Educational Psychology Review (2021) 33:1935–1964
https://doi.org/10.1007/s10648-021-09609-6

REPLICATION

Improving Elementary Grade Students' Science and Social Studies Vocabulary Knowledge Depth, Reading Comprehension, and Argumentative Writing: a Conceptual Replication

James S. Kim[1] · Jackie Eunjung Relyea[2] · Mary A. Burkhauser[1] · Ethan Scherer[1] · Patrick Rich[1]

Accepted: 22 February 2021 / Published online: 6 April 2021

Abstract
This experimental study aimed to replicate and extend a previous efficacy study of an elementary grade content literacy intervention that demonstrated positive effects on students' vocabulary knowledge depth, argumentative writing, and reading comprehension. Using a cluster (school) randomized trial design, this replication experiment was conducted with 5,494 first- and second-grade students in 30 elementary schools in an urban school district located in the southeastern USA. Teachers implemented thematic lessons (20 lessons) that provided an intellectual framework for helping students who acquire networks of related vocabulary knowledge while learning science and social studies content. Teachers integrated thematic lessons, concept mapping, and interactive read-alouds of conceptually related informational texts to enable their students to build networks of vocabulary knowledge and to transfer this knowledge to argumentative writing and collaborative research activities. Confirmatory analyses replicated positive findings on science vocabulary knowledge depth (ES = 0.50) and argumentative writing (ES = 0.24) and also extended positive findings to social studies vocabulary knowledge depth (ES = 0.56) and argumentative writing (ES = 0.44). Positive and statistically significant findings were not replicated on domain-general reading comprehension. Exploratory analyses indicated that students' vocabulary knowledge depth partially mediated the impact of content literacy instruction on domain-specific argumentative writing outcomes.

Keywords Replication · Content literacy intervention · Randomized controlled trial · Science vocabulary knowledge depth · Social studies vocabulary knowledge depth · Reading comprehension · Argumentative writing

Extended author information available on the last page of the article

To succeed in school, all children must be able to use prior domain knowledge to read and write about complex and challenging texts. Although many children acquire the procedural word reading skills to decode new and unfamiliar words, fewer than 5% of children in first and second grade can evaluate complex nonfiction texts that require prior knowledge of science and social studies topics (Reardon et al. 2012). Recent results from the National Assessment of Educational Progress (NAEP) further indicate that individual differences in reading between high- and low-performing fourth-grade children have grown wider over the past decade and writing performance remains stagnant from the elementary to secondary grades (National Center for Education Statistics 2019). These alarming descriptive statistics make it clear that many elementary grade students need opportunities to participate in evidence-based models of literacy instruction that systematically build domain knowledge. Such efforts are undoubtedly a linchpin of national literacy initiatives to help all children acquire the reading and writing abilities to succeed in college and their future careers.

Ultimately, efforts to help all students meet college and career readiness standards should begin in the elementary grades and rest on a strong foundation of replicated research. Emerging experimental evidence suggests that embedding science and social studies content in early elementary literacy instruction is a promising approach. Recently, we (Kim et al. 2021) reported findings from a multicomponent elementary grade content literacy intervention designed to help students acquire science domain knowledge and to build coherent text representations in reading and writing (Galbraith and Baaijen 2018; Kintsch 2009). In our previous efficacy study involving 10 schools, 38 first-grade teachers and their students (N = 674) were randomly assigned to treatment or control lessons. Treatment group teachers implemented a 10-day thematic unit on Arctic animal survival whereas control teachers implemented a balanced literacy program including word study, guided reading and writing activities, and leveled fiction and nonfiction texts. First-grade students in the treatment group scored statistically significantly higher than students in control classrooms on measures of science vocabulary knowledge depth (effect size [ES] = 0.30), reading comprehension (ES = 0.11), and argumentative writing (ES = 0.24).

Although these results were promising, replication studies are needed to determine whether novel findings from a single study are an anomaly or robust enough to support instructional recommendations at scale (Makel and Plucker 2014; Maner 2016; Schmidt 2009). A critical question that we sought to answer in this large-scale replication study was whether a science and social students content literacy instruction could improve first- and second-grade students' reading and writing outcomes. There are mixed findings from early grade content literacy, underscoring the rationale for replications that help build a more robust knowledge base for research and practice (Pearson et al. 2020; Strachan 2015). Therefore, this study aimed to replicate and extend our previous efficacy study in both science and social studies with a larger sample of schools, teachers, and students and over a longer program implementation period.

Theoretical and Conceptual Foundations for Content Literacy Intervention

In recent years, educational psychologists and literacy scholars have designed content literacy interventions to improve students' domain knowledge in science and social studies while building higher-order reading comprehension and argumentative writing abilities (Connor et al. 2017; Guthrie and Klauda 2014; Romance and Vitale 2001; Vaughn et al. 2013;

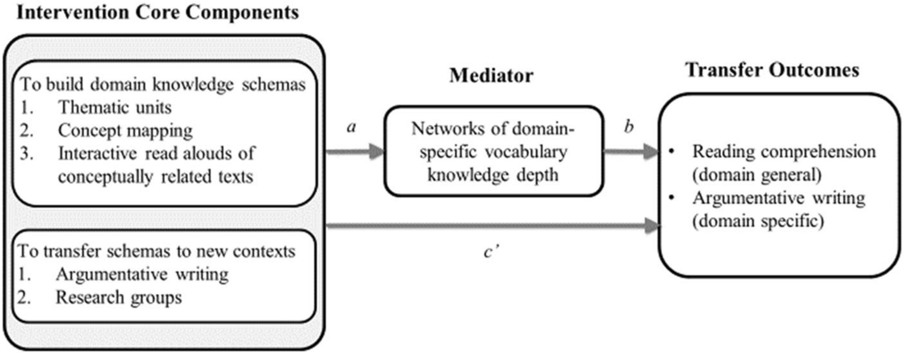
Intervention Core Components
To build domain knowledge schemas
1. Thematic units
2. Concept mapping
3. Interactive read alouds of conceptually related texts
To transfer schemas to new contexts
1. Argumentative writing
2. Research groups
a
Mediator
Networks of domain-specific vocabulary knowledge depth
b
c'
Transfer Outcomes
• Reading comprehension (domain general)
• Argumentative writing (domain specific)

Fig. 1 Hypothesized logic model for content literacy intervention: direct and indirect effects on reading comprehension and argumentative writing

Williams et al. 2016). The logic model in Fig. 1 visually displays how the core components of a content literacy intervention can lead to improvements in the mediator and, ultimately, the transfer outcomes. We hypothesize that reading comprehension and argumentative writing will improve if students can leverage schemas that build networks of vocabulary knowledge depth in science and social studies. In essence, the logic model indicates that vocabulary knowledge depth is theorized to partially mediate the effects of content literacy intervention on students' reading and writing. The logic model rests on three pillars of research, including research explaining (a) how experts represent domain knowledge in the form of schemas, (b) how networks of vocabulary knowledge depth support schema development and transfer, and (c) how the core components of content literacy can build networks of vocabulary knowledge depth in service of improved reading comprehension and argumentative writing outcomes.

Expert Readers and Writers Represent Domain Knowledge in the Form of Schemas

Domain knowledge refers to how much a student knows about a school subject like science and social studies (Alexander 2003). It is widely acknowledged that domain knowledge enables learners to build coherent text representations (Kintsch and van Dijk 1978; Minsky 1975; Rumelhart and Ortony 1977; Rumelhart 1980; Thorndyke 1984). The role of prior domain knowledge is particularly important in models of reading and writing that emphasize the two levels of cognitive processing. For example, in the construction-integration model, Kintsch (1988) proposed that domain experts leverage retrieval structures in long-term working memory to store and update the propositional textbase and the situation model. A hallmark of domain knowledge expertise in reading is the ability to build and leverage retrieval structures that support the mental instantiation of an elaborate situation model. Similarly, cognitive theories of writing emphasize two processes that are involved in learning from text, including an active knowledge-constitutive process and a reflective knowledge-transforming process in which content retrieved from long-term memory is used to satisfy rhetorical goals (Galbraith and Baaijen 2018). Fundamentally, proficient readers and writers become skilled at "building efficient retrieval structures based on accumulated domain knowledge" (Kimball and Holyoak 2000, p. 117).

There is converging evidence that experts represent accumulated domain knowledge in long-term memory in the form of schemas. Rather than defining schemas as a "monolithic building block of cognition" (Iran-Nejad 1989, p. 1), more recent scholarship has emphasized the notion that schemas are abstract knowledge structures that give experts an edge over novices in learning new content (Ericsson and Kintsch 1995; Kimball and Holyoak 2000; Kintsch 2009). In comparison with novices, for example, experts are able to instantiate a general schema as a set of related concepts in a domain (Gick and Holyoak 1983). The experts' advantage in reading and writing about complex text-based ideas depends on the application of a schema to new tasks that differ only in their superficial characteristics (Ericsson 2018; Ericsson and Pool 2016; Kintsch 1988). For example, a student who recognizes that concepts like *adaptation* and *extinction* belong to the schema for *animal survival* can read and write about how animals survive in a variety of habitats, including habitats that were not the target of direct instruction by teachers.

Networks of Vocabulary Knowledge Depth Support Schema Development and Transfer

Researchers have argued that networks of vocabulary knowledge depth are critical to supporting schema development and transfer of learning (e.g., Fitzgerald et al. 2020). The notion of a network structure implies that a single word has connections to other words with semantic overlap. As students are exposed to networks of semantically related words in science and social studies content, an individual student's vocabulary knowledge is developed incrementally by connecting and enriching word meanings over time (McKeown and Beck 2011; Stahl and Nagy 2006). Content literacy instruction may provide an ideal context for helping young children mentally instantiate and leverage networks of academic vocabulary words to further develop their domain knowledge (McKeown and Beck 2011; Perfetti 2007; Perfetti and Adlof 2012; Stahl and Nagy 2006). In essence, a student with a large network of domain-specific academic words can efficiently use this knowledge when reading and writing about new academic topics learned in school (McKeown et al. 2017; Schmitt 2014).

According to the lexical quality hypothesis, children with higher-quality mental representations of networks of vocabulary knowledge are able to automatically and flexibly access those words while reading and writing about transfer text (Perfetti 2007). As a result, students' initial reading ability could moderate the relationship between content literacy instruction and various literacy outcomes, particularly on far transfer outcomes that include many new and unknown concepts. For example, well-known Matthew effects whereby the rich get richer and the poor get poorer (Connor et al. 2017; Kellogg 2008; Stanovich 1986) could hypothetically be related to individual differences in children's ability to leverage and apply networks of vocabulary knowledge depth while reading and writing about new topics in science and social studies. Indeed, some empirical research indicates that content literacy instruction may exacerbate gaps in vocabulary, reading, and writing outcomes (Coyne et al. 2009; Stanovich 1986; Wood et al. 2020). The establishment of mental schemas and the development of domain-specific vocabulary networks may be associated with growth in reading and writing and therefore with reducing gaps in performance on transfer tasks.

Similarly, young children who are low-performing readers tend to trail their peers for word recognition and vocabulary knowledge (Kellogg 2008; Stanovich 1986). The lexical quality hypothesis (Perfetti 2007) suggests that low levels of word recognition and vocabulary knowledge reflect weak mental representations or weak representations of word networks and schemas that aid comprehension. Low-quality connections among words and concepts in semantic memory are likely to impede learners' access to connected networks of word meanings when reading or listening to complex texts (Stahl and Nagy 2006). Consequently, individual differences in young children's reading ability may moderate the impact of instruction designed to foster domain schemas and networks of vocabulary knowledge. Insofar as schemas reflect a learner's ability to instantiate networks of vocabulary knowledge, the key question then becomes: how can science and social studies words be organized to improve reading and writing?

Building Networks of Vocabulary Knowledge Depth in Service of Reading and Writing

We developed a content literacy intervention, the Model of Reading Engagement (MORE), to build students' networks of vocabulary knowledge depth in science and social studies. The components were based on models of domain and vocabulary knowledge development during content literacy instruction (Alexander 2000; Connor et al. 2017; Guthrie et al. 2004; McKeown et al. 2017; Vaughn et al. 2013). Teachers enacted thematic lessons, concept mapping, and interactive read-alouds to enable their students to build networks of vocabulary knowledge and to apply and transfer this knowledge during argumentative writing and collaborative research activities.

Thematic Lessons, Concept Mapping, Interactive Read-Alouds

To help students build domain knowledge schemas and networks of vocabulary knowledge, teachers can organize science and social studies concepts in thematic lessons (Wiske 1998). Thematic lessons provide an intellectual framework for helping students connect new learning to a domain-specific schema (Wiske 1998). As such, thematic lessons are built on a pyramid structure where big concepts help unify supporting concepts and ideas in science and social studies (Guthrie et al. 2007). This approach is based on the notion that thematic units expose children to a domain-specific schema that is a networked structured of words, concepts, and ideas. Through thematic lessons, children are afforded an opportunity to mentally construct similarly organized words, concepts, and ideas.

Concept mapping activities further build students' high-quality network connections and schemas among a few focal concepts in semantic memory (Gelman 2009; Imai et al. 1994; Levin 1988; Stahl and Fairbanks 1986; Stahl and Nagy 2006). A concept map is a graphical tool for organizing concepts in which pictures are paired with labels, and a node-link-node syntax is used to build a learners' schema (Nesbit and Adesope 2006). In other words, concept mapping of taxonomically related science and social studies words visually display the ways in which words are connected and describe the shared properties of living and non-living things (Neuman et al. 2011).

We also included interactive read-alouds of conceptually related informational texts that cohered around focal concepts (Cervetti et al. 2016) and thus provided frequent exposures to the associated words that comprised the concept maps in science and social studies. Importantly, reading and listening to conceptually related texts facilitate a student's ability to connect a pre-existing schema of related words to newly encountered words. This process of building a network of semantically associated words through connected text may facilitate comprehension of sentences with target words (Beck et al. 1982), promote incidental learning of untaught words in domains (Carlisle et al. 2000), and, in some instances, foster transfer to novel reading and writing tasks (Elleman et al. 2009; Stahl and Fairbanks 1986).

Argumentative Writing and Collaborative Research

A key aim of content literacy instruction is to help students transfer networks of vocabulary knowledge to new reading and writing tasks. Accordingly, we included two components that provided further opportunities to apply domain expertise through argumentative writing and collaborative research activities. First, we emphasized that experts develop expertise through argumentative writing, which requires novice learners to draw upon discourse and domain knowledge, as well as mentally stored academic vocabulary networks, to solve rhetorical problems (Bereiter and Scardamalia 1987; Hayes and Flower 1986). There is strong evidence that domain knowledge is a strong predictor of a student's argument writing quality (Graham et al. 2018; Olinghouse et al. 2015). These findings are consistent with schema theories that underscore how experts efficiently access and leverage "a fixed set of connection strengths between units" (Galbraith and Baaijen 2018, p. 241-242) to transform knowledge into written ideas. Furthermore, a learner's schema, as indexed by vocabulary knowledge depth, frees cognitive resources to help learners plan and evaluate one's writing and to connect text-based ideas with retrieval structures held in long-term memory (McCutchen 1996; McKeown et al. 2017).

Second, students applied learning through collaborative research activities. Research groups were designed to foster student engagement through an open intellectual task without a correct answer. Students engage in the collaborative research activities that provide autonomy to make choices, to perform authentic tasks, and to focus attention on completing appropriately challenging content-focused writing task (Guthrie et al. 2007; Parsons et al. 2015). This activity provided students with an additional context for reading (e.g., reading authentic texts including newspaper and magazine articles) and writing about words that were critical to developing domain knowledge schemas and enriching and expanding networks of vocabulary knowledge in science and social studies.

Research Aims and Questions

This study aimed to replicate and extend our previous efficacy study and to provide an initial test of the logic model for the MORE content literacy intervention. Thus, we designed this study as a conceptual replication, which systematically varies features of the original study to extend findings and analyses (Bollen et al. 2015; Makel and Plucker 2014; Nosek et al. 2015; Van Bavelet al. 2016). In particular, we systematically varied the number of domains included in the lessons, the size of the school, teacher, and student sample, and the duration of the program implementation. Thus, the study focused on three specific aims. The first confirmatory aim was to determine if findings from our previous study could replicate positive effects in science and extend effects to the domain of social studies with a larger sample of schools, teachers, and students. The second confirmatory aim was to examine whether treatment effects differed for initially low- and high-performing students. Finally, the exploratory aim was to examine whether improvements in vocabulary knowledge mediated effects on reading and writing outcomes. Thus, we addressed the following research questions:

1. To what extent do results from a previous efficacy study of the MORE content literacy intervention replicate and extend positive impact on measures of (a) science and social studies vocabulary knowledge depth, (b) reading comprehension, and (c) argumentative writing?
2. To what extent do measures of students' initial reading ability moderate the effects of MORE on reading comprehension and argumentative writing outcomes?
3. To what extent do students' science and social studies vocabulary knowledge depth mediate the effects of MORE on reading comprehension and argumentative writing outcomes?

Methods

Research Design and Participants

Thirty elementary schools in one urban school district located in the southeastern USA participated in this 2-year longitudinal study. Prior to program implementation, we preregistered the study design, conducted a power analysis to determine the minimum detectable effect size, and implemented a blocked, school-level randomized controlled trial (Kim et al. 2021). In particular, schools were blocked (i.e., stratified) by demographic and achievement characteristics and then randomized to conditions for two consecutive years.[1]

At the beginning of year 1 in spring 2019, half the schools were randomly assigned to treatment in first-grade treatment lessons and half were assigned to second-grade treatment lessons. First-grade treatment schools included controls at second grade, and second-grade treatment schools included controls at first grade. Thus, first-grade treatment schools provided valid counterfactuals for the second-grade schools, and vice versa. In this study, we report findings from the first year of program implementation. Future studies will report second-year findings and examine whether treatment effects are larger after two consecutive years of program implementation.

Table 1 displays descriptive characteristics of participating teachers and students. Participating teachers were mostly female and white, about half had earned or were working towards a master's degree, and nearly 40% had participated in a professional development program focused on the science of reading (reading research to classroom practice). Black and Hispanic students comprised the majority of the student sample. Students were from socioeconomically diverse neighborhoods and nearly 20% were designed as having limited English proficiently. Importantly, there were no statistically significant baseline differences between the treatment and control group on measured teacher and student characteristics at the individual level and the school level.

[1] Using data on the effect sizes and intraclass correlation (ICC) from our previous efficacy study, we targeted a sample size of 60 teacher/classroom clusters (30 treatment groups and 30 control groups) and 15 students per cluster. Assuming an alpha level of 0.05 (two-tailed tests) on all impact models, a pretest reading covariate (R^2 = 0.50), and 80% power, the minimum detectable effect sizes were 0.25 across the primary student outcomes

Table 1 Descriptive characteristics of treatment and control teachers and students

Characteristics	Individual level		School level	
	Treatment	Control	Treatment	Control
Teacher characteristics				
N	144	136	30	30
Female	98%	97%	98%	97%
White	67%	61%	69%	59%
Has or working toward Masters	50%	46%	51%	50%
Attended reading research to classroom practice	38%	43%	35%	39%
Student characteristics				
N	2886	2608	30	30
Female	50%	50%	50%	49%
White	18%	19%	16%	17%
Black	39%	38%	41%	39%
Hispanic	32%	32%	32%	32%
Asian	8%	8%	7%	8%
Other race	4%	3%	4%	3%
Gifted program	6%	5%	6%	4%
Limited English proficiency	22%	24%	23%	23%
Individual education plan	8%	9%	8%	9%
Low SES	40%	40%	46%	45%
Med SES	38%	39%	36%	37%
High SES	21%	20%	18%	18%
MAP reading pretest, *M* (*SD*)	175.16 (18.21)	176.76 (17.39)	174.56 (11.18)	175.71 (8.03)
DIBELS pretest, *M* (*SD*)	203.52 (113.53)	210.12 (116.08)	201.20 (56.05)	202.60 (46.09)

SES socioeconomic status, *MAP* Measures of Academic Progress, *DIBELS* Dynamic Indicators of Basic Early Literacy Skills

Intervention Description

Professional Development Before program implementation started, teachers participated in a 3-h professional development meeting. The meeting focused on the overview of the study, the MORE program principles, core components, and instructional routines. During the implementation period, the research team provided treatment teachers with ongoing support by sending daily video clips (5–10 min) that further explained key features of an online lesson guide, lesson materials, and procedures to teachers. Treatment teachers also received continuing support for lesson implementation and data collection processes from their school's literacy facilitator. The research team hosted a series of monthly meetings for literacy facilitators to communicate information on lesson implementation and to address teachers' questions about implementation.

Curriculum and Lessons Both the first- and second-grade curriculum consisted of one science and one social studies unit, and each unit focused on a single theme and organizing questions: How do animals survive in their habitat (first-grade science)? How do explorers overcome obstacles (first-grade social studies)? How do scientists study mass extinctions (second-grade science)? How do inventors solve problems (second-grade social studies)? The treatment curriculum began with the science unit followed by the social studies unit; each unit was designed to be taught over a 20-lesson sequence. Treatment teachers received a pacing calendar to reference the instructional sequence and were given substantial flexibility in the pacing of instruction to teach the 20 lessons over a 5- to 10-week window.

Each lesson included into two sections. The first section focused on building domain knowledge through interactive read-alouds and concept mapping. Interactive read-alouds involved reading thematically related informational texts (20 min), a structured equitable academic discussion and/or concept mapping activity (15 min), and mapping their progress and goal mastery status in each unit (5 min). The second section focused on transferring vocabulary knowledge to new reading and writing tasks. Students engaged in collaborative research using text features (e.g., compare and contrast) to obtain additional information on given concepts during reading and writing activities (20 min) and participated in in-depth academic group discussions while incorporating relevant details and evidence from their research (20 min). Teachers were given the option of teaching the two lesson sections over two class periods.

The instructional sequence of 10 lessons in each unit was designed to support students in expanding and deepening their networks of vocabulary knowledge in science and social studies and applying their knowledge of vocabulary to new reading and writing tasks. Teachers used a scaffolding process to demonstrate strategies for using text features, self-questioning, concept mapping, and making a claim and providing text evidence. In the first five lessons, lessons were structured to develop instructional routines that were organized in a systematic, cumulative manner. For example, in lessons 1 and 2, students read aloud informational texts and began to map a concept network using resources from text and supplemental materials. In lessons 3 and 4, students had further opportunities to analyze and think critically about the text they read and to participate in a rich academic discussion. In lesson 5, students learned an argumentative writing strategy called “A-TREE” (Graham and Harris 2005) and applied the strategy in writing an argumentative response to an open-ended question (e.g., *How do paleontologists know what dinosaurs looked like?* for second-grade science). As students progressed through the lesson, teachers gradually allowed students to independently establish group and individual learning routines. Accordingly, the second half of the unit (lessons 6–10) repeated similar instructional routines with the teachers’ gradual release of control and more emphasis on student-directed research and reading activities.

Procedures for Selecting Networks of Vocabulary in Science and Social Studies We used an iterative procedure to identify domain-specific vocabulary. First, we anchored the words to the state’s science and social studies standards and the Next Generation Science Standards (NGSS; National Research Council 2012). We conducted a content analysis of each standards document as well as our lesson texts to identify the related vocabulary. Second, we cross-validated these words against content standards that predated the state standards and NGSS to ensure that the words were relatively stable features of US school curricula over time (Hirsch Jr. 2016). Third, we created an automated concept network for each topic of unit (i.e., *Arctic animal survival*, *explorers*, *mass extinctions of dinosaurs*, *inventors*) containing the target academic words and additional semantically associated words also appearing in the lesson texts. Each target word or concept was represented as a node with weighted connections between nodes indicating the degree of similarity.

Using the concept maps for all target concepts, we identified two types of academic vocabulary words: 10 semantically associated words directly taught in the lessons of each unit, particularly in the concept mapping activity, and five related words that appeared in the lesson texts, but were not explicitly taught during the lessons. The five untaught words were in the range of lower frequency, higher age of acquisition, and/or lower concreteness at each grade level. We selected these untaught words for the purpose of assessing transfer on the vocabulary depth measure.

Because knowledge structures differ for the domain of science and social studies (Marzano 2004; Trefil et al. 2002), we clustered 10 target words in each unit in a distinctive way. Science words were organized hierarchically from more general to specific categories and concepts (Gelman 2009; Steyvers and Tenenbaum 2005). Accordingly, the Grade 1 science words (i.e., survive, habitat, physical feature) cohered around the general concept of animal survival and were designed to help students transfer and apply these words to animals in a variety of habitats. Grade 2 science words (i.e., theory, fossil, evidence) focused on the general idea of how scientists use evidence to develop theories with specific applications to the topic of mass extinctions and dinosaurs. Social studies words were organized into taxonomic groups that emphasized the shared characteristics and essential properties of explorers in Grade 1 (i.e., expedition, obstacle, persistent) and inventors in Grade 2 (e.g., ingenious, pioneering, creative).

Fidelity of Implementation

We measured four dimensions of fidelity of implementation (Dane and Schneider 1998): (a) adherence, (b) exposure to reading, science, and social studies content, (c) participant responsiveness, and (d) program differentiation. We used multiple sources of data (audio-recorded lessons, teacher surveys) to measure FOI.

Adherence to the Core Intervention Components Treatment teachers' adherence to the core components was assessed using audio-recordings. Several steps were involved in collecting and coding the audio-recordings. We first identified six school sites to collect audio recordings by randomly selecting two schools from each of low-, middle-, and high-poverty strata. We visited a total of 25 treatment classrooms from the six schools and audio-recorded a reading lesson of each classroom delivered during the implementation period. Next, we created an adherence checklist that outlined 11 indicators that were essential to the intervention core components (see Appendix S-A for the adherence checklist). Then, a research assistant listened to 25 audio-recorded lessons from the treatment teachers and tallied the presence (1 point) and absence (0 point) of each of the 11 indicators. An adherence score for each teacher was calculated by summing the total number of the indicators present in each lesson and dividing by the total number of the indicators. A percentage score was obtained by multiplying by 100 to estimate the degree of adherence for each teacher. Overall, the average adherence rate across 25 teachers was 98% (range = 80–100%), indicating a high rate of adherence to the components. After an initial coding, the second research assistant independently double-coded a randomly selected 44% subset of the lesson audio-recordings to obtain inter-rater agreement. Overall agreement was 91% (Cohen's $\kappa = 0.63$).

Exposure to Reading, Science, and Social Studies Contents We conducted a survey for all teachers to estimate the amount of instructional time that teachers spent on reading, science, and social studies contents over the course of the study. Treatment teachers spent, on average, nearly 165 min per week on science and 162 min on social studies contents which were 61 and 58 more minutes, respectively, than control teachers. Treatment and control teachers devoted an average of approximately 486 and 506 min per week to reading content, respectively, but the difference was not statistically significant ($p > 0.05$).

Participant Responsiveness To measure participant responsiveness during the lessons, we evaluated behavioral engagement with challenging literacy tasks in a given context (Lepola et al. 2016). Upon the completion of the intervention, treatment and control teachers rated a random sample of 10 students' task orientations on four items adapted from the Lepola et al.' (2005) questionnaire. The four items that tapped task orientations were (a) concentration on the task, (b) showing persistent effort when facing difficulties (i.e., not giving up easily), (c) becoming absorbed in the given task, and (d) being eager to do tasks that exceed one's competence. Teachers rated each item on a 5-point Likert scale (1 = *the behavior does not occur at all*, 2 = *very seldom*, 3 = *sometimes*, 4 = *does occur often*, 5 = *does occur very often*). Internal consistency (Cronbach's α) for task orientations was 0.92. As shown in Table 2, the average task orientation score of treatment students ($M = 3.68$, $SD = 1.03$) was slightly higher than their counterparts ($M = 3.59$, $SD = 1.1$), but the difference was not statistically significant ($p > 0.05$).

Table 2 Descriptive statistics of fidelity of implementation (FOI): program differentiation in read-aloud text Lexile level, openness of literacy tasks, and instructional time of treatment and control condition

	Treatment		Control				
FOI components	*M*	*SD*	*M*	*SD*	*t*	*df*	*ES*
Adherence	98.00%	5.70					
Exposure[a]							
ELA/reading instruction	485.94	244.02	505.30	258.91	− 0.42	243	− 0.08
Science instruction	164.51	126.84	103.25	88.50	4.97***	240	0.56
Social studies instruction	161.68	123.44	103.69	90.62	4.83***	241	0.53
Participant responsiveness							
Task orientation	3.68	1.03	3.59	1.10	1.35	2,318	0.09
Program differentiation							
Read-aloud text Lexile level							
All (science & social studies)	802.86L	140.71L	501.56L	132.30L	8.54***	58	2.21
Science	738.67L	120.35L	534.29L	137.10L	4.63***	34	1.57
Social studies	876.92L	128.67L	439.09L	100.64L	9.15***	22	3.75
Openness of literacy tasks							
Authenticity	2.65	0.39	1.64	0.54	7.89***	41	2.15
Collaboration	2.63	0.27	1.61	0.36	11.56***	41	3.23
Challenge level	2.60	0.51	1.24	0.30	13.68***	41	3.20
Student-directed work	2.59	0.46	1.59	0.44	8.84***	41	2.24
Sustained effort	2.88	0.33	1.83	0.70	7.62***	41	1.93

***$p < 0.001$

[a] Minutes per week spent on instruction

ELA English language arts

Program Differentiation: Complexity of Interactive Read-Aloud Texts We conducted a survey of all teachers on the titles of the books that they used for teacher-directed, interactive read-aloud activities during the implementation period. The list of book titles is available in Table S1 in Appendix S-B of the supplemental on-line materials. As shown in Table 2, all books that treatment teachers used were informational texts, whereas only 27% and 9% of the books used by first- and second-grade control teachers, respectively, were informational texts. The average Lexile levels for treatment-group science and social studies books were significantly higher than those for control-group books ($p < 0.001$). Particularly, the average Lexile level of social studies books in the treatment group ($M = 876.92$L, $SD = 128.67$L) was almost twice as high as that in the control group ($M = 439.09$L, $SD = 100.64$L, $p < 0.001$).

Program Differentiation: Openness of Literacy Tasks Table 2 reports program differentiation in the openness of the literacy tasks. For this measure, we first identified and documented the types of literacy tasks of 49 classrooms. Using Parsons et al. (2015) openness of literacy tasks rubric, we rated each literacy task for authenticity (simulating real-life activity), collaboration (collaborative or independent activity), challenge (intellectually stimulating activity), student-directed work (involvement of student input), and sustained effort (sustainability over time) on a 3-point Likert scale (1 = *closed*, 2 = *moderately open*, and 3 = *open task*). A higher degree of openness indicates that literacy tasks are more likely to promote student engagement and student-centered learning than tasks with a low degree of openness (Duke et al. 2006). One of the authors of the study provided a scoring training for two research assistants, in which they reviewed the rubric, conducted a scoring practice, and discussed coding discrepancies on the sample audio-recordings. They repeated the coding procedure until inter-rater agreement of 94% was reached. Then, one of the research assistants continued to code the rest of audio-recordings, while the other research assistant independently double-coded randomly selected 20% of audio-recordings to estimate inter-rater agreement. Overall agreement ranged from 81 to 97% (Cohen's $\kappa = 0.68$–0.96). Treatment-group literacy tasks were more likely characterized as authentic, collaborative, challenging, student-directed, and sustained than control-group literacy tasks.

Student Measures

Networks of Vocabulary Knowledge Depth We used the semantic association task (Read 1998, 2004; see Appendix S-C) to assess students' vocabulary knowledge depth of taught words in science and social studies units and their ability to identify semantic relations of the taught words with other words that were not explicitly targeted in the units (Collins and Loftus 1975; Schmitt 2014; Stahl and Fairbanks 1986). We developed two sets of 12-item semantic association tasks for science and social studies. Each set included seven domain-specific words taught in the treatment lessons and five associated words that were not directly taught during the lessons. The seven taught words were pre-selected words that treatment teachers directly taught via the concept mapping activity involving related vocabulary and also the argumentative writing activities. The five untaught words were not explicitly taught during the activities but were proximate and semantically associated with the taught words and were incidentally encountered through reading, listening, and discussion activities. In the semantic association task, there were four-word options for each of the 12 items that presented a target (taught or untaught) word and students were prompted to "circle two words that go with" the target word.

Each item was scored 0 to 4 (see Appendix S-D for the scoring system). Internal consistency (Cronbach's α) for science and social studies measures was 0.91 and 0.90, respectively.

Reading Comprehension The Measure of Academic Progress (MAP) Primary Grade Reading (Northwest Evaluation Association 2011) was used to assess students' domain-general reading comprehension ability. The MAP is a computer-adaptive, early literacy assessment that measures student growth in reading comprehension from kindergarten to second grade using the Rasch unit (RIT) scale. The MAP score is a composite of four strands: narrative and informational text comprehension, vocabulary use and functions, foundational skills, and language and writing. An overall RIT score was computed based on performance on the four strands and its test-retest reliability ranged from 0.89 to 0.96 (Brown and Coughlin 2007). Students' MAP RIT scores were used as pre- and posttest scores.

Basic Literacy Skills The mCLASS Dynamic Indicators of Basic Early Literacy Skills (DIBELS) was used as a measure of early literacy skills. Specifically, we used a composite score of DIBLES subtests, including sound fluency, phoneme segmentation fluency, letter naming fluency, nonsense word fluency, oral reading fluency, and retell abilities. Test-retest and inter-rater reliabilities of the composite ranged from 0.88 to 0.98 across grades. The concurrent and predictive validity, sensitivity, and specificity of DIBELS scores on end-of-year district and state standardized assessments are moderate to strong (Goffreda et al. 2009).

Argumentative Writing We conducted an argumentative writing assessment to evaluate treatment and control students' knowledge of the elements and structure of an argument (see Appendix S-C). The assessment consisted of a short source text to present background information relevant to a topic and open-ended writing prompt: *Should people be allowed to cut down trees in the rainforest?* (first-grade science); *If you had to pick just one of the women explorers to celebrate, which one would you choose - Amelia Earhart or Sally Ride? Why?* (first-grade social studies); *Do you think that an asteroid killed the dinosaurs? Why or why not?* (second-grade science); and *If you had to pick just one of the young inventors to celebrate, which one would you choose - Leonardo da Vinci or Henry Ford? Why?* (second-grade social studies). Student were asked to answer the questions—one in science and the other in social studies—by making an argument and reminded of the components of a good argument (says your opinion, says your reasons, explains your thinking, and has a conclusion). Before scoring student writing, research assistants typed all compositions by correcting for spelling and punctuation errors. They marked illegible or indecipherable words as "XXX" in the place. The purpose of this process was to reduce presentation bias stemming from poor handwriting skills (Graham et al. 2011) and to focus on the elements and structure of an argument in argumentative writing.

For scoring, we used a genre-specific rubric for argumentative writing rather than a generic rubric for generalizability and dependability (Graham et al. 2011). We first selected a random sample of 10 students from each classroom and scored their argumentative writing based on a three-dimensional rubric: *claim*, *evidence*, and *conclusion*. The *claim* dimension was scored on a scale from 0 to 2, with 0 indicating an absent claim, 1 indicating that a present claim with a lack of clarity of argument, and 2 representing that an appropriate claim was present, clear, and well developed. The *evidence* dimension was to evaluate students' ability to support their claim using their pre-existing knowledge of a given topic and/or to extract relevant knowledge from the source text. To more systematically assess the extent to which students extracted knowledge from the source text, we divided the text into several "concept units" or discrete pieces of information about a given topic. The evidence dimension was scored on a scale from 0 to 4, with 0 indicating absence of evidence statements or appropriate background knowledge; 1 indicating the inclusion of concept unit or textual evidence that was irrelevant to support the claim; 2 indicating the use of relevant background knowledge to support the claim but not found in the source text; 3 indicating the inclusion of at least one relevant concept unit from the source text to support the claim; and 4 indicating the use of at least two relevant concept units to support the claim. Finally, the conclusion dimension was ending that measured the presence of a concluding statement, scoring 0 (absent) or 1 (inclusion of a well-developed conclusion). The three-dimensional scores were summed up to yield a total score of 0 to 7 for the data analysis.

Raters received extensive training and practice before coding a large number of writing compositions. During the rater training stage, two raters first reviewed the scoring manual and anchor text and then started independent coding practice using the first batch of 100 written compositions. The raters met with one of the authors to discuss and resolve discrepancies between the two raters. To ensure a more robust and thorough training, the two raters repeated the coding practice using the second set of 100 compositions and then reached high consistency in agreement (≥ 90%). The final scores were determined by consensus reached by the raters and the author after discussion. All remaining compositions were scored by two raters and high inter-rater reliabilities were obtained (Cohen's κ ranged from 0.92 to 0.99 for total scores).

Data Analysis

Confirmatory Analyses of Treatment Effects To confirm the detected main (RQ1) and interaction effects (RQ2) of treatment from the previous study, we specified two-level hierarchical linear models (HLMs; Raudenbush and Bryk 2002) for the analysis of cluster randomized trial to account for the nested structure of the data, in which students were nested within school-level randomization blocks. Variance decomposition and intraclass correlation (ICC) estimates are provided in Appendix S-E of the supplemental on-line materials. Level 1 (within-school) model was written as follows:

$$\text{Level 1}: Y_{ij} = \beta_{0j} + \beta_{1j}\,(PREMAP)_{\mathrm{ij}} + \textstyle\sum_{p=2}^{12}\beta_{\mathrm{pj}}(COV)_{\mathrm{ij}} + \varepsilon_{\mathrm{ij}},$$

where Y_{ij}, a posttest score for student i in school j, was modeled as a function of intercept (β_{0j}), for mean posttest score in school j, MAP reading pretest (*PREMAP*; $\beta_{1\mathrm{j}}$), and within-school demographic covariates (*COV*; β_{pj}). The model also included the level 1 random effect, $\varepsilon_{\mathrm{ij}}$, assumed to be normally distributed with a mean of zero.

Level 2 (between-school) model was specified to examine the treatment effects, accounting for heterogeneity across seven randomization blocks, which was expressed as follows:

$$\text{Level 2}: \beta_{0\mathrm{j}} = \gamma_{00} + \gamma_{01}(TREATMENT)_{\mathrm{j}} + \textstyle\sum_{q=2}^{8}\gamma_{0\mathrm{q}}(RANDOM)_{\mathrm{jq}} + \tau_{0\mathrm{j}}$$

$$\beta_{1\mathrm{j}} = \gamma_{10}\ [\mathrm{RQ1}]\ \text{or}$$

$$\beta_{1\mathrm{j}} = \gamma_{10} + \gamma_{11}(TREATMENT)_{\mathrm{j}}\ [\mathrm{RQ2}]$$

$$\beta_{\mathrm{pj}} = \gamma_{\mathrm{p0}} \; (p = 2, 3, \ldots, 12)$$

where the intercept for grand-mean posttest outcome ($\beta_{0\mathrm{j}}$) was modeled as a function of the average of the school means on the posttest score (γ_{00}), treatment effect (*TREATMENT*; γ_{01}), effects of dummy-coded randomization blocks (*RANDOM*; $\gamma_{0\mathrm{q}}$), and level 2 random effect ($\tau_{0\mathrm{j}}$). The within-school covariates were specified as fixed effects ($\beta_{1\mathrm{j}}$ and β_{pj}) across schools and γ_{10} and γ_{p0} represented the pooled within-school regression coefficients for MAP reading pretest and demographic covariates, respectively. Parameter γ_{11} represented the interaction effect between the intervention treatment and MAP reading pretest on the posttest outcome. Finally, we computed an effect size (i.e., covariate-adjusted Cohen's *d*) by taking the parameter estimates for treatment variable, γ_{01}, and dividing each estimate by the unadjusted pooled within-group standard deviation. The effect size metric captures the treatment-control difference in standard deviation units and facilitates comparison of the magnitude of the estimated treatment effect to other interventions.

Exploratory Analyses of Mediation Effects We conducted multilevel mediation analyses to further explore to what extent students' vocabulary knowledge depth mediated the treatment effects on reading comprehension and basic literacy skills and science and social studies argumentative writing (RQ3). Figure 1 displays a conceptual framework of the multilevel mediation model. We estimated the direct, indirect, and total effects (Baron and Kenny 1986; Krull and MacKinnon 1999) of treatment on reading and argumentative writing outcomes. We used bias-corrected bootstrapped standard errors using 1000 draws to create our 95% confidence intervals (CIs) and included covariates for MAP reading pretest, student demographics, and school randomization blocks.

Results

Preliminary Analysis

Descriptive and Correlational Analyses Table 3 reports descriptive statistics and correlation matrix of pretest and posttest measure variables by treatment conditions. There were no statistically significant differences in MAP reading pretest ($\beta = -1.19, p = 0.63$) and DIBELS pretest ($\beta = -2.07, p = 0.88$) for the baseline sample of students in the treatment and control groups. The correlational analyses indicated that both science and social studies vocabulary knowledge depth (i.e., total, taught, and untaught words) were positively and moderately correlated with argumentative writing (*rs* = 0.18–0.41), MAP reading posttest (*rs* = 0.54–0.66), and DIBELS posttest (*rs* = 0.47–0.64). The magnitude of the correlations with vocabulary knowledge depth for the treatment group was greater than the control group. Within the treatment and control groups, the correlations of vocabulary knowledge depth with social studies argumentative writing were higher than those with science argumentative writing in magnitude.

Table 3 Pairwise correlation matrix and descriptive statistics for student assessment measures in treatment (below diagonal) and control (above diagonal) condition

Variable	1	2	3	4	5	6	7	8	9	10	11	12	*n* (control)	*M*	*SD*	Min.	Max.
Science vocabulary knowledge depth																	
1 Total words		.90	.86	.61	.52	.53	.60	.61	.54	.56	.26	.33	2362	33.83	6.63	13	48
2 Taught words	.92		.70	.56	.48	.49	.54	.56	.49	.52	.23	.28	2362	19.99	4.29	4	28
3 Untaught words	.87	.71		.54	.46	.47	.53	.54	.48	.50	.20	.31	2362	14.21	3.46	2	20
Social studies vocabulary knowledge depth																	
4 Total words	.64	.60	.54		.88	.84	.63	.64	.55	.58	.24	.36	2299	32.88	6.72	10	48
5 Taught words	.54	.50	.45	.89		.49	.53	.54	.45	.47	.24	.32	2299	19.34	4.16	6	28
6 Untaught words	.59	.56	.50	.85	.52		.56	.56	.50	.53	.18	.30	2299	13.54	3.61	2	20
Reading outcomes																	
7 MAP pretest	.66	.63	.56	.63	.54	.56		.90	.79	.80	.39	.47	2608	176.76	17.39	103	227
8 MAP posttest	.66	.63	.56	.65	.56	.58	.90		.79	.80	.41	.48	2542	181.88	16.90	115	226
9 DIBELS pretest	.60	.57	.51	.58	.50	.51	.82	.81		.92	.37	.46	2608	210.12	116.08	0	482
10 DIBELS posttest	.64	.62	.56	.59	.50	.55	.82	.82	.91		.33	.46	2573	228.00	114.91	0	521
Argumentative writing																	
11 Science	.29	.28	.22	.35	.34	.27	.34	.37	.32	.27		.39	1367	3.08	1.73	0	7
13 12 Social studies	.42	.38	.34	.41	.38	.32	.45	.46	.42	.39	.46		1278	3.23	1.89	0	7
n (treatment)	2601	2601	2601	2465	2465	2465	2886	2835	2886	2843	1381	1272					
M	36.84	21.77	15.55	36.45	22.13	14.33	175.16	180.25	203.52	220.89	3.54	4.05					
SD	6.87	4.37	3.38	7.03	4.38	3.67	18.21	17.55	113.53	116.62	1.81	2.02					
Min.	13	5	4	11	5	4	116	114	0	0	0	0					
Max.	48	28	20	48	28	20	225	228	527	518	7	7					

MAP Measures of Academic Progress, *DIBELS* Dynamic Indicators of Basic Early Literacy Skills

Attrition Analysis and Pretest Equivalence Post-randomization attrition occurred because some students were present at time of pretest assessment, but they left the school or were withdrawn from the study such that they did not complete the posttest. To examine potential threats to internal validity, we first examined attrition rates by condition on the MAP reading analyses. There was no statistically significant difference in attrition rates based on treatment condition, $\chi^2(1, N = 5494) = 0.04$, $p = 0.84$. Overall, 98% of the baseline sample of treatment students and 97% of the baseline sample of control students were included in the analyses. Thus, the final analytic sample included 2835 treatment group students and 2542 control group students.

We also conducted analyses to examine pretest equivalence for students who were included in the HLM models for the vocabulary knowledge depth and argumentative writing outcomes, which were administered only at posttest. For students with both vocabulary knowledge depth outcomes, there was no statistically significant difference on MAP reading pretest scores students in the treatment group ($n = 2352$, $M = 176.02$) and the control group ($n = 2246$, $M = 177.37$; $p = 0.49$). In addition, for students with both argumentative writing outcomes, there was no difference on MAP reading pretest scores for the treatment group ($n = 1161$, $M = 177.63$) and the control group ($n = 1206$, $M = 178.24$; $p = 0.59$). In sum, these results indicate that attrition rates were comparable by condition and pretest reading scores were equivalent for the treatment and control group students included in each analytic sample.

Research Question 1: Main Effects of MORE Intervention on Student Outcomes

Table 4 presents the results of HLM analysis estimating the main effects on science and social studies vocabulary knowledge depth, reading outcomes, and argumentative writing. The treatment significantly improved students' science and social studies vocabulary knowledge depth (ES = 0.50 and ES = 0.56, respectively), after controlling for student-level demographic covariates, MAP reading pretest, and school randomization blocks. Translating the effect sizes for science vocabulary (ES = 0.50) and social studies vocabulary knowledge depth (ES = 0.56) into raw units, the treatment group students learned approximately 3.46 more science words and 2.87 more social studies words than control group students.

The intervention increased both science and social studies vocabulary knowledge depth by similar magnitudes. The positive effects were more likely attributable to treatment students' knowledge of explicitly taught words in science and social studies (ES = 0.48 and ES = 0.64, respectively) than of the words not directly or explicitly taught (ES = 0.45 and ES = 0.28, respectively).

We did not find evidence of a statistically significant treatment effect on both reading outcomes, including the covariate-adjusted posttest measure of reading comprehension (MAP reading) and basic literacy skills (DIBELS). However, we also found the statistically significant treatment effect on students' argumentative writing in science (ES = 0.24) and social studies (ES = 0.44). The intervention impact on social studies argumentative writing was relatively larger in magnitude than the effect on science argumentative writing. In summary, the intervention significantly increased vocabulary knowledge depth and argumentative writing skills.

Table 4 Results of hierarchical linear models predicting main effect of treatment on science and social studies vocabulary knowledge depth, reading outcomes, and argumentative writing

	Coefficient (SE)									
	Science vocabulary knowledge depth			Social studies vocabulary knowledge depth			Reading outcomes		Argumentative writing	
Source	Total words	Taught words	Untaught words	Total words	Taught words	Untaught words	MAP	DIBELS	Science	Social studies
Fixed effect										
Intercept	− .30*** (.04)	− .28*** (.04)	− .27*** (.04)	− .33*** (.03)	− .37*** (.03)	− .17*** (.03)	− .03* (.02)	− .05 (.03)	− .17*** (.04)	− .28*** (.04)
Treatment[a]	.50*** (.05)	.48*** (.05)	.45*** (.06)	.56*** (.04)	.64*** (.04)	.28*** (.04)	− .01 (.02)	.02 (.04)	.24*** (.05)	.44*** (.05)
Variance components										
Level 1 (within school)	.58 (.01)	.64 (.01)	.70 (.01)	.55 (.01)	.61 (.01)	.70 (.01)	.20 (.00)	.37 (.01)	.73 (.02)	.69 (.02)
Level 2 (between school)	.04 (.01)	.03 (.01)	.04 (.01)	.02 (.01)	.02 (.00)	.01 (.00)	.00 (.00)	.02 (.01)	.03 (.01)	.03 (.01)
N	4963	4963	4963	4764	4764	4764	5377	5416	2748	2550

Note: Within-school demographic characteristics (i.e., gender, race/ethnicity, grade, limited english proficiency/individual education plan status, socioeconomic status), Measures of Academic Progress (MAP) reading pretest score, and school randomization blocks were included in HLM analyses as covariates but suppressed in the table. Standard errors in parentheses

DIBELS Dynamic Indicators of Basic Early Literacy Skills

[a] Treatment coded as a dichotomous variable (control group = 0, treatment group = 1)

$*p < 0.05$, $**p < 0.01$, $***p < 0.001$

Research Question 2: Moderating Effects of Initial Reading Ability

We tested potential Matthew effects by examining whether initial reading ability, measured by MAP reading test prior to the intervention, moderated the intervention effect on posttest outcomes. Notably, there was no evidence of a Matthew effect on the vocabulary knowledge depth outcome (total words) for either science or social studies. There was mixed evidence of Matthew effects on the taught and untaught words outcomes. For example, the MAP reading pretest was a statistically significant moderator for science vocabulary knowledge depth in untaught words (ES = − 0.09), suggesting that the treatment effect was larger among initially lower performing students than for initially higher performing students. However, there was evidence of a Matthew effect on social studies vocabulary knowledge depth in taught words (ES = 0.07); for this outcome, the treatment effect was smaller about initially lower performing students than initially higher-performing students. These results imply that Matthew effects were not replicated across the same vocabulary depth outcomes in science and social studies. In addition, there was no evidence of a treatment-by-pretest interaction effect on MAP or DIBELS measures.

We found partial evidence of a Matthew effect in the argumentative writing outcomes. Specifically, there was evidence of a treatment-by-pretest interaction effect for argumentative writing in science (ES = 0.09, $p < 0.05$), but not for social studies (ES = 0.05; $p > 0.05$). In Fig. 2, we present this interaction effect graphically for the science argumentative writing outcome. The average science argumentative writing score difference between the treatment and control groups was twice larger for initially higher-performing students (1SD above the mean; ES = 0.31) than for initially lower-performing students (1SD below the mean; ES = 0.14). This finding suggests that the treatment effect was much greater among initially higher-performing students than their counterparts. In terms of social studies argumentative writing, as shown in Fig. 3, the differences between the treatment and control groups among lower-performing

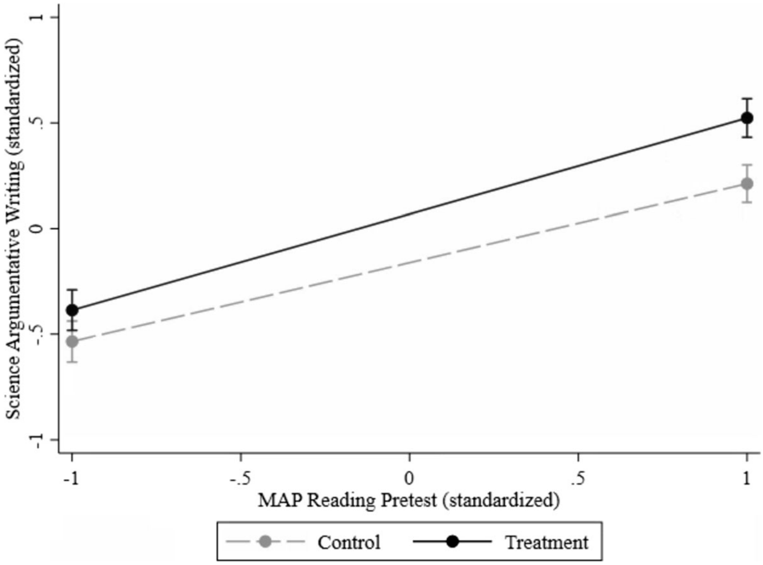

Fig. 2 Interaction effect between treatment and Measures of Academic Progress (MAP) reading pretest on science argumentative writing

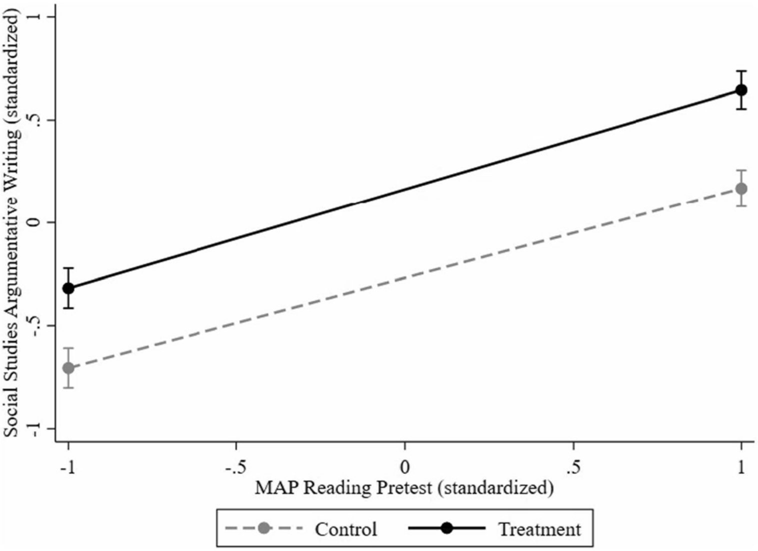

Fig. 3 Interaction effect between treatment and Measures of Academic Progress (MAP) reading pretest on social studies argumentative writing

students (ES = 0.36) and higher-performing students (ES = 0.48) were similar, which indicate a non-significant treatment-by-pretest interaction effect.

Research Question 3: Mediational Effects of Vocabulary Knowledge Depth

Table 5 shows the results of analyses for the multilevel mediation models of which a mediator was (a) domain-specific vocabulary knowledge depth (the average of science and social studies total vocabulary knowledge measures), (b) science vocabulary knowledge depth, or (c) social studies vocabulary knowledge depth. In the mediation model with overall domain-specific vocabulary knowledge depth as a mediator, the coefficients of path *a* and *b* were all positively and statistically significant (*ps* < 0.05), indicating that the treatment group children had significantly greater domain-specific vocabulary knowledge depth than the counterparts and improved vocabulary knowledge depth positively and significantly predicted MAP and DIBELS reading outcomes and science and social studies argumentative writing. The direct effects of treatment (path *c'*) were positively and statistically significantly only on science and social studies argumentative writing (ES = 0.15; 95% CI = 0.08, 0.23 and ES = 0.31; 95% CI = 0.23, 0.38, respectively) and the effect size for social studies argumentative writing was twice as large as the one for science argumentative writing. All indirect effects (path *ab*) were statistically significant (*ps* < 0.05), suggesting that overall domain-specific vocabulary knowledge depth mediated the treatment effects on MAP and DIBELS reading outcomes and science and social studies argumentative writing with the effect sizes ranging from 0.09 to 0.13. Notably, the main effects of treatment on MAP and DIBELS reading outcomes were not statistically significant (see Table 4), but when domain-specific vocabulary knowledge depth was specified as a mediator in the multilevel mediational models, the indirect effects were statistically significant (*ps* < 0.05). Similar patterns were observed in the multilevel mediation model with science or social studies vocabulary knowledge depth as a mediator.

Table 5 Results of multilevel mediation path analysis

	Reading outcomes				Argumentative writing outcomes			
	MAP		DIBELS		Science		Social studies	
Path	*β* (*SE*)	95% Bootstrap CI	*β* (*SE*)	95% Bootstrap CI	*β* (*SE*)	95% Bootstrap CI	*β* (*SE*)	95% Bootstrap CI
Mediator: Domain-specific (science & social studies) vocabulary knowledge depth								
Path *a*	.59 (.02)	.55 to .63	.59 (.02)	.55 to .63	.59 (.02)	.55 to .63	.59 (.02)	.55 to .63
Path *b*	.15 (.01)	.13 to .17	.15 (.01)	.12 to .17	.14 (.02)	.10 to .19	.21 (.02)	.17 to .26
Direct effect (path *c'*)	−.10 (.02)	−.13 to −.07	−.07 (.02)	−.11 to −.03	.15 (.04)	.08 to .23	.31 (.04)	.23 to .38
Indirect effect (path *ab*)	.09 (.01)	.08 to .10	.09 (.01)	.07 to .10	.09 (.01)	.06 to .11	.13 (.01)	.10 to .16
Total effect (*ab* + *c'*)	−.01 (.01)	−.04 to .01	.02 (.02)	−.02 to .05	.24 (.04)	.17 to .31	.43 (.03)	.37 to .50
Mediator: Science vocabulary knowledge depth								
Path *a*	.50 (.02)	.46 to .55	.50 (.02)	.45 to .55	.50 (.02)	.46 to .55	.50 (.02)	.46 to .54
Path *b*	.12 (.01)	.10 to .14	.12 (.01)	.09 to .14	.13 (.02)	.09 to .18	.18 (.02)	.14 to .22
Direct effect (path *c'*)	−.07 (.01)	−.10 to −.05	−.05 (.02)	−.09 to −.01	.17 (.03)	.10 to .24	.33 (.03)	.26 to .40
Indirect effect (path *ab*)	.06 (.01)	.05 to .07	.06 (.01)	.04 to .07	.07 (.01)	.04 to .09	.09 (.01)	.07 to .11
Total effect (*ab* + *c'*)	−.01 (.01)	−.04 to .01	.01 (.02)	−.03 to .05	.23 (.03)	.17 to .30	.42 (.03)	.36 to .49
Mediator: Social studies vocabulary knowledge depth								
Path *a*	.56 (.02)	.51 to .60	.56 (.02)	.51 to .60	.56 (.02)	.52 to .60	.56 (.02)	.52 to .60
Path *b*	.13 (.01)	.11 to .15	.12 (.01)	.10 to .15	.09 (.02)	.05 to .14	.16 (.02)	.12 to .20
Direct effect (path *c'*)	−.08 (.01)	−.11 to −.05	−.05 (.02)	−.09 to −.01	.17 (.04)	.09 to .24	.35 (.04)	.28 to .43
Indirect effect (path *ab*)	.07 (.01)	.06 to .08	.07 (.01)	.05 to .08	.05 (.01)	.03 to .08	.09 (.01)	.06 to .12
Total effect (*ab* + *c'*)	−.01 (.01)	−.04 to .02	.02 (.02)	−.02 to .05	.22 (.03)	.15 to .29	.44 (.04)	.37 to .51

Each multilevel mediation model was estimated for the following five pathways: (a) Path a to examine the intervention (level-2 predictor) effect on vocabulary knowledge depth (level-1 mediator); (b) path b to test the effect of vocabulary knowledge depth on the outcomes; (c) path c' as the direct effect of the treatment on the outcomes; (d) the coefficient of path ab (a × b) indicating the indirect effect of intervention on the outcomes through vocabulary knowledge depth; and (e) path c indicating the total effect of the treatment on the outcomes. In each model, MAP reading pretest score, student-level demographic covariates, and school randomization blocks were included as covariates. If the 95% bootstrap confidence interval (CI) does not contain the null value (zero), the effect is statistically significant at $p < 0.05$. Standard errors in parentheses. MAP = Measures of Academic Progress. DIBELS = Dynamic Indicators of Basic Early Literacy Skills

Discussion

To strengthen the evidence base for elementary-grade content literacy instruction, we conducted an earlier randomized controlled trial of the MORE content literacy intervention (Kim et al. 2021). Although our previous study of the MORE science lessons produced positive effects on students' vocabulary, reading, and writing outcomes, replication studies are needed to determine whether novel results from a single study are robust enough to support evidence-based instructional recommendations at scale (Bollen et al. 2015; Makel and Plucker 2014).

Guided by the logic model in Fig. 1, we undertook this conceptual replication (a) to replicate the positive effects of MORE in science and to extend effects to social studies, (b) to examine the moderating role of prior reading ability, and to (c) to examine the mediating role of vocabulary knowledge depth. In brief, the intervention produced positive effects on students' domain-specific measure of vocabulary knowledge depth and argumentative writing. However, there were no positive and statistically significant effects on domain-general measures of reading comprehension and mixed evidence of moderation based on prior reading. Finally, students' vocabulary knowledge depth partially mediated the treatment effects on argumentative writing. In the following sections, we discuss the broader implications of each of the main findings and suggest future research directions.

Replicating and Extending Main Effects in Science and Social Studies

The results of our confirmatory analyses suggest that a longer program implementation of MORE involving both science and social studies lesson improved students' domain-specific measures of vocabulary and writing and facilitated transfer effects. Evidence of replication was clearly observed on the science measures where the effect size on argumentative writing and vocabulary knowledge depth in our previous efficacy study were repeated in this study. These results were also extended into the social studies vocabulary and writing outcomes. Finally, our previous study did not produce positive effects on students' knowledge of transfer measures of vocabulary knowledge depth whereas this study did. How do we explain this novel result?

There are key differences between our previous study and the current study that highlight the conditions under which transfer is likely to occur. In contrast to our previous efficacy study, students in this study were taught networks of vocabulary words in both science and social studies. Thus, direct teaching of semantically related words across *two* domains may foster incidental word learning and nurture larger and more elaborate networks of vocabulary knowledge. Such opportunities are critical to facilitating incidental learning of word meanings (Jenkins et al. 1984; Swanborn and de Glopper 1999) and creating a language-rich classroom context that promotes transfer of word learning (Beck et al. 2002; Perfetti 2007; Snow et al. 2009; Stahl and Fairbanks 1986). Moving from partial to full knowledge of a word implies that students can activate connections among semantically related words and also make inferences beyond content that was directly taught (Bolger et al. 2008; Fitzgerald et al. 2020; Graves 2016; Neuman, Newman, & Dwyer, 2011). Put differently, these findings support the notion that teaching networks of vocabulary provide mental hooks that facilitate students' understanding of associated words that are learned incidentally in read-alouds and discussions of science and social studies texts (Hirsch, 2016; Perfetti 2007). Words that are a part of a dense semantic network enable students to efficiently acquire new words and to expand and differentiate knowledge schemas, leading to a virtuous cycle of ongoing vocabulary and domain knowledge growth (Borovsky et al. 2016; Graves 2016; Perfetti 2007; Steyvers and Tenenbaum 2005).

In particular, treatment group students learned six more words (3.46 science words, 2.87 social studies words), on average, than control group students. These are educationally significant gains because deep knowledge of domain-specific academic words represents the visible tip of the conceptual knowledge iceberg (Anderson and Freebody 1981). In all likelihood, the gain in students' depth of vocabulary knowledge is an indication of deeper knowledge of related science and social studies concepts that enable young children to continue learning words incidentally while reading, writing, and discussing new texts (Nagy 2007; Wright and Cervetti 2016). In essence, providing children with repeated exposures to vocabulary networks in science and social studies may facilitate transfer and generative word learning in both domains.

At the same time, the intervention had smaller effects on untaught words in social studies (ES = 0.28) than science (ES = 0.45). There are several potential explanations for this finding. It may be that learning social studies words involves reading sophisticated lexical and grammatical aspects of social studies texts (Schleppegrell 2004) and more time to master expository text structures (Williams et al. 2016). Consistent with these explanations, the social studies texts used in the MORE read-aloud lessons had higher text complexity levels (876.92L) than the science texts (738.67L). As a result, the smaller effect on social studies untaught words than science untaught words may reflect the nature and complexity of the social studies texts used in MORE read-aloud lessons. Given these findings, it remains open how best to create a language-rich classroom context that promotes transfer of word learning during social studies instruction.

Furthermore, there was evidence of transfer on domain-specific argumentative writing outcomes. Findings indicated that treatment group students enjoyed an edge over control group students on transfer measures of argumentative writing in science (ES = 0.24) and social studies (ES = 0.44). Because the writing task required students to draw on both discourse knowledge of argumentation and science and social studies domain knowledge, superior performance on this task implies that treatment group students were more skillful in retrieving schemas and then using them to solve a new problem. There are several core components in MORE that were designed to foster transfer. For example, students learned how to apply the schema for *animal survival* to many situations, ranging from writing tasks about taught topics (e.g., *Arctic animal survival*) to untaught topics (e.g., *rainforest animals*). The larger effect size on argumentative writing in social studies (ES = 0.44) than science (ES = 0.24) also suggests that content literacy instruction created a stronger treatment-control contrast in children's ability to learn a social studies schema focused on shared characteristics of *explorers* and *inventors*. Given the larger gains in taught words in social studies versus science, students were able to retrieve text-based ideas and concepts acquired during the MORE lessons while writing argumentative essays in social studies. Moreover, the stronger correlations between vocabulary knowledge depth and argumentative writing in social studies relative to science further suggest that treatment group students were able to apply their vocabulary knowledge while writing higher quality social studies essays.

Recently, researchers have shown that prior domain knowledge is a critical resource that helps students retrieve and use information in source texts as they write stronger argumentative essays (e.g., Wijekumar et al. 2019). More generally, our findings fill important research gaps on the effectiveness of content literacy instruction in first and second grades where there is virtually no experimental evidence on argumentative writing outcomes (Graham et al. 2016;

Graham et al. 2020). Therefore, our findings suggest that content literacy instruction may help young children acquire the discourse and domain knowledge needed to write superior argumentative essays that are critical to knowledge acquisition. Serious argumentative writing forms the foundation for domain knowledge acquisition and later academic success (Graham et al. 2020).

The positive impacts on argumentative writing were not echoed in the reading comprehension outcomes, however. As a result, there is currently insufficient evidence to support the claim that content literacy instruction can reliably improve reading comprehension. Our findings are consistent with other large-scale efforts to K-3 reading comprehension outcomes (e.g., Neuman et al. 2007; RAND Study Group 2002) and upper elementary-level reading achievement (e.g., Rimm-Kaufman et al. 2014). For example, situation model building during reading undoubtedly depends on prior domain knowledge, but such efforts may accumulate over time and require multiple years of intervention (Stanovich 1986).

The Moderating Role of Initial Reading Ability Our second research aim was to explore the moderating role of initial reading comprehension ability on student outcomes. These analyses yielded inconsistent evidence of Matthew effects. For example, we found no evidence of Matthew effects on science vocabulary knowledge depth, but Matthew effects were evident in social studies vocabulary knowledge depth, particularly for target words that students explicitly learned during the intervention lessons. Initially, lower performing students enjoyed larger benefits than initially higher performing students on science vocabulary, whereas initially higher-performing readers made greater gains in social studies words than initially lower-performing readers. This finding is noteworthy because Matthew effects are commonly observed in more narrowly focused vocabulary interventions (e.g., Coyne et al. 2019; Marulis and Neuman 2013). At minimum, our findings indicate that integrating vocabulary learning in the content of a whole class content literacy intervention may enhance the quantity and quality of word learning opportunities that ultimately foster transfer effects on untaught words (Apthorp et al. 2012; Coyne et al., 2010).

However, there was some evidence of Matthew effects in science argumentative writing. Our findings converge with Wood and colleague's study (2020) that found Matthew effects in writing skills among second-grade students, such that initially higher-performing readers enjoyed larger gains in writing than initially lower-performing readers. It may be that good readers are more likely than poor readers to have higher-quality lexical representations that amplify the positive effects of content literacy instruction on argumentative writing outcomes (Dobbs and Kearns 2016; Kellogg 2008; Perfetti and Hart 2002). Taken as a whole, however, the moderator results do not provide consistent evidence of Matthew effects.

The Mediating Role of Vocabulary Knowledge Depth

Our exploratory aim was to examine whether and to what extent students' vocabulary knowledge depth mediated the intervention treatment effects. The results suggest that the MORE treatment had both direct effects on writing and indirect effects via improvements on vocabulary knowledge depth; these results replicated across the entire writing measure and separately for science and social studies. Accordingly, these findings indicate that students were able to transfer their newly acquired vocabulary knowledge to a domain-specific writing task (Kendeou et al. 2003; Kimball and Holyoak 2000). These results are consistent with our theoretical proposition that access to a greater variety of domain vocabulary frees cognitive resources to help novice learners plan, organize, and write higher-quality argumentative essays (Alexander 2003; Galbraith and Baaijen 2018; McCutchen 1996).

However, these mediational effects were not replicated with reading outcomes. In short, intervention effects on reading outcomes were characterized by significant indirect effects that render the direct effects nonsignificant. These findings suggest that it may be important to develop domain-specific measures of reading comprehension in addition to domain-general measures typically used in intervention studies (Pearson et al. 2020). Therefore, one possibility is that domain-general reading comprehension measures, such as those used in this study, are not sensitive to the science and social studies lessons used in MORE classrooms.

Limitations and Future Research Directions

The study limitations suggest several fruitful areas for future research. For example, given the failure to replicate effects on reading comprehension, future research should examine whether content literacy intervention can foster transfer on both near and far transfer measures of reading comprehension. Furthermore, the mediational analyses revealed no significant direct effect of treatment on reading comprehension. Given this finding, teachers may need to include additional reading activities to help students to flexibly and rapidly access the semantic networks of words that are critical for understanding science and social studies texts (Duke et al. 2006; Hirsch Jr. 2010–2011; Vaughn et al. 2013; Williams et al. 2016). In addition, it is difficult to pinpoint whether and to what extent the MORE intervention or the additional time treatment group teachers spent on science and social studies instruction led to improvements in students' vocabulary knowledge depth and argumentative writing outcomes.

Finally, there is a clear need to determine whether a multi-year intervention can improve reading comprehension outcomes. Unconstrained competencies like reading comprehension ability accumulate slowly over time after several years of intervention (Paris 2005; Stanovich 1986). More research is needed to understand whether multi-year content literacy interventions can improve vocabulary knowledge depth and promote transfer on new reading comprehension tasks (Barnett and Ceci 2002; Pearson et al. 2020). For example, several research syntheses have underscored the need for studies that go beyond a single year and shed light on the longitudinal impact of vocabulary and content literacy interventions on children's reading comprehension (Cabell and Hwang 2020; Wright and Cervetti 2016). Given this research gap, it remains an open question whether, over time, elementary grade students can transfer their networks of vocabulary knowledge to new texts and acquire vocabulary incidentally during reading (Nagy 2005; Pressley et al. 2007). With sustained implementations of the intervention that continue for a longer period of time, it will be possible to examine transfer effects on both domain-specific and domain-general measures of reading comprehension.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10648-021-09609-6.

Funding This research was funded by the Chan Zuckerberg Initiative. The opinions expressed are those of the authors and do not represent the views of the funding agency.

Declarations Not applicable.

References

Alexander, P. A. (2000). Research news and comment: Toward a model of academic development: schooling and the acquisition of knowledge. *Educational Researcher, 29*(2), 28–44.

Alexander, P. A. (2003). The development of expertise: The journey from acclimation to proficiency. *Educational Researcher, 32*(8), 10–14. https://doi.org/10.3102/0013189X032008010.

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 77–117). Newark, DE: International Reading Association.

Apthorp, H., Randel, B., Cherasaro, T., Clark, T., McKeown, M., & Beck, I. (2012). Effects of a supplemental vocabulary program on word knowledge and passage comprehension. *Journal of Research on Educational Effectiveness, 5*(2), 160–188.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin, 128*(4), 612–637. https://doi.org/10.1037/0033-2909.128.4.612.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173.

Beck, I. L., Perfetti, C. A., & McKeown, M. G. (1982). Effects of long-term vocabulary instruction on lexical access and reading comprehension. *Journal of Educational Psychology, 74*(4), 506–521. https://doi.org/10.1037/0022-0663.74.4.506.

Beck, I., McKeown, M., & Kucan, L. (2002). *Bringing words to life*. New York: The Guilford Press.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ: Erlbaum.

Bolger, D. J., Balass, M., Landen, E., & Perfetti, C. A. (2008). Contextual variation and definitions in learning the meaning of words. *Discourse Processes, 45*(2), 122–159. https://doi.org/10.1080/01638530701792826.

Bollen, K., Cacioppo, J. T., Kaplan, R. M., Krosnick, J. A., & Olds, J. L. (2015). *Social, behavioral, and economic sciences perspectives on robust and reliable science: Report of the Subcommittee on Replicability in Science, Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences*. Retrieved from the National Science Foundation Web site: www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf

Borovsky, A., Ellis, E. M., Evans, J. L., & Elman, J. L. (2016). Lexical leverage: Category knowledge boosts real-time novel word recognition in 2-year-olds. *Developmental Science, 19*(6), 918–932. https://doi.org/10.1111/desc.12343.

Brown, R. S., & Coughlin, E. (2007). *The predictive validity of selected benchmark assessments used in the Mid-Atlantic Region* (Issues & Answers Report, REL 2007–No. 017). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. Retrieved from http://ies.ed.gov/ncee/edlabs

Cabell, S. Q., & Hwang, H. (2020). Building content knowledge to boost comprehension in the primary grades. *Reading Research Quarterly, 55*, S99–S107.

Carlisle, J. F., Fleming, J. E., & Gudbrandsen, B. (2000). Incidental word learning in science classes. *Contemporary Educational Psychology, 25*(2), 184–211.

Cervetti, G. N., Wright, T. S., & Hwang, J. (2016). Conceptual coherence, comprehension, and vocabulary acquisition: A knowledge effect? *Reading and Writing: An Interdisciplinary Journal, 29*(4), 761–779.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*(6), 407–428.

Connor, C. M., Dombek, J., Crowe, E. C., Spencer, M., Tighe, E. L., Coffinger, S., Zargar, E., Wood, T., & Petscher, Y. (2017). Acquiring science and social studies knowledge in kindergarten through fourth grade: Conceptualization, design, implementation, and efficacy testing of content-area literacy instruction (CALI). *Journal of Educational Psychology, 109* (3), 301–320. https://doi.org/10.1037/edu0000128.

Coyne, M. D., McCoach, D. B., Loftus, S., Zipoli Jr., R., & Kapp, S. (2009). Direct vocabulary instruction in kindergarten: Teaching for breadth versus depth. *The Elementary School Journal, 110*(1), 1–18. https://doi.org/10.1086/598840.

Coyne, M. D., McCoach, D. B., Loftus, S., Zipoli, R., Ruby, M., Crevecoeur, Y., & Kapp, S. (2010). Direct and extended vocabulary instruction in kindergarten: Investigating transfer effects. *Journal of Research on Educational Effectiveness, 3*(2), 93–120. https://doi.org/10.1080/19345741003592410.

Coyne, M. D., McCoach, D. B., Ware, S., Austin, C. R., Loftus-Rattan, S. M., & Baker, D. L. (2019). Racing against the vocabulary gap: Matthew effects in early vocabulary instruction and intervention. *Exceptional Children, 85*(2), 163–179.

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control. *Clinical Psychology Review, 18*(1), 23–45. https://doi.org/10.1016/S0272-7358(97)00043-3.

Dobbs, C. L., & Kearns, D. (2016). Using new vocabulary in writing: Exploring how word and learner characteristics relate to the likelihood that writers use newly taught vocabulary. *Reading and Writing: An Interdisciplinary Journal, 29*(9), 1817–1843. https://doi.org/10.1007/s11145-016-9654-8.

Duke, N. K., Purcell-Gates, V., Hall, L. A., & Tower, C. (2006). Authentic literacy activities for developing comprehension and writing. *The Reading Teacher, 60*(4), 344–355. https://doi.org/10.1598/RT.60.4.4.

Elleman, A. E., Lindo, E. J., Morphy, P., & Compton, D. L. (2009). The impact of vocabulary instruction on passage-level comprehension of school-age children: A meta-analysis. *Journal of Research on Educational Effectiveness, 2*(1), 1–44.

Ericsson, K. A. (2018). *Superior working memory in experts*. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), *Cambridge handbooks in psychology. The Cambridge handbook of expertise and expert performance (p. 696–713)*. Cambridge University Press. https://doi.org/10.1017/9781316480748.036.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102*(2), 211–245.

Ericsson, K. A., & Pool, R. (2016). *Peak: Secrets from the new science of expertise*. Boston: Houghton Mifflin Harcourt.

Fitzgerald, J., Elmore, J., Relyea, J. E., & Stenner, A. J. (2020). Domain-specific academic vocabulary network development in elementary grades core disciplinary textbooks. *Journal of Educational Psychology, 112*(5), 855–879. https://doi.org/10.1037/edu0000386.

Galbraith, D., & Baaijen, V. M. (2018). The work of writing: Raiding the inarticulate. *Educational Psychologist, 53*(4), 238–257. https://doi.org/10.1080/00461520.2018.1505515.

Gelman, S. A. (2009). Learning from others: Children's construction of concepts. *Annual Review of Psychology, 60*(1), 115–140.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*(1), 1–38. https://doi.org/10.1016/0010-0285(83)90002-6.

Goffreda, C. T., Diperna, J. C., & Pedersen, J. A. (2009). Preventive screening for early readers: Predictive validity of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS). *Psychology in the Schools, 46*(6), 539–552.

Graham, S., & Harris, K. R. (2005). Improving the writing performance of young struggling writers: Theoretical and programmatic research from the Center on Accelerating Student Learning. *Journal of Special Education, 39*(1), 19–33.

Graham, S., Harris, K. R., & Hebert, M. (2011). It is more than just the message: Analysis of presentation effects in scoring writing. *Focus on Exceptional Children, 44*(4), 1–12.

Graham, S., Harris, K. R., & Chambers, A. (2016). Evidence-based practice and writing instruction. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (Vol. 2, pp. 211–226). New York, NY: Guilford Press.

Graham, S., Harris, K., Wijekumar, K., Lei, P., Barkel, A., Aitken, A., et al. (2018). The roles of writing knowledge, motivation, strategic behaviors, and skills in predicting elementary students' persuasive writing from source material. *Reading and Writing: An Interdisciplinary Journal, 32*(6), 1431–1457. https://doi.org/10.1007/s11145-018-9836-7.

Graham, S., Kiuhara, S. A., & MacKay, M. (2020). The effects of writing on learning in science, social studies, and mathematics: A meta-analysis. *Review of Educational Research, 90*(2), 179–226. https://doi.org/10.3102/0034654320914744.

Graves, M. F. (2016). *The vocabulary book: Learning and instruction*. New York, NY: Teachers College Press.

Guthrie, J. T., & Klauda, S. L. (2014). Effects of classroom practices on reading comprehension, engagement, and motivations for adolescents. *Reading Research Quarterly, 49*(4), 387–416.

Guthrie, J. T., Wigfield, A., Barbosa, P., Perencevich, K. C., Taboada, A., Davis, M. H., Scafiddi, N. T., & Tonks, S. (2004). Increasing reading comprehension and engagement through Concept-Oriented Reading Instruction. *Journal of Educational Psychology, 96*, 403–423. https://doi.org/10.1037/0022-0663.96.3.

Guthrie, J. T., McRae, A., & Klauda, S. L. (2007). Contributions of concept-oriented reading instruction to knowledge about interventions for motivations in reading. *Educational Psychologist, 42*(4), 237–250.

Hayes, J. R., & Flower, L. S. (1986). Writing research and the writer. *American Psychologist, 41*(10), 1106–1113. https://doi.org/10.1037/0003-066X.41.10.

Hirsch Jr., E. D. (2010–2011). Beyond comprehension: We have yet to adopt a common core curriculum that builds knowledge grade by grade—but we need to. *American Educator, 34*(4), 30–36.

Hirsch Jr., E. D. (2016). *Why knowledge matters: Rescuing our children from failed educational theories*. Cambridge, Massachusetts: Harvard Education Press.

Imai, M., Gentner, D., & Uchida, N. (1994). Children's theories of word meaning: The role of shape similarity in early acquisition. *Cognitive Development, 9*(1), 45–75.

Iran-Nejad, A. (1989). A nonconnectionist schema theory of understanding surprise-ending stories. *Discourse Processes, 12*, 127–148. https://doi.org/10.1080/01638538909544723.

Jenkins, J., Stein, M., & Wysocki, K. (1984). Learning vocabulary through reading. *American Educational Research Journal, 21*(4), 767–787.

Kellogg, R. T. (2008). Training writing skills: A cognitive developmental perspective. *Journal of Writing Research, 1*(1), 1–16.

Kendeou, P., Rapp, D. N., & van den Broek, P. (2003). The influence of reader's prior knowledge on text comprehension and learning from text. In R. Nata (Ed.), *Progress in Education, Vol.13* (pp. 189–209). Nova Science Publishers, Inc: New York.

Kim, J. S., Burkhauser, M. A., Mesite, L. M., Asher, C. A., Relyea, J. E., Fitzgerald, J., & Elmore, J. (2021). Improving reading comprehension, science domain knowledge, and reading engagement through a first-grade content literacy intervention. *Journal of Educational Psychology, 113*(1), 3–26. https://doi.org/10.1037/edu0000465

Kimball, D. R., & Holyoak, K. J. (2000). Transfer and expertise. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 109–122). New York, NY: Oxford University Press.

Kintsch, W. (1988). The use of knowledge in discourse processing: A construction-integration model. *Psychological Review, 95*(2), 163–182.

Kintsch, W. (2009). Learning and constructivism. In S. Tobias & T. M. Duffy (Eds.), *Constructivist instruction: Success or failure?* (pp. 223–241). New York, NY: Routledge.

Kintsch, W., & van Dijk, T. A. (1978). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6*, 294–323.

Krull, J. L., & MacKinnon, D. P. (1999). Multilevel meditation modeling for group-based intervention studies. *Evaluation Review, 23*(4), 418–444.

Lepola, J., Poskiparta, E., Laakkonen, E., & Niemi, P. (2005). Development of and relationship between phonological and motivational processes and naming speed in predicting word recognition in grade 1. *Scientific Studies of Reading, 9*(4), 367–399.

Lepola, J., Lynch, J., Kiuru, N., Laakkonen, E., & Niemi, P. (2016). Early oral language comprehension, task orientation, and foundational reading skills as predictors of grade 3 reading comprehension. *Reading Research Quarterly, 51*(4), 373–390.

Levin, J. R. (1988). Elaboration-based learning strategies: Powerful theory = powerful application. *Contemporary Educational Psychology, 13*(3), 191–205. https://doi.org/10.1016/0361-476X(88)90020-3.

Makel, M. C., & Plucker, J. A. (2014). Creativity is more than novelty: Reconsidering replication as a creativity act. *Psychology of Aesthetics, Creativity, and the Arts, 8*(1), 27–29. https://doi.org/10.1037/a0035811.

Maner, J. K. (2016). Into the wild: Field research can increase both replicability and real-world impact. *Journal of Experimental Social Psychology, 66*, 100–106. https://doi.org/10.1016/j.jesp.2015.09.018.

Marulis, L. M., & Neuman, S. B. (2013). How vocabulary interventions affect young children at risk: A meta-analytic review. *Journal of Research on Educational Effectiveness, 6*(3), 223–262.

Marzano, R. J. (2004). *Building background knowledge for academic achievement: Research on what works in schools*. Alexandria, VA: Association for Supervision & Curriculum Development.

McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational Psychology Review, 8*(3), 299–324. https://doi.org/10.1007/BF01464076.

McKeown, M. G., & Beck, I. L. (2011). Making vocabulary interventions engaging and effective. In R. E. O'Connor & P. F. Vadasy (Eds.), *Handbook of reading interventions* (pp. 138–168). New York: Guilford.

McKeown, M. G., Deane, P. D., Scott, J. D., Krovetz, R., & Lawless, R. R. (2017). *Vocabulary assessment to support instruction: Building rich word-learning experiences*. New York, NY: Guilford Press.

Minsky, M. (1975). A framework for representing knowledge. In P. H. Winston (Ed.), *The psychology of computer vision*. New York: McGraw-Hill Book.

Nagy, W. E. (2005). Why instruction needs to be long-term and comprehensive. In E. H. Hiebert & M. L. Kamil (Eds.), *Teaching and learning vocabulary: Bringing research to practice* (pp. 27–44). Chicago, IL: Routledge.

Nagy, W. E. (2007). Metalinguistic awareness and the vocabulary-comprehension connection. In R. K. Wagner, A. E. Muse, & K. R. Tannenbaum (Eds.), *Vocabulary acquisition: Implications for reading comprehension* (pp. 52–77). New York, NY: Guilford Press.

National Center for Education Statistics. (2019). NAEP Report Card: 2019 NAEP Reading Assessment. Retrieved from https://www.nationsreportcard.gov/highlights/reading/2019/

National Research Council. (2012). *A framework for k-12 science education: Practices, crosscutting concepts, and core ideas. Committee on a Conceptual Framework for New K-12 Science Education Standards. Board on Science Education, Division of Behavioral and Social Sciences and Education*. Washington, DC: The National Academies Press.

Nesbit, J. S., & Adesope, O. O. (2006). Learning with concept and knowledge maps: A meta-analysis. *Review of Educational Research, 76*(3), 413–448.

Neuman, S., Dwyer, J., Koh, S., & Wright, T. (2007). *Instructional material: The world of words: A vocabulary intervention for low-income preschoolers*. *Ann Arbor, MI*: *University of Michigan Press*.

Neuman, S. B., Newman, E. H., & Dwyer, J. (2011). Educational effects of a vocabulary intervention on preschoolers' word knowledge and conceptual development: A cluster-randomized trial. *Reading Research Quarterly, 46*, 249–272.

Northwest Evaluation Association. (2011). *RIT scale norms study: For use with measures of academic progress (MAP) for primary grades*. Portland, OR: Author.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015). Promoting an open research culture. *Science, 348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374.

Olinghouse, N. G., Graham, S., & Gillespie, A. (2015). The relationship of discourse and topic knowledge to fifth graders' writing performance. *Journal of Educational Psychology, 107*(2), 391–406.

Paris, S. G. (2005). Reinterpreting the development of reading skills. *Reading Research Quarterly, 40*(2), 184–202.

Parsons, S. A., Malloy, J. A., Parsons, A. W., & Burrowbridge, S. C. (2015). Students' engagement in literacy tasks. *Reading Teacher, 69*(2), 223–231.

Pearson, P. D., Palincsar, A. S., Biancarosa, G., & Berman, A. I. (Eds.). (2020). *Reaping the rewards of the Reading for Understanding Initiative*. Washington, DC: National Academy of Education.

Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading, 11*(4), 357–383.

Perfetti, C. A., & Adlof, S. M. (2012). Reading comprehension: A conceptual framework from word meaning to text meaning. In J. Sabatini & E. Albro (Eds.), *Assessing reading in the 21st century: Aligning and applying advances in the reading and measurement sciences*. Rowman & Littlefield Education: Lanham, MD.

Perfetti, C. A., & Hart, L. (2002). The lexical quality hypothesis. In L. Verhoeven, C. Elbro, & P. Reitsma (Eds.), *Precursors of functional literacy* (pp. 189–213). Philadelphia, PA: John Benjamin.

Pressley, M., Disney, L., & Anderson, K. (2007). Landmark vocabulary instructional research and the vocabulary instructional research that makes sense now. In R. K. Wagner, A. E. Muse, & K. R. Tannenbaum (Eds.), *Vocabulary acquisition: Implications for reading comprehension* (pp. 205–232). New York, NY: Guilford Press.

RAND Study Group. (2002). *Reading for understanding*. Santa Monica CA: RAND.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models*. Thousand Oaks, CA: Sage.

Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. Kunnan (Ed.), *Validation in language assessment* (pp. 41–60). Mahwah, NJ: Erlbaum.

Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined. In B. Laufer & P. Bogaards (Eds.), *Vocabulary in a second language: Selection, acquisition and testing* (pp. 209–227). Amsterdam: John Benjamins.

Reardon, S. F., Valentino, R. A., & Shores, K. A. (2012). Patterns of literacy among U.S. Students. *Future of Children, 22*(2), 17–37.

Rimm-Kaufman, S. E., Larsen, R. A., Curby, T. W., Baroody, A. E., Merritt, E., Abry, T. S., Ko, M., Thomas, J., & DeCoster, J. (2014). Efficacy of the responsive classroom approach: Results from a 3-year, longitudinal randomized controlled trial. *American Education Research Journal, 52*(3), 567–603. https://doi.org/10.3102/0002831214523821.

Romance, N. R., & Vitale, M. R. (2001). Implementing an in-depth expanded science model in elementary schools: Multi-year findings, research issues, and policy implications. *International Journal of Science Education, 23*(4), 373–404.

Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Rumelhart, D. E., & Ortony, A. (1977). The representation of knowledge in memory. In R. C. Anderson, F. J. Spiro, & W. E. Montague (Eds.), *Schooling and the acquisition of knowledge* (p. 1977). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schleppegrell, M. J. (2004). *The language of schooling: A functional linguistics perspective*. Mahway, NJ: Erlbaum http://dx.doi.org/10 .4324/9781410610317.

Schmidt, W. H. (2009). *Exploring the relationship between content coverage and achievement: Unpacking the meaning of tracking in eighth grade mathematics*. East Lansing: Michigan State University, Educational Policy Center.

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning, 64*(4), 913–951.

Snow, C. E., Lawrence, J. F., & White, C. (2009). Generating knowledge of academic language among urban middle school students. *Journal of Research on Educational Effectiveness, 2*(4), 325–244.

Stahl, S. A., & Fairbanks, M. M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research, 56*(1), 72–110.

Stahl, S. A., & Nagy, W. E. (2006). *Teaching word meanings*. Mahwah, NJ: Erlbaum.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*(4), 360–401.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science, 29*(1), 41–78. https://doi.org/10.1207/s15516709cog2901_3.

Strachan, S. (2015). Kindergarten students' social studies and content literacy learning from interactive read-alouds. *The Journal of Social Studies Research, 39*(4), 207–223.

Swanborn, M. S. L., & de Glopper, K. (1999). Incidental word learning while reading: A meta-analysis. *Review of Educational Research, 69*(3), 261–285. https://doi.org/10.3102/00346543069003261.

Thorndyke, P. W. (1984). Applications of schema theory in cognitive research. In J. R. Anderson & S. M. Kosslyn (Eds.), *Tutorials in learning and memory* (pp. 167–192). San Francisco, CA: Freeman.

Trefil, J., Kett, J. F., & Hirsch, E. C. (2002). *The new dictionary of cultural literacy*. Boston: Houghton Mifflin.

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Reply to Inbar: Contextual sensitivity helps explain the reproducibility gap between social and cognitive psychology. *Proceedings of the National Academy of Sciences, 113* (34), E4935–E4936.

Vaughn, S., Swanson, E. A., Roberts, G., Wanzek, J., Stillman-Spisak, S. J., Solis, M., & Simmons, D. (2013). Improving reading comprehension and social studies knowledge in middle school. *Reading Research Quarterly, 48*(1), 77–93.

Wijekumar, K., Graham, S., Harris, K., Lei, P. W., Barkel, A., Aitken, A., Ray, A., & Houston, J. (2019). The roles of writing knowledge, motivation, strategic behaviors, and skills in predicting elementary students' persuasive writing from source material. *Reading and Writing, 32*(6), 1431–1457. https://doi.org/10.1007/s11145-018-9836-7.

Williams, J. P., Kao, J. C., Pao, L. S., Ordynans, J. G., Atkins, J. G., Cheng, R., & DeBonis, D. (2016). Close analysis of texts with structure (CATS): An intervention to teach reading comprehension to at-risk second graders. *Journal of Educational Psychology, 108*(8), 1061–1077. https://doi.org/10.1037/edu0000117.

Wiske, M. S. (1998). What is teaching for understanding? In M. S. Wiske (Ed.), *Teaching for understanding* (pp. 61–86). San Francisco: Jossey-Bass.

Wood, C., Schatschneider, C., & Wanzek, J. (2020). Matthew effects in writing productivity during second grade. Reading and Writing*:* An Interdisciplinary Journal*.* Advance online publication. https://doi.org/10.1007/s11145-019-10001-8,

Wright, T. S., & Cervetti, G. N. (2016). A systematic review of the research on vocabulary instruction that impacts comprehension. *Reading Research Quarterly, 52*, 203–226.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

James S. Kim[1] • Jackie Eunjung Relyea[2] • Mary A. Burkhauser[1] • Ethan Scherer[1] • Patrick Rich[1]

✉ James S. Kim
james_kim@harvard.edu

Jackie Eunjung Relyea
jrelyea@ncsu.edu

Mary A. Burkhauser
mary_burkhauser@gse.harvard.edu

Ethan Scherer
ethan_scherer@gse.harvard.edu

Patrick Rich
patrick_rich@gse.harvard.edu

[1] Graduate School of Education, Harvard University, Cambridge, MA 02138, USA

[2] College of Education, North Carolina State University, Raleigh, USA