outcomes.* Norwood, MA: Christopher-Gordon.

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Walpole, S., & McKenna, M. C. (2004). *The literacy coach's handbook: A guide to research-based practice.* New York: Guilford.

Wright, B. D., & Master, G. N. (1982). *Rating scale analysis.* Chicago: Mesa.