Scientific Studies of Reading

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hssr20

The Effects of Bookworms Literacy Curriculum on Student Achievement in Grades 2-5

Henry May, John Z. Strong & Sharon Walpole

To cite this article: Henry May, John Z. Strong & Sharon Walpole (03 Dec 2023): The Effects of Bookworms Literacy Curriculum on Student Achievement in Grades 2-5, Scientific Studies of Reading, DOI: 10.1080/10888438.2023.2284811

To link to this article: https://doi.org/10.1080/10888438.2023.2284811

SCIENTIFIC STUDIES OF READING
https://doi.org/10.1080/10888438.2023.2284811

The Effects of Bookworms Literacy Curriculum on Student Achievement in Grades 2-5

Henry May [a], John Z. Strong [b], and Sharon Walpole [a]

[a]School of Education, University of Delaware, Newark, Delaware, USA; [b]Graduate School of Education, University at Buffalo, Buffalo, New York, USA

ABSTRACT

Purpose: In this study, we investigated the effects of a schoolwide program, Bookworms K-5 Reading and Writing, on student achievement.

Method: The study included seven cohorts of students ($N = 8{,}806$) in grades 2–5 in 17 elementary schools across three school years. We used a comparative interrupted time-series design, conducting multilevel growth curve models of Measures of Academic Progress reading scores with up to 10 data points per student. By modeling each student's growth curve, including a time by treatment interaction term, we were able to estimate the change in students' achievement trajectories corresponding to the implementation of Bookworms.

Results: Results confirm a significant positive impact of Bookworms on achievement, with gains compounding over time and producing an overall standardized effect size of .26 by the end of 5th grade. Students who began third grade with relatively weaker achievement experienced more growth than those with average achievement, and those with average achievement experienced more growth than those with the highest achievement.

Conclusion: This study provides evidence that a comprehensive literacy curriculum that emphasizes high-volume reading of grade-level texts and the use of evidence-based instructional practices produces positive effects on student achievement for students with a range of initial reading achievement.

Recent reporting has sparked interest in elementary reading instruction (e.g., Hanford, 2018) not seen since the report of the National Reading Panel (2000). Concurrently, reviews of reading curricula and their misalignment with research have made their way into the media landscape (Adams et al., 2020). Despite a strong research base for reading and writing pedagogy (Foorman et al., 2016; Graham et al., 2012; Shanahan et al., 2010), literacy achievement remains problematic at all grade levels, especially for groups of children chronically underserved by public schools (National Center for Education Statistics, 2022). A curriculum aligned with this research base may increase use of evidence-based pedagogy and improve achievement.

The move to the Common Core State Standards (CCSS; National Governors Association Center for Best Practices National Governors Association & Council of Chief State School Officers, 2010 and its many derivations challenged teachers' content knowledge and put pressure on schools to link professional learning to curricular materials (Kane et al., 2016). Educative curricular materials are a possible support, especially if they are situated and grounded in practice (Davis et al., 2017), but curriculum

CONTACT Henry May hmay@udel.edu School of Education, University of Delaware, Newark, Delaware, USA

This research was conducted under the authority of the Institutional Review Board at the University of Delaware, in compliance with US Federal Policy for the Protection of Human Subjects. As an exempt research study involving anonymized data, it was not necessary to obtain informed consent from students and parents whose data are included in this study.

alone may not build teacher knowledge (Cohen et al., 2017). Thus, the combination of high-quality curricula and professional learning for teachers is a promising avenue for improving educational equity and student achievement (Learning Forward, 2018). For those reasons, we view the effects of research-informed curriculum design coupled with curriculum-specific professional learning as worthy of ongoing, intense empirical examination.

It is possible that strong comprehensive curricula aligned with both research and standards might bridge the longstanding research-to-practice gap. Schools adopt curricula produced by publishers. If those curricula do not produce desired outcomes, schools may layer on interventions produced by researchers. What is missing in the debate is a concerted effort to narrow the research-to-practice gap through deep connections between research and practice communities (Farley-Ripple et al., 2018), by providing evidence-based curricula designed by researchers to reduce reliance on stand-alone interventions. In this study, we investigated the effects of a comprehensive elementary literacy curriculum, *Bookworms K-5 Reading and Writing* (Bookworms; Open Up Resources, 2018), on student achievement for seven cohorts of students implemented as a whole-school program in all 17 elementary schools in one school district.

Research on whole-school literacy programs

In a Response to Intervention framework, elementary literacy curricula are tiered. High-quality, evidence-based instruction is intended to be provided in all tiers – core (tier 1) reading curricula are designed for all students, small-group (tier 2) interventions target students who need additional instruction, and intensive (tier 3) interventions target students who need more intensive support (Gersten et al., 2009). A recent review of elementary reading programs concluded that multi-tiered schoolwide approaches combining professional development and core instruction (e.g., Success for All; Cheung et al., 2021) and whole-class tier 1 approaches (e.g., Peer-Assisted Learning Strategies; Fuchs et al., 1997) were more effective for improving the achievement of students with reading difficulties than interventions delivered only at tier 2 or tier 3 (Neitzel et al., 2022). Therefore, although schools typically purchase tier 2 or tier 3 curricula to improve achievement for some students, this approach may be less effective than using a high-quality tier 1 program for all students.

Bookworms is a schoolwide program with a scope comparable to the Comprehensive School Reform projects targeted for funding in the mid-1990s. Such programs emphasize well-defined instructional practices and organizational changes to increase feasibility (Borman et al., 2003). This broad scope combines evidence-based curricular materials, specific instructional practices, ongoing assessment, and professional development and coaching (Cheung et al., 2021; Neitzel et al., 2022). Comprehensive School Reforms have been associated with improved achievement (Borman et al., 2003; May & Supovitz, 2006), but these studies were conducted prior to the shifts in instruction targeted in the CCSS and in Bookworms: (1) text complexity and academic language, (2) discussion and writing grounded in text-based evidence, and (3) knowledge building through increased emphasis on informational texts (National Governors Association & Council of Chief State School Officers, 2010).

Bookworms curriculum design and theory of change

The Bookworms design process is the result of deep connections between research and practice. Bookworms began as a Reading First research-practice partnership from 2004–2006 (Walpole et al., 2011), continued in a Striving Readers partnership from 2012–2017 (Pasquarella, n.d.), was broadened to include classroom libraries for wide reading through a U.S. Department of Education Innovative Approaches to Literacy grant in 2015, and was revised for release as an open education resource (OER) with the support of Open Up Resources in 2018. Each of these partnerships provided the opportunity to improve consistency with research and standards and to make changes to increase feasibility.

Bookworms is adopted schoolwide because it requires training, scheduling, and assessment commitments typically outside the control of classroom teachers. Bookworms requires three 45-minute

instructional blocks each day: highly structured Shared Reading and English Language Arts (ELA) blocks, and a partially scripted Differentiated Instruction block. All three blocks are implemented for the full school year and for all students in grades K-5. Together, Shared Reading and ELA form a 90-minute daily curriculum aligned with grade-level standards (National Governors Association & Council of Chief State School Officers, 2010). While students receive teacher-led, whole-class reading instruction that is similar in purpose to the whole-class portion of a traditional core reading program in Shared Reading each day, the ELA block alternates between interactive read aloud and composition units every 1–2 weeks. The Differentiated Instruction block uses a screening and diagnostic assessment protocol to place students in daily foundational skills interventions or enrichment groups. Lesson plans are offered online at no cost, but schools must purchase trade books to read in Shared Reading and ELA and a manual for Differentiated Instruction. The three instructional blocks were designed to minimize time spent on developing word recognition through a highly-specific protocol in Differentiated Instruction, allowing for more time devoted to developing academic language, knowledge, and writing in Shared Reading and ELA.

The theory of change for Bookworms is displayed in Figure 1. In the following sections, we describe how the theory of change influenced grade-level standards alignment, selection of high-quality authentic texts, inclusion of data-based differentiation of foundational skills instruction, and comprehensive lesson planning with repetitive instructional routines for reading and writing. The combined focus on reading skills, fluency, and knowledge-building read alouds make it "distinctive" in the elementary literacy curriculum landscape (Wexler, 2019, p. 250). These features help ensure that all students in a school receive grade-level, challenging, evidence-based instruction. While the curriculum covers kindergarten through fifth grade, we confine our description to grades two through five for the purposes of this study.

Grade-level standards alignment

The Bookworms theory of change posits that feasibility is enhanced when real-life requirements of teaching are addressed in the curriculum materials teachers are provided. To be acceptable to school leaders and teachers, a comprehensive elementary literacy curriculum must be aligned to standards. Bookworms was

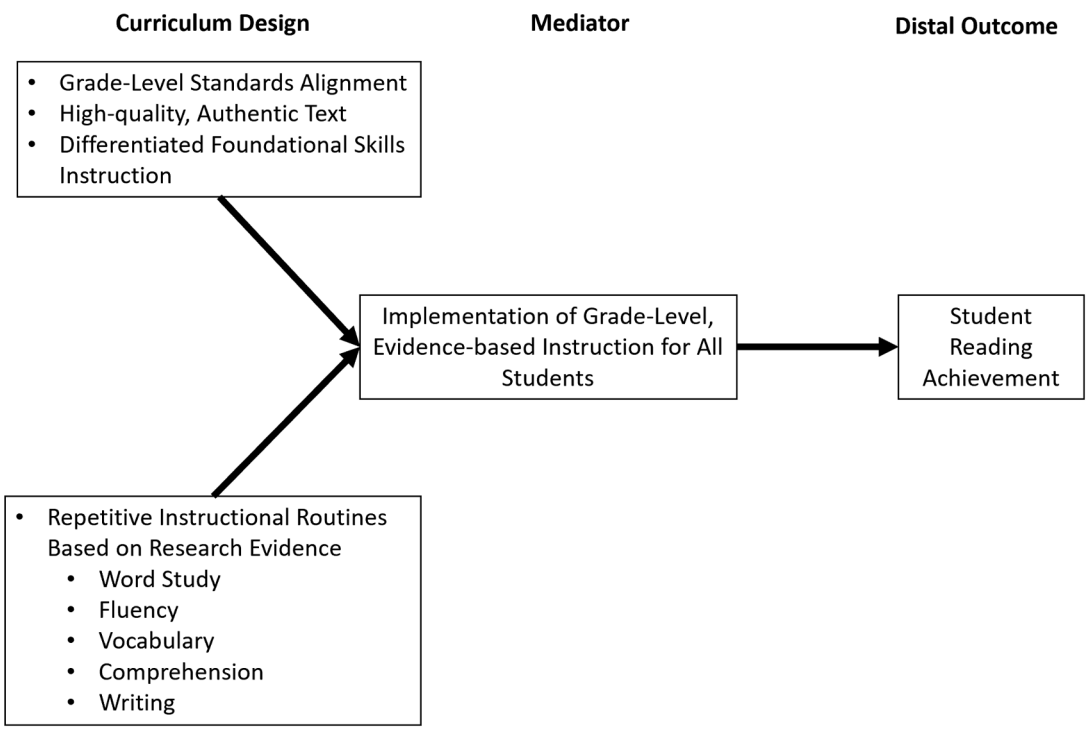

Figure 1. Bookworms curriculum theory of change.

designed to address the CCSS in reading, including foundational skills (i.e., phonemic awareness, phonics and word recognition, and fluency), writing, speaking and listening, and language (National Governors Association & Council of Chief State School Officers, 2010). These instructional targets must be integrated with one another and provided in specific doses (Gabriel, 2021). For example, because the rate at which children master foundational skills will vary (Paris, 2005), instruction may be more efficiently provided in small groups of students differentiated by assessed needs instead of in whole groups in which some students will receive instruction in skills they have already mastered (Kuhn & Stahl, 2022). Therefore, Bookworms addresses standards in small-group or whole-group instruction in relative dosages across grade levels. Foundational skills standards are addressed in larger doses during the Differentiated Instruction block and at lower grade levels, while grade-level reading, writing, and language standards are addressed in larger doses during whole-group instruction in the Shared Reading and ELA blocks.

Bookworms addresses the CCSS at each grade level with an overarching emphasis on the development of academic language, which requires students to conceptualize and discuss ideas beyond their daily experiences using the formal communication structures and words found in narrative and informational texts (Foorman et al., 2016). Daily student reading, writing, and discussion in Bookworms requires and builds academic language. Teachers name and model comprehension processes during reading to teach students to grapple with labels for cognitive and metacognitive language. Students engage in discussions during and after reading that require and develop narrative and inferential language skills. The "soft-scripted" lessons (Neuman et al., 2021, p. 387) ensure that teachers always have open-ended text-based questions to guide and prompt these discussions. In addition, structured vocabulary and grammar instruction are nested in the language of complex texts students and teachers read.

High-quality authentic text

The CCSS redefined grade-level reading to make it more challenging in elementary grades, and the Bookworms theory of change embraces this. The Shared Reading and ELA blocks are unique in their high-volume, whole-group use of challenging, grade-level text with new segments read each day. All texts are authentic, stand-alone children's books rather than excerpts or specially-written instructional materials. Table 1 displays the number of books read in Shared Reading and ELA lessons, the range and mean level of text complexity (in Lexiles), and the total number of pages read in each grade. The Lexile range in each grade is aligned with the text complexity requirements in Appendix A of the CCSS (National Governors Association & Council of Chief State School Officers, 2010). Students also read widely and independently from books of their choosing with no restriction on text type or complexity during the Differentiated Instruction block when teachers are providing small-group instruction to other groups. This focus on real book reading may motivate teachers to use the materials as designed and students to persevere with the challenging work.

The texts in Bookworms, especially the informational texts, are organized for knowledge building. Multiple texts on the same topic allow students to build and leverage background knowledge (Cervetti et al., 2016), a factor increasingly important to reading comprehension as students advance through elementary school (Willson & Rupley, 2009). For example, second grade has science text sets on cycles, animals, and physical science, and social studies text sets on economics, Native Americans, and biographies of famous

Table 1. Characteristics of texts in shared reading and interactive read aloud lessons

	Books		Lexile Range		Mean Lexile		Total Pages	
Grade	SR	IRA	SR	IRA	SR	IRA	SR	IRA
Grade 2	20	36	380 L-800 L	480 L-AD1070L	561.5 L	694.4 L	1433	1461
Grade 3	15	17	490 L-850 L	650 L-1020 L	712 L	822.5 L	1567	1115
Grade 4	8	12	680 L-890 L	660 L-1150 L	823.8 L	894 L	1500	1057
Grade 5	11	14	640 L-1050 L	770 L-1160 L	862.7 L	927 L	1251	1042
Total	54	79	380 L-1050 L	480 L-1160 L	740 L	834.5 L	5751	4675

Notes. IRA = Interactive Read Aloud; SR = Shared Reading.

Americans. Third grade has text sets on geology, patterns in physical science, American government, and biographies linked by theme. Fourth grade has text sets on natural disasters and American history. Fifth grade has text sets on animal and plant cells, earth science topics, physics, the history of science, and the civil rights era. The positive achievement effects of such integrated literacy and content area instruction on vocabulary and comprehension are well documented (Hwang et al., 2022; Kim et al., 2021). The Bookworms text sets and reading and writing tasks at each grade are positioned to cultivate content knowledge and invite truly integrated literacy and content area instruction.

Differentiated foundational skills instruction

The Differentiated Instruction block is a standard-protocol, multiple-entry foundational skills intervention, employing the concept of assessment-based differentiation (Connor et al., 2011) in a simplified assessment protocol that does not require specialized software (Walpole & McKenna, 2017; Walpole et al., 2020). The goal is that each child receives 15 minutes of teacher-managed, small-group instruction and 30 minutes to complete text-based writing assignments and engage in self-selected reading each day. Group assignment is made with an oral reading fluency screening assessment and then a diagnostic phonics inventory for students below their grade-level benchmark in fluency (McKenna et al., 2017). Based on these results, students within a classroom are grouped within one of four general lesson types with a total of ten possible lesson sets. The first two lesson types (phonemic awareness and word recognition [PAWR] and word recognition and fluency [WRAF]) comprise 165 entirely scripted, teacher-managed code-focused lessons. The next two lesson types (fluency and comprehension [FAC] with multisyllabic decoding or without multisyllabic decoding and vocabulary and comprehension [VAC]) include an open-ended number of structured, teacher-managed meaning-focused lessons. See Table 2 for the instructional focus in each group. Time allows for teachers to teach three groups each day. If data indicate that students would be better served in more than three groups, teachers collaborate with colleagues or specialists to serve additional groups within the 45-minute block. This design makes concrete the needs of individual students and provides a feasible way to address them.

The plans for these groups pair evidence-based routines with explicit and systematic instruction. PAWR and WRAF lessons target automaticity in letter-sound identification, phonemic segmentation and blending, high-frequency word recognition, and text reading (Simmons et al., 2011; Vadasy & Sanders, 2011). In PAWR lessons, phonemic awareness activities (e.g., oral segmenting and blending) are linked to letter-sound (e.g., initial sound sorting) and decoding (e.g., segmenting and blending with print) activities (Ball & Blachman, 1991; Oudeans, 2003). In WRAF lessons, phonics instruction targets synthetic decoding, onset-rime decoding, and decoding by analogy (Savage et al., 2003; White, 2005) paired with decodable text

Table 2. Instructional Ffocus in Differentiated Instruction lessons

Lesson Type and Set	Instructional Focus
PAWR	
Basic Alphabet Knowledge	Automaticity drills for letter names and sounds with finger-point reading
Using Letter Sounds	Phonemic segmentation and blending with phonemic decoding
Using Letter Patterns	Phonemic segmentation and blending with onset-rime decoding and spelling
WRAF	
Blends and Digraphs	Synthetic decoding with repeated reading of decodable texts
R-Controlled Vowels	Synthetic decoding with repeated reading of decodable texts
Vowel-Consonant-e	Analytic decoding with repeated reading of decodable texts
Long-Vowel Teams	Decoding by analogy with repeated reading of decodable texts
FAC	
With Multisyllabic Decoding	Multisyllabic decoding and spelling instruction with repeated oral reading of authentic texts with inferential discussion
Without Multisyllabic Decoding	Repeated oral reading of authentic texts with inferential discussion
VAC	Vocabulary and text structure instruction with inferential discussion

Notes. FAC = Fluency and Comprehension; PAWR = Phonemic Awareness and Word Recognition; VAC = Vocabulary and Comprehension; WRAF = Word Recognition and Fluency.

reading to build fluency. FAC and VAC lessons are structured for teachers, with authentic texts selected by them for their students. The lessons use specific procedures for repeated oral reading (Schwanenflugel et al., 2009) and silent reading and discussion (McKeown et al., 2009), which target oral reading fluency and silent reading comprehension, respectively. For all groups, teachers monitor progress every three to six weeks and regroup students for more advanced lessons or reteach them until students meet proficiency goals (Coyne et al., 2013).

Repetitive instructional routines

The Bookworms theory of change identifies a set of literacy targets (i.e., word study, fluency, vocabulary, comprehension, and writing) for which the lessons employ instructional routines from the empirical evidence base in reading and writing instruction (Foorman et al., 2016; Graham et al., 2012; Shanahan et al., 2010). Instructional routines are used consistently before, during, and after reading, listening, or writing each day so that teachers and students can focus on the text content rather than the routines. Table 3 lists the instructional routines within and across blocks along with the research that was used to select them.

Word study

Shared Reading includes 10–15 minutes of daily, whole-group word study grounded in developmental theories of word reading and spelling development (Bear et al., 2020; Ehri, 1998). Second-grade word study begins with a review of vowel-consonant-e patterns and moves through vowel teams, ending with two-syllable words and inflected endings. Third-grade word study begins with the doubling principle (Ganske, 2008), and then moves to syllable-type decoding and spelling instruction for multisyllabic words (Bhattacharya, 2006). Word study in grades four and five continues with this instruction, linking spelling to syllable types and teaching word meanings (Knight McKenna, 2008).

Table 3. Instructional routines in Bookworms

Word Study Routines	Evidence of Effectiveness	SR	ELA	DI
Automaticity drills for letter names & sounds	Oudeans (2003)			✓
Phonemic segmentation and blending	Ball and Blachman (1991)			✓
Synthetic and analytic decoding	Savage et al. (2003)			✓
Decoding by analogy	White (2005)			✓
Compare and contrast spelling instruction	Graham et al. (2002)	✓		✓
Syllable-type spelling instruction	Knight McKenna (2008)	✓		✓
Oral Reading Fluency Routines	Evidence of Effectiveness	SR	ELA	DI
Repeated reading of decodable texts	Simmons et al. (2011)			✓
Repeated reading of natural texts	Schwanenflugel et al. (2009)	✓		
Choral reading	Reutzel et al. (2008)	✓		✓
Interactive read alouds	Baker et al. (2013)		✓	
Vocabulary and Comprehension Routines	Evidence of Effectiveness	SR	ELA	DI
Rich vocabulary instruction	Beck et al. (1982)	✓	✓	✓
Setting a content-oriented purpose for reading	Guthrie et al. (2004)	✓		
Comprehension strategy modeling	Reutzel et al. (2005)	✓	✓	
Inferential discussion	McKeown et al. (2009)	✓	✓	✓
Text structure graphic organizer	Hebert et al. (2016)	✓	✓	
Text-based written response	Graham and Hebert (2010)	✓	✓	
Composition Routines	Evidence of Effectiveness	SR	ELA	DI
Sentence composing grammar instruction	Saddler and Graham (2005)		✓	
Teacher modeling of writing strategies	Troia and Graham (2002)		✓	
Planning writing with graphic organizers	Harris et al. (2006)		✓	
Use of peer revision and editing checklists	MacArthur et al. (1991)		✓	

Note. DI = Differentiated Instruction; ELA = English Language Arts; SR = Shared Reading.

Fluency

Shared Reading targets fluency growth through repeated oral reading to develop automaticity in text reading and comprehension (Schwanenflugel et al., 2009). Each trade book in the curriculum is segmented for daily lessons. The lesson begins with choral reading of a new text segment and then students immediately reread that text in pairs (Reutzel et al., 2008). Pairs are assigned by reading proficiency data with the pairing strategy from Peer-Assisted Learning Strategies (Fuchs et al., 1997).

Vocabulary

There are several times each day when lessons develop vocabulary directly through explicit instruction in word meanings and word-solving strategies (e.g., morphology), as well as opportunities for incidental word learning through multiple and repeated exposures to words during reading and discussion about text (Beck et al., 1982; Wright & Cervetti, 2017). At all grades, interactive read aloud lessons for narrative texts are followed by direct instruction in general academic words with a consistent instructional routine (Beck et al., 2013). For grades 3–5, an additional six vocabulary words are taught each week before Shared Reading of narratives. In addition to the instructional procedure designed by Beck et al. (2013), the lessons add identification of syllable types and transform the words into different parts of speech, providing explicit instruction about grammar and morphology (Bowers et al., 2010). For nonfiction texts, content-specific words are taught before students engage with interactive read alouds to build background knowledge. The teacher emphasizes connections among these terms with labeled diagrams, semantic feature analysis charts (Anders & Bos, 1986), or concept of definition maps (Schwartz & Raphael, 1985).

Comprehension

There are two routines for supporting text comprehension. Both emphasize understanding the content of the text itself rather than practicing skills or strategies (McKeown et al., 2009). For Shared Reading, teachers set a content-oriented purpose prior to each reading (Guthrie et al., 2004). Since lessons target reading fluency and stamina, student reading itself is uninterrupted except for one pause during which the teacher thinks aloud to model one of four comprehension strategies: summarizing, developing a mental image, drawing an inference, or clarifying a misconception (Reutzel et al., 2005; Shanahan et al., 2010). The bulk of the Shared Reading comprehension support comes after reading. The teacher conducts a meaning-based discussion by asking a set of open-ended, initiating and follow-up questions sequenced to highlight the important content in that day's text selection and to generate the gist of the day's segment (McKeown et al., 2009). For interactive read aloud lessons, the model is different, and more closely influenced by the design used in Questioning the Author (Beck et al., 1996). All teacher support for comprehension comes during the read aloud, with extensive oral interaction through targeted open-ended questioning, content and vocabulary explanations, and think alouds (Baker et al., 2013). In both Shared Reading and interactive read aloud lessons, the teacher brings the segment to closure by discussing the author's text structure choices and updating a shared text structure graphic organizer (Hebert et al., 2016), and students respond to text in writing (Graham & Hebert, 2010).

Writing

In addition to instruction to develop transcription skills (i.e., handwriting and spelling) during word study instruction, writing instruction focuses on composing increasingly complex sentences; brief text-based written responses; and longer narrative, informative, and persuasive texts. All grades have oral sentence composing grammar instruction with exemplars selected from the day's read aloud (Saddler & Graham, 2005). Lessons include instruction in sentence combining, unscrambling,

imitating, or expanding, all supported by direct explanation (Graham et al., 2012). In grades 3–5, students use a semantic web to plan compound and complex sentences using vocabulary words. Teachers also deliver genre-based writing strategy instruction for composing narrative, informative, and persuasive texts about the content in students' reading. Lesson plans follow a cycle of teacher-directed instruction, student work time, and group sharing. The teacher models strategies through think alouds for planning, drafting, and revising each of the three genres (Troia & Graham, 2002). Students learn to evaluate good and poor examples of each genre and to use genre-specific checklists to revise and edit their own writing both with peers and independently (MacArthur et al., 1991). Responsibility for writing strategy use during student work time is gradually released to students as they use graphic organizers in second grade (Harris et al., 2006) and construct their own organizers beginning in third grade. Students share their daily writing products with a partner or small group (Graham et al., 2012).

Prior research on Bookworms

Studies of the effects of Bookworms demonstrated feasibility and promise with school-level samples, but the present study improves upon their methodological limitations. In a quasi-experimental study, students in three Bookworms schools made significantly greater growth in fluency in grades three and five ($d = .55$) and comprehension in grades three, four, and five ($d = .42$) than students in four comparison schools (Walpole et al., 2017). Whereas this study used limited measures of achievement collected only during the first year of implementation, the present study includes a more robust measure of student achievement collected over four school years.

Next, in an external evaluation of a federal initiative to improve achievement through curriculum and professional learning, Bookworms had the largest influence on achievement of any curriculum, averaging 17% greater growth per year ("Georgia Striving Readers Comprehensive Literacy Grant: Longitudinal Evaluation 2012–2017," n.d.). However, this evaluation was correlational, examining growth in student achievement associated with teachers' self-reported curriculum and program choices, and does not provide evidence of causal impact.

Finally, a case study documented the process of adoption and implementation of Bookworms and trends in student achievement for all students in a rural district using state outcome data. The district initially underperformed the statewide average on the state's standardized outcome test during the first year of Bookworms implementation but then outperformed the state average at third, fourth, and fifth grade in the third year of Bookworms implementation (Center for Research in Education and Social Policy, 2019). Disaggregated data compared growth in the Bookworms district with growth for all subgroups of students in schools not using Bookworms. Students who are multilingual, students with disabilities, and students identifying as African American or Hispanic/Latino demonstrated stronger growth in the Bookworms district. However, this evaluation was descriptive and did not include any pre-implementation data, providing no evidence of causal impact. The present study improves on prior research with a large, longitudinal sample and a more rigorous quasi-experimental design.

Purpose and research questions

Bookworms is an ambitious attempt to curricularize a selection of findings from empirical research in reading and writing that also attends to grade-level learning standards to increase feasibility in schools. Students read a high volume of authentic and challenging books with scaffolding from evidence-based instructional routines, engaging daily with grade-level reading and writing content regardless of their previous achievement, and teachers provide skills-based differentiation (including below-grade-level skills if warranted) for 15 minutes each day. While we have small-sample evidence that the curriculum is feasible and promising, we were interested to

test its effects in a larger sample and over more than one school year. We had two research questions:

(1) Do students who receive Bookworms experience improved reading achievement trajectories compared to a business-as-usual curriculum?
(2) Are the effects of Bookworms on reading achievement different for groups of students based on their status as English language learners, free/reduced lunch status, special education status, or initial reading achievement?

Method

Design

This study supports nascent causal inference on the impact of Bookworms through a comparative interrupted time-series design, utilizing multilevel growth curve modeling of student achievement scores with up to 10 data points per student. By pooling data across seven cohorts and four school years (see Figure 2), we are able to include students who never experienced Bookworms (Cohort A; $n = 1{,}216$), students who experienced the district's business-as-usual (BAU) curriculum for at least one year prior to experiencing Bookworms and for whom we have both pre-treatment and post-treatment outcomes (Cohorts B, C, and D; $n = 4{,}125$) or only post-treatment outcomes (Cohorts E and F; $n = 2{,}393$), and students who experienced only Bookworms (Cohort G; $n = 1{,}072$).[1]

Participants

This study included all students in grades 2–5 from 17 elementary schools in one rural school district in a mid-Atlantic state. Across the seven cohorts, participants included 8,806 students in total, with 56% qualified for federal lunch subsidies, 17% with disabilities, 74% identifying as White, 10% as Black, 8% as Hispanic, and 8% as Other. The school district employs approximately 300 classroom teachers in grades K-5. Publicly available achievement data from the Partnership for Assessment of Readiness for College and Careers (PARCC), given beginning in third grade, indicating the percentage of students who met or exceeded state standards for literacy in the district and statewide from 2016–2019 are presented in Table 4. In 2016 and 2017, the year prior to Bookworms implementation and the first year of Bookworms implementation, respectively, this percentage was lower than the state average. In 2018 and 2019, the second and third years of Bookworms implementation, respectively, the percentage of fourth and fifth graders in the district who met or exceeded state standards for literacy either exceeded or was nearly equivalent to the statewide average in these grades.

	School Year / Quarter									
	2015-16			2016-17			2017-18		2018-19	
Grade	Fall	Winter	Spring	Fall	Winter	Spring	Fall	Winter	Fall	Winter
2	D	D	D	E	E	E	F	F	G	G
3	C	C	C	D	D	D	E	E	F	F
4	B	B	B	C	C	C	D	D	E	E
5	A	A	A	B	B	B	C	C	D	D

Figure 2. Seven cohorts of students by grade and treatment status **Note**. Fall 2015 through fall 2016 (shaded red) are Pre-Treatment. Winter 2017 through winter 2019 (shaded green) are Treatment. Sample sizes of unique students per cohort: $n_A = 1{,}216$, $n_B = 1{,}270$, $n_C = 1{,}395$, $n_D = 1{,}460$, $n_E = 1{,}313$, $n_F = 1{,}080$, $n_G = 1{,}072$

Table 4. Percentage of students who met or exceeded state literacy standards by year and grade

Year and Grade	District	State
2016		
Grade 3	30.5	37.5
Grade 4	30.4	40.4
Grade 5	31.3	39.4
2017		
Grade 3	35.1	39.8
Grade 4	38	41.9
Grade 5	40.1	41.4
2018		
Grade 3	36.1	38.8
Grade 4	43.6	43.1
Grade 5	42.3	42.1
2019		
Grade 3	38.7	41.2
Grade 4	43.5	43.6
Grade 5	44	43.9

Note. Based on publicly available data from the Partnership for Assessment of Readiness for College and Careers

Comparison

In 2015–16, the district implemented their BAU curriculum, serving as a baseline year in this study. The BAU curriculum was written by a district team and provided to all teachers by district ELA leads through a content management system. It was aligned to the CCSS requirements for text variety and complexity and emphasized writing strategies and text-based writing. Word study included weekly spelling lists and a vocabulary curriculum teaching Greek and Latin roots. BAU included whole-group instruction with text excerpts chosen from a commercial core program and augmented with novels, small-group guided reading instruction in which students were taught with texts matched to their instructional reading level, and a writer's workshop with strategy instruction. Teachers received initial training on the BAU curriculum and ongoing coaching from district staff tasked specifically to support them.

Treatment

The district adopted Bookworms as their only curriculum in all grade levels with phased adoption of the composition lessons to acknowledge that district leaders believed their BAU writing instruction was sufficient. Bookworms is different from BAU in that Shared Reading texts were whole texts, instructional routines targeted fluency and involved explicit instruction, vocabulary instruction was connected to texts, students were provided differentiated skills-based instruction instead of guided reading, and read alouds and grammar instruction reduced the time for writing workshop. In 2016–17, all schools implemented Shared Reading, Differentiated Instruction, and the read aloud and sentence composing portions of ELA. They continued to implement their BAU composition curriculum. In 2017–18, they revised their writing strategy instruction to be consistent with Bookworms ELA design. In fall 2018–19, they began to implement all three Bookworms instructional blocks.

School leaders and teachers received curriculum-specific training and coaching in all three years from instructional specialists associated with the curriculum. Initial training in Year 1 was day-long and grade-specific. In each year, the district contracted with a professional learning organization for curriculum coaches; these external coaches were assigned to provide support at individual or multiple schools. Principals in those schools scheduled coaching time during the day to include only a specific set of activities: classroom walkthroughs with confidential teacher debriefings, modeling lessons, facilitated individual or grade-level team planning sessions, support for scoring assessments and

grouping students for Differentiated Instruction, and before-school question and answer sessions. In year one, the district received 208 days of support distributed across 17 schools. In year two, they received 135 days. In year three, they received 133 days. Across the three years, each school received an average of 28 days of curriculum-specific implementation support and coaching, fewer than 10 days per year.

Each year the district received a summative report on implementation support and coaching from the professional learning provider. Consistent with district policies, the reports do not include teacher-level data but instead describe the services provided and identify implementation strengths and recommendations for each school. In the first year, summative reports for all 17 schools included direct statements about administrative expectations for implementation fidelity and quality. For example, Wilson (2017) stated, "Teachers and administration are receptive and open to coaching and want to implement *Bookworms* with fidelity" (n.p.). In year two, reports included evidence that all teachers were implementing lessons with fidelity and working on implementation quality improvement goals. Walpole (2018) stated, "Most teachers are well versed in the basics of how each lesson is taught in *Bookworms* and have begun shifting focus to digging deeper into more effective instruction" (p. 26). In the third year, reports included evidence of continued implementation fidelity to the Shared Reading procedures and growing independence with writing instruction. Recommendations were specific to schools (e.g., improve lesson timing, establish expectations for anchor charts), but none provided evidence that the curriculum was not being implemented fully.

Measures

Student achievement was measured using the Measures of Academic Progress (MAP) or MAP Growth assessment tool (Northwest Evaluation Association, 2019). MAP is a norm-referenced, computer-adaptive assessment used throughout the district to measure student performance and growth in reading and mathematics, with two to four administrations per year (see Figure 2). Our analyses use the Rasch Unit (RIT) scores from the MAP Growth Reading (2–5) assessment, which are vertically scaled to show growth within and across grades. Reading interventions typically have stronger effects on proximal outcomes that are more sensitive to the treatment. However, MAP Growth is a distal outcome, designed externally with no match to the intervention characteristics of Bookworms.

MAP Growth Reading consists of reading a passage and responding to multiple choice items. Items measure five goals: (1) word analysis and vocabulary development (determining multiple meanings, synonyms and antonyms, word components, and word recognition and vocabulary), (2) literary response and analysis (identifying genre characteristics, literary devices, and literary elements), (3) literal reading comprehension (determining cause and effect, locating information and reading directions, reading for main idea and details, and sequencing events), (4) interpretive reading comprehension (comparing and contrasting, drawing conclusions, inferring and predicting, interpreting author's purpose, and summarizing), and (5) evaluative reading comprehension (evaluating author's technique and viewpoint, fact and opinion, and persuasive elements). Marginal reliability coefficients range from .93 to .95 (Northwest Evaluation Association, 2011).

Data analysis

We utilize MAP RIT scores for seven cohorts of students in grades 2–5 from fall of the 2015–16 baseline school year through winter of the 2018–19 school year. Because Bookworms was first implemented in the 2016–17 school year, we have up to four baseline data points (fall 2015, winter 2016, spring 2016, and fall 2016), plus up to six data points during implementation (winter 2017, spring 2017, fall 2017, winter 2018, fall 2018, and winter 2019) for each student (see Figure 2). The district did not administer MAP in spring 2018 or spring 2019 due to the demands of the PARCC assessment as a state outcome test. By modeling each student's growth curve via a Hierarchical Linear Model (HLM), including a time by treatment interaction term, we are able to estimate the change in

students' achievement trajectories corresponding to the implementation of Bookworms at the beginning of the 2016–17 school year.[2] Thus, each student's trajectory under the treatment condition is compared to a counterfactual reflected by both the expected trajectories based on prior performance (Bloom, 2003) and the observed trajectories of students in higher grades during the 2015–16 baseline school year.

The mathematical form of the HLM growth model (in mixed model format) used to estimate impacts is as follows:

Student Level Growth: $Y_{tijk} = \beta_{0i} + \beta_{1i}\left(TIME_{tijk}\right) + \beta_{2i}\left(TIME^2_{tijk}\right) +$

Bookworms Impacts: $\beta_3\left(BW_{tijk}\right) + \beta_4\left(TIME_{tijk} \times BW_{tijk}\right) + \beta_{2i}\left(TIME^2_{tijk} \times BW_{tijk}\right) +$

Teacher & School-Level Random Coeffients: $\varphi_{jk} + \gamma_k + \eta_k\left(TIME_{tijk}\right) + \lambda_k\left(TIME^2_{tijk}\right) +$

Student-Level Random Coeffients and Residual Error: $\varphi_{ijk} + \omega_{ijk}\left(TIME_{tijk}\right) + \psi_{ijk}\left(TIME^2_{tijk}\right) + \varepsilon_{tijk}$

The TIME indicators were coded based on each student's initial grade level at the time of their first MAP data point. For example, a student whose first MAP test was administered on the first day of school as a second grader would have an initial TIME value of 2.0, a third grader would be 3.0, and so on. Subsequent values of the TIME indicator were coded by adding the time elapsed between test administration dates based on a 10-month school year (e.g., one month equals + 0.10 school years). The TIME variable was then centered around the value of 3.0 (i.e., the beginning of third grade) to position growth trajectory intercepts near the middle of the range of observed data. Furthermore, the model includes a quadratic TIME parameter to accommodate the curvilinear growth of MAP scores over time (Northwest Evaluation Association, 2019).

The model also includes a time-varying Bookworms indicator (BW), coded 1 for time points during Bookworms implementation and 0 for pre-implementation time points, with parameters representing a main effect of Bookworms (i.e., impacts at baseline) as well as interactions with TIME and $TIME^2$ to capture impacts on test score trajectories. Random effects for the growth trajectory parameters are included at both the student and the school levels, with correlations between trajectory parameters (i.e., intercept, slope, and curvature) within both levels. Random effects at the school level were cross classified in each model to account for approximately 10% of students who changed schools within the district. Random effects at the teacher level were also cross-classified, given that students change teachers each year; however, this also precludes the use of teacher-by-time interactions given the limited number of test scores for one student under a specific teacher. As such, the teacher effects can be thought of as marginal value-added effects for each teacher during the course of the study, net of any overall impacts of Bookworms and schoolwide effects on students' trajectories. All models were estimated using PROC HPMIXED in SAS 9.4 (STAT 15.1), which uses sparse matrix techniques to estimate complex mixed models containing many random effects.

To explore differentiation in impacts across student subgroups that might respond differently to instruction, the base model was elaborated to include additional interactions with student-level moderators (i.e., English language learner [ELL] status, free/reduced lunch [FRL] status, special education [Sp.Ed.] status, and baseline achievement).[3] The baseline achievement moderator variable was calculated as a z-score for each student based on the first observed MAP score during the grade and year in which they first experienced Bookworms; hence, the model of moderation by baseline achievement does not include students from Cohort A (since they never experienced Bookworms), and the design is reduced to a simple interrupted time series without a comparative cohort (i.e., the BAU condition is reflected only in the pre-treatment data for cohorts B, C, and D from Figure 2). It is important to note that our moderation analyses are preliminary in nature and do not guarantee that the differences in effects are attributable to differential effects of Bookworms because there may be other confounding factors experienced by students from different subgroups around the same time that Bookworms was implemented.

Finally, the base model was elaborated to include additional random effects at the school level to test for school-specific variation in overall impacts of Bookworms and to assess whether the impacts are significantly larger or smaller in some schools.[4] Thus, the estimated effects of Bookworms in each model reflects a marginal impact on students' RIT score trajectories, averaged across all of their teachers and schools.

Results

Results confirm a significant positive impact of Bookworms on reading achievement overall, with gains compounding over time and producing a standardized effect size of .26 by the end of fifth grade. Results also confirm a significant moderation effect for students with a special education designation, which suggests that the impacts of Bookworms for students qualifying for special education services are larger each year than for other students.

Table 5 shows simple descriptive statistics (i.e., unadjusted for cohort differences in trajectories) for the MAP scores by cohort and grade level. While the growth in scores over time for each cohort is clearly evident, the differences in baseline scores and growth trajectories take considerable effort to discern because one cannot simply compare scores across cohorts in one row due to differences at prior test points. Instead, one must recognize the differences in gains across time (i.e., down the rows) and how those gains differ across the cohorts (i.e., across the columns) before and after

Table 5. Descriptive statistics for MAP scores by cohort and grade level

			Cohort						
Grade	Period	Statistic	A	B	C	D	E	F	G
2	Fall	Mean				172.4	169.5	169.0	167.0
		SD				17.6	16.8	17.4	17.6
		N				1,188	1,139	971	1,060
	Winter	Mean				181.6	180.2	179.8	178.9
		SD				17.6	17.3	18.4	17.9
		N				1,177	1,146	985	1,040
	Spring	Mean				186.6	185.9		
		SD				17.3	17.1		
		N				1,177	1,147		
3	Fall	Mean			187.8	185.6	183.9	184.0	
		SD			17.5	17.4	16.9	17.9	
		N			1,210	1,184	1,049	1,021	
	Winter	Mean			193.4	192.9	193.0	192.7	
		SD			16.6	17.1	16.6	17.2	
		N			1,202	1,186	1,061	1,042	
	Spring	Mean			196.0	196.5			
		SD			17.0	16.9			
		N			1,199	1,183			
4	Fall	Mean		196.8	195.4	195.7	195.0		
		SD		16.3	17.1	17.1	16.8		
		N		1,143	1,220	1,078	1,108		
	Winter	Mean		202.0	201.2	202.8	201.5		
		SD		15.8	16.1	15.7	15.9		
		N		1,123	1,223	1,097	1,118		
	Spring	Mean		203.0	204.4				
		SD		16.5	16.6				
		N		1,125	1,222				
5	Fall	Mean	205.1	202.7	202.9	203.8			
		SD	16.0	16.5	16.6	16.4			
		N	1,174	1,154	1,097	1,187			
	Winter	Mean	209.4	208.5	207.9	208.6			
		SD	15.5	15.2	15.9	15.3			
		N	1,170	1,153	1,115	1,180			
	Spring	Mean	210.0	210.3					
		SD	16.5	15.8					
		N	1,170	1,136					

Note. Values above and to the left of the dotted line are pre-treatment data points, while values below and to the right of the dotted line are treatment data points.

implementation of Bookworms. For example, Cohort C started fourth grade about 1 point behind Cohort B's scores at the beginning of fourth grade. By the beginning of fifth grade, after Cohort C had experienced Bookworms for one year, Cohort C's scores caught up to, and even passed Cohort B's average score by + 0.2 points, which translates to a difference of + 1.2 points in annual gains for the treatment cohort. Further, when highlighting the apparent differences in growth rates in Table 5, it is important to acknowledge that missing test scores for a single data point are handled using listwise deletion for these descriptive statistics, which may bias the point estimate for any one cohort and test administration. As such, we implemented a multilevel growth model (with full-information maximum likelihood to address missing test scores) to minimize bias when estimating differences in growth rates.

Table 6 shows parameter estimates for the unconditional growth model (i.e., Model 1), the model of overall impacts (i.e., Model 2), models for moderation effects by ELL status, FRL status, Sp.Ed. status, and Baseline Achievement (i.e., Models 3–6, respectively), and a model of overall impacts that includes additional random effects to test for significant variation in the impacts of Bookworms across schools (i.e., Model 7). Model 1, the unconditional growth model, confirms that MAP Reading scores for students in the sample increased over time (and grade level) with a linear coefficient of 15.27 points per year on the RIT scale, and a small negative quadratic coefficient of −1.96 points per year-squared, corresponding to a deceleration of learning trajectories that mirrors that seen in the MAP norming data (Northwest Evaluation Association, 2019). Additionally, Model 1 confirms significant variance of random coefficients for trajectory intercepts, slopes, and curvature at both the student level and school

Table 6. Parameter estimates for growth curve analyses via hierarchical linear modeling (HLM)

	Model Number (Moderator)						
Fixed Effects Model Parameters	1 (none)	2 (none)	3 (ELL)	4 (FRL)	5 (Sp.Ed.)	6 (Baseline Achievement)[A]	7 (none)
Intercept	184.02***	184.45***	184.71***	185.78***	186.07***	185.57***	184.63***
Time	15.27***	15.04***	15.02***	14.95***	15.35***	15.11***	14.84***
Time^2	−1.96***	−2.26***	−2.26***	−2.08***	−2.31***	−2.97***	−2.19***
Bookworms		0.06	−0.05	0.03	0.09	−1.06***	−0.39
Time × Bookworms		0.52**	0.50**	0.33	0.26	1.11***	0.79~
Time^2 × Bookworms		0.26***	0.26***	0.20*	0.27***	0.59***	0.17
Moderator			−10.73***	−2.44***	−9.79***	14.79***	
Time × Moderator			0.02	−0.20	−1.83***	−0.32**	
Time^2 × Moderator			−0.49	−0.40***	0.36*	0.43***	
Bookworms × Moderator			0.25	−0.31	−1.20***	−0.03	
Time × Bookworms × Moderator			0.96	0.43	1.56***	−1.53***	
Time^2 × Bookworms × Moderator			−0.02	0.14	−0.14	−0.01	
Random Effects Model Parameters							
School-Level							
Intercept	1.06***	1.05***	1.05***	0.91***	1.31***	1.82***	1.41***
Time	1.36***	1.33***	1.37***	1.43***	1.22***	1.27***	2.27***
Time^2	0.10***	0.12***	0.12***	0.12***	0.11***	0.13***	0.13***
Bookworms							1.50***
Time × Bookworms							2.40***
Time^2 × Bookworms							0.29***
Teacher-Level							
Cross-Classified Teacher Effect	8.79***	9.48***	9.53***	9.37***	8.97***	8.56***	9.19***
Student-Level							
Intercept	249.59***	249.93***	243.81***	239.40***	201.40***	32.16***	249.90***
Time	18.99***	19.08***	19.13***	19.37***	19.61***	8.97***	18.92***
Time^2	1.77***	1.76***	1.79***	1.82***	1.69***	2.05***	1.76***
Residual	36.79***	36.50***	36.55***	36.63***	37.15***	36.25***	36.33***

[A]The model of moderation by baseline achievement does not include students from Cohort A, since they never experienced BW. ***p<.001,**p<.01,*p<.05,~p<.10; Estimates of correlations between random effects (representing up to 18 additional parameters) are not shown in this table due to space constraints.

level, with the majority of the variance at the student level (i.e., between 95% and 99% across the three parameters). There was also significant variation in the random teacher effects, confirming that variation in students' RIT scores around their predicted trajectories is associated with teacher-specific effects. This suggests that growth trajectories are remarkably varied for different students, but students from the same school are slightly more likely to have similar growth trajectories, and students taught by the same teacher are more likely to have similarly high or low scores relative to their predicted trajectories.

Model 2, which serves to estimate overall impacts of Bookworms, produced similar coefficients for trajectory parameters and random effects, and also revealed significant impacts of Bookworms on growth trajectory parameters. More specifically, the difference in RIT scores at baseline (i.e., the beginning of third grade) for BAU versus Bookworms was not significant ($p > .10$, $\beta = .06$), the linear effect of Bookworms during its implementation was significant ($p < .01$) and positive (+.52 RIT points per year), and the quadratic effect of Bookworms was significant ($p < .001$) and positive (+.26 RIT points per year-squared). This suggests that at the end of one year, the impact of Bookworms on third-grade MAP scores was almost 1 RIT score point (i.e., $-.06 + [.52 \times 1 \text{ year}] + [.26 \times 1 \text{ year}^2] = +.72$) and that the effect of Bookworms quickly accumulates in subsequent years (e.g., $-.06 + [.52 \times 2 \text{ years}] + [.26 \times 2 \text{ years}^2] = +2.02$) and becomes quite large during the three-year span from third grade through the end of fifth grade (e.g., $-.06 + [.52 \times 3 \text{ years}] + [.26 \times 3 \text{ years}^2] = +3.84$). Figure 3 shows predicted MAP Reading scores for third through fifth grades under Bookworms versus the BAU curriculum. Given the RIT scale cross-sectional standard deviation of approximately 15 points at the end of grades three, four, and five (see Northwest Evaluation Association, 2019), the + 3.84 additional points by the end of fifth grade correspond to a standardized effect of + .26 standard deviations (i.e., 3.84/14.9 = +.258).

Models 3 through 6 test for moderating effects of Bookworms by ELL status, FRL status, Sp.Ed. status, and Baseline Achievement, respectively. No evidence of moderation was found for ELL status or FRL status.[5] The Sp.Ed. model showed marginally significant moderation effects of the impacts of Bookworms (i.e., the 3-way interactions involving TIME or TIME^2, BW, and the moderator), where

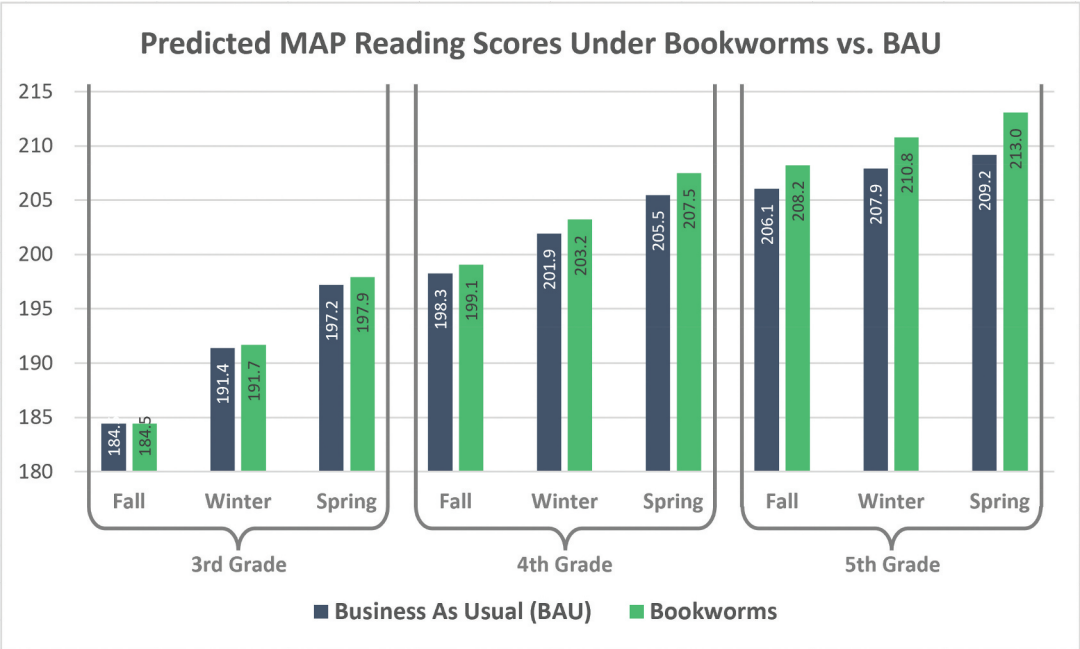

Figure 3. Predicted MAP reading scores under Bookworms vs. Business-as-usual curriculum

the initial impact of Bookworms (i.e., the linear effect) was significantly more positive (+1.56 RIT points per year). This suggests that students receiving special education services experienced greater gains in reading scores each year under Bookworms, and that the speed with which effects accumulated over time was similar to the accumulation of longer-term effects for students not receiving special education services. Based on the Model 5 fixed effects coefficients, Figure 4 shows predicted growth curve trajectories for students who received special education and students who did not.

Although the differences in MAP scores associated with Bookworms are initially small, they accumulate over time and result in trajectories of MAP reading scores that allow Bookworms students to advance further ahead of students under the baseline curriculum. When comparing to national norms for the RIT scale (Northwest Evaluation Association, 2019), the students in this study not receiving special education services scored very close to the national average under the baseline curriculum; under Bookworms, they scored nearly one-fifth of a standard deviation above the national average by the end of fifth grade. Likewise, the students receiving special education services scored .85 standard deviations below the national average at the end of fifth grade under the baseline curriculum; under Bookworms, that gap was reduced by almost half to only .47 standard deviations.

The model of moderation by Baseline Achievement showed a significant moderation effect of the impacts of Bookworms based on the 3-way interaction involving TIME, BW, and Baseline Achievement, where the initial impact of Bookworms (i.e., the linear effect) decreased by 1.53 RIT points per year for each standard deviation increase in Baseline Achievement. There was no evidence of moderation involving the quadratic effect. This suggests that students with lower Baseline Achievement experienced larger early gains than students with higher Baseline Achievement; however, all students experienced positive long-term gains while using Bookworms (i.e., based on the overall quadratic effect of Bookworms). Based on the Model 6 fixed effects coefficients, Figure 5 shows predicted growth curve trajectories for students with Baseline Achievement that is average, +1 standard deviation above average, and −1 standard deviation below average.

Predicted trajectories for students with low Baseline Achievement are similar at the beginning of third grade but diverge rapidly with large differences associated with Bookworms accumulating

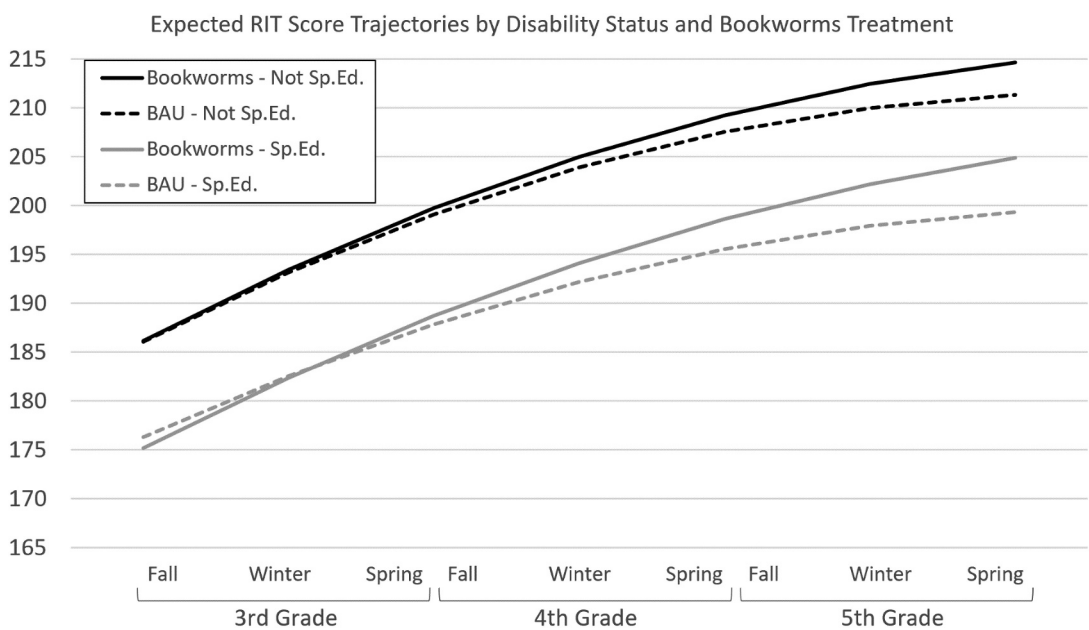

Figure 4. Expected growth curve trajectories of MAP reading scores by treatment and special education (Sp.Ed.) status

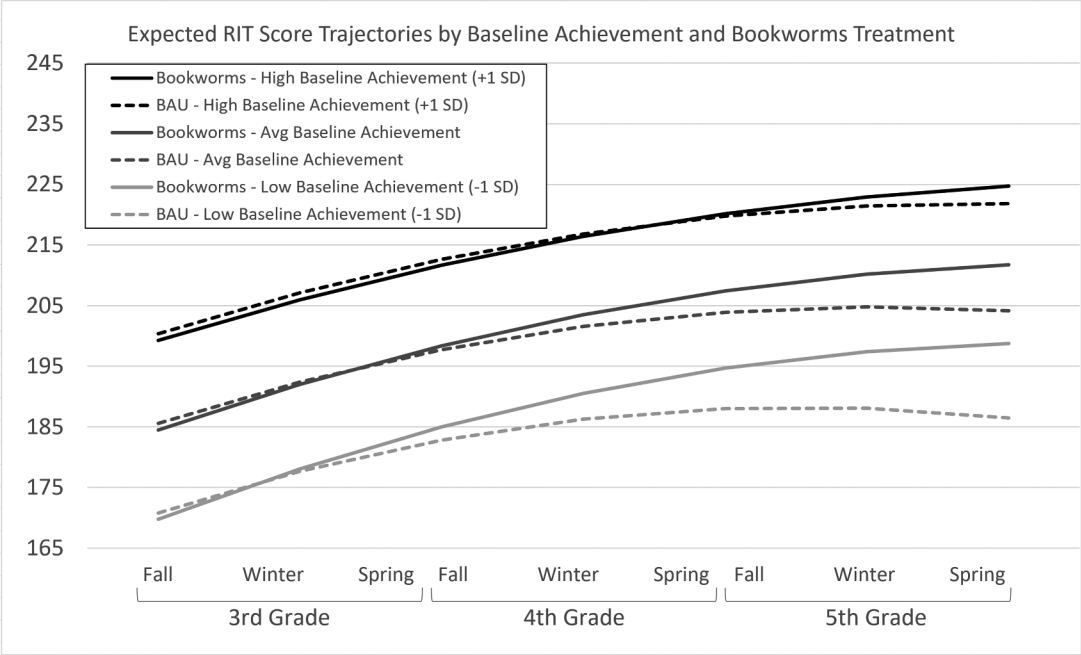

Figure 5. Expected growth curve trajectories of MAP reading scores by treatment and baseline achievement

through fifth grade. A similar, but less dramatic difference in predicted trajectories is shown for students with average Baseline Achievement. For students with high baseline achievement, the trajectories are very close from the beginning of third grade through the end of fifth grade. This suggests that students with average or low achievement appear to benefit most from Bookworms, while students with above-average achievement maintain growth trajectories that are similar to those expected under the BAU curriculum.

Because there was some confounding between special education status and baseline achievement ($r^2 = .18$), we also estimated a model including both sets of moderation effects for special education status and baseline achievement, including three-way and four-way interactions (for a total of 21 model coefficients). Results showed that the moderation effect of Baseline achievement for students not receiving special education services grew slightly larger at 2.03 RIT points per year ($t = -13.98$, $p < .001$), versus the 1.53 points per year reported above. Furthermore, the moderating effect of Sp.Ed. status persisted even after controlling for baseline achievement, now with significant linear ($p < .01$) and quadratic effects ($p < .05$), producing a total effect of + .33 standard deviations by the end of fifth grade.

Lastly, Model 7 in Table 6 presents results from an HLM that added school-level random effects for the impacts of Bookworms on all three trajectory parameters. The results suggest that the school-level variance of the impact of Bookworms is significant and substantial. Using the random effects estimates to produce plausible value intervals (see Raudenbush & Bryk, 2002, p. 78) reveals that the immediate effect of Bookworms at baseline is predicted to range from −2.8 to + 2.0 RIT points across 95% of schools, the linear effect of Bookworms is predicted to range from −2.2 to + 3.8 RIT points per year across 95% of schools, and the quadratic effect of Bookworms is predicted to range from −0.9 to + 1.2 RIT points per year-squared across 95% of schools. Therefore, while the overall effects of Bookworms on average across schools were positive and substantial, the size of these school-level random coefficients suggests that the impacts of Bookworms may be quite different from one school to another. Generalizing to a larger population of schools, it is likely that some schools implementing Bookworms will experience smaller effects than the average effects

reported here, while other schools will experience impacts of Bookworms that are potentially much larger.

Discussion

The present study tracked achievement growth for over 8,000 students in 17 schools with initially weak achievement across three years, before and during implementation of Bookworms, a whole-school program consisting of an ELA curriculum with grade-level standards alignment in reading and writing; high-quality, authentic texts; differentiated foundational skills instruction; and repetitive instructional routines derived from research for word study, fluency, vocabulary, comprehension, and writing. To test the effects of this approach, we estimated student-specific quadratic growth curves and compared them to growth under the baseline curriculum and to the national norms for MAP reading. We also explored differential effects based on student characteristics: ELL status, FRL status, special education status, and initial reading achievement.

The design of Bookworms is distinct from other popular reading curricula and schoolwide programs in several ways. It employs full-length literature selected for quality and match to grade-level text difficulty standards rather than providing students with texts primarily at their instructional or independent reading level (Adams et al., 2020). It minimizes time spent on, and increases precision for, developing word recognition to maximize time for academic language development, knowledge building, and writing development. The combination of grade-level standards alignment, high-volume reading of challenging text, skills-based differentiation, knowledge-building read-alouds, and text-based writing may work in concert to improve achievement (National Governors Association & Council of Chief State School Officers, 2010; Walpole et al., 2017). In addition, Bookworms was implemented with ongoing professional learning, including day-long, grade-level district training and then an average of 9.3 days of school-level coaching per year, a relatively low dose (and more cost effective) compared with other schoolwide programs (Borman et al., 2003; Neitzel et al., 2022).

Effects of Bookworms were not instantaneous and were not experienced similarly by all students. As with previous longitudinal studies, accumulating effects may be due to attributes of the curriculum design itself or to improvements in implementation over time (Borman et al., 2007). Other Comprehensive School Reforms found no school-level effects the first year of implementation (Borman et al., 2007), with the largest effects after three to five years (Bloom et al., 2001; May & Supovitz, 2006). In this study, effects were relatively small for third graders in year one, but they compounded to an effect size of .26, corresponding to an additional 4.9 months of school (i.e., a 16.3% increase in test score gains) across three years.[6] This effect size is comparable to the .27 to .31 found in comparable whole-school/whole-class programs (Neitzel et al., 2022) and is considerably larger than the .09 to .15 typically expected of Comprehensive School Reforms (Borman et al., 2003). Additionally, these effects were not experienced similarly by all students; students who began with weaker achievement benefited even more. The use of a distal measure of achievement rather than proximal measures strengthens the case that the curriculum was effective, even for students with initially weak achievement, as the effects of curricula are typically smaller on standardized measures (Cabell & Hwang, 2020).

Research reveals strong direction for the development of foundational skills for children in primary grades (Foorman et al., 2016). While many studies test targeted interventions at these grades, others test effects only for students with difficulties at higher grades (Donegan & Wanzek, 2021). This study measured achievement of students who received Bookworms beginning in the upper-elementary grades, including cohorts without the preventive support of Bookworms instruction in the primary grades, and including students with disabilities. While researchers have produced interventions that can support upper-elementary students' foundational skills and comprehension (Donegan & Wanzek, 2021), this study provides preliminary evidence that a comprehensive literacy curriculum aligned with evidence-based practices can produce gains in literacy achievement in the upper-elementary grades. We do not yet know the effects for children who receive Bookworms from kindergarten through fifth grade.

Researchers have recently turned their attention to investigating the impact of whole-school, content-rich literacy curricula (e.g., Cabell & Hwang, 2020) and interventions (e.g., Kim et al., 2021) that aim to improve elementary students' literacy achievement through academic language and knowledge building. While substantial research shows that interventions that integrate literacy with science and social studies instruction produce large effects on vocabulary and comprehension (Hwang et al., 2022), fewer studies have investigated the impact of building content knowledge during whole-class language arts instruction in elementary grades (Hwang et al., 2023). Research on the Core Knowledge curriculum, for example, has demonstrated positive effects on vocabulary and content knowledge in kindergarten (Cabell & Hwang, 2020) and moderate effects on general reading achievement in grades 3–6 (Grissmer et al., 2023). We view the present study as adding to the burgeoning evidence that a coherent, content-rich literacy curriculum can improve literacy achievement for elementary students in real school settings.

Schools need curricula that are useful to all students, however, meaning that they must be designed to accommodate a wide range of achievement. They must address the fact that students require different types and amounts of instruction and practice based on their diagnostic profiles (Connor et al., 2011). In this study, students receiving special education services and students with weak initial achievement experienced a stronger growth trajectory, decreasing the gap between their performance and grade-level expectations. This finding is important in that it supports the idea that skewing whole-class instructional time toward work with challenging grade-level text is a plausible intervention for students with disabilities and other students with weak achievement when they are also provided differentiated foundational skills instruction. By providing systematic phonics instruction only for those students who need it and when they need it, Bookworms teachers are able to prioritize fluent reading for meaning and vocabulary and knowledge acquisition for all students, achieving a balance that improved achievement for all students to some extent and to a greater extent for those who needed it most.

Limitations and future research

Although most of the variance in achievement was attributed to student-level differences, there were also school-level effects. We have generated nascent causal evidence to add to previous evidence of promise (Walpole et al., 2017) that a large-scale implementation of evidence-based instruction, including extensive work in grade-level text, is feasible and effective for most students, including students with disabilities and those with low initial achievement. However, we do not have teacher-level fidelity of implementation data for BAU or Bookworms to explore the differences between schools that might impact its potential effectiveness. The school district had a well-defined teacher support model that prohibited our collection of fidelity data from observations conducted by coaches. School-level summative reports from the professional learning each year provided descriptive evidence that Bookworms had been fully adopted and teachers and administrators expected it to be used as designed, but we cannot verify these claims with the underlying data. In addition, while we know the district described BAU practices as very different from Bookworms, school-level implementation descriptions of BAU are not available. When interpreting the results of this study, it may be that factors other than the Bookworms curriculum, such as generally more coherent or stronger instruction or more time spent reading and writing, were the drivers of student achievement gains.

Furthermore, as a quasi-experimental study based on an interrupted time series design, these analyses do not produce support for causal inference that is as strong as that from randomized experiments or regression discontinuity designs. The fact that the comparison cohort is not concurrent and includes a relatively small fraction of the total sample further limits the strength of inference. The number of data points available differs across cohorts, and thus extrapolation of growth trajectories (i.e., either forwards or backwards) occurs within each cohort, which would invoke "reservations" under What Works Clearinghouse (2020) standards. The Bookworms composition lessons were not fully implemented until year three, but the district plan for writing instruction was the same for all students in each year of the

study. Nevertheless, the design of this study, given its use of multiple pre-intervention and post-interventions datapoints as well as within-subjects comparisons, provides much stronger causal inference than the typical pre-post design using a non-equivalent comparison group – a research design that is quite prevalent in evaluations of education interventions (Shadish et al., 2002). We look forward to expanding the evidence base in future research using randomized experiments to test the impacts of Bookworms as it expands implementation across schools nationwide.

Future research on Bookworms should also consider additional measures to better understand its effects. First, as we continue to study its effects on student achievement, we will go beyond global measures of reading and writing achievement. Individual measures of word recognition, vocabulary, fluency, comprehension, and writing quality will add nuance to our understanding of the effects of the curriculum on literacy achievement over time. These variables will help us to better understand which instructional practices in Bookworms are most effective for different students. In addition, implementation fidelity measures and other teacher descriptors will help us to understand differences in implementation across classrooms. Finally, like Joyce and Cartwright (2020), we are interested in understanding essential structures, supports, and derailers that would help stakeholders make predictions about the likelihood of positive effects of Bookworms (or other interventions) for individual schools rather than making claims of generalized effectiveness. Mixed-methods studies including measures of school leadership, culture, teacher knowledge and expertise, and implementation fidelity are essential to unpacking these differences in the future. Because Bookworms is an OER, continued study of its effects has the potential to inform the development and implementation of curricula that employ only evidence-based instructional practices and that are fully accessible to all teachers and schools.

Conclusion

Improving student literacy achievement at scale is a complicated endeavor. Neither rigorous standards alone, teacher or principal commitment to implementation, professional development days, observation and feedback, nor collaboration are sufficient (Kane et al., 2016; Song et al., 2022). Curriculum aligned to standards and professional learning aligned to curriculum are important levers to explore in tandem. In the current study, it is possible that the combination of highly-specific curriculum supported by ongoing, confidential school-level implementation support produced the flexible specificity (Stornaiuolo et al., 2023) necessary for changes in instruction that improved achievement. We look forward to continued efforts to design and test whole-school curricula and see the current study as evidence that it can be done.

Notes

1. It is important to note that only Cohort A justifies the "comparative" component of our design, in that this is the only group of students who did not experience the treatment, and that this cohort did not experience the counterfactual at the same time as the treatment group experienced Bookworms.
2. Given that MAP testing began on the third day of school in August 2016, we treat the first two weeks of MAP scores in the 2016–17 school year as additional baseline scores, given that the implementation of Bookworms had just begun. We chose two weeks as the cutoff given that daily MAP testing rates dropped abruptly from approximately 10% of students per day between August 31 and September 13, 2016, to approximately 1.5% of students per day between September 14 and September 30, 2016.
3. We also tested moderation by gender and race/ethnicity as exploratory analyses, but we found no significant differences. Detailed results are available from the authors by request.
4. Although our models include random effects for teachers, the number of test scores for each student under a specific teacher is not sufficient to estimate variation in impacts of Bookworms across teachers as we do with schools.
5. Although each moderator had a significant main effect or two-way interaction with TIME, these suggest only differences in model intercepts and slopes, which are not associated with differential impacts of Bookworms.

Instead, these effects simply show that different subgroups have significant gaps in MAP Reading scores at the beginning of third grade or slower rates of growth in general. It is the three-way interactions that reflect moderation of Bookworms impacts.

6. Based on Northwest Evaluation Association (2019) national norms for MAP Reading scores, the expected growth for an average student in the US is + 23.5 points from the beginning of 3rd grade through the end of 5th grade. Dividing this number by 30 months of schooling (i.e., based on three 10-month school years from 3rd to 5th grade) yields an average gain of + 0.78 points per month. Dividing this into the raw impact of BW yields an average gain of 4.9 months over three years (i.e., +3.84/.78 = 4.9).

Disclosure statement

Sharon Walpole is author of the Bookworms Curriculum. Dr. Walpole contributed to the descriptions of the intervention and study context in this article, but did not influence the study design or results. All activities related to study design, data management, and statistical analyses were led by Dr. May in order to ensure independence of this evaluation and no conflicts of interest.

ORCID

Henry May http://orcid.org/0000-0002-5965-3933
John Z. Strong http://orcid.org/0000-0002-6007-9115
Sharon Walpole http://orcid.org/0000-0003-3617-7230

References

Adams, M. J., Fillmore, L. W., Goldenberg, D., Oakhill, J., Paige, D. D., Rasinski, T., & Shanahan, T. (2020). *Comparing Reading Research to Program Design: An Examination of Teachers College Units of Study*. https://achievethecore.org/page/3240/comparing-reading-research-to-program-design-an-examination-of-teachers-college-units-of-study

Anders, P. L., & Bos, C. S. (1986). Semantic feature analysis: An interactive strategy for vocabulary development and text comprehension. *Journal of Reading*, *29*, 610–616. https://www.jstor.org/stable/40029687

Baker, S. K., Santoro, L. E., Chard, D. J., Fien, H., Park, Y., & Otterstedt, J. (2013). An evaluation of an explicit read aloud intervention taught in whole-classroom formats in first grade. *The Elementary School Journal*, *113*(3), 331–358. https://doi.org/10.1086/668503

Ball, E. W., & Blachman, B. A. (1991). Does phoneme awareness training in kindergarten make a difference in early word recognition and developmental spelling? *Reading Research Quarterly*, *26*(1), 49–66. https://doi.org/10.1598/RRQ.26.1.3

Bear, D. R., Invernizzi, M., Templeton, S., & Johnston, F. (2020). *Words their way: Word study for phonics, vocabulary, and spelling instruction* (7th ed.). Pearson.

Beck, I. L., McKeown, M. G., & Kucan, L. (2013). *Bringing words to life: Robust vocabulary instruction* (2nd ed.). Guilford Press.

Beck, I. L., McKeown, M. G., Sandora, C., Kucan, L., & Worthy, J. (1996). Questioning the author: A yearlong classroom implementation to engage students with text. *The Elementary School Journal*, *96*(4), 385–414. https://doi.org/10.1086/461835

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

Bhattacharya, A. (2006). Syllable-based reading strategy for mastery of scientific information. *Remedial and Special Education*, *27*(2), 116–123. https://doi.org/10.1177/07419325060270020201

Bloom, H. S. (2003). Using "short" interrupted time-series analysis to measure the impacts of whole-school reforms: With applications to a study of accelerated schools. *Evaluation Review*, *27*(1), 3–49. https://doi.org/10.1177/0193841X02239017

Bloom, H. S., Ham, S., Melton, L., & O'Brien, J. (2001). *Evaluating the accelerated schools approach: A look at early implementation and impacts in eight elementary schools*. Manpower Demonstration Research Corporation.

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research*, *73*(2), 125–230. https://doi.org/10.3102/00346543073002125

Borman, G. D., Slavin, R. E., Cheung, A. C. K., 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

Bowers, P. N., Kirby, J. R., & Deacon, S. H. (2010). The effects of morphological instruction on literacy skills: A systematic review of the literature. *Review of Educational Research*, *80*(2), 144–179. https://doi.org/10.3102/0034654309359353

Cabell, S. Q., & Hwang, H. (2020). Building content knowledge to boost comprehension in the primary grades. *Reading Research Quarterly*, *55*(S1), S99–S107. https://doi.org/10.1002/rrq.338

Center for Research in Education and Social Policy. (January. 2019). *Bookworms case study report* (R19–001). https://www.cresp.udel.edu/wp-content/uploads/2019/01/R19-001.3-Bookworms-Seaford-Case-Study-Report.pdf

Cervetti, G., Wright, T., & Hwang, H. (2016). Conceptual coherence, comprehension, and vocabulary acquisition: A knowledge effect? *Reading & Writing*, *29*(4), 761–779. https://doi.org/10.1007/s11145-016-9628-x

Cheung, A. C. K., Xie, C., Zhuang, T., Neitzel, A. J., & Slavin, R. E. (2021). Success for all: A quantitative synthesis of U.S. evaluations. *Journal of Research on Educational Effectiveness*, *14*(1), 90–115. https://doi.org/10.1080/19345747.2020.1868031

Cohen, R., Mather, N., Schneider, D., & White, J. (2017). A comparison of schools: Teacher knowledge of explicit code-based reading instruction. *Reading & Writing*, *30*(4), 653–690. https://doi.org/10.1007/s11145-016-9694-0

Connor, C. M., Morrison, F. J., Fishman, B., Giuliani, S., Luck, M., Underwood, P. S., Bayraktar, A., Crowe, E. C., & Schatschneider, C. (2011). Testing the impact of child characteristics × instruction interactions on third graders' reading comprehension by differentiating literacy instruction. *Reading Research Quarterly*, *46*(3), 189–221. https://doi.org/10.1598/RRQ.46.3.1

Coyne, M. D., Simmons, D. C., Hagan-Burke, S., Simmons, L. E., Kwok, O., Kim, M., Fogarty, M., Oslund, E. L., Taylor, A. B., Capozzoli-Oldham, A., Ware, S., Little, M. E., & Rawlinson, D. M. (2013). Adjusting beginning reading intervention based on student performance: An experimental evaluation. *Exceptional Children*, *80*(1), 25–44. https://doi.org/10.1177/001440291308000101

Davis, E. A., Palincsar, A. S., Smith, P. S., Arias, A. M., & Kademian, S. M. (2017). Educative curriculum materials: Uptake, impact, and implications for research and design. *Educational Researcher*, *46*(6), 293–304. https://doi.org/10.3102/0013189X17727502

Donegan, R. E., & Wanzek, J. (2021). Effects of reading interventions implemented for upper elementary struggling readers: A look at recent research. *Reading & Writing*, *34*(8), 1943–1977. https://doi.org/10.1007/s11145-021-10123-y

Ehri, L. C. (1998). Grapheme–phoneme knowledge is essential for learning to read words in English. In J. L. Metsala & L. C. Ehri (Eds.), *Word recognition in beginning literacy* (pp. 3–40). Erlbaum.

Farley-Ripple, E. N., May, H., Karpyn, A., Tilley, K., & McDonough, K. (2018). Rethinking connections between research and practice in education: A conceptual framework. *Educational Researcher*, *47*(4), 235–245. https://doi.org/10.3102/0013189X18761042

Foorman, B., Beyler, N., Borradaile, K., Coyne, M., Denton, C. A., Dimino, J., Furgeson, J., Hayes, L., Henke, J., Justice, L., Keating, B., Lewis, W., Sattar, S., Streke, A., Wagner, R., & Wissel, S. (2016). *Foundational skills to support reading for understanding in kindergarten through 3rd grade* (NCEE 2016-4008). National Center for Education Evaluation and Regional Assistance (NCEE), Institute of Education Sciences, U.S. Department of Education.

Fuchs, D., Fuchs, L. S., Mathes, P. G., & Simmons, D. C. (1997). Peer-assisted learning strategies: Making classrooms more responsive to diversity. *American Educational Research Journal*, *34*(1), 174–206. https://doi.org/10.3102/00028312034001174

Gabriel, R. (2021). The sciences of reading instruction. *Educational Leadership*, *78*(8), 58–64.

Ganske, K. (2008). *Mindful of words: Spelling and vocabulary explorations 4–8*. Guilford Press.

Gersten, R., Compton, D., Connor, C. M., Dimino, J., Santoro, L., Linan-Thompson, S., & Tilly, W. D. (2009). *Assisting students struggling with reading: Response to intervention and multi-tier intervention for reading in the primary grades. A practice guide* (NCEE 2009-4045). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Graham, S., Bollinger, A., Booth Olson, C., D'Aoust, C., MacArthur, C., McCutchen, D., & Olinghouse, N. (2012). *Teaching elementary school students to be effective writers: A practice guide* (NCEE 2012- 4058). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Graham, S., Harris, K. R., & Chorzempa, B. F. (2002). Contribution of spelling instruction to the spelling, writing, and reading of poor spellers. *Journal of Educational Psychology*, *94*(4), 669–686. https://doi.org/10.1037/0022-0663.94.4.669

Graham, S., & Hebert, M. (2010). *Writing to read: Evidence for how writing can improve reading. A Carnegie Corporation time to act report*. Alliance for Excellent Education.

Grissmer, D., Buddin, R., Berends, M., Willingham, D., DeCoster, J., Duran, C., Hulleman, C., Murrah, W., & Evans, T. (2023). *A kindergarten lottery evaluation of core knowledge charter schools: Should building general knowledge have a central role in educational and social science research and policy?*. (EdWorkingPaper: 23-755). Annenberg Institute at Brown University. https://doi.org/10.26300/nsbq-hb21

Guthrie, J. T., Wigfield, A., Barbosa, P., Perencevich, K. C., Taboada, A., Davis, M. H., Scafiddi, N. T., & Tonks, S. (2004). Increasing reading comprehension and engagement through concept-oriented reading instruction. *Journal of Educational Psychology*, *96*(3), 403–423. https://doi.org/10.1037/0022-0663.96.3.403

Hanford, E. (2018, September 10). *Hard words: Why aren't kids being taught to read*. APM Reports. https://www.apmreports.org/story/2018/09/10/hard-words-why-american-kids-arent-being-taught-to-read

Harris, K. R., Graham, S., & Mason, L. H. (2006). Improving the writing, knowledge, and motivation of struggling young writers: Effects of self-regulated strategy development with and without peer support. *American Educational Research Journal*, *43*(2), 295–340. https://doi.org/10.3102/00028312043002295

Hebert, M., Bohaty, J. J., Nelson, J. R., & Brown, J. (2016). The effects of text structure instruction on expository reading comprehension: A meta-analysis. *Journal of Educational Psychology*, *108*(5), 609–629. https://doi.org/10.1037/edu0000082

Hwang, H., Cabell, S. Q., & Joyner, R. E. (2022). Effects of integrated literacy and content-area instruction on vocabulary and comprehension in the elementary years: A meta-analysis. *Scientific Studies of Reading*, *26*(3), 223–249. https://doi.org/10.1080/10888438.2021.1954005

Hwang, H., Cabell, S. Q., & Joyner, R. E. (2023). Does cultivating content knowledge during literacy instruction support vocabulary and comprehension in the elementary school years? A systematic review. *Reading Psychology*, *44*(2), 145–174. https://doi.org/10.1080/02702711.2022.2141397

Joyce, K. E., & Cartwright, N. (2020). Bridging the gap between research and practice: Predicting what will work locally. *American Educational Research Journal*, *57*(3), 1045–1082. https://doi.org/10.3102/0002831219866687

Kane, T. J., Owens, A. M., Marinell, W. H., Thal, D. R. C., & Staiger, D. O. (2016). *Teaching higher: Educators' perspectives on common core implementation*. Center for Education Policy Research, Harvard University. https://cepr.harvard.edu/publications/teaching-higher-educators-perspectives-common-core-implementation

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

Knight McKenna, M. (2008). Syllable types: A strategy for reading multisyllabic words. *Teaching Exceptional Children*, *40*(3), 18–24.

Kuhn, M. R., & Stahl, K. A. (2022). Teaching reading: Development and differentiation. *Phi Delta Kappan*, *103*(8), 25–31. https://doi.org/10.1177/00317217221100007

Learning Forward. (2018). *High-quality curricula and team-based professional learning: A perfect partnership for equity*. Author. https://learningforward.org/wp-content/uploads/2018/05/curriculaPLequity.pdf

MacArthur, C., Schwartz, S., & Graham, S. (1991). Effects of a reciprocal peer revision strategy in special education classrooms. *Learning Disability Research and Practice*, *6*(4), 201–210.

May, H., & Supovitz, J. A. (2006). Capturing the cumulative effects of school reform: An 11-year study of the impacts of America's choice on student achievement. *Educational Evaluation & Policy Analysis*, *28*(3), 231–257. https://doi.org/10.3102/01623737028003231

McKenna, M. C., Walpole, S., & Jang, B. G. (2017). Validation of the informal decoding inventory. *Assessment for Effective Intervention*, *42*(2), 110–118. https://doi.org/10.1177/1534508416640747

McKeown, M. G., Beck, I. L., & Blake, R. G. K. (2009). Rethinking reading comprehension instruction: A comparison of instruction for strategies and content approaches. *Reading Research Quarterly*, *44*(3), 218–253. https://doi.org/10.1598/RRQ.44.3.1

National Center for Education Statistics. (2022). *NAEP report card: 2022 NAEP reading assessment. Highlighted results at grades 4 and 8 for the nation, states, and districts*. Institute of Education Sciences.

National Governors Association & Council of Chief State School Officers. (2010) . *Common core state standards for English language arts and literacy in history/social studies, science, and technical subjects*. Authors.

National Reading Panel. (2000). *Report of the national reading panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*. National Institute of Child and Human Development Clearinghouse.

Neitzel, A. J., Lake, C., Pellegrini, M., & Slavin, R. E. (2022). A synthesis of quantitative research on programs for struggling readers in elementary schools. *Reading Research Quarterly*, *57*(1), 149–179. https://doi.org/10.1002/rrq.379

Neuman, S. B., Samudra, P., & Danielson, K. (2021). Effectiveness of scaling up a vocabulary intervention for low-income children, pre-K through first grade. *The Elementary School Journal*, *121*(3), 385–409. https://doi.org/10.1086/712492

Northwest Evaluation Association. (2011) . *Technical manual for measures of academic progress (MAP) and measures of academic progress for primary grades (MPG)*. Author.

Northwest Evaluation Association. (2019). *Normative data & RIT scores*. Author. https://www.nwea.org/normative-data-rit-scores/

Open Up Resources. (2018). *Bookworms K-5 reading and writing*. Author. https://openupresources.org/bookworms-k-5-reading-writing-curriculum/

Oudeans, M. K. (2003). Integration of letter-sound correspondences and phonological awareness skills of blending and segmenting: A pilot study examining the effects of instructional sequence on word reading for kindergarten children with low phonological awareness. *Learning Disability Quarterly*, *26*(4), 258–280. https://doi.org/10.2307/1593638

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

Pasquarella, A. (n.d.). *Georgia striving readers comprehensive literacy grant: Longitudinal evaluation 2012-2017*. Georgia Department of Education. https://www.gadoe.org/Curriculum-Instruction-and-Assessment/L4/Documents/SRCL_5yr_Report_FINAL%202012_2017.pdf

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

Reutzel, D. R., Fawson, P. C., & Smith, J. A. (2008). Reconsidering silent sustained reading: An exploratory study of scaffolded silent reading. *The Journal of Educational Research*, *102*(1), 37–50. https://doi.org/10.3200/JOER.102.1.37-50

Reutzel, D. R., Smith, J. A., & Fawson, P. C. (2005). An evaluation of two approaches for teaching reading comprehension strategies in the primary years using science information texts. *Early Childhood Research Quarterly*, *20*(3), 276–305. https://doi.org/10.1016/j.ecresq.2005.07.002

Saddler, B., & Graham, S. (2005). The effects of peer-assisted sentence-combining instruction on the writing performance of more and less skilled young writers. *Journal of Educational Psychology*, *97*(1), 43–54. https://doi.org/10.1037/0022-0663.97.1.43

Savage, R., Carless, S., & Stuart, M. (2003). The effects of rime- and phoneme-based teaching delivered by learning support assistants. *Journal of Research in Reading*, *26*(3), 211–233. https://doi.org/10.1111/1467-9817.00199

Schwanenflugel, P. J., Kuhn, M. R., Morris, R. D., Morrow, L. M., Meisinger, E. B., Woo, D. G., Quirk, M., & Sevcik, R. (2009). Insights into fluency instruction: Short- and long-term effects of two reading programs. *Literacy Research and Instruction*, *48*(4), 318–336. https://doi.org/10.1080/19388070802422415

Schwartz, R. M., & Raphael, T. E. (1985). Concept of definition: A key to improving students' vocabulary. *The Reading Teacher*, *39*(2), 198–205. https://www.jstor.org/stable/20199044

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage Learning.

Shanahan, T., Callison, K., Carriere, C., Duke, N. K., Pearson, P. D., Schatschneider, C., & Torgesen, J. (2010). *Improving reading comprehension in kindergarten through 3rd grade: A practice guide* (NCEE 2010-4038). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Simmons, D. C., Coyne, M. D., Hagan-Burke, S., Kwok, O., Simmons, L., Johnson, C., Zou, Y., Taylor, A. B., McAlenney, A. L., Ruby, M., & Crevecoeur, Y. C. (2011). Effects of supplemental reading interventions in authentic contexts: A comparison of kindergarteners' response. *Exceptional Children*, *77*(2), 207–228. https://doi.org/10.1177/001440291107700204

Song, M., Garet, M. S., Yang, R., & Atchison, D. (2022). Did states' adoption of more rigorous standards lead to improved student achievement? Evidence from a comparative interrupted time series study of standards-based reform. *American Educational Research Journal*, *59*(3), 610–647. https://doi.org/10.3102/00028312211058460

Stornaiuolo, A., Desimone, L., & Polikoff, M. (2023). "The good struggle" of flexible specificity: Districts balancing specific guidance with autonomy to support standards-based instruction. *American Educational Research Journal*, *60*(3), 521–561. https://doi.org/10.3102/00028312231161037

Troia, G. A., & Graham, S. (2002). The effectiveness of a highly explicit, teacher-directed strategy instruction routine: Changing the writing performance of students with learning disabilities. *Journal of Learning Disabilities*, *35*(4), 290–305. https://doi.org/10.1177/00222194020350040101

Vadasy, P. F., & Sanders, E. A. (2011). Efficacy of supplemental phonics-based instruction for low-skilled first graders: How language minority status and pretest characteristics moderate treatment response. *Scientific Studies of Reading*, *15*(6), 471–497. https://doi.org/10.1080/10888438.2010.501091

Walpole, S. (2018). Bookworms implementation and coaching: End-of-year report 2017-18. [Unpublished manuscript]. Professional Development Center for Educators, University of Delaware.

Walpole, S., & McKenna, M. C. (2017). *How to plan differentiated reading instruction: Resources for grades K-3* (2nd ed.). Guilford Press.

Walpole, S., McKenna, M. C., Amendum, S., Pasquarella, A., & Strong, J. Z. (2017). The promise of a literacy reform effort in the upper elementary grades. *The Elementary School Journal*, *118*(2), 257–280. https://doi.org/10.1086/694219

Walpole, S., McKenna, M. C., & Morrill, J. (2011). Building and rebuilding a statewide support system for literacy coaches. *Reading & Writing Quarterly*, *27*(3), 261–280. https://doi.org/10.1080/10573569.2011.532737

Walpole, S., McKenna, M. C., Philippakos, Z. A., & Strong, J. Z. (2020). *Differentiated literacy instruction in grades 4 and 5: Strategies and resources* (2nd ed.). Guilford Press.

Wexler, N. (2019). *The knowledge gap: The hidden cause of America's broken education system – and how to fix it*. Avery.

What Works Clearinghouse. (2020). *What works clearinghouse standards handbook, version 4.1*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/wwc/handbooks

White, T. G. (2005). Effects of systematic and strategic analogy-based phonics on grade 2 students' word reading and reading comprehension. *Reading Research Quarterly*, *40*(2), 234–255. https://doi.org/10.1598/RRQ.40.2.5

Willson, V. L., & Rupley, W. H. (1997). A structural equation model for reading comprehension based on background, phonemic, and strategy knowledge. *Scientific Studies of Reading*, *1*(1), 45. https://doi.org/10.1207/s1532799xssr0101_3

Wilson, J. O. (2017). Bookworms implementation and coaching: End-of-year report 2016-2017 [Unpublished manuscript]. Professional Development Center for Educators, University of Delaware.

Wright, T. S., & Cervetti, G. N. (2017). A systematic review of the research on vocabulary instruction that impacts text comprehension. *Reading Research Quarterly*, *52*(2), 203–226. https://doi.org/10.1002/rrq.163