EFFICACY

CRESST Grades 3–4

Amplify.

Evaluation of
Seeds of Science/Roots of Reading:
Effective Tools for Developing Literacy through Science in the Early Grades

Final Deliverable – May 19, 2010

Pete Goldschmidt
CRESST/University of California, Los Angeles

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported by grant number SA5415 from the SEEDS of Science/Roots of Reading Project with funding to the National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

TABLE OF CONTENTS

EVALUATION OF SEEDS OF SCIENCE/ROOTS OF READING: EFFECTIVE TOOLS FOR DEVELOPING LITERACY THROUGH SCIENCE IN THE EARLY GRADES[1]

Pete Goldschmidt and Hyekyung Jung
CRESST/University of California, Los Angeles

Abstract

This evaluation focuses on the Seeds of Science/Roots of Reading: Effective Tools for Developing Literacy through Science in the Early Grades *(Seeds/Roots)* model of science-literacy integration. The evaluation is based on a cluster randomized design of 100 teachers, half of which were in the treatment group. Multi-level models are employed to account for the clustering of students within teachers and teachers within schools. Four primary outcomes of interest are examined: Science content; vocabulary; reading; and, writing. Additional analyses focus on the impact of teacher and student background, instructional methods, and teacher-self efficacy. Quantitative results indicate that the *Seeds/Roots* intervention resulted in statistically and substantively higher student performance in science content, vocabulary, and writing. Teacher background and self-efficacy are generally unrelated to student performance. Inquiry-based teachers enhanced treatment effects. Despite *Seeds/Roots* designed integration, teachers tended to focus on the science aspect when considering time requirements to be longer than a standard unit. Qualitative results indicate that teachers overwhelmingly found the *Seeds/Roots* unit usable, effective, and engaging

This evaluation focuses on the Seeds of Science/Roots of Reading: Effective Tools for Developing Literacy through Science in the Early Grades *(Seeds/Roots)* model of science-literacy integration for Grade 4, developed and implemented by the Lawrence Hall of Science (LHS). The *Seeds/Roots* study is a multi-year project funded by the National Science Foundation. The project evaluation efforts build on previous *Seeds/Roots* evaluations (Wang & Herman, 2006) and focus on two major goals of the materials: usability and effectiveness. Formative evaluation processes (such as science assessment modification and rubric testing) provided opportunities for ongoing analysis and improvement. Summative evaluation efforts have been designed to provide evidence of usability and effectiveness. This report focuses on the summative evaluation of the Light/Energy unit. Given the experimental design (teachers randomly assigned to treatment or control groups and the abundance of data collected, the majority of the analyses reported are based on quantitative methods; however, a small random sample of teachers were interviewed to provide some qualitative perspective on the *Seeds/Roots* intervention as well. Seeds/Roots uses an integrated approach to teaching science and

[1] We would like to acknowledge important contributions from the LHS staff who provided data and clarifications for the many inquiries we made.

literacy and this evaluation will provide evidence for the benefit(s) of utilizing an integrated approach in comparison to standard instructional practices in a 4th grade Light/Energy unit.

Background on the Treatment

Seeds/Roots is an integrated science-literacy program designed for Grades 2-5, partially based on revisions of units in the Great Explorations in Math and Science (GEMS) Program. The *Seeds/Roots* unit is designed as a next generation of standards-aligned elementary inquiry science materials that advance student learning in science while meeting the challenges of: an increasingly congested school day, low levels of elementary teacher preparation and efficacy in science, the pressures of large-scale testing, and the growing diversity of our nation's classrooms. *Seeds/Roots* science -literacy integration is based on previous literature on integrated methods. The emphasis is on integrating content-area learning, reading and writing. This approach to science-literacy integration ideally fosters a synergistic relationship (Cervetti, Pearson, Bravo, & Barber, 2006). The *Seeds/Roots* model builds on previous work that has demonstrated positive effects from using an integrated approach (Guthrie & Ozgungor, 2002; Romance & Vitale, 1992). There are three approaches to instructional integration, Stoddart, Pinal, Latzke, and Canaday (2002): a thematic approach characterized by the use of overarching themes to create connections among domains; an interdisciplinary approach in which content or processes in one domain are used to support learning in another; or, an integrated approach, in which emphasis on two or more domains is balanced. Details of *Seeds/Roots* integrated curriculum and process to achieve balance are discussed in Cervetti, Barber, Dorph, Pearson, and Goldschmidt (2009).

Evaluation Design and Objectives

In order to determine whether there are statistically significant and substantively important effects from using an integrated science and literacy approach to instruction, compared to content-comparable business-as-usual science instruction, the *Seeds/Roots* unit was embedded in a curriculum unit on light, which involved students in doing, talking, reading, and writing about the characteristics of light. The unit also provided opportunities for explicit instruction of literacy abilities, such as: using the reading comprehension strategies of making predictions and summarizing; writing summaries; using nonfiction text structures to find information; and engaging in oral discourse.

During the 2007-2008 school year 100 4th grade teachers, teaching in 49 schools, in rural and urban counties in a Southern state, participated in the study. This state was selected as a study site because of the close relationship between that state's science standards (at Grade 4) addressing light and the content of the integrated science-literacy light unit, more

easily enabling a content-comparable comparison group. Teachers were randomly assigned to either: 1) present the integrated science-literacy light unit to their students (treatment group); or 2) present the content of their state science standards related to light, using whatever curriculum materials they regularly use (control group).

LHS researchers administered pre-tests and post-tests in science and literacy to students in all treatment and control classrooms, the week before and the week after a 12-week teaching window. The evaluation plan called for quantitative summative analysis of student performance, student attitudes, teacher attitudes, and teacher efficacy. The plan intended to evaluate these elements by collecting data using the following instruments for students:

1. An assessment of science knowledge;
2. An assessment of science vocabulary;
3. An assessment of reading comprehension using related and unrelated science passages;
4. A science writing assessment; and
5. An assessment of student attitudes towards science.
6. Student demographics were collected from districts as well as their results on the state standardized test results for science and English language arts[2].

   And for teachers:

7. Surveys of teacher background; and,
8. Pre- and Post-surveys of teacher attitudes and self-efficacy.

Given these data, the evaluation focused on examining two aspects related to the implementation and effectiveness of the *Seeds/Roots* unit. Evaluation of implementation relates to examining the impact of implementation on outcomes, as well, as examining teacher perceptions regarding the unit's efficacy and student engagement. Effectiveness is evaluated by examining outcomes related to student learning in science, student learning in literacy, and teacher attitudes and practices[3]. Given that students are assigned to treatments by teacher (cluster randomized design) and teachers teach within schools) a multilevel

[2] Due to the (often long) interval between *Seeds/Roots* assessments and the availability of state and (including student demographics) several districts were unable or unwilling to provide student demographic and/or state assessment results. Analyses proceeded on available data. Comparability to the full sample was examined and is discussed in the text.

[3] The initial evaluation plan also intended to utilize state assessment results; however, the subsample for which we received state assessment results substantively differed from the full sample casting doubt on inferences based on his sample. We present these analyses in an appendix. Ideally, disaggregation of results is an important aspect as it presents an opportunity to examine whether the *Seeds/Roots* unit is particularly beneficial for student at-risk. In this case this relates to low SES, free/reduced lunch, or title I students, and English Language Learners (ELL). Triangulation of results relates to using independent assessments (i.e. the *Seeds/Roots* unit assessments and state assessments, as well as teacher perceptions of efficacy).

modeling framework is used to account for the design, the lack of independence among observations within units (i.e classrooms), and to take advantage of the data structure by examining the potential impact of context on treatment effects. The MLM analyses are outlined below. The following research questions guided the data collection and choice of analyses methods. These are the focus of the quantitative evaluation of the *Seeds/Roots* unit efficacy:

- Do students who use the Seeds/Roots units make progress in science and literacy? Do they make more progress in science and literacy than students using other materials?
- Are there differences in learning outcomes by gender, ethnicity, or previous educational achievement? What learning gains are being made by students who have particular educational needs (such as English Language Learners)?
- How do the materials influence teachers' attitudes toward science teaching? Toward literacy teaching?
- How do different teachers (experience with inquiry science, years teaching, efficacy) interact with the materials?
- To what extent and how are the units implemented?
- What distinguishes successful from less successful use of these materials?
- What are teachers' reactions to the quality, usability and utility of the units?

Methods and Data

In studies of program or intervention effects in schools using pre and post tests, students are typically nested within different sites (classrooms). Ignoring the nested structure of the data gives rise to two main problems—misleadingly small standard errors for treatment effect estimates and failing to detect between-site (classroom) heterogeneity in intervention effects (Seltzer, 2004; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). The between-site heterogeneity is not surprising, because class intake can vary, teachers can vary considerably in terms of implementation, background characteristics of participants, as well as factors that are related to the treatment effects. This is both a statistically and substantively important issue. By using a three-level random effects model, we are able to divide the variation in achievement into between-student, between-teacher, and error components. This is particularly important to do because data containing multiple levels of aggregation can lead to errors in interpretation when these multiple levels are ignored (Aitkin & Longford, 1986; Burstein, 1980).

We utilize multilevel models (MLM), specifically, a three level model that includes students, teachers, and schools. This three level MLM forms the basis for analyses of the outcomes using various specifications of the model described below. The model consists of three levels and allows for a flexible specification of the covariance structure at every level of the analysis (Snijders & Bosker, 1999). MLMs are flexible, yet powerful tools for understanding the impact of a treatment on student performance (Raudenbush & Bryk, 2002). In order to examine the potential impact of the treatment, we use lagged performance in order to examine residual change in student performance. Using a three level model, students represent Level 1, teachers Level 2, and schools Level 3.

The Level 1 model is:

$$Y_{ijk} = \pi_{0jk} + e_{ijk}, \qquad (1a)$$

where Y_{ijk} is the outcome (e.g. *Seeds/Roots* Science content assessment) for student *i* in class[4] *j* in school *k*. Where π_{0jk} represents mean outcome of classroom *j* in school *k*. Finally, e_{ijk} is a random student effect.

At Level 2 (between teachers, within schools) we model the impact of the treatment, given that treatment assignment was by teacher (teacher level).

$$\pi_{0jk} = \beta_{00k} + \lambda_{01k} TRT_{jk} + r_{0j} \qquad (2)$$

In (2) β_{00k} represents the school mean performance while λ_{01k} represents the treatment effect. Both r_{0jk} and r_{1jk} are random teacher effects. Using (2) alters the interpretation of π_{0jk}. Now π_{0jk} is the mean class performance of control classrooms and $\pi_{0jk} + \lambda_{01k}$ is the mean performance of treatment classrooms.

$$\beta_{00k} = \gamma_{000} + u_{00k}$$

$$\lambda_{01k} = \gamma_{010} \qquad (3)$$

In (3) γ_{000} is the grand mean of student performance. γ_{010} is the overall treatment effect.

The Level 1 model represented in (1a) can be further specified to account for differences in classroom intake characteristics—e.g. pre-test performance or student background characteristics. The Level 1 model then, becomes:

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk} (Y_{ijk} - Y_{..k}) + e_{ijk}, \qquad (1b)$$

Hence, π_{0jk} becomes the adjusted mean outcome of control[5] classroom *j* in school *k*.

[4] We use the term class and teacher interchangeably. It is natural to consider a group of students sitting in a classroom, but each classroom is taught by a single teacher. Moreover student performance is considered to be impacted by the teacher.

$$\pi_{1jk} = \beta_{10k} + \gamma_{11k} TRT_{jk} + r_{1jk} \qquad (2b)$$

Given the extension (or possible extension) in 1b, the Level 2 model can be specified to include treatment indicators. Hence, β_{10k} represents the mean class relationship between the pre-test and the post-test in control classrooms. γ_{11k} represents the cross-level interaction between the treatment and pre-test scores. Whereas γ_{01k} represents the main effect of the treatment; i.e. did treatment classrooms outperform control classrooms, given pre-test performance, γ_{11k} estimates whether the treatment is differentially effective for students with different levels of preparedness—i.e. pre-test scores. This cross-level interaction tests whether the student preparedness moderates the treatment effect. This becomes an important mechanism for testing the differential impact of the treatment on specific subgroups of students. The example above uses prior student knowledge which allows for the evaluation of *Seeds/Roots* unit impact on low/high achievers, but additional student characteristics can be added to 1b and tested by expanding 2b (e.g. including ELL status in model 1b and adding a $\gamma_{11k} TRT_{jk}$ into 2b).

At Level 3 we account for the fact that classrooms are nested within schools. Using an average pre-test for the classroom tests the impact of the classroom average achievement, or context, on individual student post-test performance. An interaction between the treatment and control indicator and the average classroom performance tests whether the impact of average classroom performance affects individual student performance differently in control and treatment classrooms.

Data

Given that teachers are the unit of assignment, we first present in Tables 1a and 1b descriptive results for teachers. These include teacher background characteristics as well as pre and post treatment survey results related to practices, perception of student engagement, unit efficacy, and self efficacy. Results indicate that treatment teachers were less experienced (4.5 years of teaching) than comparison (control) teachers (5.5 years of teaching). Control teachers were also more educated, with 51% vs 34% having an advanced degree. The natural log of salary was roughly equal. Salary is a potentially interesting covariate because it combines tenure and education in a specific way (determined by the district) and provides an additional indicator of the potential impact of the combination of education and experience. Class size was roughly equal across conditions although comparison classes consisted of about twice as many ELL students.

[5] Control classroom given (2).

Table 1a also presents indicators of teacher practices prior to the treatment period. This includes

Table 1a

Teacher Background, Practices, and Perceptions

	Total			Comparison classrooms			Treatment classrooms		
	Mean	*N*	*SD*	Mean	*N*	*SD*	Mean	*N*	*SD*
Treatment teacher							.50	94	.503
Teacher education and experience									
Two or more certifications	0.43	94	0.50	0.43	47	0.50	0.43	47	0.50
Teach math and science	0.77	94	0.43	0.72	47	0.45	0.81	47	0.40
Years teaching	5.00	94	4.18	5.55	47	4.61	4.45	47	3.67
BA Degree	0.17	94	0.38	0.15	47	0.36	0.19	47	0.40
MA Degree	0.38	94	0.49	0.45	47	0.50	0.32	47	0.47
PhD. Degree	0.04	94	0.20	0.06	47	0.25	0.02	47	0.15
Other Degree	0.28	94	0.45	0.19	47	0.40	0.36	47	0.49
Advanced Degree	0.43	94	0.50	0.51	47	0.51	0.34	47	0.48
Ln (Salary	10.57	94	0.15	10.61	47	0.16	10.53	47	0.13
Classroom characteristics									
Number of student in class	22.40	94	4.25	22.15	47	4.45	22.64	47	4.08
Number of ELL in class	1.80	94	4.51	2.41	47	6.04	1.20	47	2.01
Percent ELL	8.0	94	2.10	11.0	47	28.0	5.00	47	8.00

both time and instructional mix. A key element of these practices is whether a teacher used inquiry-based teaching practices. Inquiry-based is dichotomized by defining an inquiry-based teacher as one who used hands-on practices at least 50% of the time. According to teacher responses, 34% of comparison teachers as compared to only 23% of treatment teachers would be considered inquiry-based, a priori. Another important pre-treatment teacher indicator is potentially the number of times a teacher has previously taught LE. Results indicate that comparison teachers have, in fact, taught LE more often than treatment teachers prior to this study.

Table 1b

	Total			Comparison Classrooms			Treatment Classrooms		
	Mean	*N*	*SD*	Mean	*N*	*SD*	Mean	*N*	*SD*
Pre-study teacher practices									
Hrs sci instruct	3.66	94	1.11	3.74	47	1.19	3.59	47	1.04
Hrs Lit Instr.	9.71	94	4.68	9.57	47	4.39	9.84	47	5.00
Inquiry-based 0-24%	0.33	94	0.47	0.28	47	0.45	0.38	47	0.49
Inquiry-based 25-49%	0.38	94	0.49	0.38	47	0.49	0.38	47	0.49
Inquiry-based 50-74%	0.23	94	0.43	0.26	47	0.44	0.21	47	0.41
Inquiry-based 75-100%	0.05	94	0.23	0.09	47	0.28	0.02	47	0.15
Inquiry-based teacher	0.29	94	0.45	0.34	47	0.48	0.23	47	0.43
Minutes teaching science/wk	172.4	94	68.7	188.7	47	69.0	156.1	47	65.22
Number times taught LE	3.26	94	3.46	3.69	47	4.16	2.84	47	2.56
Teacher Self Perceptions									
Science Efficacy	43.98	82	7.47	43.98	42	7.28	43.98	40	7.77
Literature Efficacy	50.62	78	6.20	51.74	39	7.03	49.49	39	5.30
During study teacher practices									
Minutes teaching science/wk	201.1	94	89.7	182.3	47	71.1	219.8	47	102.6
Percent of time with									
Hands-on Inquiry:	25.77	94	14.72	26.85	47	18.0	24.70	47	10.50
Read from Books/Txt	21.24	94	12.69	22.55	47	15.6	19.92	47	8.88
Class Discussions	24.78	94	10.21	24.68	47	12.5	24.89	47	7.26
Writing	13.55	94	7.14	11.49	47	6.09	15.61	47	7.57
Science Vocabulary	14.76	94	7.61	14.74	47	9.20	14.78	47	5.71
Teacher Perceptions Related to Unit									
Implementation very Successful	0.11	94	0.31	0.09	47	0.28	0.13	47	0.34
Implementation for ELL	0.17	94	0.38	0.19	47	0.40	0.15	47	0.36
Implement for low achv	0.12	94	0.32	0.11	47	0.31	0.13	47	0.34
Implement for high	0.55	94	0.50	0.49	47	0.51	0.62	47	0.49

achv										
unit	Spent more time on	0.52	47	0.50	0.17	47	0.38	0.87	47	0.33

Teachers also indicated, in Table 1b, how they perceived the implementation of their LE unit (business as usual for controls and *Seeds/Roots* for treatment). According to teachers, only about 11% (9% control and 13% treatment) thought the lesson was implemented very successfully. Consistent with this perception are the perceptions that the unit went "very well" for ELL and low achievers, 17% and 12%, respectively. There appeared to be a difference however, in teacher perceptions in how well the unit went for high achievers –with 49% of control teachers indicating the unit went very well for high achievers as compared to 62% of treatment teachers who indicated that the unit went very well for high achievers (n.s.). Treatment teachers were significantly more likely to indicate that they (87%) increased their time on teaching the unit over previous efforts compared to control teachers (17%).

Teachers were also asked whether they thought students were engaged with the lesson. Seventy-seven percent of treatment teachers compared to 66% of control teachers thought that students were engaged or very engaged in the unit. About 2/3 of the *Seeds/Roots* users thought that it supported state standards well (or very well).

Treatment teachers were asked additional questions that related to the *Seeds/Roots* unit. Overall, the responses were positive towards materials and virtually all of the teachers thought that the *Seeds/Roots* materials provided more literacy support than the standard state LE unit. The final descriptives displayed in tables 1a and 1b summarize post-LE unit self-efficacy in science and literacy. Results indicate that self-efficacy was quite similar in both conditions and very similar to pre-efficacy levels.

In order to examine the impact of the *Seeds/Roots* curriculum on student performance, the dataset used for analysis also contains individual student observations on the measures noted above, including both pre and post treatment results. Table 2 presents the reliabilities of the pre and post treatment science assessments. An assessment's reliability represents score consistency for individual students. However, the reliability of classroom or teacher assessment means provides an indication of how well we can distinguish among classrooms in true student performance. A low reliability for an assessment is generally substantially higher when aggregated to the classroom level. However, low assessment reliability significantly impacts the reliability of gain scores. For example, the reliability of the gain between pre and post vocabulary scores is approximately 0.27. Hence, gain scores

potentially obfuscate the impact of the treatment. The reliabilities displayed in Table 2 are acceptable except for the vocabulary pre-tests, which is moderate, at best.

Two reliabilities are displayed for the Seeds/Roots science content assessment. One reliability for the original 42 item assessment and one for a reduced 23 item score. The original assessment included 42 items, but preliminary 3-parameter item response theory models indicated that several of the items did not perform well. The moderate reliability implies that, potentially, the items assessed more than a single construct. This was in fact the case and additional exploratory analysis by LHS partitioned the Seeds/Roots assessment into its appropriate components, based on the state grade-level standards. Student scores based on the subset of 23 items are more reliable than the original assessment, more closely linked to the state grade level science content standards, and provide for a more accurate comparison between treatment and control classrooms as the outcome scores are more closely related to content that students had the opportunity to learn. Preliminary analyses indicate that results are robust to test specification (whether 42 or 23 items). All models use outcomes based on the 23 item scores.

Table 2

Reliabilities of Science Assessments

	n items	Cronbach's α
Pre- treatment		
Reading	15	.77
Vocabulary	20	.43
Science content	42	.50
Science content	23	.84
Post-treatment		
Reading	15	.76
Vocabulary	20	.69
Science content	42	.75
Science content	23	.81

The means and standard deviations of the three components of the Science assessment are presented in Table 3. Table 3 presents the overall means and standard deviations as well as the comparison and treatment classrooms means and standard deviations separately. The descriptives in Table 3 indicate that pre-test scores across all three domains (science, vocabulary and reading) are quite similar between the treatment and control groups.

Preliminary Multilevel models using pre-tests as outcomes indicated that pre-science did not vary significantly among teachers, and there was no difference in mean pre-science performance between treatment and control classrooms. However, both vocabulary and reading pre-test results indicated significant between-teacher variability in scores, and also significant differences between treatment and control classrooms. Control classroom intake (pre-test scores) in reading was about 0.10 standard deviations higher, and control classroom intake was about 0.30 standard deviations higher in vocabulary. Given pre-tests are related to post results, it is important to account for intake differences when comparing whether the treatment was effective.

Table 3

Descriptive Results for Science Assessment

	Total			Comparison classrooms			Treatment classrooms		
Science assessment	Mean	*N*	*SD*	Mean	*N*	*SD*	Mean	*N*	*SD*
Vocabulary Pre-test	11.5	2019	2.59	11.67	992	2.55	11.33	1027	2.62
Vocabulary Post-test	13.3	1913	3.21	12.89	939	2.79	13.72	974	3.51
Reading Pre-test	9.9	2018	3.38	10.21	992	3.28	9.59	1026	3.46
Reading Post-test	10.5	1905	3.19	10.72	936	3.06	10.3	969	3.29
Science Pre-test (all items)	23.3	2018	4.00	23.63	992	4.00	22.99	1026	3.97
Science Post-test (all items)	27.04	1913	5.54	26.08	937	4.63	27.95	976	6.15
Science Pre-test (23 items)	12.50	1913	2.133	12.59	937	2.149	12.42	976	2.116
Science Post-test (23 items)	14.74	1913	3.126	14.05	937	2.576	15.41	976	3.448

Table 4 presents results related to the writing assessment and the consistency of scores based on raters' scores. The results in Table 4 are based on a generalizeabiilty study that moves beyond simply examining agreement of raters and carefully identify sources of error (Shavelson & Webb, 1991). Ideally, the majority of the variability in raters' scores would be due to variability in true student performance. The writing sample consisted of scores on seven dimensions: introduction; clarity; conclusion; evidence; vocabulary use; vocabulary count; and, science content. The results in Table 4 suggest that the largest sources of error are related to variation in true student performance on the writing task; comprising approximately 43% and 35% of the total variability on observed pre and post writing scores, respectively. The next largest source of variability was due to the student by dimension interaction, 27% and 36% for pre and post writing, respectively. This indicates that students'

performance differed substantially across the seven dimensions scored by the raters. Importantly, however, variability due to raters was virtually zero. The variation attributable to overall rater stringency was less than or equal to about 0.2% while the rater by student and the rater by dimension variability accounted for only about 0.3% to 2.8%, indicating that raters were fairly well calibrated. The standard deviations presented in Table 4 indicate that post-test results are more variable than pre-test results. A 95% confidence interval around the true score would include a range of +/- 0.51 for the pre-test and +/- 0.78 for the post-test. Similar to the reliability coefficient presented for the *Seeds/Roots* science assessment results, we can calculate an index of dependability, ϕ, which indicates the consistency of rater scores. The results in Table 4 are based on two raters[6], but we can estimate ϕ for a single rater (that was used to score a subset of writing results). In either case, results are sufficiently reliable for use in evaluating treatment effects.

Table 4

Writing Score Consistency Across Dimensions

Component	Pre	Post
Student	42.9%	34.6%
rater	0.2%	-0.1%
dimension	7.9%	11.9%
Student * rater	1.8%	2.8%
Student* dimension	29.6%	36.0%
rater * dimension	1.4%	0.3%
Error	16.2%	14.5%
Variance	0.26	0.38
Index of Dependability		
Two raters, ϕ =	0.85	0.79
One rater, ϕ =	0.81	0.75

Overall pre-test writing results indicate that the treatment and control students were virtually identical in performance. The descriptive results in Table 5 indicate that scores on writing varied among domains and that average classroom scores favored control classrooms

[6] A subset of scores (155) were initially scored by two raters. Results based on only the initial sample of four dimensions demonstrated consistent variance partitioning patterns as those presented in Table 4. The smaller number of dimensions reduce ϕ, for that sample to .70 (2 raters) and .63 (one rater).

in most instances. In several instances the differences are statistically significant. Also, average scores improved on all domains in both treatment and control classrooms.

Table 5

Descriptive Results for 7 Writing Domains

	Total			Comparison classroom			Treatment classroom		
	Mean	*N*	*SD*	Mean	*N*	*SD*	Mean	*N*	*SD*
Concepts – pre	1.71	537	0.71	1.77	274	0.68	1.65	263	0.74
Concepts – post	2.36	464	0.93	2.06	248	0.80	2.69	216	0.97
Vocab. use – pre	1.46	536	0.91	1.55	274	0.94	1.36	262	0.86
Vocab. use – post	2.10	463	1.11	2.00	247	1.13	2.20	216	1.08
Vocab. count – pre	2.28	530	1.32	2.39	269	1.35	2.16	261	1.29
Vocab. count – post	3.40	461	1.87	2.70	246	1.63	4.21	215	1.79
Evidence use – pre	1.55	538	0.84	1.63	275	0.85	1.47	263	0.82
Evidence use – post	2.00	475	1.13	1.82	255	1.00	2.20	220	1.23
Introduction – pre	2.04	538	0.81	2.09	275	0.79	1.98	263	0.83
Introduction – post	2.48	475	0.94	2.28	255	0.89	2.72	220	0.94
Conclusion – pre	1.89	538	0.67	1.89	275	0.59	1.88	263	0.74
Conclusion – post	2.00	474	0.60	1.95	254	0.63	2.06	220	0.57
Clarity – pre	1.67	538	0.76	1.68	275	0.74	1.65	263	0.79
Clarity – post	1.98	475	0.77	1.84	255	0.76	2.15	220	0.75

For a subset of students[7] there exist student background characteristics and State assessment information. These descriptive results are presented in Appendix A.

Table 6 summarizes the number of workbooks completed. Additional analyses were conducted with a subset of teachers for whom there existed diary (or student workbook) information.

[7] Demographic and state assessment data are available for approximately half of the original sample (n=1,000). The descriptive results for this subset are presented in appendix a in table A2.

Table 6

Descriptive Results for Teacher Diaries (sessions)

	Minimum	Maximum	Mean	S.D.
Completed	1.07	4.10	3.52	0.83
N 44				

Results

We present in detail below the results for each of the research questions presented above. Overall, the *Seeds/Roots* unit demonstrated statistically significant and substantively important treatment effects in science, vocabulary, and writing, but not in reading[8]. Teacher background was generally not an important factor. Teacher perceptions are generally not systematically related to the impact of the lesson, except in reading and in science for high achievers. Teacher practices are important in science as inquiry-based teachers, when teaching in the treatment classrooms, provide substantial incremental impact to the *Seed/Roots* unit. The impact of student background is somewhat uncertain given the limited sample for analysis. However, it is important to note that the significant treatment effects are robust to model specification. The following addresses each of the research questions in detail.

Student academic outcomes

1.a.i.) Do students who use the *Seeds/Roots* units make progress in science?

The results in Table 7 indicate students in both conditions demonstrated statistically significant gains ($p < .01$).

[8] When subjects are tested on multiple outcomes within a domain, corrections for multiple t-tests are utilized (e.g. Benjamini-Hochberg (BH) correction, which uses only significant results as a basis for correction). Clearly Science and literacy are an integration of two domains, and the omnibus tests for treatment effects for these require no correction. Within these domains, however – e.g. the individual writing constructs – utilize multiple t-tests within a domain. The BH correction places no theoretical order on tests; however, we first conduct an omnibus test on a single latent indicator of writing and determine whether there is a significant treatment effect on writing and then continue with exploratory analyses of the individual constructs.

Table 7

	Gain	S.E.	
Treatment	2.99	0.12	***
Control	1.46	0.10	***

*** p < .01, ** p <.05, * p< .10.

1.a.ii.) Does the *Seeds/Roots* treatment result in higher student performance compared to the comparison, business-as-usual, condition in science content[9]?

The following results address the question of whether or not there were treatment effects. Also, this question is addressed in other sub-sections as related questions concerning student background, teacher background, and teacher processes are examined. These results all demonstrate the robustness of the most parsimonious results presented here. Taking advantage of the available data and data structure we not only evaluate whether, on average, the treatment had a significant impact on student performance, but also whether there are specific conditions under which the treatment effect was either exacerbated or mitigated. In this way we can begin to establish when the treatment might be most beneficial.

The results in Table 8 summarize the two models examining the *Seeds/Roots* science assessment results. Model 1 tests the main effect of the treatment and answers the question whether students in treatment classrooms scored higher on the post-test, accounting for the fact that the treatment was assigned at the classroom level and classrooms were nested within schools. The results indicate that treatment classrooms scored about 1.5 points higher on the science post-test, which is an effect size of about 0.65. Model 2 tests whether there is a joint effect between the pre-test and the treatment, that is whether the relationship between the pre-and the post-test is different in treatment and control classrooms. If this effect is significant it provides evidence that the treatment is more/less effective for high/low achieving students. These results, based on the 23 item scores, are very similar to results obtained from the full 42 item science test – substantively all interpretations would be the same.

The results for model two imply that there is no joint effect (essentially the pre-post slopes are parallel in treatment and control classrooms), hence there is no change in the performance gap between high and low achievers due to the treatment.

[9] The analyses of LHS assessments are based on a student sample size of approximately 1,950 (of the 2,144 in the data set), except where explicitly noted. The sample size varies somewhat +/- 50 students, by content area.

Table 8

Estimated Treatment Effects on Student Post-Test Results

Model[1]	Science content			
	1		2	
Fixed Effects				
Mean Post-test				
Control classroom	14.06		14.06	
Treatment classroom	15.50	***	15.50	***
Treatment effect size[2]	0.65		0.65	
Treatment interaction			0.05	
Treatment effect size[3]				
Random Effects				
Post-Tests				
Student	2.66		2	
Classroom	1.12	***	1.12	***
School	0.97	***	0.97	***

Notes. *** p < .01, ** p <.05, * p< .10. (1) Odd numbered models include only unconditional treatment effects. Even numbered models estimate conditional treatment effects, conditioned on pre-tests and pre-test by treatment joint (2) Effect size estimated as δ, Treatment Control)/s.d.(outcome).

We also examined the original 42 item test as part of the analyses to check the robustness of the results by utilizing different metrics, and different specifications using subsets of the entire sample that have different data elements available for analysis. Table 9 re-examines the effect of the treatment on Science content, but utilizes IRT scores[10]. The IRT results take item difficulty into account, as not all science content items demonstrated the same performance. However, the results in Table 9 indicate that using IRT-based assessment scores doe not appreciably change the results, nor the inferences about the effectiveness of the treatment on science content. The model used to create results takes advantage of the conditional standard errors of measure (SEM) generated by the IRT analysis. The model is similar to (1b) except scores are weighted by their precision and true gains can be modeled. By modeling true gains in student performance, we eliminate the spurious negative correlation between pre-test and gains (and potential regression to the

[10] IRT scores based on 3-paraeter model using al 42 items.

mean effects). This model is based on Bryk, Thum, Easton, and Luppescu (1998) who used a similar approach to examine school effects[11].

Overall, the results are similar to the covariance model results presented above. The effect size differ somewhat, but this is likely due to reduced overall variability due to the weighting of student scores by their estimated precision.

Table 9

Estimated Treatment Effects on Student Post-Test Results (IRT scores)[1]

	Science content	
Fixed Effects		
Mean Post-test		
Control classroom	0.86	
Treatment classroom	1.13	**
Treatment Effect Size[2]	0.23	
Random Effects		
Post-Tests		
Student	1.18	
Classroom	0.46	***
School	0.43	***

*Notes.*** p < .01, ** p <.05, * p< .10. (1) Based on all 42 items; (2) Effect size estimated as δ, (Treatment -Control)/s.d.(treatment).

[11] More detail is available from the author.

1.b.i.) Do students make progress in vocabulary and reading?

The results in Table 10 summarize the student progress in vocabulary and reading.

Regardless of condition students demonstrated gains in both vocabulary and reading. For both vocabulary and reading the treatment students demonstrated gains about twice as large as the control group. The following analyses address whether the differences in gains by the treatment and control students are statistically significant.

Table 10

Student Gains

	Gain	S.E.	
Vocabulary			
Treatment	0.69	0.086	***
Control	0.39	0.079	***
Reading			
Treatment	2.38	0.104	***
Control	1.18	0.090	***

*** p < .01, ** p <.05, * p< .10

1.b.ii.) Does the *Seeds/Roots* treatment result in higher student performance compared to the comparison, business-as-usual, condition in vocabulary and reading?

Table 11 presents results for both vocabulary and reading. Models three and five present results testing only the treatment condition and the control condition, accounting for student intake (i.e. pre-tests). The results indicate that students in the treatment condition score scored significantly higher than students in the control condition. The effect size is approximately 0.23. The results for reading indicate that treatment and control students did equally well on the post-test. The results for models four and six test whether there are joint effects. There are no joint effects for either vocabulary or reading.

Table 11

Estimated Treatment Effects on Student Post-Test Results

	Vocabulary				Reading			
Model[1]:	3		4		5		6	
Fixed Effects								
Mean Post-test								
Control classroom	12.97		12.97		10.69		10.69	
Treatment classroom	13.72	***	13.67	***	10.33		10.35	
Treatment Effect Size[2]	0.23		0.22		-0.11		-0.11	
Treatment Interaction			0.11				-0.06	
Treatment Effect Size[3]								
Random Effects								
Post-Tests								
Student	2.89		2.62		3.01	***	2.32	***
Classroom	0.84	***	0.87		1	***	1.1	***
School	1.15	***	1.15		0.22	**	0.23	**

Notes. *** p < .01, ** p <.05, * p< .10. (1) Odd numbered models include only unconditional treatment effects. Even numbered models estimate conditional treatment effects, conditioned on pre-tests and pre-test by treatment joint effects. (2) Effect size estimated as δ, (Treatment -Control)/s.d.(outcome). (3) Effect size estimated comparing effect at (+/- 1 S.D. mean of pre-test)/s.d.(outcome).

The next outcome this evaluation considers is student writing and the potential impact of the *Seeds/Roots* unit on differences in writing performance between treatment and control classrooms. The following analysis are based on a subset of student who participated in the study (n=550). Table 12 presents the correlations among the seven writing dimensions assessed in each essay. It is important to reiterate that ratings were subject to a generalizeability analysis that determined that there is sufficient precision in scores to use them for additional analyses. The results in Table 12 indicate that the correlations among assessed domains are moderate at best—indicating that, in general, they tap into different aspects of student writing.

Table 12

Correlations Among Writing Dimensions

	Pre-Test					
	Vocab. use	Vocab. count	Evidence	Introduction	Conclusion	Clarity
Science concepts	0.54	0.48	0.68	0.48	0.35	0.31
Vocabulary use	1.00	0.46	0.64	0.42	0.28	0.33
Vocabulary count		1.00	0.31	0.45	0.38	0.31
Evidence			1.00	0.38	0.29	0.36
Introduction				1.00	0.62	0.24
Conclusion					1.00	0.28
Clarity						1.00
	Post-Test					
Science concepts	0.55	0.66	0.63	0.56	0.30	0.52
Vocabulary use	1.00	0.37	0.67	0.37	0.25	0.36
Vocabulary count		1.00	0.33	0.57	0.31	0.46
Evidence			1.00	0.33	0.17	0.39
Introduction				1.00	0.46	0.44
Conclusion					1.00	0.28
Clarity						1.00

There are two possible avenues to proceed: one, to examine the underlying latent writing achievement based on the observed scores on the seven dimensions; and two, to examine student achievement based on each domain separately. Ultimately, in order to determine whether the treatment had a significant effect on student writing, the former is more appropriate as it controls for the intra-person correlation of scores; however, the latter provides more information in that different results for separate domains can provide additional formative information.

In order to test the global research hypothesis as to whether the *Seeds/Roots* unit results in statistically significant and substantively higher outcomes than the control, the former model is tested. The results are presented in Table 13. The results indicate that at the pre-test, there was suggestive evidence ($p < .10$) that control students had higher writing achievement. The results in Table 13 also indicate that at the post-test students in the treatment group had higher latent writing achievement ($p < .05$). The treatment effect size is 0.40.

Table 13

Estimated Treatment on Latent Student Writing Results

	Writing	
Fixed Effects		
Mean Pre-test		
Control classroom	1.80	*
Treatment classroom	1.70	
Mean Post-test		
Control classroom	2.02	
Treatment classroom	2.36	***
Treatment Effect Size[1]	0.40	
Random Effects		
Heterogeneous random effects		

Note. *** p < .01, ** p <.05, * p< .10. (2) Effect size estimated as δ, (Treatment -Control)/s.d.(outcome).

Based on the latter approach discussed above, Table 14 presents results from examining each of the seven dimensions independently. Overall, the results in Table 14 corroborate the results presented in Table 15. Among the seven writing dimensions, only vocabulary use and conclusion demonstrated no treatment effect. The remaining five dimensions demonstrated statistically significant treatment effects with effect sizes ranging from 0.33 (evidence) to 0.80 (vocabulary count). The models in Table 14 also examined whether there were any effects on writing associated with science content knowledge; under the research hypothesis that content knowledge has a positive effect on writing scores. The results indicate that in some instances it was the pre-post gain in science content knowledge that was related to better writing scores, and some instances it was overall science content knowledge that was associated with higher writing scores. Both vocabulary count and clarity were associated with gains in science content knowledge. The effect sizes were quite large, 0.85 and 0.77 for vocabulary count and clarity, respectively. Both the evidence and conclusion dimensions were impacted by overall science content knowledge, as represented by post-test scores (but not pre-tests or gains). The effect sizes were moderately large, 0.66 and 0.44, for evidence and conclusion, respectively. These results are consistent with the supposition that content knowledge is positively associated with writing performance. It is interesting to note that the effects of science content knowledge were independent effects of the treatment; and in one

case (conclusion) occurred without a significant treatment effect. It is important to note that the subset of students for whom we have both writing, pre and post test science content scores (n= 458) scored similarly on the *Seeds/Roots* science pre- and post-tests to the entire sample (23.3 vs. 23.3 on the pre-test and 27.4 vs 27.04 on the post test, for the writing sample students and the entire sample, respectively). Hence, these results are not attributable to performance of students who were exceptional on science performance.

Table 14

Estimated Treatment on Student Writing by Dimension

		Writing		
				Effect Size[1]
Fixed Effects				
Science concepts				
	Control classroom	2.10	***	
	Treatment classroom	2.69	***	0.63
Vocabulary use				
	Control classroom	2.01	***	
	Treatment classroom	2.23		
Vocabulary count				
	Control classroom	2.74	***	
	Treatment classroom	4.23	***	0.80
Pre-Post Science GAIN		0.08	***	0.85
Evidence				
	Control classroom	1.84	***	
	Treatment classroom	2.22	**	0.33
Post Science score		0.03	***	0.66
Introduction				
	Control classroom	2.36	***	
	Treatment classroom	2.71	***	0.38
Conclusion				
	Control classroom	1.97	***	
	Treatment classroom	2.05		
Post Science score		0.01		0.41
Clarity				
	Control classroom	1.81	***	
	Treatment classroom	2.14	***	0.43
Pre-Post Science GAIN		0.02	***	0.77

Notes. *** p < .01, ** p <.05, * p< .10. (1) Treatment effect sizes as in Table 10, note (2); GAIN and score effect sizes as in Table 11, note (3).

To further examine the concept of integrating science and literacy we evaluate whether a) overall preparedness in the three assessed domains (science, vocabulary and reading[12]) and b) whether there is any transfer between reading and science. Hence, the following analyses examine the effect of including all pre-test scores and all gain scores. The pre-test scores capture a broader picture of student intake, while gains (focusing on science and reading) capture the extent to which students can transfer skills and knowledge from one domain to another. Table 15 presents results of a MLM that examine the impact of student intake measured by science, vocabulary and reading. Model 1 includes only the pretest measures and an indicator for the treatment effect. The results indicate that, consistent with expectations, there are positive relationships between the three intake measures and student science post-tests. Importantly, however, the effect of the treatment is consistent with that reported above. Model two includes treatment by pre-test interactions. Consistent with results presented in Table 9, there is no science pre-test by treatment joint effect. However, there is a reading by treatment joint effect. This implies that students, with better pre-treatment reading achievement benefited more from the treatment than students with lower pre-treatment reading achievement. Table 16 examines the potential effect of student intake as measured by the

Table 15

Science Post-Test Outcome

Fixed effect	Estimate 1		Estimate 2	
Treatment main effect	1.67	***	1.65	***
Science pre-test effect	0.09	***	0.09	***
Treatment effect				
Vocabulary pre-test effect	0.16	***	0.16	***
Treatment effect			0.09	
Reading pre-test effect	0.25	***	0.24	***
Treatment effect			0.14	***

Note. *** p < .01, ** p <.05, * p< .10.

[12] Writing results are excluded as only a subset of students has all four sets of scores.

three pre-tests for the reading outcome. Unlike the reading effect for science, there is no science effect for reading (Model 1). The main effect for vocabulary is significant. Model two indicates that there is no science main effect, nor is there a joint effect. This implies students with better pre-treatment knowledge in science do not demonstrate better performance on the post-treatment reading assessment—this was equally true in the treatment and control classrooms.

While Tables 15 and 16 provide some insight as to how reading and science performance might be related to pre-treatment achievement, Tables 17 and 18 provide additional results pertaining to transfer and the potential for integration. Table 18 presents results for the post-treatment science outcomes and Table 18 provides results for the post-treatment reading assessment.

Table 16

Reading Post-Test Outcome

Fixed effect	Estimate 1		Estimate 2	
Treatment main effect	0.07		0.07	
Science pre-test effect	0.04		0.01	
Treatment effect			0.06	
Vocabulary pre-test effect	0.17	***	0.13	***
Treatment effect			0.07	
Reading pre-test effect	0.59	***	0.62	***
Treatment effect			-0.06	

*** p < .01, ** p <.05, * p< .10.

Model 1, in Table 17 tests the main effects of gains in vocabulary and reading performance, accounting for pre-treatment science achievement. The results indicate students gains in reading achievement are related to higher science post-test scores. The main treatment effect (p< .05) is consistent with previous estimates. The results in model two are consistent with those in model one and also indicate that the joint reading gain by treatment effect is significant (p < .05) and substantively important. It implies that for control students, every five points gained in reading is related to an additional point on the science content outcome. For treatment students, the results imply that three points gained in

reading achievement is related to an additional point on the science content outcome. The average reading gain in the treatment group was about 2.2 points.

Table 17

Science Post-Test Outcome

Fixed effects	Estimate 1		Estimate 2	
Treatment main effect	1.13	***	1.13	***
Science pre-test effect	0.18	***	0.18	***
Treatment effect				
Vocabulary gain effect	0.01		0.07	***
Treatment effect			-0.11	***
Reading gain effect	0.28	***	0.20	***
Treatment effect			0.14	***

*** p < .01, ** p <.05, * p< .10.

Table 18 presents the results for the same analyses using the reading outcome. The results

Table 18

Reading Post-Test Outcome

Fixed effects	Estimate 1		Estimate 2	
Treatment main effect	-0.01		-0.01	
Science gain effect	0.00		0.00	
Treatment effect			0.00	
Vocabulary gain effect	1.00	***	1.00	****
Treatment effect			0.00	
Reading pre-test effect	0.99	***	0.99	****
Treatment effect			0.00	

*** p < .01, ** p <.05, * p< .10.

indicate that science gains have no relationship with post reading outcomes. Hence, in this context, the reading gains transfer to improved science performance, but science gains do not relate to improved reading (although it should be reiterated that all students demonstrated statistically significant reading gains, across both the treatment and control condition).

1c.i.) Are there differences in learning outcomes by gender, ethnicity, or previous educational achievement; and,

1c.ii.) What learning gains are being made with students who have particular educational needs (such as English Language Learners)?

Both of these research questions are substantively important. To the extent that the *Seeds/Roots* unit can close existing achievement gaps, the intervention would be effective not only as a main effect for students who are in classrooms using these materials, but also a mechanism through which at-risk and lower achieving students might close exiting achievement gaps with higher achieving classmates.

As noted in the data description section, student background and state assessment information is available for only a subset of students that participated in the study; and given the lack of representatives of this subsample, we present these results in Appendix B (tables B1 through B6).

Other student outcomes

2) What are the effects of using *Seeds/Roots* units on students' engagement and interest in science and literacy?

Based on survey results, treatment teachers perceived students to be more engaged than teachers in the control group. Teachers indicated that 38% of the treatment as opposed to 11% of the control students were "very engaged" in the LE unit ($p < .01$). Open ended teacher responses also indicated that student "felt like scientists" in the *Seeds/Roots* classrooms, and that students enjoyed investigating and keeping track of data. On the other hand, some teachers indicated that outside of hands-on activities, the unit was sometimes repetitive and lengthy, thus sometimes, losing students' attention.

Exploratory MLM models revealed that student engagement (as perceived by the teachers) was not predictive of student performance on post-tests. To clarify, these models tested whether average classroom engagement had an either a direct impact on student performance, or whether it mediated the treatment effect.

Teacher outcomes

3a) How do the *Seeds/Roots* materials (the treatment) influence teachers' attitudes toward science and literacy teaching?

Teachers in both conditions were given a self-efficacy survey designed to assess each teacher's perceived self-efficacy in teaching science and literacy. The survey was administered prior to the LE unit and after the LE unit. Overall, at the time of the pre-unit assessment teachers rated their self-efficacy moderately high (44/60 in science and 51/65 in literacy). There was no significant difference between treatment and control teachers in either self-efficacy rating before the LE unit. Teachers demonstrated a significant increase in science self-efficacy ($p < .01$), but no change in literacy self efficacy over the treatment period. However, as the results in Table 19 indicate, the difference in pre-post changes between treatment and control teachers was not

Table 19

Change in Teacher Self-Efficacy

Content	Group	*N*	Mean Change	*SD*	*S.E.*	Difference	$Diff_{S.E.}$
Science	Control	40	2.33	5.32	0.84		
	Treatment	42	3.40	4.70	0.73	-1.1	1.10
Literacy	Control	39	0.51	4.27	0.68		
	Treatment	39	1.13	4.41	0.71	0.62	0.98

statistically significant. There was not a significant increase in reported teacher self-efficacy in literacy and there was also no difference in literacy self-efficacy between the treatment and control teachers.

Addressed below is the impact of teacher self-efficacy on student science content, vocabulary and reading outcomes.

3b) Does teacher education, training, experience, experience with inquiry science, and self-efficacy impact student outcomes and do they moderate/mediate treatment effects?

Using teacher survey responses and linking these to the student outcomes, the evaluation next examines the potential effects of teacher background and process on science, vocabulary, and reading outcomes. The sample for the following analyses is based on 90 teachers due to missing data. However, student performance is consistent with the full sample; hence, likely representative of the entire sample under study. It is important to note

that preliminary analyses considered several specifications and tested teacher and classroom variables; including:

Background
- Credential type
- Number of credentials
- Certification level
- Years of teaching experience
- Salary
- Number of certifications
- Number of times taught LE
- Degree earned
- Self-efficacy (appropriate for outcome – science or literacy)

Teacher practices;
- Percent of time spent on hands on experiences
- Percent of time spent on reading
- Percent of time spent on writing
- Percent of time spent on class discussions
- Percent of time spent on vocabulary
- Hours of science instruction
- Hours of literacy instruction
- Minutes taught science previously
- Minutes of science instruction this unit
- Responsible for Science and literacy

Classroom composition:
- Class size
- Pct ELL

Teacher perceptions
- Students engagement
- Implementation success
- Implementation for high achievers
- Implementation for low achievers
- Implementation for ELLs

Interaction with *Seeds/Roots* materials
- Inquiry-based teachers
- Percent of time spent on hands on experiences
- Percent of time spent on reading
- Percent of time spent on writing
- Minutes teaching science
- Teaching experience

Additional joint effects

Inquiry-based teachers and Percent of time spent on hands on experiences (for current LE unit)
Inquiry-based teacher classification and
minutes of science instruction

It is important to note that among the various specifications, the (main) treatment effects remained consistent with the original models reported above for all three outcome measures. Table 20 summarizes the reduced form models that best capture teacher background and process effects on student outcomes. Overall, the results are consistent with previous research that fails to consistently link specific teacher characteristics with student performance. Among teacher background variables, only two remained in a parsimonious specification. The results indicate that, in general, teacher experience has no impact on any of the three outcomes (in either condition). Teacher certification bears some relationship to student outcomes in that teachers not majoring in Early Childhood Education tend to have higher student performance.

Table 20

Effect of Teacher Self-Efficacy

Fixed Effects	Science estimate		Vocabulary estimate		Reading estimate	
Control classroom	13.93		12.85		10.41	
Mean class performance	0.56	***	0.55	***	0.17	***
Treatment effect (+/-)	1.71	***	0.77	***	0.08	
Teacher experience	-0.14	***	-0.09	***	-0.02	
Class size	-0.10		0.12	**	-0.03	
Teacher certification not ECE	0.88	*	0.40		0.03	
Inquiry-based teacher	0.29		0.16		-0.15	
Percent of time hands on	-0.01		0.01	*	0.00	
Self efficacy[2]	0.06	***	-0.03	*	0.00	
Pre-test performance	0.17	***	0.54	***	0.64	***

Notes: (1) *** p < .01, ** p <.05, * p< .10 (2) model with self-efficacy based on 80 teachers.

The primary interest in Table 20 was the impact of teacher self-efficacy (as measured before the LE unit). Teacher self-efficacy[13] is positively related to science performance ($p < .01$). There is suggestive evidence that it is negatively related to vocabulary and is not related to reading.

Implementation

4a) To what extent and how are the units implemented?

An important aspect that helps specify potential treatment effects is the fidelity with which the treatment was implemented. In this case we use student workbooks and teacher diaries to proxy implementation. The proxy does not account for quality, but does provide a measure of quantity, as the workbooks and diaries provide information regarding the session

[13] Science self-efficacy is used for the science outcome and literacy self efficacy is used for reading and vocabulary.

that was completed. One aspect we intonated above was the potential relationship between classrooms with higher pre-test scores and teachers' ability to teach at a quicker pace due to higher baseline knowledge. We test this proposition by correlating classroom average pre-test results and teacher sessions completed. The correlation r=.325 is substantively moderate to low, indicating that teachers tended not to take advantage of pre-existing content knowledge. Table 21 presents results for students in treatment classrooms[14]. The results indicate that students who had workbook/diary information score slightly higher than all treatment students. The average impact of completing sessions was significantly and substantively positive – indicating that the more sessions that were completed the higher students would score on the science post-test. The effect size estimate is approximately 0.60. It should be noted, of course, that any unobserved teacher or student characteristics not captured by the pre-test, associated with session completion biases session completion estimates upward.

Table 21

Estimated Effect of Sessions on Student Post-Test Results

Fixed effects	Science content			
	1		2	
Mean post-test				
At mean session	15.41		15.4	
Sessions completed	1.26	***	1.28	***
Inquiry-based teacher			0.97	
Treatment Effect Size[1]	0.37			
Random effects				
Post-tests				
Student	2.82		2.82	
Classroom	0.8	***	0.82	***
School	1.38	***	1.26	***

Notes. *** p < .01, ** p <.05, * p< .10. (1) Effect size estimated as δ, (Treatment - Control)/s.d.(treatment).

Focusing on implementation, we re-specify the model generating the results in Table 20 and eliminate teacher self-efficacy which does not impact results presented in Table 22, but does allow for the inclusion of an additional 10 teachers. The results in Table 22 provide

[14] Only treatment classrooms had session completed information as this related to the treatment.

some insight into implementation effects. The model generating the results in Table 22 also tested all of the teacher variables noted above. Again, the treatment effect is consistent with previous results. The key finding in Table 22 is that inquiry-based teaching methods work jointly with the treatment to generate an effect. That is, students in control classrooms do not benefit from teachers' (pre-existing) inquiry-based instructional methods; while students in treatment classrooms, who also have inquiry-based teachers, gain about twice as much from the treatment as students who are in treatment classrooms and do not have inquiry-based teachers. This effect was not preset for the vocabulary or reading outcomes (Tables 23 and 24).

Table 22

Effect of Teacher Characteristics and Processes on Science Outcome

Fixed effect	Estimate	S.E.	
Mean control classroom	14.23	0.11	
Effect of treatment (+/-)	1.12	0.30	***
Effect size	0.48		
Class size	-0.06	0.03	*
Teacher certification not ECE	0.99	0.24	***
Minutes of science instruction	0.00	0.00	*
Inquiry-based teacher/control	0.24	0.23	
Inquiry-based teacher/treatment	1.32	0.59	**
Science Pre-test	0.18	0.04	***

*** p < .01, ** p <.05, * p< .10.

In general the teacher background and process variables listed above had no impact on Vocabulary outcomes (table 23). Consistent with Science results, students, whose teachers did not have an ECE certification performed better than students whose teacher did have an ECE certification (p <.05); this result was consistent across treatment conditions.

Table 23

Effect of Teacher Characteristics and Processes on Vocabulary outcome

Fixed effect	Estimate	S.E.	
Mean control Classroom	12.83	0.15	
Effect of treatment (+/-)	1.03	0.24	***
Class size	-0.06	0.04	
Teacher Certification not ECE	0.58	0.26	**
Vocabulary pre-test	0.53	0.03	***

*** p < .01, ** p <.05, * p< .10.

As noted above, the analyses focusing on teachers also examined teacher perceptions related to how well teachers thought the LE unit was implemented, this included whether the unit went well for various subgroups and how successfully the unit was implemented overall. There was no relationship between teacher perceptions about implementation on science and vocabulary outcomes, but there was a positive relationship between teacher perceptions about unit implementation success and the reading outcome ($p < .01$). This result is displayed in table 24. Students, whose teachers thought the lesson was implemented very successfully scored about 0.62 points greater than students whose teachers did not hold such a belief ($p > .01$).

Table 24

Effect of Teacher Characteristics and Processes on Reading outcome

Fixed effect	Estimate	S.E.	
Mean control classroom	10.48	0.09	
Effect of treatment (+/-)	0.04	0.13	
Lesson implemented very successfully	0.62	0.18	***
Reading pre-test	0.66	0.02	***

*** p < .01, ** p <.05, * p< .10.

4b) What distinguishes successful from less successful use of these materials?

Except for inquiry-based teacher experience, there is little objective information identifying "successful" implementation. Although teachers were given opportunities to provide some insight, none of these responses are systematically related to outcomes. Although teachers felt fairly comfortable with the *Seeds/Roots* unit, this did not translate into changes in self-efficacy, nor to improved student performance. Teachers thought the unit was more successful for high achievers and less successful for low achievers, but these perceptions were unrelated to student outcomes, but were borne out to some degree by other MLM analyses that indicated that the most prepared students did better and that this was not mediated by treatment.

4c) What are teachers′ reactions to the quality, usability and utility of the units?

Overall teachers liked the *Seeds/Roots* unit and thought it met state standards fairly well. As noted they were comfortable (60% were comfortable or very comfortable) using the unit. They indicated that they spent more time than in the past (87%) on the LE unit and some indicated that this left them with little time to teach the other units. Again teachers liked the materials, but some wanted to "pick and choose." Several teachers indicated that they did not know that science and literacy could be integrated so well and thought that the *Seeds/Roots* units had several good ideas. A consistent theme was that while the inquiry elements were engaging, other elements were "repetitive and long." Only four of the 47 respondents indicated that they would not use the *Seed/Roots* unit again in the following year. A few of the teachers indicated they would use it again, but at a slower pace or with modification.

Qualitative teacher interview results are based on three respondents. The response rate was very low, but these responses are, it is important to note, consistent with open-ended short answers on the end-of unit survey that teachers completed.

The teachers interviewed expressed positive reviews of the unit. For example, to the prompt "*Tell me about your experience teaching the unit."* Teachers found the units well organized. For instance, one LE teacher offered that the "*Books were well designed (Graphics, Charts, Topics,)… fit nicely with the standards. Liked the journals"* Another offered that that the unit was, "*Well-thought out, solid unit and enjoyable*," however that the unit was "*long and hard to get through everything. Although was also, teacher friendly.*"

All three participants suggested that their students really liked the hands-on aspects and working in pairs. For example, "*Students really enjoyed the hands-on and experiments. The students loved the small readers and that provided them with a good understanding of the Light unit. They liked the focus on the scientific process (Hypothesis, Prediction, Observe, Collect data, error analysis). Could Social Studies and Reading be combined?*" Another participant offered that, "*In particular the hands-on and paired work. They like demonstrations. Students liked to take control of their learning. Also they liked to read and work with partners and in groups.*"

The length of the unit seemed to challenge the teachers implementing the unit. In addition, one participant offered that the scripting was a challenge. For example, "*The hands on worked well and the Convex and Concave lessons too. There was too much explanation in many cases. Too much reading in many cases and the materials were leveled great for the high achievers. The short stories were interesting. The Scripts were challenging. Too much repetition and either compacted and over blown planning that didn't always work.*"

For the most part, the participants seemed to find the materials fairly easy to use. Yet the length was too long. For instance, one teacher offered that "*The materials were easy to use, surprised with the number of books that come with the unit.*" Or that, "*Some materials were too high tech and weren't visible to all students in the lab. Couldn't always spend 8-12 weeks needs to complete when only allotted 6 weeks by the curriculum.*" In addition, one participant found the materials hard to implement with a blind student, and a student with Down syndrome. Furthermore, students with behavioral problems can act up in the groups.

Although, one teacher did not have a response to the question "*How did your use of the unit influence your thinking about teaching and science instruction?*" one respondent suggested that they had always practiced interdisciplinary lessons, however that the unit helped to inspire creativity. The other respondent offered that "*It is great as we're beginning to have to teach both content areas. The journaling was great and really helped the students with writing.*"

One respondent suggested that the unit impacted their science instruction. For example, to the prompt "*As a result of using the unit has your science instruction changed?*" offered, "*Definitely changed towards hands-on: eager to re-use the materials and conduct lots of experiments.*" While another felt the unit engendered a greater appreciation of science, finally, one respondent seemed to be more encouraged with respect to interdisciplinary instruction offered that they s/he "*Always has integrated. The unit was helpful and creative in the combined content areas.*" Furthermore that, interdisciplinary

lessons *"should be implemented in all areas*." Two respondents were happy with the materials sent, would possibly like interactive technology. For example one offered the following list "*Tech support. Links to reviews, CD-Roms, Game Type activities, Stream science lessons through computer. Interactive technologies.*" In addition, a respondent offered that the unit "*Should have included copied students sets. Took too much school printing, paper ink, and time. Would like more materials*" furthermore that, " *Everything sent was fine. Not sure if online materials would help or hinder. Maybe the journal should be shorter. Teachers are limited to a specific number of copies per month. Overhead may be suboptimal for the class. Should be differentiated better*."

To the prompt, "*Having taught this unit, what do you think about integrating science and literacy?*" All LE respondents suggested that the content areas work well together and expressed interest in learning of other interdisciplinary lessons. Finally, a respondent offered that their "*School has began to departmentalize and this provides a well thought out alternative to teach things across the curriculum in interdisciplinary framework.*"

Conclusion

This evaluation of the *Seeds/Roots* unit was multi-faceted and examined posited effects in two general areas: one, on student outcomes; and two, on teacher outcomes. Student outcomes consisted of outcomes in two domains: Science and literacy. Science content served as the primary outcome in the Science domain, while writing served as the primary outcome in the literacy domain. Student outcomes also include engagement with the lesson (although measured by teachers). Teacher outcomes included self efficacy and perceptions about students and the *Seeds/Roots* unit. Teachers also represented an important input that potentially either moderated or mediated the effect of the treatment. Teacher background, practices, perceptions, and self-efficacy were all examined. Teacher processes were also examined In order to refine when/how the treatment might be effective. Another dimension under study was the hypothesized benefit of an integrated approach to teaching science and literacy – implying that skill transfer was likely between content domains.

Overall, students in classrooms using the *Seeds/Roots* unit demonstrated statistically significant and substantively higher performance than students in control classrooms. That is, teacher classrooms, randomly assigned to the treatment or control conditions demonstrated significantly different performance on post test results on three of the four assessments. These results were robust to model specification, including, for example, whether or not student preparedness (pre-test) scores were included or not. Results were also robust to specification changes that included teacher background and other teacher

variables[15]. Specifically, the results indicate that there was a significant positive treatment effect on science content and vocabulary, and no effect on reading. There was also a significant treatment effect on writing for the subsample for which there were writing results. It should be noted that students demonstrated significant pre - post gains on all domains over the unit period. In the case of reading, all students gained equally – irrespective of the treatment condition. The results also imply that the *Seed/Roots* unit is equally beneficial for low and high achievers since the treatment shifts science and vocabulary performance up equally among parallel pre-post slopes.

It is important to consider context, and this reveals that treatment teachers spent less time on reading (although not significantly so) and more time on writing (significantly so). Despite somewhat decreased attention to reading, the control and treatment groups performed about equally well on the reading post-test. Treatment teachers spent significantly more time on writing and writing results did vary by treatment condition. Given that the unit was longer and 87% of teacher said they took longer than normal, it might be the case that the treatment does not provide any substantive benefits other than increasing time on task. However, exploratory analyses include the minutes taught by treatment interaction and found an insignificant interaction term, indicating that results are unlikely to be solely due to differences in time on task.

Although the treatment had a significant impact on the *Seeds/Roots* science and vocabulary assessment, generalizability would be broadened if the results were consistent using state assessment results. Preliminary analyses with data received by June 8 2009, are inconclusive as this sample of students demonstrated effects inconsistent from the broader sample under study. Suggestive results indicate that the *Seeds/Roots* science assessment post-test was related to the state science assessment, suggesting that processes impacting the *Seeds/Roots* science assessments plausibly impact state assessments in a similar fashion.

There was no evidence that the *Seeds/Roots* unit was either exceptionally beneficial for low achievers or students at-risk (e.g. low SES, SWD, or ELL). The evidence for low achievers comes from both empirical results indicating that low achievers did not close gaps, and from teacher perceptions that the lesson was not as successful for low achieving students as it was for high achieving students. Results for at-risk students are based on the limited

[15] Given that teachers were randomly assigned, some schools had both treatment and control teachers—which improves power, but may lead to diffusion of results. Preliminary analyses, limited the sample to schools that had both treatment and control teachers, found no differences in results—indicating that the treatment effects are consistent with those based on the full sample.

subsample discussed above and should be viewed with the significant caveat that there were no treatment effects for this subsample.

Still the results imply that the treatment is most effective for prepared students. This is evidence by the unaltered pre-post relationship in treatment classrooms compared to control classrooms and by teachers' perception that the lesson tended to be successful for high achievers.

The supposition that an integrated balanced approach to science and literacy instruction results in increased performance in both domains was tested by examining the impact of gains in one domain as predictors of another. Focusing on reading and science revealed that there is some transfer, but that it is unidirectional. That is, students demonstrating gains in reading during the unit demonstrated higher post science performance. However, student gains in science had no impact on post-test reading performance. However, some elements of science content knowledge (and gains) did positively influence several dimensions of student writing. For example, greater science content knowledge was related to better conclusions (conclusion domain) on the writing assessment. This is consistent with the notion that students with deeper content knowledge will be better able to summarize and express their thoughts and ideas.

Given that teachers are the key mechanism through which these units are delivered, several hypotheses were related to teachers – both as outcomes and as potential moderating or mediating factors. Consistent with previous research on teacher effects, few teacher background characteristics played a major role in determining student outcomes. Moreover, teacher perceptions tended to by unrelated to actual class performance. On exception is how successfully teachers perceived the unit to have been implemented and students' post-test reading performance. Subjectively, it may be easier for teachers to evaluate success on student reading on an ad-hoc basis than other content areas.

Teacher self–efficacy matters in student science content performance. However, the *Seeds/Roots* unit did not promote changes in teacher self-efficacy.

Another important process is the extent to which the LE unit was inquiry based (hands on) and the extent to which teachers were experienced with inquiry-based instruction. Several hypotheses were tested in relation to both of these similar, yet different notions. Teachers with inquiry-based instruction experience tended to outperform teachers without such experience (although the measure was not precise, effects were still observed). There were no effects related to the amount of inquiry-based instruction and exploration that was occurring in the LE unit. However, these findings raised the questions of whether there were

differential effects of these two elements in the treatment and control groups, and whether inquiry-based teachers using more hands on instruction during the LE unit were more effective. The results of these examinations indicate that teachers that had inquiry-based instructional experience only had positive effects on student outcomes when teaching in treatment classrooms. That is, the treatment was statistically effective whether or not an inquiry-based teacher was teaching, but was substantially more effective when taught by an inquiry-based teacher. Inquiry-based teachers had no impact in control classrooms. Other teacher qualifications or background did not impact this relationship. Also, the effect of using more hands-on instructional strategies was not impacted by a teacher's inquiry-based experience.

Overall, the *Seeds/Roots* unit would be considered an effective intervention, but additional data are needed to more carefully examine effects on student subgroups. Teachers are generally very happy with the Seeds/Roots materials and an over-whelming majority would use unit again. However, empirically, their perceptions of the usability and quality do not systematically relate to their students' performance. Moreover, despite substantively positive responses to the unit, teacher self efficacy in science and literature did not significantly improve due to using the unit. The only drawback, as seen by teachers using the materials was the unit's length and time commitments, which in some instances had effects on student engagement and potentially on other units teachers need to teach. This drawback was clearly viewed with respect to teaching science and not in relation to the integrated nature of the curriculum which is designed to supplant and note merely supplement literacy instruction. Additional research is warranted in identifying ways in which teachers can be assisted in utilizing the integrated approach in conjunction with other literacy curriculum and where specifically the standard literacy could be supplanted by the *Seeds/Roots* curriculum. The science curriculum, itself, was effective, well received, and was generally effective irrespective of teacher background and experience[16] – implying strong scalability potential.

[16] The one major exception being previous hands on experience, which significantly enhanced treatment effects – but was not a necessary condition to achieve statistically significant treatment effects.

References

Aitkin, M., & Longford, N. (1986). Statistical modeling issues in school effectiveness studies. *Journal of the Royal Statistical Society, 149*(1), 1-43.

Bryk, A. S., Thum, Y. M., Easton, J. Q., & Luppescu, S. (1998). Assessing School Academic Productivity: the Case of Chicago School Reform, *Social Psychology of Education*, *2*, 103-142.

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. *Review of Research in Education, 8*, Washington, DC: American Educational Research Association, 158--233

Cervetti, G. N., Barber, J., Dorph, R., Pearson, P. D., & Goldschmidt, P. (2009). *Integrating science and literacy: A value proposition?* Paper presentation, Annual meeting of the American Educational Research Association, San Diego, CA.

Cervetti, G., Pearson, P.D., Bravo, M.A., & Barber, J. (2006). Reading and writing in the service of inquiry-based science. In R. Douglas, M. Klentschy, and K. Worth (Eds.), *Linking science and literacy in the K-8 classroom*. Arlington, Virginia, NSTA.

Guthrie, J.T., & Ozgungor, S. (2002). Instructional contexts for reading engagement. In C.C. Block & M. Pressley (Eds.), *Comprehension instruction: Research-based best practices.* New York: Guilford Press.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models – Applications and Data Analysis*, 2nd ed. Thousand Oaks, CA: Sage.

Romance, N. R., & Vitale, M. R. (1992). A curriculum strategy that expands time for in-depth elementary science instruction by using science-based reading strategies: Effects of a year-long study in grade four. *Journal of Research in Science Teaching*, *29*(6), 545–554.

Seltzer, M. (2004). The use of hierarchical models in analyzing data from experiments and quasi-experiments conducted in field settings. In D. Kaplan (Ed.), *The Handbook of Quantitative Methods for the Social Sciences* (pp. 259-280). Thousand Oaks, CA: Sage Publications.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory: A primer*. Newbury Park: Sage.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage Publications.

Stoddart, T., Pinal, A., Latzke, M., & Canaday, D. (2002). Integratins inquiry science and language development for English language learners. *Journal of Research in Science Teaching*, *39*(8), 664-687.

Wang J., & J. Herman (2006). *Evaluation of Seeds of Science/Roots of reading Project: Shoreline Science and Terrarium Investigations*, CSE Technical Report 676, National Center for Research on Evaluation, Standards, and Student Testing, Center for the Study of Evaluation (CSE) Graduate School of Education & Information Studies, University of California, Los Angeles.

Appendix A

The subsample represented in Appendix A is those students for whom we have additional individual student information. These students form the basis for analyses that examine moderating factors potentially impacting the treatment effect. Given this subset representative of the original full sample (both treatment and control), we could with reasonable certainty maker inferences related to the efficacy of the treatment on state assessment results and for students with specific demographic characteristics. However, given this subset of students does not represent a random sample, we first must compare students on observable characteristics. Based on this, and the fact that, unlike for the entire sample, the treatment effect was not significant, we conclude that this subset of students is not representative of the original sample. However, for exploratory purposes we present descriptive information and model results. We first present the descriptive results. The students represented in Table A1 are predominantly White and predominantly native English speakers. The treatment sample consists of 15% ELL students, while student in the control group consist of only 5% ELL. Approximately 50% of the sample is classified as low SES. The pre and post *Seeds/Roots* science assessment results are consistent with those of the larger sample (Table 3). These students also have data on state assessments in science including Criterion Referenced Competency Tests (CRCT). The CRCT can also be used an outcome to evaluate the impact of the treatment.

Table A1

Student Characteristics: Proportion of Students with Various Background and Classifications

	Total			Comparison classroom			Treatment classroom		
Characteristics	Mean	*N*	*SD*	Mean	*N*	*SD*	Mean	*N*	*SD*
Girl	0.50	1006	0.50	0.49	475	0.50	0.51	531	0.50
Asian	0.04	970	0.20	0.05	458	0.21	0.04	512	0.18
White	0.44	970	0.50	0.42	458	0.49	0.46	512	0.50
African American	0.39	970	0.49	0.41	458	0.49	0.38	512	0.49
Hispanic	0.10	970	0.29	0.10	458	0.30	0.09	512	0.29
other	0.03	970	0.16	0.02	458	0.15	0.03	512	0.17
low SES	0.50	703	0.50	0.50	353	0.50	0.51	350	0.50
Student w/ disabilities	0.09	742	0.29	0.11	369	0.31	0.08	373	0.27
GATE	0.14	985	0.34	0.16	475	0.37	0.11	510	0.32
ELL	0.11	530	0.31	0.05	214	0.21	0.15	316	0.36

Table A2

Assessment Results for Students in Table A1

	Total			Comparison classroom			Treatment classroom		
Assessment	Mean	*N*	*SD*	Mean	*N*	*SD*	Mean	*N*	*SD*
Science Pre Test[1]	12.56	1044	2.12	12.63	487	2.15	12.51	557	2.10
Science Post Test	14.88	1044	3.24	14.08	487	2.69	15.57	557	3.52
Vocabulary Pre Test	11.54	1072	2.57	11.70	507	2.57	11.40	565	2.55
Vocabulary post Test	13.37	1043	3.23	12.85	487	2.83	13.82	556	3.49
Reading Pre test	9.90	1072	3.39	10.10	506	3.34	9.73	566	3.43
Reading Post test	10.58	1037	3.11	10.75	484	3.07	10.43	553	3.15
CRCT Reading 06-07	828.43	402	32.15	833.19	222	31.71	822.56	180	31.81
CRCT Reading 07-08	824.74	459	27.72	826.46	248	27.61	822.71	211	27.76
CRCT ELA 06-07	822.25	421	26.35	823.06	222	25.68	821.35	199	27.11
CRCT ELA 07-08	821.12	480	29.42	822.64	248	27.86	819.49	232	30.98
CRCT Science 06-07	817.39	667	35.79	820.33	331	35.56	814.50	336	35.84
CRCT Science 07-08	820.96	727	39.63	824.96	358	37.37	817.07	369	41.40

Notes. 1) Descriptives based on 23 item test that aligns with state standards.

Appendix B

Tables B1 through B5 summarize results addressing both of the aforementioned questions. As noted above and highlighted in Table A1, this subset of students is similar to the complete sample, at least as indicated by observable performance on the LHS Science, Vocabulary, and Reading assessments. Consistent, with the full sample results, the SR treatment was statistically significant in Science and Vocabulary, but not in Reading. Effect sizes are similar. These results are presented in Table B1. The results in Table B1 also indicate that there are no performance gaps in science among any student background or classification indicators. There is a gender gap in Vocabulary and low SES student performance about 0.40 points below their non-low SES classmates. It is important to note that Table B1 does not provide results for ELL students. ELL classification information was missing on a significant number of students and was therefore excluded from the analysis presented in Table B1.

In order to examine the performance of ELL students, the sample was further subdivided into students who had this additional information and the analysis was performed on this reduced sub-set. There is some indication that this subset is not representative of the overall sample, as performance, was lower on all three LHS assessments. Further, the results presented in Table B3 indicate that the treatment effect both smaller in absolute value, and more heterogeneous. The results indicate that there were not performance differences between ELL and English-Only students, but this result may not generalize to the entire sample, given, what appears to be the biased sample.

Consistent with expectations, the pre test is related to the post test for all three outcome measures.

Table B1

Effect of Student Background and At-Risk Indicators

Fixed effects	Science			Vocabulary			Reading		
	Effect	*S.E.*	p-value	Effect	*S.E.*	p-value	Effect	*S.E.*	p-value
Control classroom	13.93	0.37		12.51	0.28		10.46	0.15	
Treatment effect (+/-)	1.41	0.51	0.009	1.42	0.40	0.001	0.11	0.21	
Mean-pretest (subject specific)	1.09	0.42	0.012	0.50	0.19	0.010	0.22	0.07	0.003
Pre-test effect: Science	0.13	0.13	0.013	0.07	0.05		0.07	0.04	
Pre-test effect: Vocabulary	0.11	0.05	0.028	0.23	0.05	0.000	0.15	0.04	0.000
Pre-test effect: Reading	0.27	0.04	0.000	0.29	0.04	0.000	0.55	0.03	0.000
Girl	-0.20	0.22		-0.82	0.20	0.000	-0.06	0.18	
Asian Vs. White	0.49	0.60		-0.52	0.57		-0.13	0.51	
African American Vs. White	0.34	0.35		-0.14	0.32		0.06	0.24	
Hispanic Vs. White	-0.53	0.41		-0.24	0.38		-0.12	0.34	
Others Vs. White	0.04	0.71		0.43	0.67		0.53	0.60	
Low SES Vs. Non-Low	-0.06	-0.06		-0.41	0.23	0.075	-0.40	0.21	0.049
SWD Vs. non-SWD	-0.54	0.40		-0.57	0.38		-0.51	0.34	0.136
DF for level-1 variables		567			563		560		

Notes: (1) *** p < .01, ** p <.05, * p< .10

Table B2

Effect of Student Background and At-Risk Indicators

Fixed effects	Science			Vocabulary			Reading		
	Effect	*S.E.*	p-value	Effect	*S.E.*	p-value	Effect	*S.E.*	p-value
Control classroom	14.48	0.91		13.55	0.61		11.00	0.36	
Treatment effect (+/-)	1.37	1.18		0.48	0.78		-0.36	0.44	
Pre-test effect	0.17	0.08	0.036	0.42	0.06	0.000	0.55	0.05	0.000
Girl	-0.60	0.33	0.069	-0.94	0.31	0.003	-0.26	0.29	
Asian Vs. White	1.16	0.97		-1.64	0.86	0.058	-0.49	0.77	
African American Vs. White	0.91	0.64		0.86	0.60		0.73	0.54	
Hispanic Vs. White	0.01	0.59		-0.07	0.53		-0.26	0.48	
Others Vs. White	0.19	1.00		-0.53	0.90		-0.14	0.84	
Low SES Vs. Non-Low	-1.29	0.39	0.001	-0.71	0.36	0.051	0.84	0.34	0.056
SWD Vs. non-SWD	-1.06	0.57	0.064	-0.75	0.52		0.18	0.48	
ELL Vs. Non-ELL	-1.20	0.81		0.74	0.70		-0.07	0.58	
DF for level-1 variables			252		242			242	

Notes: (1) *** p < .01, ** p <.05, * p< .10

The results in Table B3 summarize the effects of student background on LHS assessments for student in treatment classrooms. Preliminary analyses examined the impact of student risk-factors (i.e. SWD, Low SES, ELL), but these factors are unrelated to outcomes. One reason may be that the sample size is significantly reduced when including these risk factors. The results in Table B3 focus on student background. The results indicate that only in vocabulary are there statistically significant difference in post-test performance ($p < .05$). Girls are expected to perform about 0.86 points below boys and Hispanics are expected to score about 0.88 points below Whites (the corresponding effect sizes are approximately -0.34) . However, it is likely that part of the performance gap between Hispanics and Whites is due to language status, which is not explicitly included in the model, but is likely correlated with Hispanic status.

Table B3

Effect of Student Background on Outcomes in Treatment Classrooms

Fixed effects	Science			Vocabulary			Reading		
	Effect	*S.E.*	p-value	Effect	*S.E.*	p-value	Effect	*S.E.*	p-value
Control classroom Mean-pretest (subject specific)	15.75	0.35		14.00	0.26		10.46	0.12	
	1.08	0.62	0.090	0.67	0.30	0.032	0.22	0.09	0.023
Pre-test effect: Science	0.20	0.06	0.001	0.08	0.06		0.13	0.05	0.010
Pre-test effect: Vocabulary	0.16	0.06	0.005	0.27	0.06	0.000	0.18	0.05	0.001
Pre-test effect: Reading	0.32	0.04	0.000	0.38	0.04	0.042	0.54	0.04	0.037
Girl	-0.39	0.23	0.094	-0.86	0.24	0.001	0.09	0.21	
Asian Vs. White	0.40	0.67		-0.35	0.68		0.67	0.58	
Afrian American Vs. White	-0.49	0.44		-0.37	0.43		0.00	0.26	
Hispanic Vs. White	-0.81	0.43	0.061	-0.88	0.44	0.045	-0.14	0.36	
Others Vs. White	-0.22	0.67		-0.25	0.68		0.75	0.58	
DF for level-1 variables		446			444			442	

Tables B4, B5, and B6 provide results on the impact of the SR treatment on student performance on State CRCT science and ELA assessment results. In each case the pre-CRCT is included in the analyses. The results in Tables B4 and B5 indicate that the students in treatment classrooms did not score significantly differently than did students in control classrooms.

Table B4

Treatment Effect on State Science Assessment

Fixed effects	Estimate	S.E.	p-value
Control classroom	821.79	2.48	
Treatment effect	-0.17	3.51	0.962
Mean CRCT pretest	0.14	0.09	0.147
CRCT pretest	0.82	0.03	0.000

*** p < .01, ** p <.05, * p< .10; df=735.

Table B5

Treatment Effect on State ELA Assessment

Fixed effects	Estimate	*S.E.*	p-value
Control classroom	821.17	2.08	
Treatment effect	-0.35	3.01	0.910
Mean CRCT(ELA) pretest	0.21	0.13	0.114
CRCT(ELA) pretest	0.83	0.04	0.000

The results in Table B6 include student background and indicate that accounting for student background the student performance was not statistically different between treatment and control classrooms. It is interesting to note that unlike the LHS results, there is a small gender gap in CRCT Science, with girls scoring about 0.16 SD below boys. It is important to note that LHS assessment results are related to CRCT results, which provides evidence of criteria related validity for inferences based on LHS assessment results.

Table B6

Impact of Student Background on State Assessment Results

Fixed effects	SCI			ELA		
	Effect	*S.E.*	*approx p*	Effect	*S.E.*	*approx p*
Control classroom	820.90	2.40		820.34	1.98	
Treatment effect (+/-)	0.40	3.38		-0.44	2.87	
Pre-test effect: CRCT ELA07				0.50	0.06	0.01
Pre-test effect: CRCT Science07	0.65	0.04	0.01	0.16	0.16	0.01
Pre-test effect: Vocabulary	1.23	1.23	0.01	0.88	0.38	0.02
Pre-test effect: Reading	1.59	1.59	1.59	1.27	0.36	0.01
Pre-test effect: Science	0.93	0.46	0.05			
Girl	-4.56	1.94	0.02	1.00	1.83	
Asian Vs. White	-5.51	5.22		0.93	6.10	
African American Vs. White	-4.84	2.97		-2.72	3.38	
Hispanic Vs. White	-1.35	3.64		-2.31	4.72	
Others Vs. White	-2.23	6.35		2.28	7.33	
Low SES Vs. Non-Low	-2.50	2.21		0.18	2.01	
SWD Vs. non-SWD	-6.57	3.54	0.06	-8.48	3.43	0.01
DF for level-1 variables		525		327		