ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 23-769

Time to Transfer: Long-Term Effects of a Sustained and Spiraled Content Literacy Intervention in the Elementary Grades

James S. Kim
Harvard University

Joshua B. Gilbert
Harvard University

Jackie E. Relyea
North Carolina State University

Patrick Rich
American Institutes for Research

Ethan Scherer
Harvard University

Mary A. Burkhauser
Harvard University

Johanna N. Tvedt
Harvard University

We investigated the effectiveness of a sustained and spiraled content literacy intervention that emphasizes building domain and topic knowledge schemas and vocabulary for elementary-grade students. The Model of Reading Engagement (MORE) intervention underscores thematic lessons that provide an intellectual structure for helping students connect new learning to a general schema in Grade 1 (animal survival), Grade 2 (scientific investigation of past events like dinosaur mass extinctions), and Grade 3 (scientific investigation of living systems). A total of 30 elementary schools (N = 2,870 students) were randomized to a treatment or control condition. In the treatment condition (i.e., full spiral curriculum), students participated in content literacy lessons from Grades 1 to 3 during the school year and wide reading of thematically related informational texts in the summer following Grades 1 and 2. In the control condition (i.e., partial spiral curriculum), students participated in lessons in only Grade 3. The Grade 3 lessons for both conditions were implemented online during the COVID-19 pandemic school year. Results reveal that treatment students outperformed control students on science vocabulary knowledge across all three grades. Furthermore, intent-to-treat analyses revealed positive transfer effects on Grade 3 science reading (ES = .14), domain-general reading comprehension (ES = .11), and mathematics achievement (ES = .12). Treatment impacts were sustained at 14-month follow-up on Grade 4 reading comprehension (ES = .12) and mathematics achievement (ES = .16). Findings indicate that a content literacy intervention that spirals topics and vocabulary across grades can improve students' long-term academic achievement outcomes.

VERSION: December 2023

Suggested citation: Kim, James S., Joshua B. Gilbert, Jackie E. Relyea, Patrick Rich, Ethan Scherer, Mary A. Burkhauser, and Johanna N. Tvedt. (2023). Time to Transfer: Long-Term Effects of a Sustained and Spiraled Content Literacy Intervention in the Elementary Grades. (EdWorkingPaper: 23-769). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/t3c6-xh48

Time to Transfer: Long-Term Effects of a Sustained and Spiraled Content Literacy Intervention in the Elementary Grades

James S. Kim[1]

Joshua B. Gilbert[1]

Jackie Eunjung Relyea[2]

Patrick Rich[3]

Ethan Scherer[1]

Mary A. Burkhauser[1]

Johanna N. Tvedt[1]

[1]Graduate School of Education, Harvard University

[2]College of Education, North Carolina State University

[3]American Institutes for Research

Forthcoming in *Developmental Psychology*

Acknowledgments: 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. We thank Sara Hiebert Burch (Edward Hicks Magill Professor Emerita of Mathematics and Natural Science at Swarthmore College), Catherine Crouch (Professor of Physics at Swarthmore College), and Eric Klopfer (Professor and Director of the Sheller Teacher Education Program and Education Arcade at MIT) for providing valuable feedback on the Grade 3 science content reading comprehension assessment.

Corresponding Author:
Jackie Eunjung Relyea (jrelyea@ncsu.edu)
Assistant Professor
College of Education
North Carolina State University
Raleigh, NC

Abstract

We investigated the effectiveness of a sustained and spiraled content literacy intervention that emphasizes building domain and topic knowledge schemas and vocabulary for elementary-grade students. The Model of Reading Engagement (MORE) intervention underscores thematic lessons that provide an intellectual structure for helping students connect new learning to a general schema in Grade 1 (animal survival), Grade 2 (scientific investigation of past events like dinosaur mass extinctions), and Grade 3 (scientific investigation of living systems). A total of 30 elementary schools (N = 2,870 students) were randomized to a treatment or control condition. In the treatment condition (i.e., full spiral curriculum), students participated in content literacy lessons from Grades 1 to 3 during the school year and wide reading of thematically related informational texts in the summer following Grades 1 and 2. In the control condition (i.e., partial spiral curriculum), students participated in lessons in only Grade 3. The Grade 3 lessons for both conditions were implemented online during the COVID-19 pandemic school year. Results reveal that treatment students outperformed control students on science vocabulary knowledge across all three grades. Furthermore, intent-to-treat analyses revealed positive transfer effects on Grade 3 science reading (ES = .14), domain-general reading comprehension (ES = .11), and mathematics achievement (ES = .12). Treatment impacts were sustained at 14-month follow-up on Grade 4 reading comprehension (ES = .12) and mathematics achievement (ES = .16). Findings indicate that a content literacy intervention that spirals topics and vocabulary across grades can improve students' long-term academic achievement outcomes.

Impact Statement

This experimental study illustrates how sustaining and spiraling science schemas (background knowledge) and vocabulary from Grades 1 to 3 can improve students' ability to comprehend passages in science, English language arts, and mathematics. Furthermore, findings suggest that systematically building background and vocabulary knowledge can sustain positive gains in elementary-grade students' reading comprehension ability through the end of Grade 4, 14 months after the conclusion of the intervention activities.

Keywords: schema theory, spiral curricula, content literacy intervention, far transfer, domain and topic knowledge, reading comprehension, randomized controlled trial

Time to Transfer: Long-Term Effects of a Sustained and Spiraled Content Literacy Intervention in the Elementary Grades

Can learning science help elementary school children read better throughout their schooling careers? Can a third grader apply their life science knowledge of the human body's muscular, skeletal, and nervous system to a passage about how the anatomy of a skyscraper is like a human body? Over time, does building science knowledge lead to broader improvements in reading across other domains like literature, history, and mathematics? And do multi-year interventions designed to promote depth of science knowledge and breadth of reading proficiency produce effects that are ephemeral or enduring? In essence, these questions are about far transfer. Far transfer (Barnett & Ceci, 2002) has been conceptualized along two dimensions—the content of learning (what is transferred) and the context of learning (when and where learning is transferred to and from). In this study, we examined whether a multi-year literacy intervention from Grades 1 to 3 can systematically build background and vocabulary knowledge, sustain positive gains in students' reading comprehension, and promote reading transfer to other academic domains (i.e., mathematics).

Indeed, a timeless concern for developmental psychologists and educators is whether sustained, high-quality educational opportunities that complement prior learning can promote far transfer and sustained impacts on academic achievement (Agodini et al., 2009; Ceci & Barnett, 2002; Watts et al., 2018). One promising approach to maintaining long-term impacts on far-transfer measures of learning is to sustain and align content and practices across consecutive elementary grades. In recent years, numerous researchers and practitioners have established partnerships in several U.S. school districts to better align curriculum and instruction across grades and to maintain long-term impacts on students' reading and mathematics achievement

(Mattera et al., 2021; Stipek et al., 2017). Such partnerships are often inspired by the *sustaining environments hypothesis* (Bailey et al., 2020) which posits that early gains are more likely to persist if subsequent learning environments maintain continuity in learning and avoid the repetition of already mastered content (Engel et al., 2013; Newmann et al., 2001). Although research supporting this hypothesis is mixed, emerging evidence indicates that multi-year whole-class Kindergarten to Grade 3 literacy interventions (e.g., Borman et al., 2007; Language and Reading Research Consortium [LARCC], 2019) can produce positive impacts on reading comprehension outcomes following a third and final year of program implementation. To date, however, few studies have measured far transfer impacts of a sustained content literacy intervention beyond Grade 3.

To generate long-term reading comprehension gains, scholars have emphasized the need for content literacy interventions that sustain and spiral content and practices across grades and across school and home contexts (e.g., Bronfenbrenner & Morris, 2006; Bruner, 1960; Pressley et al., 2007). Sustained and spiraled content literacy interventions may provide an ideal context for helping children build schemas (i.e., intellectual structures) that make background knowledge transferable to new, related topics. As children develop greater expertise, they represent knowledge in the form of schemas, and schema representation plays a critical role in a learner's comprehension of text (Anderson & Pearson, 1984; Hirsch, 1988, 2016). In all likelihood, it takes time for young children to develop generalized schemas that can be accessed and applied when reading about related new topics in science, social studies, and mathematics.

The purpose of this study was to investigate whether a sustained and spiraled content literacy intervention, Model of Reading Engagement (MORE), could produce cumulative impacts on students' (a) vocabulary knowledge, (b) domain-specific (i.e., science) and domain-

general reading comprehension, and (c) mathematics. Students in the treatment group learned science content through a well-structured spiral curriculum that gradually introduced complex topics and vocabulary (Bruner, 1960). Students were provided recurrent exposure to thematically related topics through wide reading of informational texts during the summer months following the Grade 1 and 2 lessons. In this study, we define a sustained and spiraled content literacy intervention as an instructional approach in which teachers implement thematically connected lessons from Grades 1 to 3, thus helping students to build and transfer their background and vocabulary knowledge while reading about new, related topics across academic subjects. Using a longitudinal cluster (school) randomized trial, we examined the impacts on far transfer measures of reading comprehension and mathematics in Grades 3 and 4.

Conceptual Foundations for a Sustained and Spiraled Content Literacy Intervention

To read complex nonfiction texts with understanding, novice learners must acquire domain and topic knowledge. Importantly, domain knowledge refers to how much a student knows about an academic subject like life science of American history, and topic knowledge refers to how much a student knows about topics within a specific domain (Alexander, 2003; Hirsch, 2016). Over the past decade, numerous psychologists have suggested that elementary grade content literacy instruction in science and social studies may provide an ideal context for helping young children acquire the domain and topic knowledge and language and literacy skills to comprehend complex nonfiction texts (e.g., Connor et al., 2017; Duke et al. 2021; Williams et al., 2016). To date, content literacy interventions have largely been designed to supplement and augment the core English Language Arts (ELA) curriculum used by teachers, to target improvement in whole-classroom Tier I instruction, and to improve the quantity and quality of domain and topic knowledge as a lever for improving students' reading comprehension ability.

However, content literacy interventions that build knowledge beyond a single school year are rare and evidence of far transfer to new topics is even rarer.

This study contributes to the research base through the conceptualization of a sustained and spiraled approach to content literacy curriculum and instruction. Bruner (1960) first theorized that knowledge forms a "metamorphic spiral" whereby students revisit, reconstruct, and transform knowledge, thus leading to "eventual mastery of the connexity and structure of a large body of knowledge" (p. 32). In other words, a spiral curriculum is not simply the repetition of ideas but the iterative revisiting of topics throughout a course of study, with each encounter leading to a further deepening of knowledge.

The heuristic in Figure 1 highlights the foundational principles of sustained and spiral curricula and a specific instantiation of those principles. The four principles emphasize the importance of fostering schema awareness, gradually introducing complex topics, developing academic vocabulary networks, and promoting and measuring far transfer on unconstrained outcomes like reading comprehension (Bruner, 1960; Fitzgerald et al., 2020; Harden & Stamper, 1999).

Fostering Schema Awareness

The first aim of a spiral curriculum is for students to foster schema awareness. Schemas are intellectual structures that help novice learners build expertise within a given domain (i.e., science) by making it easier to acquire, organize, connect, and transfer knowledge (Alexander, 2003; Graesser & Nakamura, 1982; Rumelhart & Norman, 1981). In many ways, schemas can be "linked to abstract, generalizable features of situations" (Kintsch, 2009, p. 231), enabling novices to navigate new topics in a variety of academic domains and make sense of new situations across related topics. Schemas are generalizable to multiple topics within a given domain and are

strengthened when students establish topic-specific schemas learned in previous grades as they read and listen to complex texts about these topics. Although developing schema awareness takes time, it is particularly vital for novice learners because it will help them build and organize knowledge into larger, interconnected, and robust schemas in the long term (Anderson & Pearson, 1984; Kimball & Holyoak, 2000; Rist, 1989).

By design, the principles of spiral curricula emphasize the importance of teaching children about the structure of knowledge and promoting the continuity of learning over time. Bruner (1960) argued that the spiral curriculum was inherently forward-looking by helping learners transfer knowledge beyond the specific context in which learning originally occurred. For example, a student who understands how the musculoskeletal and nervous systems help the human body stay healthy has learned not only about a specific science topic but also a more general understanding of the schema for how living systems function properly. Over time, then, the spiral curriculum can make learning more efficient by helping students build knowledge that can be transferred beyond the original context where knowledge was first acquired.

Gradually Introducing Complex Topics

The second aim of a spiral curriculum is for students to master simpler topics which lay the foundation for learning more complex topics. Bruner's (1960) notion of a spiral curriculum was based on the cognitive theory that young children should be gradually exposed to complex topics over time. The theory behind the spiral curriculum is that well-structured schemas enable learners to connect knowledge across grades and to build a general schema that can be deployed when learning about novel topics. For example, first graders in this study started with concrete examples of living animals such as polar bears and what they look like, how they behave, where they live, and how they survive and adapt. Students then leveraged the schema for animal

survival to independently read thematically related books on this topic at home during summer months. As second graders, students used their prior knowledge to study the topic of how paleontologists study the fossils of dinosaurs to understand their physical characteristics and behaviors. Finally in Grade 3, students encountered the most complex topic in the spiral curriculum involving how living systems function properly.

The recurrent induction of rudimentary generalizations about living systems acquired from concrete topics can be organized, internalized, and employed by students to understand newly encountered topics in a particular text or task (Alexander, 1992). Once students have internalized the general schema about living systems through repeated exposure to various topics (i.e., Artic animal survival, dinosaur survival and extinction, our human body) over time, they can expand existing topic knowledge and acquire more advanced information (i.e., how scientists investigate the muscular and skeletal system of animals such as monkeys and birds and how skyscrapers are like a human body) which eventually represent science domain knowledge. Because academic domains share common properties, it is also possible that general schemas established in one domain (e.g., living systems) could be applied to build new knowledge and comprehend texts in another domain (e.g., non-living systems).

Developing Academic Vocabulary Networks

The third aim of a spiral curriculum is for students to know words deeply by the company they keep (Firth, 1957)—that is, to situate semantically related words in academic vocabulary networks. The lexical quality hypothesis holds that learners, who master the form and meaning of words, as well as the semantic networks in which they are situated, can efficiently access this knowledge while reading connected texts (Perfetti, 2007). For instance, over time, as children connect the words to the general schema, they deepen their knowledge of words in contexts that

reveal the meaning of the words and the associated network. The schema fosters a cohesive instructional context that facilitates a learner's ability to store and retrieve words from memory.

Establishing students' mental networks of semantically associated academic words is vital for their academic content learning and schema development (e.g., Fitzgerald et al., 2020; McKeown & Beck, 2011). As shown in Figure 1B, students deepened their knowledge of academic vocabulary networks that were crucial to understanding passages about the topic of living systems in third grade. For example, if a third grader has a deep understanding of the concept of a *living system like the human body*, their mental network for its meaning includes semantically related words like *structure, function, skeletal, muscular, nervous*, and *system*. As students repeatedly encounter the concept of *a living system* through thematically related texts, the academic vocabulary network for it would gradually grow and expand, adding more words related to the meaning of *system* within and across grade levels. Over time, recurrent exposures to vocabulary networks provide retrieval cues that activate the schema and related concepts (Alexander, 1997; Gelman, 2009; Gelman & O'Reilly, 1988; Kimball & Holyoak, 2000).

Promoting Far Transfer in Reading Comprehension

The ultimate aim of a spiral curriculum is to promote far transfer on unconstrained outcomes. Specifically, a spiral curriculum is designed to foster a growing awareness of schemas, to gradually introduce complex topics, and to develop students' academic vocabulary networks, which are triggered by repeated induction of the schema. As students build schemas in their minds, study topics that rest on those schemas, and master the vocabulary that tends to co-occur when the topics are studied, they are more likely to transfer this knowledge to far transfer measures of general reading comprehension tasks that require background and vocabulary knowledge (Adams et al., 1995; Kendeou & O'Brien, 2016; Nagy, 2005; Pearson et al., 2010;

Perfetti & Stafura, 2014). As shown in Table 1, we conceptualized transfer (cf. Figure 1; Barnett & Ceci, 2002) on our Grade 3 domain-specific (science) reading comprehension tests along two factors: the content (i.e., the number of directly taught vocabulary words) and the context (i.e., the similarity between the passage scenario and the instructional scenario). More specifically, the retrieval cues for the general schema were activated by the inclusion of the academic vocabulary networks in the near and mid transfer passages. However, to isolate the role of schema awareness in promoting far transfer, the passage about how the anatomy of a skyscraper is like a human body included no exposure to the directly taught words in the vocabulary network and was different from the instructional context.

Spiral curricula target improvement in students' unconstrained outcomes that gradually improve over time through extensive learning experiences at school and home (e.g., background and vocabulary knowledge). It is well-established that unconstrained outcomes like reading comprehension are fundamental to school learning but develop at a slower rate in the business-as-usual, non-spiral curricula (Bailey et al., 2020; Paris, 2005; Snow & Matthews, 2016). The targeting of unconstrained outcomes by spiral curricula may be key to sustaining gains over time. Indeed, recent research indicates that early interventions are more likely to show patterns of fade out when the performance of treatment and control students converge over time on constrained outcomes (e.g., word recognition; Ansari et al., 2020; Bailey et al., 2017).

In addition to within-domain (i.e., reading) transfer, it is also possible that spiral curriculum can promote cross-domain (i.e., reading to math) transfer given recent correlational and causal research findings. For example, children's reading and mathematics ability tap similar skills such as working memory, vocabulary, problem solving, and non-verbal reasoning (Bailey et al., 2020, Duncan et al., 2007; Zhang & Peng, 2023). The reciprocal relation between reading

and mathematics is also well established in longitudinal research (e.g., Bailey et al., 2020; Cirino et al., 2018; Zhang & Peng, 2023). Emerging causal evidence also indicates that literacy interventions produce cascading effects that lead to immediate direct effects on reading and subsequent spillover effects in math. For example, quasi-experimental research evidence from the middle school Expeditionary Learning literacy intervention (Nichols-Barrer & Haimson, 2013) reveals a pattern of causal effects that begin with early impacts on reading (Year 1) that eventually transfer to improved mathematics performance in later years (Year 2 and 3). These findings are consistent with the idea that multi-year literacy interventions may have spillover effects on mathematics outcomes, which rely heavily on schema acquisition and problem-solving (Cooper, H. et al., 1996; Cooper & Sweller, 1987; Geary, 1995). In this study, we used an experimental design to test a model of cross-domain transfer that emphasized the increasingly important role of reading comprehension in learning academic content in the elementary grades.

MORE Research and Development

Foundational Principles of MORE

The MORE intervention served as a specific instantiation of a sustained and spiraled content literacy intervention in Grades 1 to 3 (Figure 1.B.). Accordingly, the spiral curriculum emphasized fostering schema awareness over learning discrete facts, inquiry-driven lessons that gradually introduced more complex questions and topics, and mastery of academic vocabulary networks that appeared in the texts and tasks across three grades. To maintain continuity of practices across grades, teachers enacted practices to build and transfer knowledge. Thus, to build students' knowledge of schemas, topics, and vocabulary, teachers interleaved practices involving direct teaching of academic vocabulary networks, interactive read-aloud to provide recurrent exposures to schema-related vocabulary, concept mapping to illustrate semantic

relations among words in academic vocabulary networks, and wide reading of thematically related informational texts with autonomy and competence supports. Then, teachers provided further opportunities for students to transfer their knowledge to collaborative research activities and structured word inquiry activities. Instruction was organized to help students use their literacy skills (listening, reading, writing, discussion) to develop greater domain and topic knowledge expertise. Teachers infused each practice with activities that fostered students' engagement, which included their affective motivational, cognitive, and behavioral engagement (Fredricks et al., 2004; Sinatra et al., 2015).

Research-Practice Partnership Context for Aligning Grades 1 to 3 Content

The MORE intervention was developed in the context of a research-practice partnership that focused on the twin goals of problem-solving and knowledge generation (Donovan et al. 2021). As part of a 10-year partnership, we collaborated with practitioners in an urban school district located in the southeastern United States to improve Grade 3 reading comprehension while keeping in mind broader knowledge gaps on the long-term effectiveness of early literacy interventions (Pearson et al., 2020). In early partnership meetings, district leaders supported the development of school-based Instructional Leadership Teams that aligned content and instruction in core academic subjects. However, given the absence of a consistent core curriculum during the content literacy instruction block, the partnership began to focus on creating a spiraled curriculum that could align content and instruction across grades and enhance continuity in science and social studies learning. Finally, we also identified limited teacher time as a critical barrier to scale. Given this constraint, we designed MORE as a low-cost intervention that spaced learning (30 hours of teacher-directed lessons) across school years and the summer months (including wide reading of thematically related informational texts).

The MORE Evidence Base

Previous research provides validating evidence that the MORE intervention can produce short-term impacts following one- and two-year program implementations. In an early efficacy study of the MORE 10-day thematic science lesson on the topic of animal survival in Grade 1 (Kim et al., 2021a), students in the treatment group outperformed control students on vocabulary knowledge (ES = .30), argumentative writing (ES = .24), and domain-general reading comprehension (ES = .11). Despite these promising findings, the implementation of the partial spiral curriculum (Grade 1 only) left open the question of whether sustaining MORE through the summer following Grade 1 and into the Grade 2 school year lessons could promote far transfer in reading.

This intervention study was designed to test the long-term impact and costs of the full Grades 1 to 3 spiral using a cluster (school) randomized controlled trial (RCT) with a large sample of students (N = 2,870) across 30 schools. In our most recent study (Kim et al., 2023), we reported findings from the Grades 1 to 2 implementation of the MORE spiral, in which we conducted a conceptual replication of the MORE spiral in Grade 1 science and social studies in spring 2019, wide reading of thematically related informational texts in summer 2019, and continued Grade 2 science implementation through spring 2020. We found positive impacts on growth of domain-general reading comprehension from the winter of Grade 1 to the fall of Grade 2 and positive impact on end of Grade 2 science reading comprehension (ES = .18), thus replicating and extending findings from the original Grade 1 study (i.e., Kim, Burkhauser et al., 2021).

Furthermore, the findings provided a more precise understanding of the mechanisms of transfer in reading and raised key questions. There was evidence of near- and mid-transfer on

domain-specific (i.e., science) reading comprehension that was measured using reading passages that were closer to the instructional science context in Grade 2, included more directly taught vocabulary, and provided retrieval cues that activated the general schema about the scientific investigation of past events (see Figure 1B). However, there was no significant impact on the far transfer reading test that included no exposure to the words in the academic vocabulary network in the lessons about how scientists (e.g., paleontologists) study past events. This finding left open the possibility that it takes more time for novice learners to spontaneously retrieve and apply the general schema to better understand far transfer passages (Barnett & Ceci, 2002; Ericsson & Kintsch, 2000; Kimball & Holyoak, 2000). In this study, we examined whether the full MORE spiral curriculum through Grade 3 could promote far transfer on both a domain-specific and domain-general measure of reading comprehension.

Current Study Context

To address this research aim, we continued the longitudinal implementation of the full MORE spiral curriculum through the end of Grade 3. However, the COVID-19 pandemic triggered school closings in spring 2020, leading to modifications to the study design. First, it precluded us from implementing the Grade 2 social studies lessons and thus this study focused on the implementation of full spiral curriculum in science. Second, it forced us to move the Grade 3 implementation online as we shifted the intervention to a hybrid instructional model, including synchronous Zoom lessons and asynchronous digital app activities. Given the education emergency, we also worked with educators in our sites to provide the Grade 3 MORE curriculum to both treatment and control schools. Thus, the experimental contrast was between the full spiral curriculum relative to a counterfactual condition involving the partial spiral curriculum. Finally, this study provided novel evidence on the implementation fidelity of Grade

3 MORE when delivered remotely and cumulative impacts on vocabulary knowledge and far-transfer effects on domain-specific and domain-general reading in Grades 3 and 4.

Research Questions and Study Aims

Although emerging evidence indicates that MORE can improve short-term outcomes measured immediately after the program period (Kim et al., 2023; Kim, Burkhauser, et al., 2021; Kim, Relyea, et al., 2021), there are four research questions (RQs) that remain unanswered by previous research.

RQ1: What is the effect of the full MORE spiral curriculum (Grades 1 to 3) on domain-specific vocabulary knowledge, compared to a control condition that received the partial MORE spiral curriculum (Grade 3 only)? Our first aim was to examine whether a three-year implementation of the full MORE spiral curriculum can promote far transfer on longitudinally assessed domain-specific vocabulary knowledge through Grade 3. Given the design of the full MORE spiral curriculum, we hypothesized that students would be able to build and instantiate academic vocabulary networks that gradually develop from Grades 1 to 3. Theoretically, the concept of preferential attachment suggests that students can learn words more easily if they are situated within an existing network of related vocabulary (Barabási & Albert, 1999; Borovsky et al., 2016; Carey, 2009, Fitzgerald et al., 2017; Neuman & Dwyer, 2011). Within the context of the full MORE spiral curriculum, teachers gradually introduced networks of academic vocabulary that were interleaved in practices to help students build and then transfer domain and topic knowledge schemas and vocabulary knowledge to novel literacy tasks.

RQ2: What is the effect of the full MORE spiral curriculum (Grades 1 to 3) on far-transfer measures of (a) Grade 3 domain-specific and domain-general reading comprehension and (b) Grade 3 mathematics? Our second aim was to examine how far the

intervention effects can travel both within and across domains (reading and mathematics) when students participate in the treatment condition with the full MORE spiral curriculum. To examine far transfer within the domain of reading, we hypothesized that the implementation of the full MORE spiral curriculum would lead to improvements in both domain-specific and domain-general reading comprehension outcomes. Because the grade 3 science passages (scientific investigation of living systems) required prior knowledge of grade 1 (animal survival) and grade 2 (scientific investigation of past events like dinosaur extinctions), we hypothesized that students participating in the full spiral curriculum would be able to leverage their schemas and vocabulary knowledge to better understand the transfer passages. We examined cross-domain transfer on Grade 3 mathematics, which primarily assessed conceptual word problems requiring strong comprehension ability. The cross-domain transfer effects from reading to mathematics are well-identified and literacy-focused activities are likely to be key active ingredients driving any observed cross-domain transfer effects (Bailey et al., 2020; Cirino et al., 2018).

RQ3: Are there long-term effects on Grade 4 reading comprehension and mathematics outcomes? Our third aim was to examine long-term impacts at a 14-month follow-up at the end of Grade 4. A timely and timeless question in developmental psychology is whether and to what extent early interventions produce long-term impacts. A fundamental aim of the full MORE spiral curriculum is to create sustaining environments from Grades 1 to 3 that promote longer-term positive impacts within and across domains. The question of sustained impact is particularly important because early intervention programs may not produce enduring long-term impacts and initially positive impacts can become negative over time (Durkin et al., 2022; Lipsey et al., 2018; May et al., 2022; Sirinideas et al, 2018). For both Grades 3 and 4

student outcomes, we examined treatment-on-the-treated (TOT) effects that provide causal estimates of the impact of participating in the full MORE spiral curriculum on student outcomes.

RQ4: What is the annual per pupil cost of the full MORE spiral curriculum? Our fourth aim was to examine the per pupil annual costs that allow for an assessment of the feasibility of scaling and sustaining the full MORE spiral curriculum beyond the study site. To date, scholars who have conducted RCTs of elementary grade content literacy interventions (e.g., Connor et al., 2017) have yet to generate cost analyses, thus making it difficult to determine the cost-effectiveness of existing programs.

Methods

Research Design

For this longitudinal cluster (school) RCT study, 30 elementary schools in one urban school district located in the southeastern United States were recruited. Table 2 compares the demographic and baseline achievement characteristics of the students in the RCT sample to students in the non-RCT sample. The difference between the RCT and non-RCT samples indicates that the RCT sample had significantly fewer White students ($p < .05$), more students receiving individual education plan (IEP; $p < .01$) and attained lower reading and mathematics scores at Grade 1 baseline ($ps < .05$) compared to the non-RCT sample. Schools were blocked by three levels of school size (i.e., small, medium, and large schools based on total student enrollment), two levels of prior reading proficiency (i.e., above and below the median on end-of-Grade 3 reading performance), and prior experience with MORE and then randomized to a treatment (full MORE spiral curriculum from Grades 1 to 3) or a control condition.

Description of Treatment and Control Conditions

Figure 2 displays the timeline for implementing the intervention activities and outcome measures from Grades 1 to 4. In the spring of Grades 1 (2019) and 2 (2020), students in the treatment condition participated in the full MORE spiral curriculum while students in the control condition received business-as-usual (BAU) instruction (Our previously published studies [Kim et al., 2023; Kim, Burkhauser, et al., 2021; Kim, Relyea, et al., 2021] have provided detailed descriptions of the Grades 1 and 2 intervention curriculum and lesson activities). A notable difference between the two conditions was that the read-aloud books used in the MORE intervention consisted of informational texts with significantly higher readability levels as measured in Lexiles (M = 696.25L, SD = 82.80L), while the books used in the BAU classrooms were predominantly narrative texts with lower Lexiles on average (M = 522.63L, SD = 138.72L). Since Lexiles capture the semantic and syntactic complexity of texts, these descriptive results indicate that MORE read aloud texts included more complex vocabulary and syntax.

In the spring of Grade 3 (2021), which was the COVID-19 pandemic school year, we provided both treatment and control conditions with the MORE intervention lessons and associated professional development for teachers to ensure *all* students had equal access to learning opportunities. Accordingly, the Grade 3 MORE lessons originally designed for in-person instruction were modified in two major ways. First, we adapted the Grade 3 lesson delivery using a hybrid model of instruction involving (a) shorter (10 hours) online synchronous lessons via Zoom, (b) asynchronous activities that were delivered through a digital educational App and print books, and (c) paper-based home activities on a trifold (8.5 by 11-inch paper, folded into 3 sections) based on science informational texts used for lessons. Second, all teachers were given an opportunity to participate in a 60-minute, researcher-facilitated online MORE lesson training held on Zoom. Teacher training was interactive, providing participants with the

chance to experience lesson activities adapted for online instruction. We recorded synchronous Zoom lessons to explore how teachers adapted the MORE implementation and a future study will explore the relationship between the quality of implementation and Grade 3 outcomes.

Study Sample

Figure 3 presents a consort diagram of student participants and attrition across Grades 1 through 4. A total of 30 elementary schools were randomized to a treatment and control group, involving 2,870 students (n = 1,587 from the 15 treatment schools and n = 1,283 from the 15 control schools) who received active parental consent to participate in the study. In spring Grade 3 (May 2021), an attrition rate of approximately 30% was observed as a total of 2,001 students (treatment n = 1,130; control n = 871) remained, and they were included in the analysis of the Grade 3 outcome measures (research question 2). However, the differential attrition rates were low, approximately 3.3%, which is below the boundary of 4.1% identified by What Works Clearinghouse (WWC) Standards (2022). The attrition rates were higher (55%) on the winter Grade 3 (March 2021) domain-specific reading comprehension test for treatment (n = 712) and control (n = 580) group students. Given the higher attrition rates on the winter Grade 3 domain-specific reading comprehension test, we conducted sensitivity analyses (see Online Supplementary Materials [OSM]) to examine whether impacts on the first follow-up outcomes were robust for the subsample of students who completed the winter assessments. In the second follow-up in spring Grade 4, there was no statistically significant difference in attrition rates between treatment (n = 1,123) and control (n = 902) conditions ($p > .05$). Attrition rates were lower at Grade 4 follow-up because some students who missed the Grade 3 assessments during the COVID-19 school year in spring 2021 completed the assessments when schools were fully open in spring 2022. Institutional Review Board (IRB) approval was obtained from Harvard

University, and informed written consent was secured from parents/guardians of student participants and teachers. Details on the original study design and analysis are in our preregistration plan (Kim, 2021).

Baseline Equivalence Between Conditions

We assessed baseline equivalence on student pretest measures for the analytic sample of students in the 30 RCT schools. For the balance tests, we fit Ordinary Least Squares (OLS) regression models in which the school average pretest (Grade 1) Measure of Academic Progress (MAP) reading and mathematics test scores and demographic variables were predicted by the treatment indicator and randomization block. We found a statistically significant difference of -0.16 SDs (*SE* = 0.07, z = -2.39) on the pretest MAP reading test score, -0.13 SDs (*SE* = 0.06, z = -2.08) on the pretest MAP mathematics score, and -0.02 (SE = 0.01) on the proportion of students with an individual education plan (see Table 3). There were no other significant baseline differences. For both analytic samples, the baseline difference between treatment and control conditions on pretest scores was below the threshold of 0.25 SDs established by WWC Standards (2022), and all statistical models included pretest scores and demographic variables to adjust for these baseline differences and improve the precision of the impact estimates.

Fidelity of Implementation: Program Differentiation

Fidelity of implementation (FOI) for the treatment group in each grade level was assessed through audio-recorded sessions and teacher surveys, particularly focusing on adherence and program differentiation (Dane & Schneider, 1998). In Grade 1 (spring 2019), as reported in our previous studies (i.e., Kim et al., 2023; Kim, Relyea, et al., 2021), treatment group teachers' adherence to the core components assessed using audio recordings and researcher-raters' adherence checklist demonstrated an average adherence rate of 98% (inter-rater agreement:

91%). The evaluation of program differentiation was performed using a teacher survey on the amount of instructional time that teachers in both conditions spent on reading, science, and social studies content. The results showed a significantly higher amount of instructional time in science and social studies in treatment group than control group, but no significant difference was found in ELA/reading instruction time. Likewise, for Grade 2 (spring 2020), treatment group teachers' adherence rate was relatively high (ranging between 87% and 94%) when it was assessed using a teacher survey in which treatment group teachers reported the frequency of the MORE lesson components implemented in their classrooms and whether they used the MORE lessons books. The group difference in instructional time devoted to ELA/reading and science instruction was not statistically significant (*ps* > .05), suggesting that treatment group teacher implemented science MORE lessons not at the expense of instructional time for the district's ELA curriculum (see more details in Kim et al., 2023).

In Grade 3, we assessed adherence and program differentiation by conducting a teacher survey upon completion of the intervention implementation. For the adherence assessment, we documented the instructional adherence checklist in the teacher survey in which teachers reported the extent to which they taught MORE instructional components during the MORE lesson implementation. There were 59 items ($\alpha = .88$) organized by 10 core components (see Table 4). All teachers responded to the question, "*Consider how you implemented the MORE unit. How characteristic were the following statements of your MORE implementation*?" on a 5-point Likert scale (1 = *not at all characteristic* to 5 = *extremely characteristic*). Table 4 shows descriptive statistics of the adherence scores for each component by treatment conditions. For most components (components 0 to 7), there was no statistically significant difference (*ps* ≥ .05) in adherence to MORE implementation procedures. The two groups were significantly different

in two components related to MORE App use and asynchronous App activities (*ps* < .01), although the practical significance of these differences was small (average difference between a teacher who reported "*very characteristic*" versus "*moderately characteristic*" of MORE implementation).

To assess program differentiation between the groups, we asked teachers to report in the survey the amount of instruction time spent on ELA/reading, science, and vocabulary- and content-focused lessons. As shown in the bottom of Table 4, teachers in both groups devoted a largely similar amount of instruction time to ELA/reading and vocabulary-focused and content-focused lessons (*ps* > .05). Finally, the control group teachers were more likely to spend instructional time in science class than the treatment group teachers ($p < .001$).

Procedures for Selecting Domain-Specific Vocabulary in MORE Lessons

To select target domain-specific vocabulary words for Grades 1 to 3 MORE lessons, we first performed a content analysis of the state's science standards and the Next Generation Science Standards (NGSS; National Research Council, 2012) and anchored a set of science words from our lesson texts to identify the related vocabulary. We cross-validated these words against the state standards and NGSS to ensure that the words were relatively stable features of U.S. school curricula over time (Hirsch, 2016; National Research Council, 2012). With those words, we created an automated concept network for each grade-level lesson that contained target words and semantically associated words (see OSM for a Sample Automated Concept Network). Each target word was represented as a node with weighted connections between nodes indicating the degree of similarity. The automated concept network was used to identify taught words and untaught words. Taught (or target) words were explicitly taught or discussed during the lessons, while untaught words were not directly taught but students incidentally encountered

those untaught words through interactive read-alouds and discussion activities. The untaught words were in the range of lower frequency, higher age of acquisition, and/or lower concreteness at each grade level. For example, the Grade 3 taught words were: *skeletal, muscular, nervous, diagnosis, structure, system,* and *function,* whereas untaught words were *signal, repair, organ, fracture,* and *sensory*. We selected these untaught words to assess transfer on the domain-specific vocabulary knowledge measure.

Student Measures

To examine the effects of the full MORE spiral curriculum in RQs 1 to 3, we assessed three types of student outcomes: (a) domain-specific vocabulary knowledge from Grades 1 to 3, (b) Grade 3 domain-specific reading comprehension, and (c) domain-general reading comprehension and mathematics achievement in spring Grades 3 and 4.

Domain-Specific (Science) Vocabulary Knowledge

We administered the domain-specific (i.e., science) vocabulary knowledge assessments upon the completion of the intervention in spring Grades 1, 2, and 3 (all in person). We used a semantic association task (see details in Kim et al., 2023; Kim, Burkhauser, et al., 2021; Kim, Relyea, et al., 2021) to assess students' ability to identify semantically related words and their knowledge of how words are networked to each other. The prompt asked students to "circle two words that go with the word…" (e.g., *signal*) and presented four options (e.g., *metal, messenger, transmit, similar*). There were 12 items (12 target words: seven taught words and five untaught words) in each grade level and each item was scored 0 to 4. Cronbach alpha reliabilities for the total items, taught items, and untaught items were .90, .82, and .80, respectively.

Domain-Specific (Science) Reading Comprehension

Domain-specific reading comprehension was measured only in Grade 3 after the intervention implementation. We developed near-, mid-, and far-transfer domain-specific (science) reading comprehension passages and 29 multiple-choice questions to assess students' ability to read and understand main ideas and scientific concepts in those science passages. As noted in Table 1, transfer was conceptualized along a continuum from near to far passages, with passages varying along the dimension of content (extensive to no exposure to directly taught academic vocabulary networks) and context (similarity to the instructional scenario on human body systems). Thus, the three passage topics included scientists studying how monkeys recover from heart attacks (near transfer) and how North American migratory birds' skeletal and muscular systems are adapting over time (mid transfer). Finally, the far-transfer passage focused on non-living systems (e.g., the anatomy of a skyscraper) without including any directly taught words in the academic vocabulary network. Cronbach's alpha reliabilities for the full assessment were .86, and for the near-, mid-, and far-transfer passages were 72, .63, and .68, respectively. The reading passages had similar readability levels (610L to 800L). The OSM includes detailed psychometric information.

Domain-General Reading Comprehension and Mathematics

We measured students' domain-general reading comprehension and mathematics ability with post-tests in Grades 3 and 4, using statewide end-of-grade (EOG) standardized assessments, during the last week of the school year. For each of the domains, we used the IRT-scaled EOG test score to estimate the overall ability in respective domains. Internal consistencies are approximately .90 for Grades 3 and 4 across demographic subgroups (North Carolina Department of Public Instruction, 2020).

Baseline (Pretest) Reading and Mathematics

We used the Measure of Academic Progress (MAP; NWEA, 2019) reading and mathematics assessment scores (test-retest reliabilities = .79 to .86) obtained in Grade 1 to control for students' baseline reading and mathematics abilities. The MAP assessment is a vertically scaled and computer-adaptive assessment that measures student growth in reading and mathematics using the Rasch unit (RIT) scale.

Participation in the Full MORE Spiral Curriculum

We created a dummy variable indicating whether students participated in the full MORE spiral curriculum regardless of their initial random assignment status. Because students transfer between schools, there was imperfect compliance with the initial randomization plan. As a result, some students randomized to the treatment schools moved to either control schools or non-study schools with the district, and some control students did the same. Of the 2,870 students randomized to treatment or control schools (see Figure 3), 90% (n = 1,430 of 1,587) for treatment group students and 98% (n = 1,261 of 1,283) for control group students remained in their originally assigned conditions for two or more years. Therefore, in addition to examining the impact of being randomized to MORE treatment schools (intent-to-treat impacts), we used a student-level measure indicating whether students received two or more years of the full MORE spiral condition to conduct our TOT analyses.

Data Analytic Plans

Intent-to-Treat (ITT) Analysis

To estimate the ITT effects of being randomly assigned to participate in the full MORE spiral curriculum schools on the Grades 3 and 4 EOG outcomes, we specified a series of multiple linear regression models as follows:

$$Y_{ij} = \alpha_0 + \beta(MORE)_j + X_{ij} + \phi_b + \epsilon_{ij},$$

where Y_{ij} is the respective outcome for student *i* in school *j*; *MORE* is a treatment indicator; *X* is a vector of student-level covariates (i.e., a cubic specification of Grade 1 MAP reading and mathematics pretest scores, and demographic characteristics); ϕ is a set of school randomization block fixed effects; and ϵ is the error term. Due to the cluster-randomized nature of the design, we applied clustered standard errors at the school level using the "cluster" option in Stata 17.

Treatment-on-the-Treated (TOT) Analysis

We further estimated the TOT effects on the Grades 3 and 4 EOG outcomes, using two-stage least squares (2SLS) instrumental variables estimation, to examine the impact of treatment for students who participated in the full MORE spiral curriculum. Our prior efficacy work has indicated that additional years of MORE participation could lead to larger positive impacts (Kim et al., 2023; Kim, Relyea, et al., 2021). We leveraged random assignment as our instrument to isolate the exogenous variation in whether students participate in two or more years (i.e., the "full spiral") of MORE treatment.

The first stage model is:

$$FullSpiral\_MORE_{ij} = \alpha_0 + \beta(MORE)_j + X_{ij} + \phi_b + \epsilon_{ij},$$

where *FullSpiral_MORE* is the potential mediator; and *MORE,* the excluded instrument, is the treatment assignment indicator. The second stage model is:

$$Y_{ij} = \alpha_0 + \beta(\widehat{FullSpiral\_MORE})_{ij} + X_{ij} + \phi_b + \epsilon_{ij},$$

By estimating TOT effects, we accounted for crossovers where there was imperfect compliance with the original random assignment and the resulting estimates can be interpreted as the effect of participating in the full spiral of MORE for students who were induced to participate by the offer of treatment.

Cost and Cost-Effectiveness Analysis

To calculate the per pupil program cost for the full MORE spiral curriculum, we gathered cost estimates using the ingredients method (Levin et al., 2017; Levin & McEwan, 2001). The ingredients method requires researchers to identify the key elements needed to replicate the program in a new context and in the case where they are non-monetary, value these elements using the fair market value. Key ingredients for MORE included: personnel, books, computer & technology (e.g., iPads), travel (e.g., site visits, a convening for participants), and other miscellaneous expenses (e.g., equipment, printing, shipping). See OSM for additional details on the cost and cost-effectiveness analysis.

Results

Descriptive and Correlational Analyses

Table 5 reports descriptive statistics of key measures by treatment-control conditions and a pairwise correlation matrix. Descriptive statistics for the MAP total reading RIT scaled scores were below the national norm at baseline in the winter of Grade 1 for both conditions (Thum & Hauser, 2015). As expected, the pretest MAP reading administered in winter Grade 1 was positively and strongly correlated with all Grade 3 and 4 outcomes (r range = .63 to .72). Domain-specific vocabulary knowledge measures (at all grades) showed moderate-to-strong correlations with Grade 3 and 4 domain-general reading comprehension (r range = .46 to .72) and mathematics outcomes (r range = .42 to .63), indicating the positive role of vocabulary knowledge predicting concurrent and future academic performance. Additionally, we found positive and strong within- and cross-domain correlations as well as concurrent and longitudinal correlations. For example, Grade 3 domain-general reading comprehension (EOG reading) was strongly correlated with Grade 3 domain-specific (science) reading comprehension (r = .75),

Grade 3 mathematics (r = .73), Grade 4 reading (r = .79), and G4 mathematics (r = .68).

Research Question 1: Effects on Domain-Specific Vocabulary Knowledge

Table 6 reports the ITT effects on domain-specific vocabulary knowledge outcomes (total words, taught words, and untaught words) measured in Grades 1, 2, and 3. We observed significant positive treatment effects in science and social studies vocabulary knowledge (total words) in Grade 1 (ES = .33 and .64, respectively) and science vocabulary knowledge in Grade 3 (ES = .14). However, there was no significant difference between treatment and control groups on science vocabulary knowledge (total words) in Grade 2. Across all grades, treatment group students consistently outperformed control group students on domain-specific vocabulary words in Grade 1 science (ES = .22), Grade 2 science (ES = .15), and Grade 3 science (ES = .16) that were not directly taught by teachers but acquired incidentally through oral language activities during the MORE lessons (e.g., interactive read-alouds and follow-up discussions).

Research Question 2: Effects on Grade 3 Reading Comprehension and Mathematics

Table 7 reports both ITT and TOT effects on Grade 3 domain-specific reading comprehension, domain-general reading comprehension, and mathematics outcomes. The ITT analyses show that treatment students achieved a significantly higher total score (ps < .05) in Grade 3 domain-specific reading comprehension (ES = .14) than control students. ITT impacts were also comparable in magnitude for the near- (ES = .11), mid- (ES = .11), and far-transfer (ES = .14) passage about how the anatomy of a skyscraper is like the human body. Moreover, the magnitude of the effect size on the far-transfer science passage was statistically indistinguishable from the effect sizes for the near- and mid-transfer tests. The results for the TOT analyses also replicated the results of the ITT analyses. Therefore, the TOT results suggest that non-compliance with the initial random assignment condition was relatively small and that larger

treatment effects were observed for students who participated in MORE for two or more years compared to those who did the partial spiral.

Research Question 3: Long-Term Effects at 14-Month Follow-Up in Grade 4

We found consistent evidence that long-term impacts were maintained at 14-month follow-up. As shown in Table 8, there were statistically significant ITT effects (*ps* < .001) on Grade 4 domain-general reading comprehension (ES = .12) and mathematics (ES = .16). The TOT analyses also reveal that the treatment effects were maintained for students who participated in the MORE intervention with the full spiral curriculum. In sum, these findings for Grade 4 indicate that the sustained and spiraled content literacy intervention had positive and statistically significant transfer effects on domain-general reading comprehension and cross-domain (mathematics) outcomes that were sustained through Grade 4.

Research Question 4: Cost and Cost-Effectiveness Analyses

As shown in Table 9, we spent approximately $411.85 (in 2019 dollars) per student using nationally representative prices. Thus, these per pupil cost estimates represent what it would cost to implement the full MORE spiral curriculum as it was in 2018-19 through 2020-21 when this study was implemented. The cost-effectiveness of MORE (see Results of Cost Analysis in OSM) is also comparable to other literacy interventions that have measured costs and program impacts using a randomized controlled trial design (Kim et al., 2016; Jacob et al., 2015; Gray et al., 2022).

Robustness and Sensitivity Analyses

We undertook several analyses to check the robustness of the impact analyses for research questions 1 to 3 and the results are presented in the OSM. First, we found no evidence that the treatment main effects on Grades 3 and 4 reading comprehension and mathematics

outcomes reported in Tables 7 and 8, respectively, were moderated by (a) students' pretest reading comprehension ability in first grade, (b) students' racial ethnic background, (c) student's English language proficiency, and (d) school poverty levels (see Table S1 in OSM), suggesting that MORE was equally beneficial for all student subgroups. Second, the results were not sensitive to alternative modeling strategies using multilevel models with random intercepts for schools for ITT and TOT models instead of cluster-robust standard errors (see Tables S2 and S3 in OSM). Third, we examined the impacts on student outcomes by addressing the missing data on Grade 3 domain-specific vocabulary knowledge and reading comprehension. When we restricted analyses to the smaller subsample of students who completed the Grade 3 domain-specific vocabulary knowledge and science reading assessments, the magnitude and statistical significance of the ITT impacts on domain-specific vocabulary knowledge remained unchanged (see Tables S4 to S7 in OSM). Fourth, we estimated ITT impacts on a binary indicator for whether students met or exceeded the cut score for being "college and career ready" (CCR) using logistic regression with clustered standard errors, which indicate that the treatment increased CCR rates (see Figure S1 in OSM). Fifth, we estimated ITT impacts on alternative, standardized reading and mathematics measures using the Grade 3 NWEA MAP assessments, which replicate the ITT impacts on the statewide Grade 3 EOG domain-general reading and mathematics assessments (see Table S8 in OSM). In sum, these additional analyses support the robustness of the impact analyses findings reported in the results.

Discussion

The chief aim of this longitudinal experimental study was to examine the long-term impacts of a sustained and spiraled content literacy intervention. Three major findings emerged. First, students who participated in the full MORE spiral curriculum (Grades 1 to 3) enjoyed

larger gains on the transfer test of vocabulary knowledge across three years (acquisition of untaught domain-specific vocabulary knowledge) compared to students who received the partial spiral curriculum in Grade 3. Second, the full MORE spiral curriculum produced significant and positive effects on third graders' science reading comprehension (ES = .14), domain-general reading comprehension (ES = .11), and mathematics (ES = .12) outcomes. Finally, these positive effects in Grade 3 were sustained at a 14-month follow-up, with significant improvements in Grade 4 domain-general reading comprehension (ES = .12) and mathematics (ES = .16). The results are also practically significant because the impact of the MORE intervention on Grade 4 for reading and mathematics could eliminate nearly 15% of the reading and mathematics gap between low- and high-income students (i.e., students eligible and not eligible for free lunch) and nearly 29% of the reading gap and 45% of the mathematics gap between English learners and non-English learners (National Center for Education Statistics, 2022).

More broadly, our findings underscore the value of incorporating sustained and spiraled curricula into core school subjects (Agodini et al., 2009; Bruner, 1960) to help students foster schema awareness, gradually learn complex science concepts, and make connections between newly acquired and previously learned domain-specific vocabulary networks. We discuss implications related to each question and conclude with study limitations and future research.

Sustaining and Spiraling a Content Literacy Intervention to Build Vocabulary Knowledge

Our first research aim was to examine the intervention impact on students' ability to build domain-specific vocabulary networks that were anchored to the schemas in the MORE lessons. Consistent with the heuristic for the full spiral curriculum (Figure 1), there were positive effects on domain-specific vocabulary knowledge across all three grades (ES = .07 to .76), with stronger performance observed for the treatment students compared to their control group peers on the

untaught words (ES = .15 to .25). The impact on the untaught, incidentally acquired words was consistently positive in science as students progressed from studying simpler topics in Grades 1 and 2 to the more complex Grade 3 topic of how living systems function properly.

Theoretically, the results support the hypothesis that a spiral curriculum can positively impact networks of semantically related vocabulary, including words that students acquired largely through incidental exposures in the oral language and independent reading activities. Prior developmental research on children's vocabulary acquisition has emphasized the notion of preferential attachment of newly acquired concepts to existing networks (Barabási & Albert, 1999; Borovsky et al., 2016; Steyvers & Tenenbaum, 2005), underscoring the idea that prior word knowledge facilitates learning new words. In the full MORE spiral curriculum, instruction was organized to develop word schemas—that is, knowledge about semantic patterns that can be applied to learn new vocabulary (Nagy & Scott, 1990). Indeed, the treatment effects on the untaught transfer words indicate that children were not only learning the meaning of individual words but were also acquiring a deeper knowledge of the academic vocabulary networks that support thinking and reading (Gelman & O'Reilly, 1988; Neuman & Dwyer, 2011, Stahl & Nagy, 2006). In sum, a spiral curriculum creates a fertile ground for growing students' vocabulary knowledge and their independent word learning ability.

From a practical standpoint, the results suggest that a full three-year spiral curriculum is superior to the more common practice of implementing only a partial one-year spiral curriculum. In short, it takes time to foster students' schema awareness and to build the background knowledge needed to understand complex science text. It is also noteworthy that positive effects on Grade 3 domain-specific vocabulary knowledge were observed even when both treatment and control groups participated in the Grade 3 MORE intervention. These results imply that spiral

curricula may act as a bootstrapping mechanism as children first develop schema awareness, learn foundational concepts, and develop partial knowledge of vocabulary networks. The unique affordance of the full MORE spiral is that students appear to benefit from repeated and gradual exposure to the network of semantically related vocabulary. In this study, students first learned about the topic of animal survival (Grade 1) and the scientific investigation of past events, such as dinosaur mass extinctions (Grade 2). This prior knowledge enabled novice learners to acquire a deeper understanding of the more complex vocabulary related to the concept of a living system.

The Full Spiral Curriculum Promotes Far Transfer in Grade 3 Reading and Mathematics

Our second research aim was to examine far-transfer effects on Grade 3 reading and mathematics outcomes. Regarding reading transfer, we observed a cascading series of effects on students' ability to comprehend domain-specific science passages (ES = .14) which extended to domain-general reading comprehension (ES = .11). In particular, the far-transfer science passage on skyscrapers contained unfamiliar topics and words that were not directly taught or discussed in the Grade 1 to Grade 3 MORE lessons. Furthermore, the magnitude of the impact on the far-transfer science passage was statistically indistinguishable from the near and mid transfer tests, providing strong convergent evidence of transfer in reading.

The positive impact on the far transfer science passage is consistent with the hypothesis that the development of schema awareness is an active ingredient in MORE. Importantly, the far transfer passage included no directly taught words that were part of the academic vocabulary networks included in the MORE lessons. As a result, students were provided no retrieval cues to activate the general schema in the far transfer passage. Thus, the positive ITT effect size of .14 suggests the students in the full spiral curriculum were able to leverage their knowledge of the schema for living systems to understand the far transfer passage about the anatomy of a

skyscraper. Consistent with research on the science of learning (Gick & Holyoak, 1983; National Research Council, 2000), this finding suggests that schema awareness facilitated transfer by helping learners analogically map knowledge from a general schema about living systems to a new topic specific schema about non-living systems like skyscrapers. This finding also suggests that a spiral curriculum can help students build knowledge of simpler topics (e.g., animal survival, scientific investigation of past events), which provide a foundation for learning about more complex topics in the later grades (e.g., scientific investigation of the survival of living and non-living systems).

We also found cross-domain transfer effects from reading to mathematics, underscoring the hypothesis that improvement in children's reading comprehension can spill over to other domains. These results imply that a combination of domain-specific and domain-general reading abilities contribute to subsequent improvement in mathematics performance. Specifically, the development of domain-specific vocabulary networks plays a critical role in enhancing comprehension of science passages with varying exposure to taught words. These results are consistent with the hypothesis that background and vocabulary knowledge are key determinants of reading comprehension across domain-specific and domain-general texts (Anderson & Freebody, 1985; Anderson & Pearson, 1985; Kintsch, 2009). In addition, performing well on Grade 3 mathematics tests depends on strong reading comprehension abilities (Fuchs & Fuchs, 2002; Joyner & Wagner, 2020). Our study revealed strong correlations between reading and mathematics in Grade 3 (r = .73) and Grade 4 (r = .80), suggesting that general language and literacy skills may be a latent factor driving these strong associations (Purpura et al., 2017).

Notably, the mathematics tests used in our study consisted mostly of word problems requiring strong reading and language skills such as knowledge of quantitative and spatial

language. For example, there were moderate-to-strong correlations between the Grades 1 to 3 vocabulary knowledge scores and Grade 4 reading (*rs* = .46 to .72) and mathematics (*rs* = .42 to .63). In many ways, vocabulary knowledge is "the exposed tip of the conceptual iceberg" (Anderson & Freebody, 1981, p. 82) and conceptual knowledge may support both reading and mathematics. Given the emphasis on conceptual knowledge in the full MORE spiral curriculum, it is plausible that the intervention activities enhanced children's ability to comprehend and develop conceptual understanding rather than factual recall (Geary, 1995). In sum, future research should focus more specifically on the vocabulary, language demands, and cognitive processes involved in both reading and mathematics abilities.

Potential Mechanisms Driving Long-Term Effects through Grade 4

How might we explain the long-term impacts on the end of Grade 4 student outcomes? As a spiral curriculum, MORE has the built-in advantage of providing "educational continuity of supports" (Ramey & Ramey, 2006, p. 455) by facilitating students' continuous acquisition of domain, topic, and vocabulary knowledge that are crucial for success in core academic subjects. In other words, the results imply that the initial positive effects on Grade 1 and 2 vocabulary knowledge depth can be sustained by targeting skills and knowledge that are malleable, fundamental, and do not appear to improve rapidly in the absence of the MORE spiral intervention (Bailey et al., 2020). In terms of malleability, the development of robust schemas and domain-specific vocabulary networks are unconstrained outcomes that are sensitive to intervention efforts but cannot be improved quickly through short-term, one-year interventions (Paris, 2005; Snow & Matthews, 2016). In the current study, there were nearly 36 months between the beginning and end of the full MORE spiral curriculum. The long-term impact of a spiral curriculum may depend on *dynamic complementarities* (Agodini et al., 2010; Bailey et al.,

2020; Johnson & Jackson, 2019) between high-quality learning in the early and upper elementary grades. In essence, a spiral curriculum strengthens the schemas that provide a mental home for students to remember, retrieve, and transfer knowledge, facilitating their subsequent learning across academic subjects.

Furthermore, we identified fundamental vocabulary networks based on disciplinary textbooks used in the U.S. school curriculum (Fitzgerald et al., 2022). Early elementary grade students encounter increasingly challenging vocabulary in disciplinary textbooks, with the volume and complexity of domain-specific academic words peaking in Grades 3 or 4 (Fitzgerald et al., 2022). As a result, early implementation of a sustained content literacy instruction over multiple years for advancing domain and topic knowledge is critical to building the foundational schematic structures and semantic networks necessary for content learning. More generally, our results support the hypothesis that sustaining learning environments is crucial for maintaining enduring impacts and facilitating the continuity of educational practices (Stipek et al., 2017).

Practical Implications for Curriculum Design and Scale-Up

In many ways, knowledge can be viewed as a form of intellectual capital whereby knowledge begets more knowledge, as students acquire, organize, and transfer knowledge to increasingly more complex topics in academic domains. Unfortunately, however, the principles underlying spiral curricula are missing in action from many literacy curricula, which have yet to demonstrate longer-term effects beyond Grade 3 (Pearson et al., 2020; Quint et al., 2015). Yet, the current longitudinal study and evaluations of mathematics curricula (Agodini et al., 2009) clearly show that spiral curricula produce positive impacts on student outcomes. Thus, it is crucial to examine how the integration of spiral curricula principles into curricula in a variety of academic disciplines (e.g., economics) may support learning.

Another important practical concern is the cost-effectiveness and scalability of MORE compared to other Tier 1 content literacy interventions. Our cost and cost-effectiveness analyses in Table 9 yield an estimated annual cost of MORE at $411.85 per student (in 2019 U.S. dollars), which is comparable to Zoology One, another Tier I content literacy intervention in the early elementary grades ($480 per student cost) (Gray et al., 2022). The MORE intervention is also cost-effective relative to other elementary grade literacy interventions involving mostly moderate-to-high-poverty schools (Jacob et al., 2015; Kim et al., 2016). Furthermore, a meta-analysis of causal research on education interventions (Kraft, 2020) suggests a median effect size of .10 SDs on broad standardized achievement assessments, which are in line with our ITT impacts of the full MORE spiral on state reading and mathematics tests. Considering both effect sizes and per pupil costs, MORE is an evidence-based and cost-effective program that is likely to scale with fidelity.

Limitations and Future Research Directions

The novel findings from this study merit future replication and extension. First, future research should explore whether students' acquisition of domain-specific vocabulary knowledge is the most immediate indicator of intervention effectiveness. Although prior research on both content literacy and language-based interventions (Connor et al., 2017, LARCC et al., 2019) has emphasized the importance of vocabulary for improving Grade 3 reading, models of reading comprehension have demonstrated that background knowledge, including both domain-specific and domain-general knowledge, is a separable construct that facilitates students' reading comprehension (Cromley & Azevedo, 2007; Kintsch, 2009). Future research should examine the contributions of vocabulary and background knowledge to students' reading comprehension.

Our findings suggest that improving whole-class instruction through a sustained and spiraled curriculum can promote reading success by Grades 3 and 4. However, achieving this goal will require policymakers to enact a systemic approach to sustaining and aligning content and practices across grade levels. Thus, it remains to be seen whether such a systemic approach to building a sustained and spiraled content literacy intervention is feasible and scalable after the conclusion of a longitudinal randomized trial that was supported by a research-practice partnership. To examine the external validity of our findings, it will be critical to study the implementation of the full MORE spiral curriculum in new school districts to determine if the treatment effects replicate in other contexts. Despite the impact of the COVID-19 pandemic on our study design and implementation, the findings from this study indicate that systematically building background and vocabulary knowledge causes lasting improvements in elementary grade students' ability to read for understanding across academic subjects.

References

Adams, B. C., Bell, L. S., & Perfetti, C. A. (1995). A trading relationship between reading skill and domain knowledge in children's text comprehension. *Discourse Processes, 20*(3), 307–323. https://doi.org/10.1080/01638539509544943

Agodini, R., Harris, B., Atkins-Burnett, S., Heaviside, S., Novak, T., & Murphy, R. (2009). *Achievement effects of four early elementary school math curricula: Findings from first graders in 39 schools*. NCEE 2009-4052. National Center for Education Evaluation and Regional Assistance.

Alexander, P. A. (1992). Domain knowledge: Evolving themes and emerging concerns. *Educational psychologist, 27*(1), 33-51.

Alexander, P. A. (1997). Mapping the multidimensional nature of domain learning: The interplay of cognitive, motivational, and strategic forces. In M. L. Maehr & P. R. Pintrich (Eds.), *Advances in motivation and achievement* (Vol. 10, pp. 213–250). JAI Press.

Alexander, P. A. (2003). The development of expertise: The journey from acclimation to proficiency. *Educational researcher, 32*(8), 10-14.

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

Anderson, R.C., & Pearson, P.D. (1984). A schema-theoretic view of basic processes in reading. In R. Barr, M. L. Kamil, & Mosenthal (Ed.), *Handbook of Reading Research* (pp.255- 291).

Ansari, A., Pianta, R. C., Whittaker, J. V., Vitiello, V. E., & Ruzek, E. A. (2020). Persistence and convergence: The end of kindergarten outcomes of pre-K graduates and their nonattending peers. *Developmental Psychology*, *56*(11), 2027–2039. https://doi.org/10.1037/dev0001115

Bailey, D. H., Duncan, G. J., Cunha, F., Foorman, B. R., & Yeager, D. S. (2020). Persistence and fade-out of educational-intervention effects: Mechanisms and potential solutions. *Psychological Science in the Public Interest, 21*(2), 55–97. https://doi.org/10.1177/1529100620915848

Bailey, D., Duncan, G. J., Odgers, C. L., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness, 10*(1), 7-39.

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509-512.

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

Borman, G. D., Slavin, R. E., Cheung, A. C., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2007). Final reading outcomes of the national randomized field trial of Success for All. *American Educational Research Journal, 44*(3), 701-731. https://doi.org/10.3102/0002831207306743

Borovsky, A., Ellis, E. M., Evans, J. L., & Elman, J. L. (2016). Semantic structure in vocabulary knowledge interacts with lexical and sentence processing in infancy. *Child Development, 87*(6), 1893-1908.

Bruner, J. S. (1960). *The process of education*. Cambridge, MA: Harvard University Press.

Carey, S. (2009). *The origin of concepts*. New York, NY: Oxford University Press.

Cirino, P. T., Child, A. E., & Macdonald, K. T. (2018). Longitudinal predictors of the overlap between reading and math skills. *Contemporary Educational Psychology, 54*, 99–111. https://doi.org/10.1016/j.cedpsych.2018.06.002

Compton, D. L., & Pearson, P. D. (2016). Identifying robust variations associated with reading comprehension skill: The search for pressure points. *Journal of Research on Educational Effectiveness, 9*(2), 223-231.

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

Connor, C. M., Morrison, F. J., Fishman, B., Crowe, E. C., Al Otaiba, S., & Schatschneider, C. (2013). A longitudinal cluster-randomized controlled study on the accumulating effects of individualized literacy instruction on students' reading from first through third grade. *Psychological Science, 24*(8), 1408–1419. https://doi.org/10.1177/0956797612472204

Cooper, H., Nye, B., Charlton, K., Lindsay, J., &Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research, 66*, 227–268.

Cooper, G., & Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Education Psychology, 79*, 347-362

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.

Donovan, M. S., Snow, C. E., & Huyghe, A. (2021). Differentiating research-practice partnerships: Affordances, constraints, criteria, and strategies for achieving success. *Studies in Educational Evaluation, 7*1, Article 101083.

Duke, N. K., Halvorsen, A. L., Strachan, S. L., Kim, J., & Konstantopoulos, S. (2021). Putting PjBL to the test: The impact of project-based learning on second graders' social studies and literacy learning and motivation in low-SES school settings. *American Educational Research Journal, 58*(1), 160-200.

Duncan, G.J, Dowsett, C.J, Claessens, A., Magnuson, K., Huston, A.C., Klebanov, P. & Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43*(6), 1428-1446. https://doi.org/10.1037/0012-1649.43.6.1428

Durkin, K., Lipsey, M. W., Farran, D. C., & Wiesen, S. E. (2022). Effects of a statewide pre-kindergarten program on children's achievement and behavior through sixth grade. *Developmental Psychology, 58*(3), 470–484. https://doi.org/10.1037/dev0001301

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. https://doi.org/10.1080/19345740802539200

Engel, M., Claessens, A., & Finch, M. A. (2013). Teaching students what they already know? The (mis)alignment between mathematics instructional content and student knowledge in kindergarten. *Educational Evaluation and Policy Analysis, 35*(2), 157-178. https://doi.org/10.3102/0162373712461850

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

Firth, J. R. (1957). *Papers in linguistics*. 1934–1951. London: Oxford University Press.

Fitzgerald, W. J., Elmore, J., Kung, M., & Stenner, A. J. (2017). The conceptual complexity of vocabulary in elementary-grades core science program textbooks. *Reading Research Quarterly, 52*(4), 417–442. https://doi.org/10.1002/rrq.184

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, 12*(5), 855-879. https://doi.org/10.1037/edu0000386

Fitzgerald, J., Relyea, J. E., & Elmore, J. (2022). Academic vocabulary volume in elementary grades disciplinary textbooks. *Journal of Educational Psychology, 114*(6), 1257–1276. https://doi.org/10.1037/edu0000735

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, *74*(1), 59–109. https://doi.org/10.3102/00346543074001059

Fuchs, L. S., & Fuchs, D. (2002). Mathematical problem-solving profiles of students with mathematics disabilities with and without comorbid reading disabilities. *Journal of Learning Disabilities, 35*(6), 564-574.

Geary, D. C. (1995). Reflections of evolution and culture in children's cognition. *American Psychologist, 50*, 24-37.

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

Gelman, S. A., & O'Reilly, A. W. (1988). Children’s inductive inferences within superordinate categories: The role of language and category structure. *Child Development*, 876-887.

Graesser, A. C., & Nakamura, G. V. (1982). The impact of a schema on comprehension and memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 16, pp. 59–109). New York: Academic Press.

Gray, A. M., Sirinides, P. M., Fink, R. E., & Bowden, A. B. (2022). Integrating literacy and science instruction in kindergarten: Results from the efficacy Study of Zoology One. *Journal of Research on Educational Effectiveness, 15*(1), 1-27. https://doi.org/10.1080/19345747.2021.1938313

Hirsch Jr., E. D. (1988). *Cultural literacy: What every American needs to know*. New York, NY: Vintage Books.

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

Jacob, R., Goddard, R., Kim, M., Miller, R., & Goddard, Y. (2015). Exploring the causal impact of the McREL balanced leadership program on leadership, principal efficacy, instructional climate, educator turnover, and student achievement. *Educational Evaluation and Policy Analysis, 37*(3), 314–332. https://doi.org/10.3102/0162373714549620

Johnson, R. C., & Jackson, C. K. (2019). Reducing inequality through dynamic complementarity: Evidence from Head Start and public school spending. *American Economic Journal: Economic Policy, 11*(4), 310-349.

Joyner, R. E., & Wagner, R. K. (2020). Co-occurrence of reading disabilities and math disabilities: a meta-analysis. *Scientific Studies of Reading, 24*(1), 14-22.

Kendeou, P. & O'Brien, E. J. (2016). Prior knowledge, acquisition and revision. In P. Afflerbach (Ed.), *Handbook of Individual Differences in Reading: Reader, Text, and Context* (pp. 151-163). New York, NY: Routledge Taylor & Francis.

Kim, J. S. (2021). A randomized controlled trial to replicate and personalize with technology the effects of a model of reading engagement (MORE) on first- and second-graders' science and social studies domain knowledge, reading engagement, reading comprehension, and basic literacy skills. *AEA RCT Registry*. https://doi.org/10.1257/rct.3489-1.2000000000000002

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

Kim, J. S., Burkhauser, M. A., Relyea, J. E., Gilbert, J. B., Scherer, E., Fitzgerald, J., Mosher, D., & McIntyre, S. (2023). A longitudinal randomized trial of a sustained content literacy intervention from first to second grade: Transfer effects on students' reading comprehension outcomes. *Journal of Educational Psychology, 115*(1), 73–98. https://doi.org/10.1037/edu0000751

Kim, J. S., Guryan, J., White, T. G., Quinn, D. M., Capotosto, L., & Kingston, H. C. (2016). Delayed effects of a low-cost and large-scale summer reading intervention on elementary school children's reading comprehension. *Journal of Research on Educational Effectiveness, 9*(sup1), 1-22. https://doi.org/10.1080/19345747.2016.1164780

Kim, J. S., Relyea, J. E., Burkhauser, M. A., Scherer, E., & Rich, P. (2021). Improving elementary grade students' science and social studies vocabulary knowledge depth, reading comprehension, and argumentative writing: A conceptual replication. *Educational Psychology Review, 33*, 1935-1964. https://doi.org/10.1007/s10648-021-09609-6

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). Oxford University Press.

Kintsch, W. (2009). Learning and constructivism. In *Constructivist instruction* (pp. 223-241). Routledge.

Language and Reading Research Consortium, Jiang, H., & Logan, J. (2019). Improving reading comprehension in the primary grades: Mediated effects of a language-focused classroom intervention. *Journal of Speech, Language, and Hearing Research, 62*(8), 2812-2828.

Lipsey, M. W., Farran, D. C., & Durkin, K. (2018). Effects of the Tennessee Prekindergarten Program on children's achievement and behavior through third grade. *Early Childhood Research Quarterly, 45*(1), 155–176. https://doi.org/10.1016/j.ecresq.2018.03.005

Mattera, S. K., Jacob, R., MacDowell, C., & Morris, P. A. (2021). Long-term effects of enhanced early childhood math instruction: The impacts of making pre-k count and high 5s on third-grade outcomes. *MDRC*. Retrieved from https://www.mdrc.org/sites/default/files/MPC_High5-3yr_Impact_Report.pdf

McCormick, M. P., Weissman, A. K., Weiland, C., Hsueh, J., Sachs, J., & Snow, C. (2020). Time well spent: Home learning activities and gains in children's academic skills in prekindergarten years. Developmental Psychology, 56(4), 710-721. http://dx.doi.org/10.1037/dev0000891

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 Press.

Mol, S. E., & Bus, A. G. (2011). To read or not to read: a meta-analysis of print exposure from infancy to early adulthood. *Psychological bulletin, 137*(2), 267-296.

Nagy, W. E., & Scott, J. A. (1990). Word schemas: Expectations about the form and meaning of new words. *Cognition & Instruction, 7*(2), 105-127.

National Center for Education Statistics. (2022). *National Assessment of Educational Progress (NAEP), 2022 Reading and Mathematics Assessments.* U.S. Department of Education, Institute of Education Sciences.

National Research Council. (2000). *How People Learn: Brain, Mind, Experience, and School: Expanded Edition*. Washington, DC: The National Academies Press. https://doi.org/10.17226/9853

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.

Neuman, S. B., & Dwyer, J. (2011). Developing vocabulary and conceptual knowledge for low-income preschoolers: A design experiment. *Journal of Literacy Research, 43*(2), 103-129.

Newmann, F. M., Smith, B., Allensworth, E., & Bryk, A. S. (2001). Instructional program coherence: What it is and why it should guide school improvement policy. *Educational Evaluation and Policy Analysis*, *23*(4), 297-321.

Nichols-Barrer, I., & Haimson, J. (2013). Impacts of five Expeditionary Learning Middle Schools on Academic Achievement. Retrieved July 10, 2023 from: https://files.eric.ed.gov/fulltext/ED618299.pdf

North Carolina Department of Public Instruction (2020). *The North Carolina Department of Public Instruction Mathematics 3-8 End of Grade (EOG) NC Math 1 and NC Math 3 End of Course (EOC) Technical Report 2018-2019*. Retrieved from https://www.dpi.nc.gov/media/10219/open

NWEA. (2019). *MAP® Growth™ technical report.* Portland, OR: Author.

Paris, S. G. (2005). Reinterpreting the development of reading skills. *Reading Research Quarterly, 40*(2), 184–202. https://doi.org/10.1598/RRQ.40.2.3

Pearson, P. D., Moje, E., & Greenleaf, C. (2010). Literacy and science: Each in the service of the other. *Science, 328*(5977), 459–463. https://doi.org/10.1126/science.1182595

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

Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading, 11*(4), 357–383. https://doi.org/10.1080/10888430701530730

Perfetti, C. A., & Stafura, J. (2014). Word knowledge in theory of reading comprehension. *Scientific Studies of Reading, 18*(1), 22–37. https://doi.org/10.1080/10888438.2013.827687

Purpura, D. J., Logan, J. A., Hassinger-Das, B., & Napoli, A. R. (2017). Why do early mathematics skills predict later reading? The role of mathematical language. *Developmental Psychology, 53*(9), 1633-1642. http://dx.doi.org/10.1037/dev0000375

Quint, J., Zhu, P., Balu, R., Rappaport, S., & DeLaurentis, M. (2015). *Scaling up the success for all model of school reform: Final report from the investing in innovation (i3) evaluation*. MDRC.

Ramey, S. L., & Ramey, C. T. (2006). *Early educational interventions: Principles of effective and sustained benefits from targeted early education programs*. Guilford Press.

Rist, R. S. (1989). Schema creation in programming. *Cognitive Science, 13*(3), 389-414.

Rumelhart, D. E., & Norman, D. A. (1981). Analogical processes in learning. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 335-360). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Sinatra, G. M., Heddy, B. C., & Lombardi, D. (2015). The challenges of defining and measuring student engagement in science. *Educational Psychologist, 50*(1), 1–13. https://doi.org/10.1080/00461520.2014.1002924

Snow, C. E., Matthews, T. J. (2016). Reading and language in the early grades. *The Future of Children, 26*(2), 57–74. https://doi.org/10.1353/foc.2016.0012

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

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.

Stipek, D., Clements, D., Coburn, C., Franke, M., & Farran, D. (2017). PK-3: What does it mean for instruction?. *Social Policy Report, 30*(2).

Thum, Y. M., & Hauser, C. H. (2015). *NWEA 2015 MAP norms for student and school achievement status and growth*. Portland, OR: NWEA.

Watts, T. W., Duncan, G. J., & Quan, H. (2018). Revisiting the marshmallow test: A conceptual replication investigating links between early delay of gratification and later outcomes. *Psychological Science, 29*(7), 1159–1177. https://doi.org/10.1177/0956797618761661

What Works Clearinghouse. (2022). *What Works Clearinghouse procedures and standards handbook, version 5.0.* U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance (NCEE). Retrieved from https://ies.ed.gov/ncee/wwc/Handbooks

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

Zhang, Z., & Peng, P. (2023). Co-development among reading, math, science, and verbal working memory in the elementary stage. *Child Development.* https://doi.org/10.1111/cdev.13962

Table 1

A Framework for Measuring Transfer Effects of Grade 3 Domain-Specific (Science) Reading Comprehension Passages

Dimensions of Transfer	Instructional Topic	Science Content Reading Comprehension Passage		
		Near Transfer Topic	Mid Transfer Topic	Far Transfer Topic
General schema: *Scientific investigation of living systems*	*How do astronauts keep their muscular, skeletal, nervous system healthy in space?*	*Scientific investigation of how monkeys survive and recover from a heart attack*	*Scientific investigation of how birds' skeletal and muscular systems adapt over time*	*How the anatomy of a skyscraper is like a human body system*
Content: Academic Vocabulary Networks (N of words in the networks)	Full Exposure (12 words: *system, structure, function, skeletal, muscular, diagnosis, fracture, sensory, organ, repair, signal*)	Limited Exposure (4 words: *systems, function, diagnosis, muscular*)	Limited Exposure (4 words: *system, skeletal, muscular, diagnosis*)	No Exposure (0 words)
Context: Similarity to the Instructional context	n/a	Similar	Different	Different

Table 2

Demographic and Baseline Achievement Comparisons of Students in the Longitudinal RCT Full Sample and non-RCT sample

Variable	*N*	RCT sample mean	RCT v. non-RCT difference (*SE*)
White	13,047	0.18	-0.11 (.05)*
Black	13,047	0.38	0.05 (.04)
Hispanic	13,047	0.32	0.07 (.04)
Asian	13,047	0.08	-0.01 (.02)
Other	13,047	0.04	0.00 (.01)
Male	13,047	0.50	-0.02 (.01)
Limited English proficiency	13,047	0.23	0.05 (.03)
Individual education plan	13,047	0.10	0.02 (.01)**
Low SES	13,047	0.41	0.05 (.07)
Mid SES	13,047	0.39	0.06 (.06)
High SES	13,047	0.20	-0.11 (.07)
Baseline MAP reading (G1)	11,069	168.42	-3.42 (1.59)*
Baseline MAP mathematics (G1)	11,159	169.90	-3.19 (1.59)*

Note. RCT = randomized controlled trial. SES = socioeconomic status. MAP = Measure of Annual Progress. RCT/non-RCT differences are regression coefficients from OLS regression models at baseline, that include the sample indicator and clustered standard errors at the school level.
†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

Table 3

Balance Checks for Analytic Sample of Students Remaining in the Long-Term Impact Analysis (School-Level Averages)

	Treatment schools		Control schools		
Characteristics	*n*	*M* (*SD*)	*n*	*M* (*SD*)	Difference (*SE*)
White	15	0.14 (.22)	15	0.17 (.21)	-0.04 (.07)
Black	15	0.41 (.24)	15	0.39 (.20)	0.02 (.08)
Hispanic	15	0.37 (.22)	15	0.33 (.20)	0.03 (.07)
Asian	15	0.06 (.07)	15	0.07 (.07)	-0.02 (.03)
Other	15	0.03 (.03)	15	0.03 (.02)	.004 (.01)
Male	15	0.48 (.05)	15	0.51 (.07)	-0.03 (.02)
Limited English proficiency	15	0.25 (.18)	15	0.22 (.14)	0.03 (.06)
Individual education plan	15	0.07 (.03)	15	0.10 (.03)	-0.02 (.01)*
Low SES	15	0.53 (.37)	15	0.39 (.38)	0.14 (.10)
Mid SES	15	0.29 (.26)	15	0.42 (.37)	-0.12 (.11)
High SES	15	0.17 (.30)	15	0.19 (.32)	-0.02 (.10)
Baseline MAP reading (G1)	15	-1.10 (.28)	15	-0.94 (.22)	-0.16 (.07)*
Baseline MAP mathematics (G1)	15	-1.08 (.25)	15	-0.95 (.23)	-0.13 (.06)*

Note. SES = socioeconomic status. MAP = Measure of Annual Progress. Differences derived from an OLS regression model with treatment indicator, fixed effects for randomization block and standard errors clustered at the school level. As an alternative balance specification, we conducted an omnibus test of all variables predicting student-level treatment status in a logistic regression with standard errors clustered at the school level. The result is non-significant, with a model omnibus chi-square value of 18.7 with *p*-value at 0.1.

†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

Table 4

Descriptive Statistics for Fidelity of Implementation (FOI) for Treatment and Control Teachers in Grade 3

FOI Components	Treatment teachers (n = 52) *M* (*SD*)	Control teachers (n = 43) *M* (*SD*)	Difference (*SE*)	ES
Adherence to MORE instruction core components[a]				
Component 0: teachers' lesson preparation	3.67 (.65)	3.59 (.53)	0.10 (.14)	0.16
Component 1: building student learning interests and engagement	4.11 (.64)	4.12 (.55)	0.08 (.11)	0.14
Component 2: taught the MORE vocabulary-focused lessons	0.98 (.09)	0.99 (.04)	-0.02 (.01)	-0.25
Component 3: vocabulary and morphology instruction	3.25 (.51)	3.08 (.39)	0.21 (.16)	0.45
Component 4: deepening and applying vocabulary knowledge	3.72 (.80)	3.95 (.70)	-0.17 (.18)	-0.22
Component 5: taught the MORE content-focused lessons	0.98 (.10)	.98 (.07)	-0.004 (.01)	-0.05
Component 6: collaborative research	3.95 (.75)	4.11 (.73)	-0.10 (.20)	-0.14
Component 7: modeling and motivating MORE trifolds use	2.10 (1.14)	2.05 (1.19)	-0.21 (.31)	-0.18
Component 8: modeling and motivating MORE App use	3.82 (.87)	3.47 (.79)	0.39 (.14)**	0.46
Component 9: modeling and motivating MORE asynchronous App activities	3.94 (.89)	3.54 (.81)	0.42 (.14)**	0.48
Time spent on lessons (minutes)				
ELA/reading	520.19 (168.81)	472.67 (159.97)	52.49 (39.10)	0.32
Science	117.73 (88.65)	143.14 (82.95)	-35.91 (9.51)***	-0.42
Vocabulary-focused lessons	35.19 (11.71)	34.65 (11.77)	1.40 (2.07)	0.12
Content-focused lessons	37.88 (9.57)	37.21 (10.87)	0.81(2.17)	0.08

Note. Differences derived from an OLS regression model with treatment indicator, fixed effects for randomization block and standard errors clustered at the school level. ES = effect size (Cohen's *d*). ELA = English Language Arts.

[a]How characteristic were the following statements of your MORE implementation: 1 = *not at all characteristic*; 2 = *a little bit characteristic*; 3 = *moderately characteristic*; 4 = *very characteristic*; 5 = *extremely characteristic*.

†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

Table 5

Descriptive Statistics for Study Measures for Treatment and Control Groups and Correlation Matrix

	Treatment		Control		Correlation								
Measures	*M* (*SD*)	*n*	*M* (*SD*)	*n*	1	2	3	4	5	6	7	8	9
1. Baseline MAP reading	166.64 (15.97)	1,123	171.01 (15.68)	902	--								
Domain-specific vocabulary knowledge													
2. G1 science vocabulary	33.48 (5.91)	1,349	32.46 (5.27)	1,082	.52	--							
3. G1 social studies vocabulary	34.78 (6.88)	1,355	31.25 (5.91)	1,064	.53	.57	--						
4. G2 science vocabulary	33.3 (9.73)	1,159	34.27 (9.06)	967	.55	.34	.39	--					
5. G3 science vocabulary	20.18 (7.75)	739	21.18 (7.67)	602	.69	.44	.51	.52	--				
Domain-specific reading comprehension													
6. G3 science reading comp. (total)	13.02 (6.55)	712	13.56 (6.54)	580	.62	.42	.46	.45	.67	--			
Domain-general reading comprehension													
7. G3 EOG reading	433.15 (9.74)	1,130	434.37 (9.77)	871	.72	.46	.50	.53	.72	.75	--		
8. G4 EOG reading	540.62 (10.36)	1,123	541.23 (11.21)	902	.71	.46	.50	.53	.72	.69	.79	--	
Mathematics													
9. G3 EOG mathematics	540.96 (9.70)	1,125	541.87 (9.93)	862	.64	.43	.45	.43	.63	.69	.73	.68	--
10. G4 EOG mathematics	545.55 (10.44)	1,120	545.58 (10.98)	900	.63	.42	.44	.44	.63	.63	.68	.80	.75

Note. MAP = Measure of Annual Progress. G = Grade. Comp. = comprehension. EOG = end-of-grade

Table 6

Intent-to-Treat (ITT) Intervention Effects on Domain-Specific Vocabulary Knowledge (Scores for Total, Taught, and Untaught Words) from Grades 1 to 3

	Grade 1			Grade 2			Grade 3		
Source	Total	Taught	Untaught	Total	Taught	Untaught	Total	Taught	Untaught
ITT Science	.33 (.06)***	.33 (.06)***	.22 (.06)***	.07 (.07)	-.01 (.06)	.15 (.07)*	.14 (.05)**	.12 (.05)*	.16 (.06)**
ITT Social Studies	.64 (.05)***	.76 (.05)***	.25 (.05)***						
N[a]	2,419	2,419	2,419	2,126	2,126	2,126	1,341	1,341	1,341

Note. The sample size was attenuated by missing data from students who did not take the vocabulary knowledge during the pandemic 2020-2021 school year. ITT models include controls for demographics, randomization block, and cubic baseline reading and mathematics scores. Standard errors were clustered at school level, the unit of randomization. Social studies vocabulary knowledge was assessed only in Grade 1.

[a]*N* reported is for science vocabulary knowledge.

†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

Table 7

Intent-to-Treat (ITT) and Treatment-on-the-Treated (TOT) Intervention Effects on Grade 3 Domain-Specific Reading Comprehension, Domain-General Reading Comprehension, and Grade 3 Mathematics Outcomes

Source	G3 Domain-specific reading comprehension				G3 Domain-general reading comprehension	G3 Mathematics
	Total	Near-transfer passage	Mid-transfer passage	Far-transfer passage		
ITT	.14 (.06)*	.11 (.06)	.11 (.05)*	.14 (.06)*	.11 (.04)**	.12 (.04)**
2SLS	.14 (.07)*	.12 (.06)	.11 (.06)*	.15 (.06)**	.12 (.04)**	.13 (.04)**
N	1,292	1,292	1,261	1,210	2,001	1,987
1st Stage F	12092.89	12092.89	16181.21	20622.15	4570.87	5186.16

Note. 2SLS = two-stage least squares. ITT and 2SLS analysis models included covariates for demographics, school randomization block, and baseline MAP reading and mathematics scores. Standard errors were clustered at school level, the unit of randomization. †$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$. There were no statistically significant differences in pairwise comparisons of the effect sizes for the domain-specific reading comprehension subtests.

Table 8

Intent-to-Treat (ITT) and Treatment-on-the-Treated (TOT) Intervention Effects on Grade 4 Domain-General Reading Comprehension and Mathematics Outcomes

Source	G4 Domain-general reading comprehension	G4 Mathematics
ITT	.12 (.03)***	.16 (.04)***
2SLS	.13 (.03)***	.18 (.05)***
N	2,025	2,020
1st Stage F	3530.88	3583.02

Note. 2SLS = two-stage least squares. ITT and 2SLS analysis models included covariates for demographics, school randomization block, and baseline MAP reading and mathematics scores. Standard errors were clustered at school level, the unit of randomization. †$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

Table 9

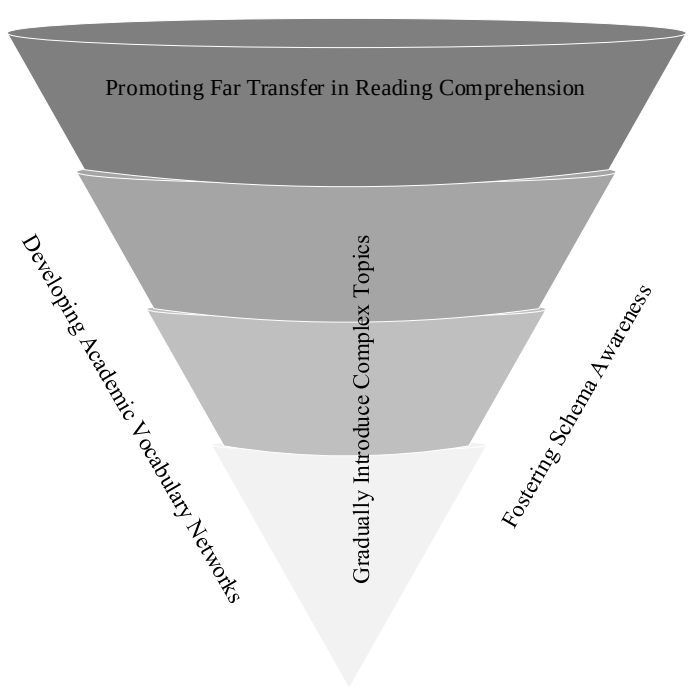
MORE Per Pupil Costs (in 2019 US Dollars)

Ingredients	Nationally representative inflation adjusted prices, with ingredient method approach	Proportion of Cost
Non-Teacher/Coach Personnel	$97.49	24%
Teacher/Coach Personnel	$18.62	5%
Materials/Equipment	$86.02	21%
Books	$113.10	27%
Other Inputs	$96.62	23%
Total	$411.85	100%

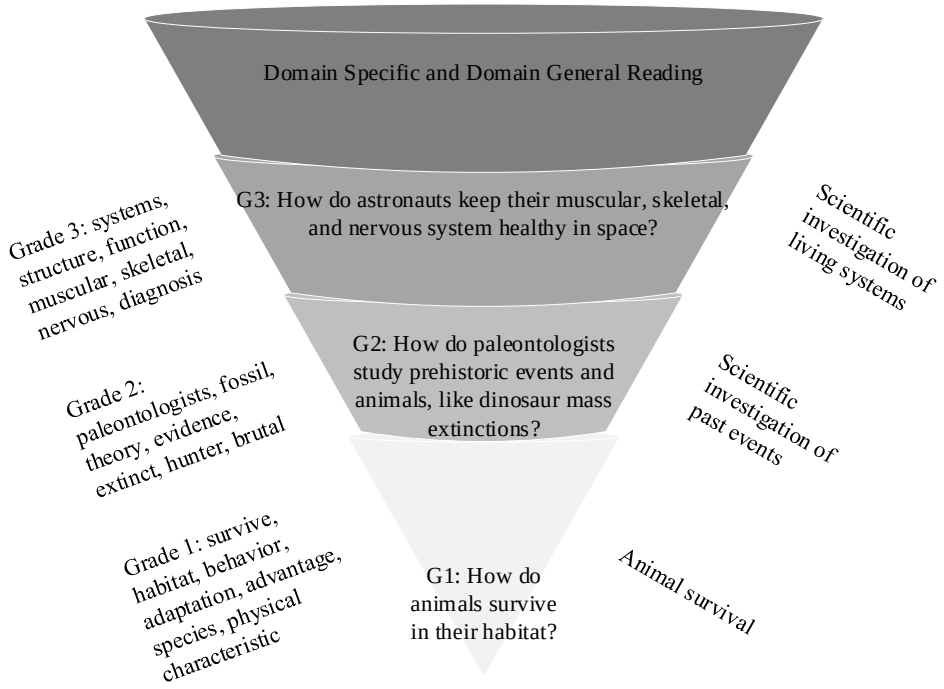
Note. We used the E$timator© Tool (https://www.costtoolkit.org/) developed by Teachers College (Hollands et al., 2015) to calculate nationally representative prices for all our costs (in 2019 dollars). As explained in the text, we made several assumptions to calculate these per student costs. Our intervention took place during the regular school day as part of the district's regular scope and sequence. As such, we have not added the facilities cost. In each year, the MORE intervention program hired schoolteachers and literacy facilitators (i.e., school coaches) to help with implementation. Teachers were required to complete 3 hours of professional development, attend a 1-hour meeting with the grade-level team, and provided 1 hour for general administrative tasks related to the program. Facilitators attended a 3-hour training, attended or led six hours of meetings, and also were allocated about an hour of administrative tasks. We used the district's compensation rate for these teachers and then adjusted them to nationally representative wage rates using E$timator©. Finally, as the research staff traveled, prepared, and led the training of the teachers and facilitators in addition to preparing and mailing curriculum materials and paper assessments in the first two years, we include 50% of a coordinator and assistant, as well as 10% of all other staff time to implement the intervention.

Figure 1

Schematic Representation of a Sustained and Spiraled Content Literacy Intervention



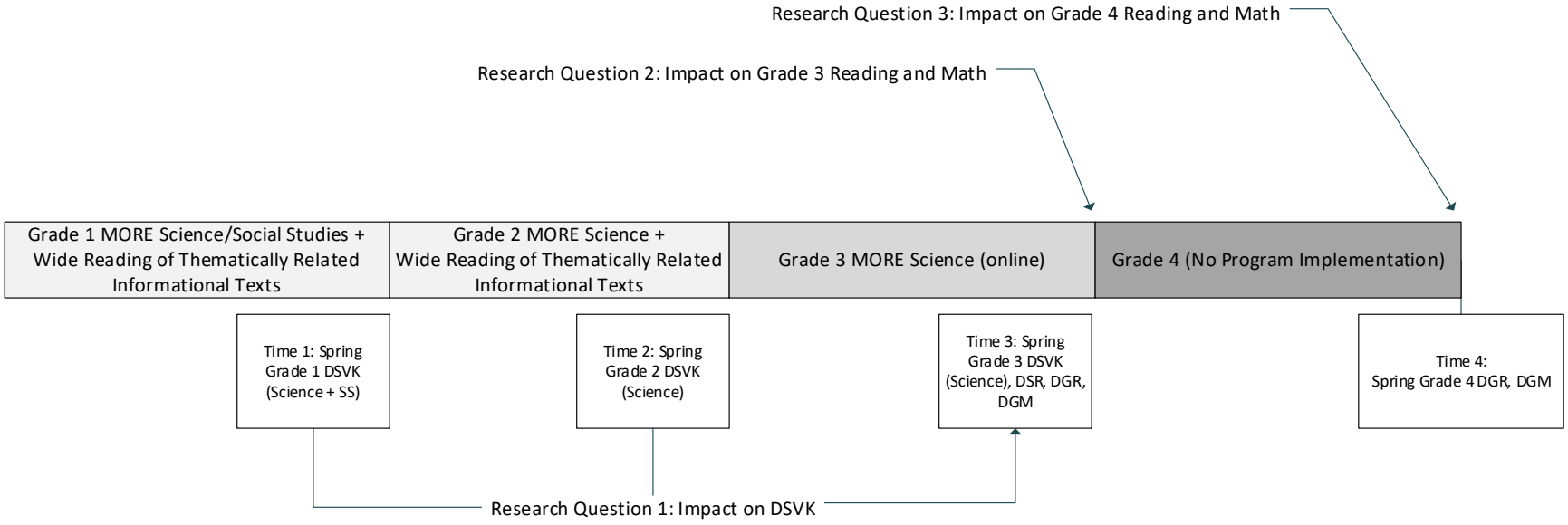
1.A. Foundational Principles



1.B. Instantiation of Foundational Principles in MORE

Figure 2

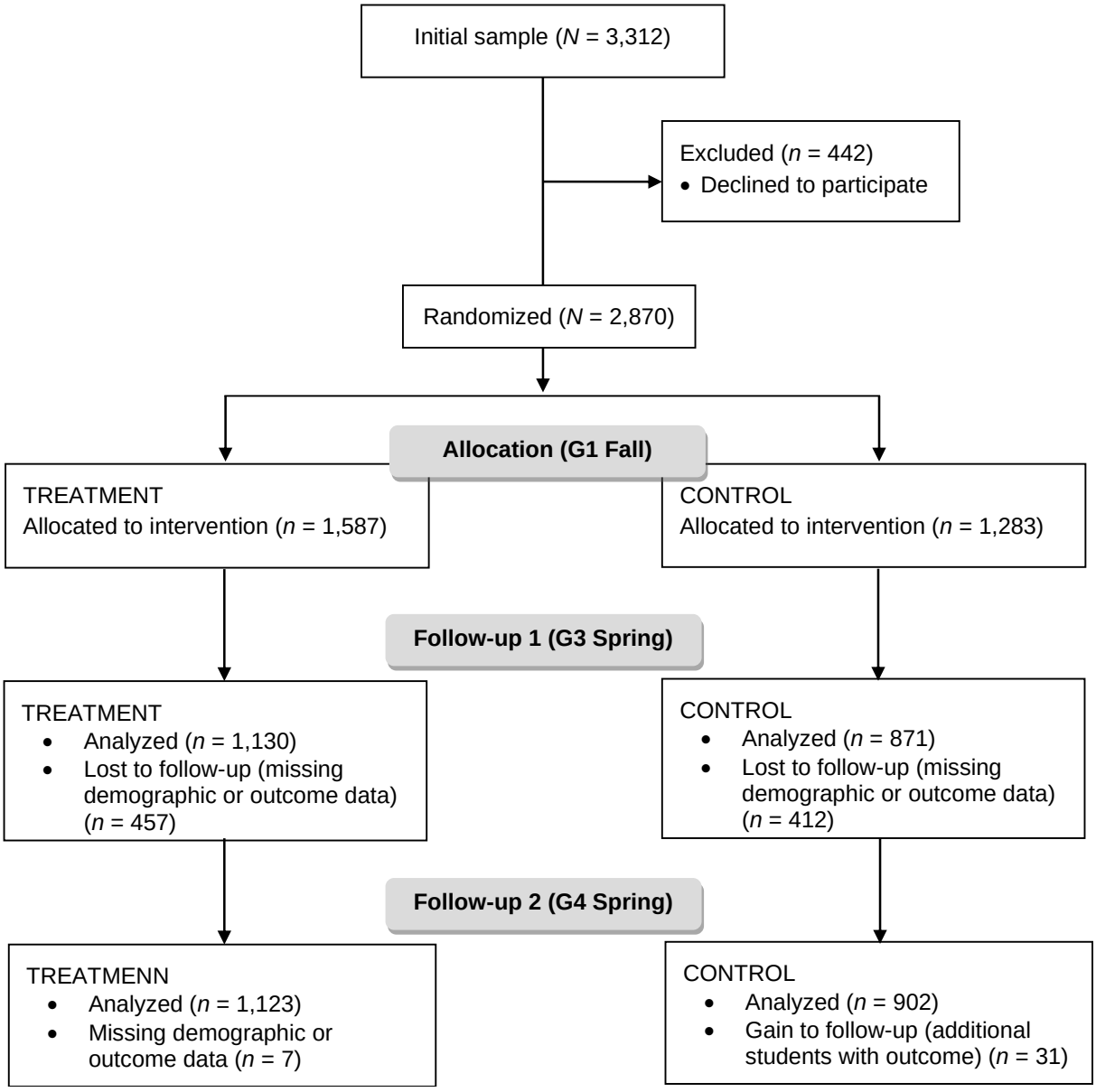
Visual Displaying Study Design for Implementation and Intervention Activities and Evaluation of Long-Term Impact



Note. DSVK = domain-specific vocabulary knowledge. DSR = domain-specific reading, DGR = domain-general reading, DGM = domain-general mathematics

Figure 3

MORE Intervention Consort Diagram for Study Sample



Time to Transfer: Long-Term Effects of a Sustained and Spiraled Content Literacy Intervention in the Elementary Grades

Online Supplementary Materials

Table S1

Moderation Analysis by Student Characteristics on Grade 3 and 4 Domain-General Reading Comprehension and Mathematics

Table S-A.1 presents results from moderation analyses involving interactions between MORE treatment and student demographic characteristics (i.e., baseline MAP reading in Grade 1, Black or Hispanic student, English learner, and low or middle socioeconomic status [SES] neighborhoods) on reading comprehension and mathematics outcomes for Grades 3 (see Table 7 for main-effect results) and 4 (see Table 8 for main-effect results)

	Grade 3 domain-general reading comprehension				Grade 3 Mathematics			
	Baseline MAP reading	Black or Hispanic	English learners	Low or mid SES (school)	Baseline MAP reading	Black or Hispanic	English learners	Low or mid SES (school)
Main MORE effect	.13 (.07)*	.21 (.07)**	.10 (.04)*	.21 (.11)*	.12 (.053)*	.11 (.05)*	.10 (.04)*	.17 (.07)**
Interaction effect	.02 (.05)	-.13 (.07)	.04 (.06)	-.12 (.11)	.00 (.04)	.01 (.06)	.07 (.04)	-.07 (.08)
N	2,001	2,001	2,001	2,001	1,987	1,987	1,987	1,987

	Grade 4 domain-general reading comprehension				Grade 4 Mathematics			
	Baseline MAP reading	Black or Hispanic	English learners	Low or mid SES (school)	Baseline MAP reading	Black or Hispanic	English learners	Low or mid SES (school)
Main MORE effect	.10 (.07)	.13 (.06)*	.12 (.04)**	.15 (.09)	.13 (.06)*	.15 (.05)**	.13 (.04)	.23 (.05)***
Interaction effect	-.02 (.06)	-.02 (.06)	-.002 (.06)	-.05 (.10)	-.03 (.06)	.02 (.07)	.14 (.05)**	-.09 (.08)
N	2,025	2,025	2,025	2,025	2,020	2,020	2,020	2,020

Note. Models include controls for demographics, randomization block, and cubic baseline reading and mathematics scores. Standard errors were clustered at school level, the unit of randomization. The results for the domain-general reading comprehension and mathematics in Grades 3 and 4 reveal no evidence of moderation by student characteristics, suggesting that MORE was equally effective across a wide range of subgroups.
†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

Table S2

Sensitivity Analysis: Intent-to-Treat (ITT) Intervention Effects on Domain-Specific Vocabulary Knowledge (Scores for Total, Taught, and Untaught Words) from Grades 1 to 3

The impact analyses presented in the main text (Table 6) utilized clustered standard errors at the school level. Below, we replicate the analysis using random effects for schools in multilevel models.

	Grade 1			Grade 2			Grade 3		
	Total	Taught	Untaught	Total	Taught	Untaught	Total	Taught	Untaught
ITT Science	.34 (.06)***	.33 (.06)***	.23 (.06)***	.07 (.06)	-.002 (.06)	.15 (.06)*	.14 (.05)**	.12 (.05)**	.15 (.05)**
ITT Social Studies	.63 (.06)***	.75 (.06)***	.25 (.05)***						
N[a]	2,419	2,419	2,419	2,126	2,126	2,126	1,341	1,341	1,341

Note. The sample size was attenuated by missing data from students who did not take the vocabulary knowledge during the pandemic 2020-2021 school year. ITT models include controls for demographics, randomization block, and cubic baseline reading and mathematics scores. We include random effects for schools, the unit of randomization. Social studies vocabulary knowledge was assessed only in Grade 1.

[a]*N* reported is for science vocabulary knowledge.

†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

Table S3

Sensitivity Analysis: Intent-to-Treat (ITT) and Treatment-on-the-Treated (TOT) Intervention Effects on Grade 3 Domain-Specific Reading Comprehension, Grades 3 and 4 Domain-General Reading Comprehension, and Grades 3 and 4 Mathematics Outcomes

The impact analyses presented in the main text (Tables 7 and 8) utilized clustered standard errors at the school level. Here, we replicate the analysis using random effects for schools for both the ITT and 2SLS models in multilevel models.

Source	Grade 3 Domain-specific reading comprehension				Grade 3 Domain-general reading comprehension	Grade 3 Mathematics
	Total	Near-transfer passage	Mid-transfer passage	Far-transfer passage		
ITT	.13 (.06)*	.11 (.05)*	.11 (.05)*	.14 (.05)**	.10 (.04)*	.10 (.04)*
2SLS	.13 (.10)	.11 (.08)	.10 (.06)	.13 (.10)	.11 (.05)*	.12 (.05)**
N	1,292	1,292	1,261	1,210	2,001	1,987

Source	Grade 4 Domain-general reading comprehension	Grade 4 Mathematics
ITT	.11 (.03)**	.15 (.05)**
2SLS	.12 (.04)**	.17 (.06)**
N	2,025	2,020

Note. 2SLS = two-stage least squares. ITT and 2SLS analysis models included covariates for demographics, school randomization block, and baseline MAP reading and mathematics scores.
†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

Table S-A.4

Sensitivity Analysis: Intent-to-Treat (ITT) Intervention Effects on a Reduced Sample of Domain-Specific Vocabulary Knowledge (Scores for Total, Taught, and Untaught Words) from Grades 1 to 3

The results below replicate Table 6 but limit the sample to n = 1,341 students who took the domain-specific vocabulary knowledge assessments in Grade 3 to ensure that results were not driven by non-random attrition.

	Grade 1			Grade 2			Grade 3		
Sources	Total	Taught	Untaught	Total	Taught	Untaught	Total	Taught	Untaught
ITT Science	.36 (.07)***	.35 (.06)***	.25 (.07)***	.15 (.07)*	.09 (.07)	.19 (.06)**	.14 (.05)**	.12 (.05)*	.16 (.06)**
ITT Social Studies	.69 (.06)***	.80 (.05)***	.29 (.06)***						
N[a]	1,225	1,225	1,225	1,286	1,286	1,286	1,341	1,341	1,341

Note. The sample size was attenuated by missing data from students who did not take the vocabulary knowledge during the pandemic 2020-2021 school year. ITT models include controls for demographics, randomization block, and cubic baseline reading and mathematics scores. Standard errors were clustered at school level, the unit of randomization. Social studies vocabulary knowledge was assessed only in Grade 1.

[a]*N* reported is for science vocabulary knowledge.

†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

Table S5

Sensitivity Analysis: Intent-to-Treat (ITT) Intervention Effects on a Reduced Sample of Domain-Specific Vocabulary Knowledge (Scores for Total, Taught, and Untaught Words) from Grades 1 to 3

The results below replicate Table 6 but limit the sample to n = 1,292 students who took the domain-specific reading comprehension assessments in Grade 3 to ensure that results were not driven by non-random attrition.

	Grade 1			Grade 2			Grade 3		
Sources	Total	Taught	Untaught	Total	Taught	Untaught	Total	Taught	Untaught
ITT Science	.35 (.07)***	.35 (.06)***	.23 (.07)***	.13 (.07)	.07 (.07)	.18 (.06)**	.12 (.05)*	.09 (.05)*	.14 (.06)**
ITT Social Studies	.68 (.06)***	.78 (.05)***	.29 (.06)***						
N[a]	1,181	1,181	1,181	1,240	1,240	1,240	1,278	1,278	1,278

Note. The sample size was attenuated by missing data from students who did not take the vocabulary knowledge during the pandemic 2020-2021 school year. ITT models include controls for demographics, randomization block, and cubic baseline reading and mathematics scores. Standard errors were clustered at school level, the unit of randomization. Social studies vocabulary knowledge was assessed only in Grade 1.

[a]*N* reported is for science vocabulary knowledge.

†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

Table S6

Sensitivity Analysis: Intent-to-Treat (ITT) Intervention Effects on a Reduced Sample of Domain-Specific Vocabulary Knowledge (Scores for Total, Taught, and Untaught Words) from Grades 1 to 3

The results below replicate Table S-A.4 but limit the sample to n = 1,341 students who took the domain-specific vocabulary knowledge assessments in Grade 3 to ensure that results were not driven by non-random attrition. Clustered standard errors at the school level were used. Here, we replicate the analysis using random effects for schools for both the ITT and 2SLS models.

	Grade 1			Grade 2			Grade 3		
Sources	Total	Taught	Untaught	Total	Taught	Untaught	Total	Taught	Untaught
ITT Science	.36 (.07)***	.35 (.06)***	.25 (.07)***	.15 (.07)*	.09 (.07)	.19 (.06)**	.14 (.05)**	.12 (.05)*	.16 (.06)**
ITT Social Studies	.69 (.06)***	.80 (.05)***	.29 (.06)***						
N[a]	1,225	1,225	1,225	1,286	1,286	1,286	1,341	1,341	1,341

Note. The sample size was attenuated by missing data from students who did not take the vocabulary knowledge during the pandemic 2020-2021 school year. ITT models include controls for demographics, randomization block, and cubic baseline reading and mathematics scores. Social studies vocabulary knowledge was assessed only in Grade 1.
[a]*N* reported is for science vocabulary knowledge.
†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

Table S7

Sensitivity Analysis: Intent-to-Treat (ITT) and Treatment-on-the-Treated (TOT) Intervention Effects on Grade 3 Domain-Specific Reading Comprehension, Grades 3 and 4 Domain-General Reading Comprehension, and Grades 3 and 4 Mathematics Outcomes

The results below replicate Table S5 but limit the sample to the sample of n = 1,292 students who took the domain-specific reading comprehension assessments in Grade 3 to ensure that results were not driven by non-random attrition using random effects for schools for both the ITT and 2SLS models.

Source	Grade 3 Domain-specific reading comprehension				Grade 3 Domain-general reading comprehension	Grade 3 Mathematics
	Total	Near-transfer passage	Mid-transfer passage	Far-transfer passage		
ITT	.14 (.06)*	.11 (.06)	.11 (.06)*	.14 (.06)**	.10 (.06)	.09 (.06)
2SLS	.14 (.07)*	.12 (.06)	.11 (.06)	.14 (.06)	.11 (.06)	.10 (.06)
N	1,292	1,292	1,259	1,208	1,195	1,186

Source	Grade 4 Domain-general reading comprehension	Grade 4 Mathematics
ITT	.10 (.05)*	.14 (.06)*
2SLS	.11 (.05)*	.14 (.06)*
N	1,157	1,155

Note. 2SLS = two-stage least squares. ITT and 2SLS analysis models included covariates for demographics, school randomization block, and baseline MAP reading and mathematics scores.
†p < .10, *p < .05, **p < .01, ***p < .001.

Table S8

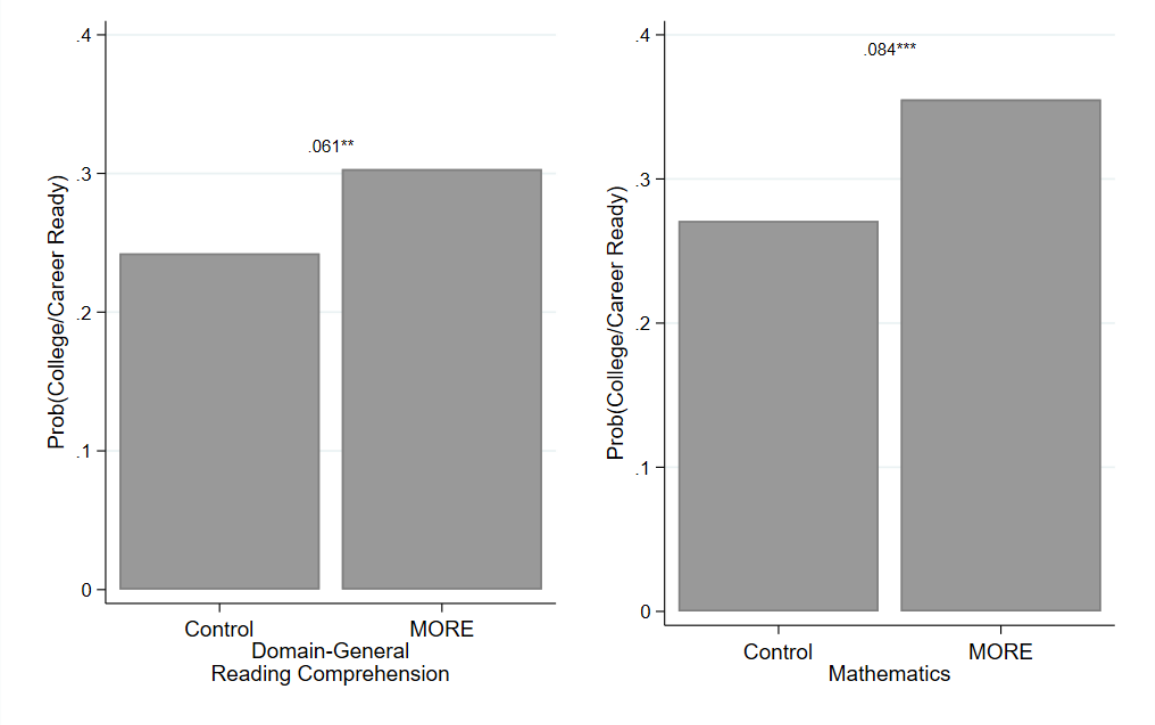
Sensitivity Analysis: Intent-to-Treat (ITT) Effects on Alternative Standardized Reading and Mathematics Assessments (Measure of Academic Progress, MAP) in Grade 3

	G3 MAP Reading	G3 MAP Mathematics
ITT	.07 (.034)*	.083 (.032)**
N	2,065	2,082

Note. ITT models included covariates for demographics, school randomization block, and baseline MAP reading and mathematics scores.
†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

Figure S1

Results of Logistic Regression on College and Career Ready (CCR) Predicted Probability of Exceeding CCR Standard for G4 Domain-General Reading Comprehension and Mathematics



Note. Marginal probability differences derived from a logistic regression model with clustered standard errors at the school level. The marginal probabilities are 24.2% (Control) and 30.3% (Treatment) in Reading, and 27.1% (control) and 35.5% (treatment) in math.

Results of Cost Analysis: Description of Methods for Calculating Program Costs and Cost-Effectiveness

We assess the per pupil annual costs to determine the feasibility of scaling MORE beyond the study site. Using the ingredients method (Lewin & McEwan, 2001), we identified the resources needed to implement MORE for a single year (e.g., personnel, materials, equipment). To date, very few scholars who have conducted RCTs of elementary grade content literacy interventions generate cost analyses, thus making it difficult to determine the scalability and cost-effectiveness of existing programs. Our analysis describes the money and time needed to implement a sustained and spiraled content literacy intervention. Such cost analyses are critical to understanding the feasibility of scaling interventions. To calculate MORE's cost, the research team compiled all expenses for the project and identified and aggregated the costs into key areas: personnel, materials/equipment, etc. In addition, after the completion of the intervention, the authors also collected compensation data on district personnel to reflect more accurately what the program would cost the district.

We used the E$timator© website developed by Teachers College (Hollands et al., 2015) to calculate our cost per student. E$timator© has several advantages over simply identifying the ingredients and their associated value for the project. First, since our experiment took place over three years, E$timator© calculates the present-day value of future expenses, taking into account the fact that expenses in the future will be less burdensome than in the current years. For example, if we know that we need to incur $1,000 worth of expenses two years from now, we could invest the money in a bank account and earn interest, making it worth *more* in the second year than the $1,000 we have today. Since interest rates were at historic lows during the period of intervention, we utilized a relatively low discount rate of 1%. We also explored other rates to understand how it would affect the calculations and the changes were relatively small. Second, E$timator© adjusts the costs for inflation using the CPI-U and translates them into 2019 dollars (the first year of our intervention). Third, E$timator© also takes geographical differences in cost into account and ensures that the prices are nationally representative. For example, as personnel in our context were hired in both Massachusetts and other states, the program will adjust the personnel costs to better reflect the national average. Having estimated the cost of the intervention ingredients, we then leverage the number of students that were served during the program to calculate a cost per student. We also break these costs down by personnel, materials/equipment, books, and other types of expense categories.

Several assumptions in our cost estimate should be highlighted. First, our intervention took place during the school day as part of the district's regular scope and sequence. As such, we have not added the facilities cost to the school. Second, in each year of program implementation, we hired teachers and literacy facilitators (i.e., school coaches) to support implementation. Teachers were required to complete 3 hours of professional development, attend a 1-hour meeting with the grade-level team, and provide 1 hour of general administrative tasks related to the program. Facilitators attended a 3-hour training, attended or led six hours of meetings, and were also allocated about an hour of administrative tasks. In the years after the intervention, we've worked with the district to institutionalize compensation for these costs and have used the hourly rate for teachers/facilitators and adjusted them to the national averages. These costs were included as personnel charges in our cost assessment. Third, the research staff prepared for and helped facilitate interaction with district leadership, other central office personnel (e.g., specialists, area superintendents), literacy facilitators, and teachers. The research staff also had

the primary responsibility for printing and shipping formative assessments and materials, for example. As such we included the staff time and travel costs associated with these ingredients of the program. Finally, as noted in the main paper, during the 2020-21 school year, we decided to offer MORE to all 3rd grade students in the study schools. For the purposes of these cost estimates, we only include the cost of the treatment group. It is also worth noting that many pieces of the intervention structurally changed during the course of implementation. For example, in the context of COVID-19, all formative assessments were provided online during the 2020-21 school year, where the two prior years the assessments were printed and shipped and returned to the team by mail. Similarly, the iPads that are included in our cost estimates were not needed in the last year of the intervention because our web-based app operated on school computers. These changes would likely have had little effect on the intervention itself but would have resulted in significant cost reductions if introduced at the beginning of the program. As such, we view our cost estimates as conservative and it is likely that if a new district were to implement the intervention, it could be done for much less.

Results of the Costs Analysis

As shown in Table 9, in our trial we spent $411.85 per student (in 2019 dollars) using nationally representative prices. These estimates represent what it would cost to implement MORE as it was in 2018-19 through 2020-21. The cost-effectiveness of MORE is comparable to other literacy interventions that have measured cost (Kim et al., 2016; Jacob et al., 2015; Gray et al., 2022). We believe these costs are an upper estimate of the potential future cost given innovations and changes to the program that happened in the third and final year of the intervention that will continue in future implementation. For example, the largest proportion of the cost is related to Non-Teacher/Coach Personnel. However, much of these costs were reduced in the final year as a significant portion of personnel time in prior years was spent printing, shipping, and organizing lesson materials and formative assessments that are now all online.

Results of Cost Effectiveness Analysis

MORE is relatively low in cost when compared to other interventions with similar effect sizes. We estimate a cost of the intervention of $411.85 per student (in 2019 US dollars) including personnel, books, computer & technology (e.g., iPads), travel (e.g., site visits, and convening for participants), and other miscellaneous expenses. The closest comparison we can find to MORE is *Zoology One*, an integrated science and literacy program for kindergarten. Program expenses were similar at $480 per student. Given the age of the children, effects were assessed for more constrained skills like passage comprehension and letter naming fluency, ranging from 0.16 to 0.28 standard deviations (Gray et al., 2022) There are at least two other studies related to literacy that also analyzed cost. A cost analysis of Project READS, a summer book program that provided free books and paper activities for elementary school students, found that program expenses were relatively low (between $250-$480 per student) for the observed ITT effects of 0.04 standard deviations (Kim et al., 2016). Similarly, Reading Partners, a program that uses volunteers to provide one-on-one tutoring to struggling readers in elementary school, found effects of 0.10 standard deviations for costs between $480-$1,270 per student (Jacob et al., 2015). Finally, a more general meta-analysis on school spending across grade levels, focus areas, and context found that increasing spending by $1,000 per pupil for four years improved test scores by 0.03 standard deviations (Jackson & Mackevicius, 2023). Thus, MORE is comparatively similar or more cost-effective than other evidence-based, large-scale literacy interventions, which have demonstrated long-term impacts.

Domain-Specific (Science) Reading Comprehension Passages and Psychometric Properties

A. Near Transfer Passage (Underlined Words Appear in the Academic Vocabulary Networks)

(1) A human is a primate. In primates, the heart sends oxygen in the blood all around the body. The body must have oxygen to function properly. The strong heart muscle contracts to pump blood all around the body. A healthy heart never rests.

(2) But what happens when a person's heart gets weaker? People with weak hearts need help. Scientists have to study a medical mystery. Then they might be able to help people.

(3) First, scientists need to diagnose the reasons that a human heart gets weak. They need to do tests. But sometimes it would be hard on a sick person to have tests done. So, scientists study animals that have body systems like humans. Scientists knew that the macaque monkey's heart is similar to the human heart. So, they did tests with macaque monkeys. The scientists learned what happens when a monkey has a heart attack. After a heart attack, the monkey's heart muscle has scar tissue where it was damaged. The scar tissue cannot contract like a strong heart muscle can. Then the monkey's heart cannot pump enough blood to give the body the oxygen it needs. Without enough oxygen, the monkey's body cannot function properly.

(4) Second, scientists wanted to try out new solutions that would keep the heart working. Sometimes the new solutions might fail. So, scientists didn't want to try the solutions with people right away. They tried a solution with macaque monkeys.

(5) Scientists developed an idea. Then, they tested an idea. They injected human stem cells from the human heart into the monkey's heart. After four weeks the human heart cells grew where the monkey's scar tissue was. After three months, the monkey's heart got stronger.

(6) Now scientists are using what they found in this experiment to help humans who have had heart attacks.

Text statistics:
Lexile® Measure: 610L – 800L
Mean Sentence Length: 10.07
Mean Log Word Frequency: 3.50
Word Count: 299

B. Mid Transfer Passage (Underlined Words Appear in the Academic Vocabulary Networks)

(1) Some birds that are here today are smaller than they were many years ago. Why is this happening? It is a mystery. How do scientists study a mystery?

(2) First, scientists try to understand the bird's skeletal and muscular systems. Those systems help birds to fly. One thing scientists know is that birds' bones are hollow inside. So, they don't weigh much. Birds' skeletons weigh less than their feathers. They have also learned that a bird wing works like a human arm. Birds contract their chest muscles to make their wings flap.

(3) Second, scientists try to diagnose the reasons that the birds' bodies might have changed over time. Some scientists thought that birds' skeletal or muscular systems might have changed. So, they studied migratory birds in North America. Migratory birds fly long distances to warmer places for the winter.

(4) Third, scientists measured the physical characteristics of the birds. They measured the size of the birds' bones. They found some of the birds' bones were shorter than they were 40 years ago. For example, the lower leg bone of many birds is smaller.

(5) Fourth, scientists measured the migratory birds' wings. They found the birds' wings are longer than they used to be.

(6) Many scientists asked, "Why are the birds' bones smaller than they used be. But their wings are longer?" They thought about the long distance the birds fly in winter. They thought that the birds' bodies changed to adapt to the long winter flight. They would need a lot of energy for the flight. A smaller body would have smaller muscles and less fat. But smaller bones and longer wings would make flying easier.

(7) Over time, these characteristics of birds have become more common.

(8) The scientists came up with some very good ideas about why these skeletal changes might be occurring.

Text statistics:
Lexile® Measure: 610L - 800L
Mean Sentence Length: 10.00
Mean Log Word Frequency: 3.48
Word Count: 299

C. Far Transfer Passage

(1) In many ways, skyscrapers are like the human body. They have bones that keep it tall and strong and skin to protect the body. They have a command center that is like a brain.

(2) How do engineers design and build strong and smart skyscrapers? It is very challenging to build a skyscraper because there are many parts that have to work together. If one part is broken, the skyscraper will not work properly.

(3) First, engineers need a strong and tall frame that supports the floors and walls. This frame is like the bones of a skyscraper. The bones of a skyscraper are made of concrete, steel and other materials. The columns are like the backbone of a human body. The columns go up. The steel columns hold up each floor. There are also steel beams that go across each floor. The steel columns and beams must be strong to resist the force of gravity. The steel columns and beams must work together to form a strong backbone.

(4) Second, engineers need to design a command center. The command center is a computer that sends signals through wires in a skyscraper. For example, it sends signals to help bring fresh air into rooms, just like your lungs. The command center also controls the heating and air conditioning. It helps the building cool down when it’s too hot and warm up when it’s too cold.

(5) A skyscraper has many different parts that work together. The bones make the building tall and strong. The skin protects the inside of the building. The skin of a skyscraper can be made of metal, stone, or glass. The brain controls the building temperature.

(6) The anatomy of a skyscraper is like a human body. All the parts have to work together. That's how a city skyscraper works properly.

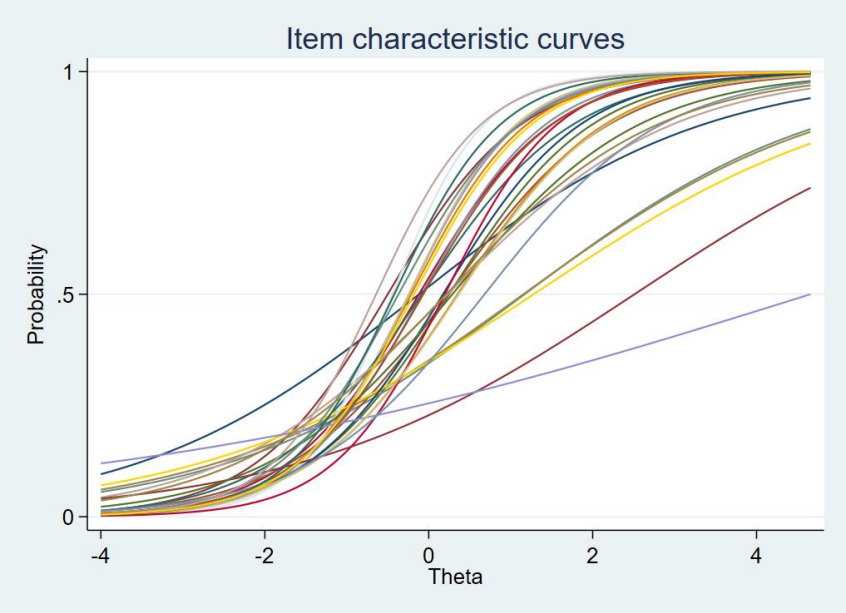
Text statistics:
Lexile® Measure: 610L – 800L
Mean Sentence Length: 10.71
Mean Log Word Frequency: 3.61
Word Count: 300

Psychometric Properties of the Grade 3 Content Comprehension and Vocabulary Assessments

1) Descriptive Statistics for Item Properties based on a 2PL Model

Test	type	mean	min	max	SD
Content Comp.	Diff	0.34	-0.64	4.65	1.06
Content Comp.	Disc	1.09	0.23	1.79	0.4
Vocab	Diff	-0.16	-1.38	4.12	1.2
Vocab	Disc	1.36	0.4	3.22	0.63

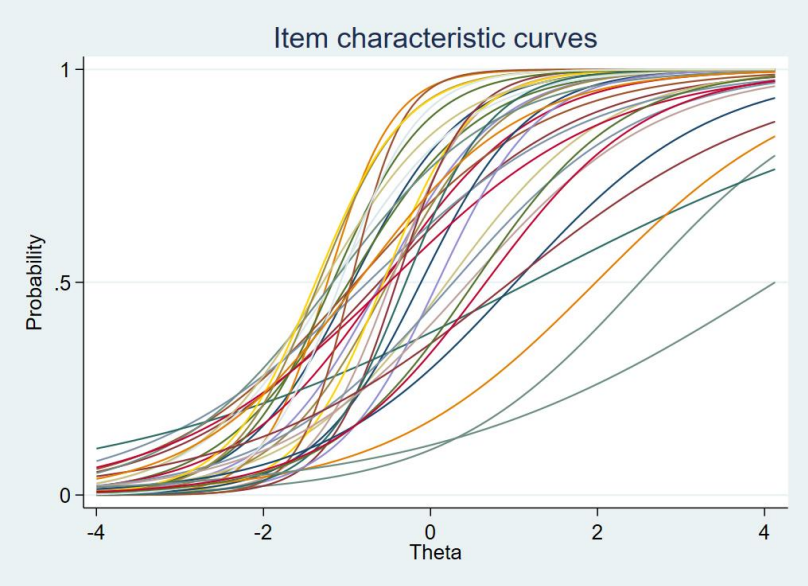
2) 2PL Item Characteristic Curves



item	Disc.	Diff.
1	0.58	-0.12
2	1.23	-0.5
3	1.15	0.25
4	1.42	-0.21
5	0.55	1.16
6	1.24	-0.12
7	1.49	-0.24
8	1.53	-0.24
9	1.03	0.24
10	1.33	-0.08
11	1.09	-0.08
12	0.53	1.15
13	0.73	0.23
14	0.49	1.27
15	1.79	-0.45
16	1.2	0.17
17	0.49	2.49
18	0.88	0.28
19	1.12	0.35
20	1.36	-0.37
21	1.47	0.2
22	0.23	4.65
23	1.11	0.36

24	1.28	-0.07
25	0.93	0.68
26	1.55	-0.42
27	0.78	0.22
28	1.57	-0.64
29	1.4	-0.18

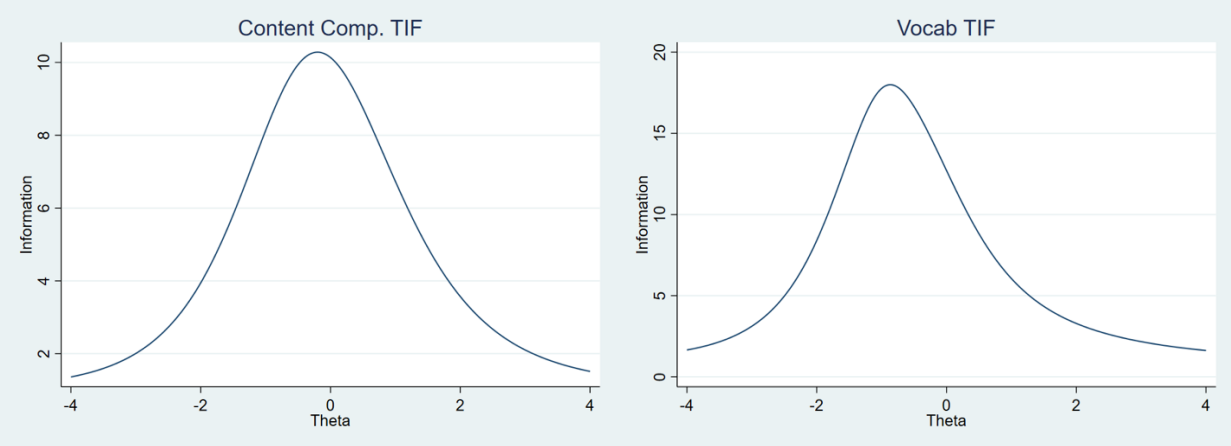
G3 Vocabulary



item	Disc.	Int.
1	1.53	-0.93
2	0.85	-0.61
3	1.29	-0.98
4	2.56	-1.23
5	0.85	2.51
6	1.12	-0.56
7	1.44	-0.59
8	1.05	0.2
9	0.88	-0.9
10	0.75	-0.76
11	0.4	1.2
12	1.47	-0.5
13	2.08	-0.48
14	1.95	-0.56
15	1.26	-1.17
16	1.57	-0.11

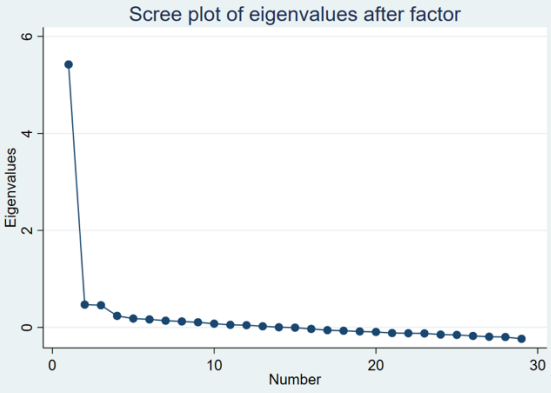
17	2.4	-0.4
18	1.76	-1.17
19	0.78	1.99
20	1.03	-1.17
21	0.76	-0.5
22	1.65	0.08
23	1.32	-1.29
24	3.22	-0.95
25	0.89	0.27
26	1.95	-0.3
27	1.96	-1.33
28	0.87	0.46
29	1.88	-1.38
30	2.15	-1.08
31	0.85	1.02
32	0.62	0.95
33	1.14	0.53
34	1.04	-0.88
35	0.49	4.12
36	1.04	0.66

3) Test Information Functions

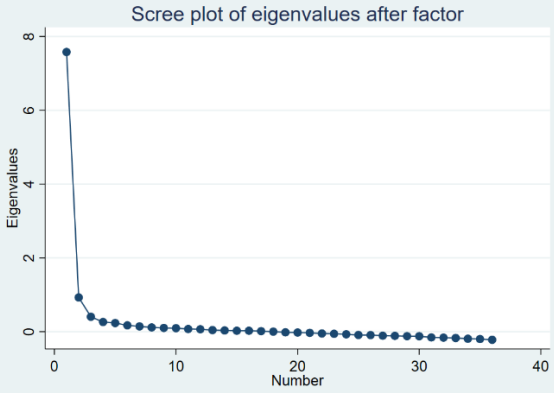


4) EFA Scree Plots

Content Comprehension



Vocabulary



5) CFA Fit Statistics

Content Comprehension: RMSEA = 0.03, CFI = 0.922, TLI = 0.916, SRMR = 0.032

Vocabulary: RMSEA = 0.031, CFI = 0.921, TLI = 0.916, SRMR = 0.034

6) Correlations between 2PL Theta Estimates and Sum Scores used in the analysis

Content Comprehension: 0.945
Vocab: 0.967

Sample Automated Concept Network for the Word "systems"

The following automated concept map was developed for the word "*systems*." As noted in purple, the concept "*body systems*" appears in the adjacent neighborhood to system, as does "*organ*," an untaught word, but one that connects the *systems* (*skeletal*, *muscular*, *nervous)* that were the focus of the Grade 3 MORE lessons.