DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY

Building a Science of Teaching Reading and Vocabulary: Experimental Effects of Structured Supplements for a Read Aloud Lesson on Third Graders' Domain-Specific Reading Comprehension

Citation

Mosher, D. M., & Kim, J. S. (2024). Building a Science of Teaching Reading and Vocabulary: Experimental Effects of Structured Supplements for a Read Aloud Lesson on Third Graders' Domain-Specific Reading Comprehension. Scientific Studies of Reading, 29(1), 7–31.

Published version

https://doi.org/10.1080/10888438.2024.2368145

Link

https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37380255

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Open Access License Articles (IOAL), as set forth at

https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFlOWI2M2EwYTg

Accessibility

https://accessibility.huit.harvard.edu/digital-accessibility-policy

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story

Building a Science of Teaching Reading and Vocabulary: Experimental Effects of Structured Supplements for a Read Aloud Lesson on Third Graders' Domain-Specific Reading Comprehension

Douglas M. Mosher

James S. Kim

Harvard University, Graduate School of Education

Mosher, D. M., Kim, J. S. (2024). Building a science of teaching reading and vocabulary: Experimental effects of structured supplements for a read aloud lesson on third graders' domain-specific reading comprehension. *Scientific Studies of Reading, 0*(0), 1-25. https://doi.org/10.1080/10888438.2024.2368145

© 2024, Taylor & Francis. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors' permission. The final article will be available, upon publication, via its DOI: 10.1080/10888438.2024.2368145

Abstract

Purpose: This study contributes to the science of teaching reading and vocabulary by illustrating how a ubiquitous classroom practice – read alouds – can be enhanced by structured supplements. This experimental study examines whether and to what extent providing *structured supplements* can improve student comprehension outcomes by helping teachers foster discussions about academic vocabulary that support schema transfer as students make connections between known and new topics.

Method: A total of 80 third-grade teachers and their students (N = 965; 32% Black, 31% Hispanic, 25% white, 9 Asian, 48% Male) were randomly assigned to treatment or control conditions. Treatment students received a single social studies read aloud on the story of Apollo 11 with structured supplements while control students received the same read aloud story but without structured supplements. Students were from linguistically, economically, and ethnically diverse backgrounds.

Results: Effect sizes from hierarchical linear models indicated that students in the treatment condition outperformed students in the control condition on four measures of domain-specific reading comprehension: recall (ES=.17), near-transfer (ES=.17), mid-transfer (ES=.18), and content comprehension (ES=.18). Further exploratory analyses using structural equation modeling revealed that teacher language scaffolds—that is, temporary dialogic supports in which teachers went above and beyond the intervention script—explained 66% of the treatment effect on domain-specific reading comprehension.

Conclusion: Results from this study suggest that read alouds, when enhanced with structured supplements designed to facilitate schema transfer, can increase the amount of academic

vocabulary teachers use during classroom instruction and improve their students' ability to comprehend disciplinary texts.

Introduction

Attempting to improve reading comprehension outcomes is no small task, yet incremental research that bends the knowledge-seeking arc and provides teachers with practical and actionable practices is exactly what is needed to improve literacy outcomes for children. Indeed, students in the United States continue to struggle comprehending grade-level texts, as evidenced by the stagnant growth among fourth-grade reading achievement for a decade prior to the COVID-19 pandemic (U.S. Department of Education, 2019). With the science of reading in the forefront of public discourse, there is a timely need to highlight ways in which teachers can positively impact students' ability to comprehend texts. This study contributes to the science of teaching reading and vocabulary by illustrating how a ubiquitous classroom practice – read alouds – can be enhanced by structured supplements that are designed to foster teachers' language scaffolds and to improve students' reading comprehension.

Thus, our aim was to examine the causal effects of embedding a number of *structured supplements* into a third-grade read aloud lesson on various measures of reading comprehension. Structured supplements for a read aloud lesson are designed to help teachers foster more discussions about academic vocabulary during classroom instruction. More specially, structured supplements include an integrated set of instructional practices such as providing concise definitions and meaningful examples of target vocabulary, asking stimulating discussion questions to engage students in using the target vocabulary to talk about larger topics related to the topic schema, making connections between known and new schemas, and discussing key vocabulary words using a concept map. While there is ample research documenting the importance of the numerous components of the science of reading – developing phonological

awareness, explicit phonics instruction, activating background knowledge, and building vocabulary knowledge – there is a need to better understand how teachers can use read aloud lessons to help students leverage, connect, and build the necessary knowledge to instantiate schemas efficiently in order to improve reading comprehension outcomes (Shanahan, 2020). Put differently, understanding how teachers can efficiently integrate schematic knowledge while introducing new content to students is an area of research that needs further investigation.

One type of approach that has had positive effects on knowledge building and reading outcomes more broadly is content literacy, which privileges opportunities for students to build domain knowledge in science and social studies while also developing literacy abilities specific to English Language Arts (ELA) (Connor et al., 2017; Duke et al., 2021; Guthrie et at., 2004 Williams et al., 2016). In fact, meta-analytic evidence suggests that content literacy programs can improve elementary school standardized comprehension by a quarter of a standard deviation (Hwang et al., 2021). In the recently completed Reading for Understanding (RfU) Initiative, a $125M effort by the U.S. Department of Education to improve reading comprehension from preschool to Grade 12, Pearson and colleagues (2020) found that the most promising whole-class interventions provided curricular resources that fostered teacher language and peer discussion around engaging science and social studies topics and vocabulary. Building on findings from the RfU projects, our aim in this study was to develop curricular resources for teachers called "structured supplements" that are usable and effective in whole class contexts (Tier I instruction).

Theory of Change

The theory of change shown in Figure 1 illustrates how the core components of structured supplements, which complement an existing content literacy program, can lead to improvements

in students' reading comprehension as well as the language scaffolds that teachers provide for their students. The theory of change rests on three pillars of research, including research on how structured supplements can help teachers foster discussions about academic vocabulary, go above and beyond the lesson script to support vocabulary learning, and contribute to students' reading comprehension.

Research Supporting Structured Supplements

The first component of structured supplements emphasizes the importance of domain-specific vocabulary instruction. There is growing evidence that teaching domain-specific vocabulary in semantic networks has been shown to impact proximal measures of vocabulary knowledge (Kim et al., 2020; 2021) and that networks of domain-specific vocabulary have mediated treatment effects on domain-specific reading comprehension measures (Mosher et al., 2024). When teachers create rich semantic representations, it helps students store word meanings in their long-term memory (Ericsson & Kintsch, 1995), thereby allowing students to retrieve word meanings efficiently to aid in understanding various texts (Fitzgerald et al., 2020). Furthermore, creating "an extended network of meanings for a given focal word represents instantiation of stronger and richer meaning" (Fitzgerald et al., 2020, p. 857), and visualizing networks of vocabulary with concept maps can help students see the connections between words and concepts (Karpicke & Blunt, 2011; Novak, 1990).

As teachers aid students in building networks of domain-specific vocabulary knowledge, they are ultimately providing the necessary instructional supports for students to acquire and retain domain and topic knowledge in disciplines such as social studies and science (Kim et al., 2020; Douglas & Albro, 2014; Pearson et al., 2020). These networks of word knowledge signify

the various interconnections between concepts within a student's memory (Kendeou & O'Brien 2016). Put differently, deep understanding of semantic word meanings can serve as a proxy for content knowledge, suggesting that depth of vocabulary knowledge should be privileged over breadth of vocabulary knowledge (Anderson & Freebody, 1981; Stahl & Nagy, 2006).

The second component of structured supplements emphasizes the importance of making explicit connections between known and new schemas. As teachers' instruction focuses on domain-specific vocabulary words and students acquire content-specific knowledge, there is a need for students to develop supportive intellectual structures on which to hang their knowledge in a systematic and structured way. As such, teachers guiding students towards instantiating schemas can help learners acquire, access, and retain relevant information (Anderson & Pearson, 1984). Essentially, "a schema functions as a unified system of background relationships whose visible parts stand for the rest of the schema" (Hirsch, 1987, p. 54). Schemas are also essential to comprehending texts (Anderson, 2013). In order to comprehend a text, an individual needs to have "found a mental 'home' for the information in the text" or make the necessary adjustments to integrate any new information with that which is already established (Anderson & Pearson, 1984, p. 225). Thus, schemas are malleable, constantly evolving, and hierarchical (Kimball & Holyoak, 2000).

As they teach different topics within content literacy programs, teachers can help students develop *topic schemas*. For example, teaching an in-depth unit on various human body systems – i.e., muscular, skeletal, and nervous – provides opportunities for students to develop a topic schema on how human body systems function to help the body stay alive. If teachers follow this unit with another that centers around a system of individuals working together to achieve a goal,

they will likely help students develop a new but similar topic schema about the system of individuals. Both topic schemas share a common overarching *general schema* that centers around the concept of systems functioning together to thrive. As students instantiate different topic schemas that have an underlying connection, they begin to develop a general schema. While background knowledge and schemas are sometimes referenced interchangeably, we argue that even though background knowledge is important, it does not necessarily refer to knowledge that is systematically organized. Schema instantiation suggests a deeper understanding of concepts – concepts that are linked through semantic, conceptual, and topical elements that organize knowledge in meaningful ways.

In the present study, we rely on a science topic schema previously covered in depth earlier in the school year – human-body systems functioning together to help the human body stay alive – to provide context for teachers to help students build a new topic schema in social studies (a system of individuals working to achieve a goal of sending humans to the moon). Both topic schemas are nested within the general schema, systems that function together. Figure 2 shows how domain-specific vocabulary words serve as kernels of content knowledge, with both topic schemas sharing the words *system*, *function*, and *diagnose*, thereby helping to highlight conceptual similarities between schemas.

The third component of structured supplements emphasizes the importance of fostering discussions about topic schemas using academic vocabulary. In order to build networks of vocabulary, which are integral to instantiating topic schemas, teachers need to create opportunities for students to engage with the vocabulary words they are learning. Teachers can support this by engaging students in discussion about the topic, with an emphasis on using the

domain-specific vocabulary words they have been learning. One way to facilitate discussions among students is to embed discussion prompts into an actual read aloud lesson. In fact, scripting lessons has been shown to lead to higher-quality teaching (LARRC; Language and Reading Research Consortium et al., 2014). Furthermore, engaging students in scaffolded and inferential discussions is an important element of read aloud lessons that develops students' oral language abilities, impacts vocabulary acquisition (LARRC et al., 2016), and correlates with comprehension scores (van Kleeck, 2006; Collins, 2016). Teachers providing instructional scaffolds for key vocabulary words and interactive discussions have improved students' vocabulary knowledge and indirectly impacted content knowledge (Relyea et al., 2022). Read alouds including conversational turn-taking and open-ended questions (i.e., dialogic reading; Ezell & Justice, 2005; Swanson et al., 2011) are among the most common type of read aloud lesson and have yielded positive effects. Indeed, shifting from students listening to a read aloud and answering low-level questions to a more interactive and cognitively engaging read aloud experience is essential to helping students acquire new knowledge.

Research Supporting Teacher Language Scaffolds

In this study, we define teacher language scaffolds as temporary dialogic supports that are designed to increase the quantity of exposure to key words (input), to provide additional examples and word meanings (instruction) through language extensions (Neugebauer et al., 2017), and to promote more discussions (interaction; Snow, 2014). While teacher language has been theorized to be a central mechanism for building students' background and vocabulary knowledge, particularly in content literacy programs (Conner et al., 2017; LARRC et al., 2022), little is known about causal effects of teacher language scaffolds on student outcomes. However,

there is convergent correlational evidence that teacher language scaffolds could be an important pathway for improving students' comprehension. For example, in a study of third-grade classrooms in high poverty schools, Carlisle and colleagues (2013) found that teacher language scaffolds that foster cognitively challenging dialogic actions – discussions about words and asking questions about word meanings – were positively associated with third graders' reading comprehension, controlling for prior student achievement and teacher reading knowledge. Additionally, there is experimental evidence indicating that targeted inferential and literal questioning can positively impact comprehension outcomes among elementary school children (McMaster et al., 2012), and that providing inferential questioning during read alouds promotes increases in inferencing ability (J. Kim et al., 2023). In another recent study, Al-Adeimi and O'Connor (2021) found that teacher discourse features that were more dialogic and fostered "synchronous exchanges about thinking and reasoning" about open-ended questions (e.g., "Should everyone learn a second language?") predicted students' persuasive writing outcomes, controlling for students' academic language ability and background characteristics (p. 2). These empirical findings are broadly consistent with the idea that dialogic language scaffolds (teacher discussions) more effectively spark and sustain student engagement during classroom instruction than authoritative scaffolds (teacher telling) (Bakhtin, 1981; Wood et al., 1976).

Teacher language scaffolds may be a particularly important mechanism for supporting vocabulary learning during content literacy instruction when students are learning about science and social studies topics (Anderson et al., 2023). For example, what makes vocabulary learning so difficult is that the meaning of words like "system" are indeterminate and unconstrained. When "system" is used in life science, it may refer to the human body system, which has parts

that move or work together to keep us healthy. When "system" is used in history, however, it may refer to a social system, such as the team of engineers, designers, and builders who worked together to send Apollo 11 to the moon. Vygotsky (1986) emphasized that a word's meaning is often a "dynamic, fluid, complex whole, which has several zones of unequal stability" (p. 245). A practical challenge facing teachers is stimulating questions that make learning new vocabulary challenging but not frustrating. Vygotsky's notion of the Zone of Proximal Development (ZPD) suggests temporary dialogic supports are needed to reduce the word learning burden (McKeown et al., 2017) by helping navigate challenging cognitive tasks above their current level of development (Vygotsky, 1978). As students acquire new knowledge via domain-specific vocabulary words and work to master the form and meaning of words, they need numerous opportunities to hear vocabulary in varied contexts and settings (Perfetti, 2007). Thus, Vygotsky's (1978) assertion that "what a child can do with assistance today she will be able to do by herself tomorrow" (p. 87) suggests that the teacher language scaffolds enable students to begin to internalize word meanings and concepts – and ultimately apply this knowledge in relevant contexts.

In the absence of adequate teacher scaffolding, students will encounter a large volume of unknown academic words, thus taxing working memory, slowing down word-to-text integration processes, and disrupting the cognitive processes that support the building of coherent text representations (Fitzgerald et al., 2022; Perfetti & Stafura, 2014). While scripting lessons can create a desired baseline-level of instructional language, it does not take into account students' current development levels and the scaffolds they may need to work within their ZPD (Vygotsky,

1978). Thus, it seems likely that teachers providing additional language scaffolds might mediate the relation between a scripted lesson and reading-comprehension outcomes.

How Structured Supplements and Teacher Language Scaffolds Contribute to Students' Reading Comprehension

Because it is impossible to teach every concept, reading comprehension requires that readers be able to "transfer knowledge from one situation to another by a process of *mapping*" (Gick & Holyoak, 1983, p. 2). More specifically, establishing a set of one-to-one correspondences that map from one concept to another – i.e., one topic schema mapping onto another schema – helps students ultimately transfer and apply knowledge when they encounter an unfamiliar topic. This process allows students to leverage their existing knowledge organized in schemas as students encounter knew situations that vary along a continuum from familiar to unfamiliar (Kimball & Holyoak, 2000). To move past lower-level comprehension tasks such as recalling facts and to progress towards a deeper understanding of knowledge and how it applies to other concepts, acquired knowledge "must be actively linked to semantic retrieval cues" to make acquiring knowledge from texts possible in unfamiliar situations (Kintsch, 2013, p. 812). Put differently, words that convey conceptual meaning that are organized in networks are essential to reading and comprehending informational texts.

In the construction-integration model, readers must build a robust situation model that merges a reader's literal understanding of a text with that reader's existing background knowledge (Kintsch, 1993; 2009). First, all learners must formulate propositions by accessing word meanings, utilizing working memory, and assessing the coherence of the various propositions (Kintsch & Kintsch, 2005). During the integration phase, however, readers must

incorporate the literal propositions from the text with their existing knowledge and experiences – i.e., robust schemas – to form a mental model of the text and its meaning.

Many read aloud lessons, including those found in content literacy programs, include multiple lessons over numerous days of instruction. There has yet to be a study examining whether and to what extent supports for teachers – that is, structured supplements for a single read aloud lesson – can impact reading comprehension outcomes for elementary school students. Thus, while teachers may have content knowledge around the science of reading, there is a need to better understand what supports teachers need to actually teach the science of reading.

Study Hypotheses and Research Questions

Our primary goal for the study was to investigate the effect of structured supplements for a single read aloud lesson on assessments of reading comprehension transfer to provide theoretical, empirical, and usable knowledge of how to measure and promote classroom practices that could advance the science of teaching reading.

As shown in the theory of change in Figure 1, we hypothesized that structured supplements for teachers would aid their students' ability comprehend grade-level texts. Furthermore, we hypothesized that structured supplements would impact teacher language scaffolds – language not in the script that supports acquisition of word meanings and concepts – and that teacher language scaffolds would mediate the proposed treatment effect.

The research questions were as follows:

(1) Compared to a read aloud lesson without structured supplements, to what extent does embedding structured supplements into a read aloud lesson improve: (a) third graders' recall of basic topics discussed in the text, (b) their domain-specific reading

comprehension as measured by near- and mid-transfer assessments, and (c) their domain-general reading comprehension as measured by an End-of-Grade (EOG) reading comprehension assessment?

(2) To what extent do teacher language scaffolds mediate effects of the read aloud lesson with structured supplements on student comprehension scores (recall, near, and mid transfer)?

Method

Design

To address these questions, we conducted a cluster randomized controlled trial where classrooms were randomly assigned to either the treatment (read aloud with structured supplements) or control conditions (read aloud without structured supplements). The experimental design for this study blocked on schools with random assignment occurring at the classroom level. Randomization was achieved using Stata 17 (Stata Corp, 2021). Balance tests at the student level showed no significant differences between treatment and control groups in standardized measures of baseline reading and math ability or in baseline student demographic characteristics, except for students classified as Other (two or more races/Native American), which was slightly higher in the treatment condition. Balance tests at the teacher level showed no significant differences in years teaching or past experience teaching previous intervention units. (See Supplemental Appendix for balance tables S-A.1-4)

Transparency Statement

We preregistered the study at the Registry of Efficacy and Effectiveness Studies prior to lesson implementation in May of 2022 (Mosher, 2022). The data and analysis code to replicate

all results found in this study are available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/LUKGIA .

Participants

We recruited teachers from the larger 2022 Model of Reading Engagement (MORE) science intervention that had desirable compliance rates – i.e., teachers who implemented most MORE lessons (Figure 3). In total, 82 teachers agreed to participate, with a prospective total of 1,467 third-grade students from 20 elementary schools. In preparation for this initial efficacy trial, we piloted the read aloud lesson in multiple third-grade classrooms resulting in adjustments to the story and activities prior to implementation.

Of the original 1,467 students at the time of randomization, 502 students dropped out of the study due to the transient nature of the district, absences, lack of returned consent forms, and students opting out of the study. The final analytic sample included 965 students from 80 classrooms across 19 schools with an attrition rate of 34 percent. There was no evidence of differential attrition between treatment and control conditions ($\gamma_{01} =. -.14, SE = .18, p = .43$). In the final analytic sample, there were no statistically significant differences between the treatment and control group on the baseline MAP reading ($\gamma_{01} = .52, SE = 1.31, p = .69$) or math ($\gamma_{01} = .52, SE = 1.03, p = .62$) tests. The two groups were also equivalent on most demographic characteristics, suggesting that there were no major threats to the internal validity of the findings.

Procedure

In March of 2022, all third-grade students within a large, urban district in the Southeastern United States participated in the MORE science intervention for three weeks where

they learned about three important human body systems: the muscular, skeletal, and nervous systems. Students also learned key domain-specific vocabulary words that helped build their domain and topic knowledge about human body systems. As lessons progressed, students constructed a schema around how each system works together to help the body function. Thus, we hypothesized that if students had instantiated a robust topic schema from the science lessons about how human body systems function together to keep the body alive, students should be able to leverage the existing schema to construct a new but closely related topic schema in the domain of social studies about a system of individuals working together to send astronauts to the moon (Anderson & Pearson, 1984; Gick & Holyoak, 1983). Yet, what scaffolds are needed to help students instantiate this closely related schema?

Our principal aim of this study was to see if a single lesson delivered two months after the conclusion of the MORE intervention could impact reading comprehension outcomes. To test the impact of a single read aloud lesson with structured supplements on comprehension outcomes, we wrote a nonfiction text called *The True Story of the Apollo 11 Moon Landing*. The text focuses on how a team of individuals – Katherine Johnson, Chuck Lowry, Neil Armstrong, and John F. Kennedy – worked together as a system to overcome obstacles and accomplish their goal of sending astronauts to the moon and returning them safely to Earth. Drawing on Lin-Siegler and colleagues' (2016) study of how students who read stories about scientists overcoming challenges resulted in improved learning, we wrote our own struggle story about the challenges that a group of individuals faced when trying to land on the moon and how they overcame them. Teachers read the story aloud in May 2022, well over a month after the MORE science intervention had ended. Teachers received the lesson slides one week prior to

implementation. We did not provide any professional development. The present study functioned as a stand-alone intervention. Structured supplements are shown in Table 1 and included vocabulary instruction and discussion activities designed to help students build their vocabulary, domain, and topic knowledge while drawing on the schema developed during the MORE science intervention. Structured supplements also included schema-mapping and concept-mapping activities.

We included everything treatment teachers needed on the lesson slides to keep lesson preparation (creating two anchor charts) at a minimum. Otherwise, we instructed teachers to use the slides to teach the lesson and did not give explicit guidelines about whether teachers could adapt the lesson. Control teachers received the same Apollo 11 read aloud story but without any of the structured supplements designed to identify and explain key vocabulary words, pose discussion questions, or help students link the science and social studies topic schemas. In essence, the read aloud text that control teachers received was a blank slate that gave full autonomy to control teachers to decide which, if any, words to introduce and what questions to ask. We indicated to control teachers that while there was no set lesson plan, teachers were welcome to stop at various points and ask questions. We stressed that there was no "right" way to use the read aloud. Students in both conditions completed identical assessments at the conclusion of the lesson. In sum, all students were exposed to the social studies read aloud and random assignment to the treatment condition allows for an estimate of the effectiveness of structured supplements on student comprehension outcomes.

The read aloud text included three vocabulary words from the MORE science intervention students had already learned: *function*, *system*, and *diagnosis*. The text also

introduced four new social studies words: *contribute*, *engineer*, *persist*, and *ingenious*. The treatment version of the text provided embedded explanations of the four social studies words along with relevant examples and specific questions designed to prompt students to access the science schema and make connections between it and the read aloud content. For example, after learning about different members of the team that helped send astronauts to the moon, treatment teachers asked students, "How is the moon team *system* similar to the different human body *systems*?" Students subsequently completed a schema-mapping activity designed to help students see the analogic link between science and social studies schemas, thereby engaging students in a 1:1 mapping of both topic schemas (Gick & Holyoak, 1983). In this activity, teachers asked students about the similarities between the "moon team system" and the "human body system". Teachers then led students through a "madlibs" activity (Figure 4) where students had to fill in the missing words scattered across two paragraphs using key vocabulary from both the previously taught science unit (paragraph 1) and the social studies read aloud lesson (paragraph 2). Both paragraphs had similar language and the main difference between them centered around which target words completed the paragraph. We wanted this activity to make the 1:1 schema mapping explicit. Elsewhere in the lesson, treatment teachers posed numerous questions for students to discuss with a partner – e.g., "How was Chuck Lowry an *ingenious engineer*?" – which required students to use target vocabulary in their discussions (see "must haves/amazing" in Table 1).

Student Measures

Pretests

We used winter reading and math scores from the Measure of Academic Progress (MAP) assessment (Northwest Evaluation Association, 2011) as measures of baseline equivalence. The district used the MAP assessment at three times points throughout the year to assess student achievement. The winter assessment date represents the mid-year assessment of student achievement and was the most recent standardized measure of student reading and math ability prior to the start of the lesson.

Posttests

Students took four assessments following the conclusion of the read aloud lesson. We developed three measures of domain-specific reading comprehension, all of which were scored dichotomously. The fourth measure, the state-wide end-of-grade (EOG) assessment, was administered one month later.

Recall. To measure basic recall, students took a 5-item multiple choice assessment (with 4 options per item) measuring basic understanding of key details from the text immediately following the lesson. The measure had a Cronbach's alpha reliability estimate of 0.63. Results from a 2-paramenter logistic (2PL) item response theory (IRT) model showed the recall measure was relatively easy with location parameters ranging from -1 to -.53 and information parameters ranging from 1.41 to 1.81.

Near- and Mid-Transfer. Near- and mid-transfer domain-specific reading comprehension measures each were comprised of 13 multiple-choice items with four options per item. Recent research suggests measuring a continuum of transfer rather than simply assessing proximal and distal measures (Kim et al., 2022). The near-transfer passage discussed the creation of the lunar module that enabled astronauts to explore the moon. While students had learned about sending

astronauts to the moon, the read aloud lesson had not covered the lunar module or what astronauts actually did once they had landed on the moon. Nevertheless, the topic of the near-transfer passage was similar to that of the read aloud. The mid-transfer passage focused on the creation of the Empire State building, a topic much more removed from that of the read aloud content centered around the moon. We instructed teachers to administer these assessments within a week of completing the lesson. Because of the length of the lesson and needing the recall assessment to be administered immediately following the read aloud, we knew that even if there was time, students would likely face test-exhaustion. To accommodate shifting school schedules, we had to be flexible with when teachers would administer both transfer assessments. Near- and mid-transfer assessments yielded a Cronbach's alpha reliability of 0.82 and 0.76 respectively. Results from 2PL IRT models indicated location parameters ranging from -1 to 3 and item information parameters ranging on average from 1 to 2.

Content Comprehension. Content comprehension represents both near- and mid-transfer measures combined ($\alpha = .88$).

End-of-Grade (EOG). Domain-general reading comprehension was measured via the North Carolina third grade EOG assessment, which includes multiple-choice items (4 options) that are scaled using a 3PL IRT model (North Carolina Department of Public Instruction, 2020).

Teacher Measures

Fidelity of Implementation

We gave teachers two recorders to wear around their necks while delivering the read aloud lesson. Of the 80 participating teachers, 75 teachers returned usable recorders. The five unusable recorders were either broken, preventing us from accessing the recording, or the

recorder was never activated. Fidelity analyses use data from the 75 teachers who returned usable recorders.

To assess fidelity of implementation, we used two of Dane and Schneider's (1998) facets of fidelity: adherence and dosage. We examined teacher adherence to the lesson by comparing the number of times a teacher said each target word within each condition. For example, the read aloud with structured supplements used the word *contribute* 31 times while the lesson without structured supplements used *contribute* 10 times. When listening to the audio recordings, we would expect at a minimum to hear treatment and control teachers say *contribute* 31 and 10 times respectively. We created dichotomous variables for the seven words to indicate if each of the target words was said the prescribed number of times. We also created a dichotomous variable for whether or not the number of scripted questions included in the lesson were asked. We then summed the eight dichotomous variables to create a composite with a maximum score of eight. Results from these comparisons are listed in Table 2 where percentage of adherence is also reported. There was not a significant difference in adherence between treatment and control conditions ($\beta = -.46, SE = .33, p = .16$).

To assess dosage, we report the length of audio measured in minutes for reach recording. Unfortunately, the recorders of two different teachers inadvertently shut off mid-lesson, thus skewing both teachers' adherence and dosage and bringing down the overall treatment average. Because the read aloud lesson with structured supplements had more text, we expected the lesson to be longer. Results indicate that on average, the treatment lesson was roughly 19 minutes longer ($SE = 2.20, p < .001$).

Teacher Language Scaffolds.

We conceptualized teacher language scaffolds as the number of temporary dialogic supports in which teachers went above and beyond the intervention script. We specified a confirmatory factory analysis (CFA) measurement model for *teacher language scaffolds* where we included an indicator variable for each target word with values indicating the number of times a teacher said each target word above and beyond the number of times it appeared in the lesson script. For example, if a treatment teacher had a value of 15 for the word *system*, this would indicate that after factoring the number of times *system* was mentioned in the treatment lesson, the teacher said the word 15 additional times. We included an indicator variable for language extensions (Neugebauer et al., 2017) – the number of times a teacher provided additional explanations or examples of a word's meaning above and beyond the script. Language extensions were coded in *Dedoose* (2016) and 20% of the recorded audio transcripts were double coded with an inter-rater-reliability kappa estimate of $\kappa = .88$. The final indicator variable, teacher questions, included the number of times a teacher asked the class a question related to the text that did not appear in the lesson script.

Data Analyses

To assess the causal impact of structured supplements on five comprehension outcomes, we fit a series of 3-level hierarchical linear models (HLMs) with teacher random effects at level 2 and school fixed effects at level 3 (Raudenbush & Bryk, 2002). The model is as follows:

$$Y_{ijk} = \gamma_{00} + \alpha_k + \gamma_{01} TREATMENT_{jk} + \sum_{p=1}^{10} \gamma_{pjk}\, L1COV_{ijk} + \sum_{q=2}^{3} \gamma_{0qk}\, L2COV_{jk} + \varepsilon_{ijk} + u_{0jk}$$

$$\varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2) \qquad u_{0jk} \sim N(0, \sigma_u^2)$$

where Y_{ijk} represents the 5 outcomes (recall, near transfer, mid transfer, content comprehension, and EOG) for student *i* in classroom *j* in school *k*. γ_{00} is the intercept for the school reference group and α_k represents each school's average deviation from the reference school intercept. γ_{01} is the adjusted causal effect of structured read aloud supplements, $\gamma_{pjk\ (p=1,\ldots,10)}$ is a vector of 10 student-level covariates including winter MAP pretests in reading and math as well as student demographic variables (race/ethnicity, gender, English Language Learner status, measure of neighborhood poverty), and $\gamma_{0qk\ (q=2,3)}$ is a vector of two teacher-level covariates (years of teaching experience, past experience with MORE) to improve the precision of the treatment effect estimates. ε_{ijk} represent the student-level residuals and u_{0jk} are the classroom-level random intercepts. Because we had five outcomes, we tested our results to the sensitivity of false discoveries using the Benjamini-Hochberg procedure with a false discovery rate set to 0.05 by outcome domain (Benjamini-Hochberg, 1995).

Some teachers administered the recall, near-, and mid-transfer assessments on different days, and as a result, we had varying numbers of students that completed each part of the assessment due to student absences. We used 2PL IRT-stored estimates for recall, near- and mid-transfer, and content comprehension outcomes. For the EOG assessment, we used district reported scale scores. Because of the discrepancy in student completion, we tested for a potential treatment effect on missing assessment data to determine if there was an imbalance in treatment and control missingness. HLMs revealed that there were no significant treatment/control differences on recall ($\gamma_{01} = -.01, SE = .01, p = .23$) near- ($\gamma_{01} = -.05, SE = .05, p = .32$), or mid-transfer ($\gamma_{01} = -.03, SE = .06, p = .65$) measures. Because of missing pretest data, we conducted Little's missing completely-at-random (MCAR) test (Little, 1988). Results revealed

that missing values were MCAR ($\chi^2 = 1.74, df = 2, p = .42$). Subsequent HLMs use multiple imputation by simulating 20 data sets with potential values in place of the missing observations (StataCorp, 2021).

To address our second research question on the extent to which teacher language scaffolds mediated the treatment effect, we used coded data from classroom recordings to specify a structural equation model (SEM) using *Mplus 7* (Kline, 2016; Muthén & Muthén, 2012). As our mediation model included latent variables for *Teacher Language Scaffolds* and *Domain-Specific Reading Comprehension Transfer*, we specified two different confirmatory factor analysis (CFA) measurement models to verify each latent variable's properties. *Teacher Language Scaffolds* were measured by nine indicator variables (Figure 5). The latent factor *Domain-Specific Reading Comprehension* was measured by three indicator variables: stored IRT estimates for recall, near-transfer, and mid-transfer questions (Figure 6). We assessed adequate model fit using cutoffs specified by Hu & Bentler (1999): RMSE <.06, CFI & TLI >.90, and SRMR <.08. To account for the clustered nature of the data, we used the Maximum Likelihood Robust (MLR) estimator where we clustered standard errors at the teacher level using the TYPE = COMPLEX command. Because the key predictor (treatment) is binary and the outcome continuous, we used the STANDARDIZE command to generate effect size estimates (STDY), allowing use to interpret the treatment effect of structured supplements on teacher language scaffolds. To account for missing pretest data, we used Full Information Maximum Likelihood (FIML).

Results

Descriptive and correlational analyses

Table 3 displays descriptive statistics for the full sample and for the treatment and control groups separately on key pretest and posttest variables. The correlations in Table 4 indicate that the recall and transfer tests (near and mid) were strongly correlated (*r* range = .58 to .73) with the EOG domain general reading test. Table 5 displays descriptive statistics for the full teacher sample and Table 6 displays correlations between teacher-level variables. The language scaffolds include temporary dialogic supports that go above and beyond the intervention script. For example, treatment group teachers used the word ingenious 7 more times than what appeared in their scripts whereas control group teachers used ingenious 2 more times. The raw mean difference between conditions was 5.5 (*d* = .93). The pattern of results suggests that, in contrast to control teachers, treatment group teachers used academic vocabulary more frequently and asked more questions during their read aloud lessons. The indicators in Table 5 comprised the latent variable for *Teacher Language Scaffolds* for research question 2.

Both latent variables were assessed using CFA. All indicators were significantly correlated with the latent construct *Teacher Language Scaffolds* (Figure 5), with factor loadings for eight of the nine indicators above .70. One of the indicators had a weak correlation with the factor because the word (*diagnose*) only appeared in the text a few times (unlike the other six words), and consequently, there was minimal variation of teachers mentioning the word beyond the lesson script. Additionally, we allowed three different sets of indictors to covary as each set of words were often mentioned within the same utterance. For example, when the word *system* was mentioned, the word *function* was often used in the same phrase. The model fit for *Teacher Language Scaffolds* was adequate (RMSEA = .016 [CI=.00, .03], CFI = .977, TLI = .966, SRMR = .051). For *Domain-Specific Reading Comprehension* (Figure 6), all three indicators were

significantly correlated with the factor, with factor loadings above .60. There were no fit statistics for this factor given that the model was just identified.

Research Question 1: Effect of Structured Supplements on Comprehension Outcomes

Table 7 displays the HLM results for recall, near-transfer, mid-transfer, content comprehension (near-/mid-transfer combined) and EOG. Model estimates are standardized (M = 0, SD = 1) and indicate that there were significant treatment effects on recall ($\gamma_{01} = .17, SE = .07, p < .05$), near-transfer ($\gamma_{01} = .17, SE = .06, p < .01$), mid-transfer ($\gamma_{01} = .18, SE = .07, p < .05$), and content comprehension ($\gamma_{01} = .18, SE = .07, p < .05$) measures. Put differently, if control students' average was at the 50th percentile, treatment students performed at the 57th percentile (calculation from: Hippel, 2023). Thus, there is evidence that structured read aloud supplements positively impacted domain-specific reading comprehension[1]. The intervention, however, did not improve student outcomes on the domain-general EOG reading comprehension measure ($\gamma_{01} = .01, SE = .02, p = .65$).

We tested for false discoveries following the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) with a false discovery rate set to .05 and confirmed the significant treatment effects on recall, near-, and mid-transfer passages (see Appendix S-A.5).

Research Question 2: Teacher Language Scaffolds as a Mediator for Comprehension Outcomes

We used SEM to determine whether and to what extent teacher language scaffolds mediated the treatment effect. Figure 7 shows the specified model where we included freely

[1] Effect sizes based on standardized outcome variables are conservative because they do not adjust for measurement error. The following calculation, $\frac{ES}{\sqrt{\alpha}}$ gives the unbiased estimate (Gilbert, 2023; Hedges, 1981). Recall ES = .21; near-transfer ES = .19; mid-transfer ES = .21; content comprehension ES = .19.

estimated paths between random assignment to structured read aloud supplements and students' domain-specific reading comprehension outcomes as well as the latent variable teacher language scaffolds. The model fit to the data was adequate (RMSEA = .025 [CI=.02, .03], CFI = .918, TLI = .906, SRMR = .054). Results indicated that the language teachers used that exceeded the language included in the intervention script explained 66% of the treatment effect on domain-specific reading comprehension with a significant total indirect effect ($b = .230, SE = .10, p < .05, \beta = .114$) and total effect ($b = .351, SE = .15, p < .05, \beta = .174$). Furthermore, random assignment to read aloud structured supplements significantly predicted an increase in teacher language scaffolds ($b = 1.196, SE = .29, p < .001, \beta = 1.026$), which in turn significantly predicted domain-specific reading comprehension ($b = .192, SE = .08, p < .01, \beta = .111$).

Sensitivity Analyses

We conducted sensitivity analyses to confirm the robustness of our findings. For research question 1, we aggregated our data at the classroom level and fit a regression model with school fixed effects using the same covariates. Results were slightly larger, with effect sizes ranging from .21 to .25 (Table 8). We also used SEM with clustered standard errors at the classroom level to estimate the treatment effect. Consistent with the HLMs, SEM revealed a significant treatment effect of .17, identical to the main findings (see Appendix S-A.6).

As a sensitivity check for determining the extent to which teacher language scaffolds mediated the treatment effect, we employed the bootstrap method (1,000 draws) (MacKinnon et al., 2004; Preacher & Hayes, 2008). Results were stable. We ran two additional sensitivity checks to account for the fact that the duration of the treatment lesson was longer than the control lesson. To ensure teacher language scaffolds were not a proxy for additional time, we fit a

mediation model with time as the mediator and found no evidence of an indirect effect. A second sensitivity check included time as an additional covariate for the model shown in Figure 7. Time was not a significant predictor of domain-specific reading comprehension and teacher language scaffolds continued to mediate the treatment effect.

Discussion

Using a within-school cluster randomized controlled design, we examined the causal effects of structured supplements for teachers in a read aloud lesson, including (a) concise definitions of vocabulary words with meaningful examples and a visual representation of a word network, (b) explicit connections between known and new schemas, and (c) discussion questions to get students talking about topics related to topic schemas using vocabulary. While all students had participated in three weeks of the MORE science intervention and had instantiated a robust schema on the human body systems two months prior, they had not yet studied the Apollo 11 mission to the moon. We purposely posed a question at the beginning of the treatment lesson to help students access the instantiated science schema (first structured supplement in Table 1) and also included directions for teachers to record student responses on an anchor chart. This activity likely provided students with the necessary time and prompting to retrieve the relevant background knowledge (McCarthy et al., 2018; O'Reilly et al., 2019; Kaefer, 2020). With a familiar topic schema activated, teachers then read aloud the Apollo 11 story focusing on a group of people working together to send astronauts to the moon. Thus, we leveraged a familiar topic schema about the human body with structured supplements in the treatment lesson, likely helping students map the science schema onto an unfamiliar social studies topic schema about a system of individuals trying to send astronauts to the moon (Anderson & Pearson, 1984; Gick &

Holyoak, 1983). Structured supplements in the treatment read aloud positively impacted student outcomes with effect sizes ranging between .17 and .18 (a seven percentage-point increase) on measures of domain-specific reading comprehension: recall, near-transfer, mid-transfer, and content comprehension. Findings from this study provide strong evidence that structured supplements are a critical lever of change that can enhance the benefits of read aloud lessons.

First, the causal findings from this study suggest that structured supplements can directly impact and improve students' domain specific reading comprehension. Importantly, the magnitude of the effect sizes was comparable across the recall, near-, and mid-transfer reading tasks. These findings are consistent with the idea that the structured supplements supported students' ability to construct a literal understanding of the texts and integrate this knowledge with a familiar schema to build a coherent situation model (Kintsch, 1998). In particular, treatment students outperformed control students on their ability to locate and recall key textual ideas (recall text), to integrate and interpret ideas across passages in both known topics (exploring the moon) and new topics (Making the Empire Statement Building). Although we cannot pinpoint precisely which components of the structured supplements were critical to supporting comprehension gains, a critical aim of the structured supplements was to increase the amount of academic vocabulary that teachers used in their classrooms to foster schema transfer—that is, helping students make connections between known topics like the human body and moon team system and new topics that appeared on the transfer task (i.e., Making the Empire State Building).

More specifically, providing an explicit 1:1 schema-mapping activity highlighted the connection between the human body systems working together and a system of individuals

working together (Gick & Holyoak, 1983). In this activity, students were asked about the similarities between both systems and tasked with using both science and social studies vocabulary to fill in missing words (Figure 4). We designed this activity to aid students in seeing the links between both topic schemas so that they might be able to leverage these similarities when reading passages on topics that vary in context (Barnett & Ceci, 2002) – i.e., the creation of the lunar module and the construction of the Empire state building. Our findings confirm that "expertise can sometimes be transferred to novel tasks within and beyond the initial domain" (Kimball & Holyoak, 2000, p. 119), which support similar results from a different implementation of the MORE intervention (Gilbert et al., 2023).

Second, the mediation results suggest that teacher language scaffolds can function as temporary dialogic supports that go above and beyond the intervention script and support students' reading comprehension. In our theory of change (Figure 1), we hypothesized that teachers providing additional scaffolds for their students – exposures to hearing key words, explanations of word meanings, and questions, all of which exceeded what was included in the lesson script – would mediate the treatment effect. We found that on average, treatment teachers used the seven target words in their instruction 46 times beyond the number of times those words appeared in the lesson script, compared to the control teachers' 13 times. Similarly, treatment teachers asked around 80 additional questions compared to 52 additional questions asked by control teachers. Structured supplements likely resulted in treatment teachers providing language extensions for the seven target words essential to instantiating the social studies schema, while control teachers who did provide language extensions did so for a myriad of words such that there were no unifying explanations linking words and concepts together. Findings suggest that

providing structured supplements for teachers during a read aloud can serve as a catalyst for expanded opportunities for word learning and schema instantiation. Providing teachers with clear word explanations and probing questions may have helped direct their attention to key concepts, thus providing students with meaningful scaffolds as students work within their ZPD on difficult cognitive tasks (Vygotsky, 1978). Furthermore, the sheer quantity of questions, language extensions, and utterances of key domain-specific vocabulary words provided students with opportunities to work towards mastering the form and meanings of words in varied contexts (Perfetti, 2007). Thus, by embedding structured supplements within the lesson, we created a baseline level of instruction so that treatment teachers increased the questions posed to students (interactions) as well as the exposures and explanations of words (inputs) (Snow, 2014).

Finally, inspection of treatment and control lesson transcripts further suggests that procedural fidelity is necessary but not sufficient to increase the quantity of academic vocabulary teachers use orally during content area instruction. Importantly, both teachers in the treatment and control groups largely adhered to the lesson scripts, reaching high levels of procedural fidelity (above 80%, as noted in Table 2). However, the scripted supplements facilitated teachers' use of language scaffolds – i.e., temporary dialogic actions – that went above and beyond the lesson scripts. The Appendix (S-A.6) provides annotations of selected transcripts from treatment and control groups. For example, control teachers typically read their scripts, rarely stopped to provide child-friendly definitions of the word meanings, or to discuss the target academic vocabulary connected to the topic schema. In contrast, treatment group teachers were guided by the supplements to define academic words, use them more frequently, and foster discussion about word meanings as they related to the social studies topics. The positive correlations

between the frequency with which teachers use academic vocabulary and ask questions (see Table 6) further supports the patterns in the lesson transcripts in which treatment teachers talked about target vocabulary to foster schema transfer by helping students make connections between known science topics (i.e., human body) and a novel social studies topics (i.e., Apollo 11 mission). In essence, treatment group teachers provided more opportunities for students both to hear and use academic vocabulary by engaging in discussions to make connections between known and new topics. Having discussions about words and asking students to provide word meanings are dialogic actions rarely observed in elementary classrooms (Gamse et al., 2008; Wright & Neuman, 2014) but are crucial determinants of vocabulary learning and reading comprehension (Carlisle et al., 2013; Snow, 2014). Furthermore, the number of academic words used orally by teachers in classrooms may be a simple and useful indicator of language input that predicts stronger student literacy outcomes.

Limitations and Future Directions

While the current findings are promising, several limitations should be addressed in future research.

First, our measure of teacher language scaffolds captured the amount of academic vocabulary used by teachers but not the quality of academic language more generally. As noted by Snow (2014), academic language is not a separate register, but "a set of features that can be used to greater or lesser degree, defining a continuum from highly academic to minimally academic language" (p. 120). Our narrow measure of academic vocabulary used by teachers is just one feature of academic language, albeit a critical one, and thus future research should aim

to capture the quality of academic language more generally within classrooms (Snow & Uccelli, 2008; Uccelli et al., 2015).

Second, our mediation analysis could be expanded to include other child-level mechanisms that underlie the treatment effects. In other words, we focused on teacher language scaffolds as a key mediator, but other research on content literacy interventions has emphasized the role of children's reading engagement (Guthrie, Wigfield, & You, 2012) and children's vocabulary and oral comprehension (Connor et al., 2017). We focused primarily on teacher language scaffolds given limited attention to this construct in existing Tier I and Tier 2 language focused interventions (LARCC et al., 2022; Connor et al., 2018), and the need to identify malleable instructional levers of change to build a more robust and useful science of teaching reading and vocabulary. However, future research should include multilevel SEM models that include both child- and teacher-level mediators.

Third, the challenge of identifying schemas is both an art and science. Thus, there is a clear need to replicate and extend this study design to other domains. The topic schemas of the life science (human body systems working together to keep the body alive) and the social studies (system of individuals working together to send astronauts to the moon) lessons were both related to the general schema of *systems*. To be useful to young children, schemas must be "linked to abstract, generalizable features of situations" (Kintsch, 2009, p. 231), and thus neither too specific nor too general. Extending structured supplements to other topics and read aloud lessons in social science domains like economics or civics would help determine the generalizability of the current study findings.

Fourth, the transfer assessments were designed to be given the day after the read aloud lesson. We instructed teachers to teach the lesson and have students complete all assessment activities within a 10-day window. Due to time constraints at schools, many teachers administered the assessments multiple days after the lesson and sometimes spread across multiple days, thereby explaining incomplete assessment data for certain students. For example, some students may have taken the recall and near-transfer assessments but were absent when the mid-transfer assessment was administered. In other instances, some students did not complete the assessment, leaving it unclear if time constraints or test exhaustion were factors.

Finally, future research should explore the impact of this type of lesson over the span of a unit. Does embedding structured supplements in multiple read aloud lessons positively impact comprehension outcomes to a greater extent than the effects detected in the present study? And how far do those effects travel along a continuum of transfer? We did not include a far-transfer passage due to time constraints, and including one without any of the taught words would provide a more stringent test of whether the schema alone or in combination with the vocabulary networks supports text comprehension.

Conclusion

In conclusion, findings from this study provide causal evidence for including structured supplements for teachers in read aloud lessons. Structured supplements emphasize domain-specific vocabulary instruction, questioning, and include activities to promote schema instantiation and highlight similarities between topic schemas. Interleaving these dialogic actions cultivate the kind of language-rich classroom environment that supports children's literacy development. As noted by Perfetti (2007), a key source of individual differences in children's

reading comprehension ability originates in children's "literacy and language experiences" and "engagement with concepts and their language forms" (p. 380), all of which are critical determinants of reading comprehension. Importantly, the structured supplements used in the present study are curriculum agnostic. Thus, they could be easily incorporated into a range of classroom curricula and instructional practices, particularly content literacy programs that aim to improve content knowledge, vocabulary, and comprehension in elementary grades (Cabell & Hwang, 2020). As the science of reading continues to gain traction within public discourse, it is important to provide teachers with tools, strategies, and usable knowledge that helps them integrate principles of the science of teaching reading and vocabulary into their daily instructional practices. This study provides a strong proof of concept that read alouds, when enhanced with structured supplements to foster schema transfer, can increase the amount of academic vocabulary that teachers use during classroom instruction and help their students read disciplinary texts with greater understanding.

Acknowledgements
We thank the teachers and students for their participation in this study. We also thank Joshua B. Gilbert, Ethan Scherer, Mary A. Burkhauser, Jackie Relyea, and Johanna Tvedt for their insights and expertise as well as Taylor Harrison for her transcriptions of the audio recordings and assistance coding the transcripts.

Data availability statement.
The data and analysis code to replicate all results found in this study are available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/LUKGIA

Contact
Correspondence concerning this article should be addressed to Douglas M. Mosher, Graduate School of Education, Harvard University, Room Q-409, READS Lab, 50 Church Street, Cambridge, MA 02138, United States. E-mail: douglasmosher@gmail.com

References

Ahmed, Y., Francis, D. J., York, M., Fletcher, J. M., Barnes, M., & Kulesz, P. (2016). Validation of the direct and inferential mediation (DIME) model of reading comprehension in grades 7 through 12. *Contemporary Educational Psychology*, *44–45*, 68–82. https://doi.org/10.1016/j.cedpsych.2016.02.002

Al-Adeimi, S., & O'Connor, C. (2021). Exploring the relationship between dialogic teacher talk and students' persuasive writing. *Learning and Instruction*, *71*, 101388. https://doi.org/10.1016/j.learninstruc.2020.101388

Anderson, B. E., Wright, T. S., & Gotwals, A. W. (2023). Teachers' vocabulary talk in early-elementary science instruction. *Journal of Literacy Research*, *55*(1), 75–100. https://doi.org/10.1177/1086296X231163117

Anderson, R. (2013). Role of the reader's schema in comprehension, learning, and memory. In D.E. Alvermann, N.J. Unrau, & R.B. Ruddell (Eds.), *Theoretical Models and processes of reading* (6th ed., pp. 476-488). Newark, DE: International Reading Association.

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

Anderson, R. C., & Pearson, P. D. (1984). A schema-theoretic view of basic processing in reading. In P. D. Pearson (Ed.), *Handbook of Reading Research* (pp. 255–291). New York: Longman.

Ard, L. M., & Beverly, B. L. (2004). Preschool word learning during joint book reading: Effect of adult questions and comments. *Communication Disorders Quarterly*, *26*(1), 17-28,56-58. http://dx.doi.org/10.1177/15257401040260010101

Bakhtin M. M. (1981). *The dialogic imagination: Four essays* (M. Holquist, Ed.; C. Emerson & M. Holquist, Trans.). University of Texas Press.

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

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

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

Carlisle, J. F., Kelcey, B., & Berebitsky, D. (2013). Teachers' support of students' vocabulary learning during literacy instruction in high poverty elementary schools. *American Educational Research Journal*, *50*(6), 1360–1391. https://doi.org/10.3102/0002831213492844

Collins, M. F. (2016). Supporting inferential thinking in preschoolers: Effects of discussion on children's story comprehension. *Early Education and Development*, *27*(7), 932–956. https://doi.org/10.1080/10409289.2016.1170523

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

Connor, C. M., Phillips, B. M., Kim, Y. G., Lonigan, C. J., Kaschak, M. P., Crowe, E., Dombek, J., & Al Otaiba, S. (2018). Examining the efficacy of targeted component interventions on language and literacy for third and fourth graders who are at risk of comprehension difficulties. *Scientific Studies of Reading*, *22*(6), 462–484. https://doi.org/10.1080/10888438.2018.1481409

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

Dedoose Version 7.0.23, web application for managing, analyzing, and presenting qualitative and mixed method research data (2016). Los Angeles, CA: SocioCultural Research Consultants, LLC www.dedoose.com

Douglas, K. M., & Albro, E. R. (2014). The progress and promise of the reading for understanding research initiative. *Educational Psychology Review*, *26*(3), 341–355. https://doi.org/10.1007/s10648-014-9278-y

Duke, N. K., Halvorsen, A.-L., Strachan, S. L., Kim, J., & Konstantopoulos, S. (2021). Putting PjBL to the test: The impact of project-based learning on second graders' social studies

and literacy learning and motivation in low-ses school settings. *American Educational Research Journal*, *58*(1), 160–200. https://doi.org/10.3102/0002831220929638

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*(2), 211–245. https://doi.org/10.1037/0033-295X.102.2.211

Ezell, H., & Justice, L. (2005). *Shared storybook reading: Building young children's language & emergent literacy skills*. Baltimore: Paul H. Brookes.

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

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

Gamse, B. C., Jacob, R. T., Horst, M., Boulay, B., & Unlu, F. (2008). Reading first impact study. Final report. Executive summary. Ncee 2009-4039. In *National Center for Education Evaluation and Regional Assistance*. National Center for Education Evaluation and Regional Assistance. https://eric.ed.gov/?id=ED503345

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

Gilbert, J. B. (2023). How measurement affects causal inference: Attenuation bias is (usually) more important than scoring weights. *Edworkingpapers.com*.

Gilbert, J. B., Kim, J. S., & Miratrix, L. W. (2023). Modeling Item-Level Heterogeneous Treatment Effects With the Explanatory Item Response Model: Leveraging Large-Scale Online Assessments to Pinpoint the Impact of Educational Interventions. *Journal of Educational and Behavioral Statistics*, *48*(6), 889–913. https://doi.org/10.3102/10769986231171710

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

Guthrie, J. T., Wigfield, A., & You, W. (2012). Instructional Contexts for Engagement and Achievement in Reading. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of Research on Student Engagement* (pp. 601–634). Springer US. https://doi.org/10.1007/978-1-4614-2018-7_29

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *journal of Educational Statistics*, *6*(2), 107-128.

Hippel, P. T. von. (2023). Multiply by 37: A surprisingly accurate rule of thumb for converting effect sizes from standard deviations to percentile points. In *EdWorkingPapers.com*. Annenberg Institute at Brown University. https://www.edworkingpapers.com/ai23-829

Hirsch, E. D. (1988). *Cultural literacy : What every American needs to know*. Houghton Mifflin.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hwang, H., Cabell, S. Q., & Joyner, R. E. (2021). Effects of integrated literacy and content-area instruction on vocabulary and comprehension in the elementary years: A meta-analysis. *Scientific Studies of Reading*, 1–27. https://doi.org/10.1080/10888438.2021.1954005

Kaefer, T. (2020). When did you learn it? How background knowledge impacts attention and comprehension in read-aloud activities. *Reading Research Quarterly*, *55*(S1), S173–S183. https://doi.org/10.1002/rrq.344

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*(6018), 772–775. https://doi.org/10.1126/science.1199327

Kendeou, P., & O'Brien, E. J. (2016). Prior knowledge, acquisition and revision. In P. Afflerbach (Ed.), *Handbook of individual differences in reading: reader, text, and context* (pp. 151-163). Routledge/Taylor & Francis.

Kim, J. S., Burkhauser, M. A., Mesite, L. M., Asher, C. A., Relyea, J. E., Fitzgerald, J., & Elmore, J. (2020). Improving reading comprehension, science domain knowledge, and reading engagement through a first-grade content literacy intervention. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000465

Kim, J. S., Burkhauser, M. A., Relyea, J. E., Gilbert, J. B., Scherer, E., Fitzgerald, J., Mosher, D., & McIntyre, J. (2022). A longitudinal randomized trial of a sustained content literacy intervention from first to second grade: Transfer effects on students' reading comprehension. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000751

Kim, J. S., Relyea, J. E., Burkhauser, M. A., Scherer, E., & Rich, P. (2021). Improving elementary grade students' science and social studies vocabulary knowledge depth,

reading comprehension, and argumentative writing: A conceptual replication. *Educational Psychology Review*. https://doi.org/10.1007/s10648-021-09609-6

Kim, J., Burey, J., Hwang, H., McMaster, K., & Kendeou, P. (2023). Supporting inference-making during COVID-19 through individualized scaffolding and feedback: A natural experiment. *Reading and Writing*, *36*(2), 467–490. https://doi.org/10.1007/s11145-022-10391-2

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

Kintsch, W. (1993). Information accretion and reduction in text processing: Inferences. *Discourse Processes*, *16*(1–2), 193–202. https://doi.org/10.1080/01638539309544837

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York, NY: Cambridge University Press.

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

Kintsch, W. (2013). The construction-integration model of text comprehension and its implications for instruction. In D.E. Alvermann, N.J. Unrau, & R.B. Ruddell (Eds.), *Theoretical Models and processes of reading* (6th ed., pp. 807-839). Newark, DE: International Reading Association.

Kintsch, W., & Kintsch, E. (2005). Comprehension. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 71–92). Routledge.

Kline, R. B. (2016). *Principles and practice of structural equation modeling, fourth edition: Vol. Fourth edition*. The Guilford Press. http://ezp-prod1.hul.harvard.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1078917&site=ehost-live&scope=site

Language and Reading Research Consortium. (2016). Use of the curriculum research framework (CRF) for developing a reading-comprehension curricular supplement for the primary grades. *The Elementary School Journal*, *116*(3), 459–486. https://doi.org/10.1086/684827

Language and Reading Research Consortium, Lo, M.-T., & Xu, M. (2022). Impacts of the let's know! Curriculum on the language and comprehension-related skills of prekindergarten and kindergarten children. *Journal of Educational Psychology*, *114*(6), 1205–1224. https://doi.org/10.1037/edu0000744

Language and Reading Research Consortium, Pratt, A., & Logan, J. (2014). Improving language-focused comprehension Instruction in primary-grade classrooms: Impacts of the Let's Know! Experimental curriculum. *Educational Psychology Review*, *26*(3), 357–377. https://doi.org/10.1007/s10648-014-9275-1

Lin-Siegler, X., Ahn, J. N., Chen, J., Fang, F.-F. A., & Luna-Lucero, M. (2016). Even Einstein struggled: Effects of learning about great scientists' struggles on high school students' motivation to learn science. *Journal of Educational Psychology*, *108*(3), 314–328. https://doi.org/10.1037/edu0000092

Little, R.J.A. (1988) A Test of Missing Completely at Random for Multivariate Data with Missing Values. Journal of the American Statistical Association, 83, 1198-1202. http://dx.doi.org/10.1080/01621459.1988.10478722

MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, *39*(1), 99–128. https://doi.org/10.1207/s15327906mbr3901_4

McCarthy, K. S., Guerrero, T. A., Kent, K. M., Allen, L. K., McNamara, D. S., Chao, S.-F., Steinberg, J., O'Reilly, T., & Sabatini, J. (2018). Comprehension in a scenario-based assessment: Domain and topic-specific background knowledge. *Discourse Processes*, *55*(5–6), 510–524. https://doi.org/10.1080/0163853X.2018.1460159

McKeown, M. G., Deane, P. D., Scott, J. A., Krovetz, R., & Lawless, R. R. (2017). *Vocabulary Assessment to Support Instruction: Building Rich Word-Learning Experiences*. Guilford Publications. http://ebookcentral.proquest.com/lib/harvard-ebooks/detail.action?docID=4880718

McMaster, K. L., den Broek, P. van, A. Espin, C., White, M. J., Rapp, D. N., Kendeou, P., Bohn-Gettler, C. M., & Carlson, S. (2012). Making the right connections: Differential effects of reading intervention for subgroups of comprehenders. *Learning and Individual Differences*, *22*(1), 100–111. https://doi.org/10.1016/j.lindif.2011.11.017

Mosher, D. M. (2022). Embedding scripted teacher language into a read aloud lesson and its effect on measures of comprehension transfer: A randomized controlled trial. *Registry of Efficacy and Effectiveness Studies*. https://sreereg.icpsr.umich.edu/sreereg/subEntry/13800/pdf?action=view

Mosher, D. M., Burkhauser, M. A., & Kim, J. S. (2024). Improving second-grade reading comprehension through a sustained content literacy intervention: A mixed-methods study

examining the mediating role of domain-specific vocabulary. *Journal of Educational Psychology*, *116*(4), 550–568. https://doi.org/10.1037/edu0000868

Muthén, L. K., & Muthén, B. O. (2012). Mplus User's Guide (7th ed.). Los Angeles, CA: Muthén & Muthén.

Neugebauer, S., Coyne, M., McCoach, B., & Ware, S. (2017). Teaching beyond the intervention: The contribution of teacher language extensions to vocabulary learning in urban kindergarten classrooms. *Reading and Writing*, *30*(3), 543–567. https://doi.org/10.1007/s11145-016-9689-x

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

Northwest Evaluation Association. (2011). *Technical manual for measures of academic progress and measures of academic progress for primary grades*. Lake Oswego, OR: Author.

Novak, J. D. (1990). Concept mapping: A useful tool for science education. *Journal of Research in Science Teaching*, *27*(10), 937–949. https://doi.org/10.1002/tea.3660271003

O'Reilly, T., Wang, Z., & Sabatini, J. (2019). How much knowledge is too little? When a lack of knowledge becomes a barrier to comprehension. *Psychological Science*, *30*(9), 1344–1351. https://doi.org/10.1177/0956797619862276

Pearson, P. D., Palinscar, A., Biancarosa, G., & Berman, A. (Eds.). (2020). *Reaping the rewards of the reading for understanding initiative*. National Academy of Education. https://doi.org/10.31094/2020/2

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

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

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*(3), 879–891. https://doi.org/10.3758/BRM.40.3.879

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

Read, J. (2004). *Plumbing the depths: How should the construct of vocabulary knowledge be defined* (P. Bogaards & B. Laufer-Dvorkin, Eds.; pp. 209–227). John Benjamins Publishing Company.

Relyea, J. E., Zhang, J., Wong, S. S., Samuelson, C., & Wui, Ma. G. L. (2022). Academic vocabulary instruction and socio-scientific issue discussion in urban sixth-grade science classrooms. *The Journal of Educational Research*, *115*(1), 37–50. https://doi.org/10.1080/00220671.2021.2022584

Shanahan, T. (2020). What constitutes a science of reading instruction? *Reading Research Quarterly*, *55*(S1), S235–S247. https://doi.org/10.1002/rrq.349

Snow, C. E. (2014). Input to interaction to instruction: Three key shifts in the history of child language research. *Journal of Child Language*, *41*(S1), 117–123. https://doi.org/10.1017/S0305000914000294

Snow, C. E., & Uccelli, P. (2008). The challenge of academic language. In D. Olson & N. Torrance (Eds.), The Cambridge handbook of literacy (pp. 112-133). New York: Cambridge University Press.

Stahl, S. A., & Nagy, W. E. (2006). *Teaching word meanings*. L. Erlbaum Associates.

StataCorp. 2021. *Stata Statistical Software: Release 17*. College Station, TX: StataCorp LLC.

Swanson, E., Vaughn, S., Wanzek, J., Petscher, Y., Heckert, J., Cavanaugh, C., Kraft, G., & Tackett, K. (2011). A synthesis of read-aloud interventions on early reading outcomes among preschool through third graders at risk for reading difficulties. *Journal of Learning Disabilities*, *44*(3), 258–275. https://doi.org/10.1177/0022219410378444

Uccelli, P., Galloway, E. P., Barr, C. D., Meneses, A., & Dobbs, C. L. (2015). Beyond vocabulary: Exploring cross-disciplinary academic-language proficiency and its association with reading comprehension. *Reading Research Quarterly*, *50*(3), 337–356. https://doi.org/10.1002/rrq.104

U.S. Department of Education, Institute of Education Sciences, National center for Education Statistics, National Assessment of Educational Progress (NAEP), 2019 Reading Assessments & Achievement Levels.

van Kleeck, A., Vander Woude, J., & Hammett, L. (2006). Fostering literal and inferential language skills in head start preschoolers with language impairment using scripted book-sharing discussions. *American Journal of Speech-Language Pathology*, *15*(1), 85–95. https://doi.org/10.1044/1058-0360(2006/009)

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

Vygotsky, L. S. (1986). Thought and language. Cambridge, MA: MIT Press.

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

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, *17*(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Wright, T. S., & Neuman, S. B. (2014). Paucity and disparity in kindergarten oral vocabulary instruction. *Journal of Literacy Research*, *46*(3), 330–357. https://doi.org/10.1177/1086296X14551474

Figure 1. Study Theory of Change

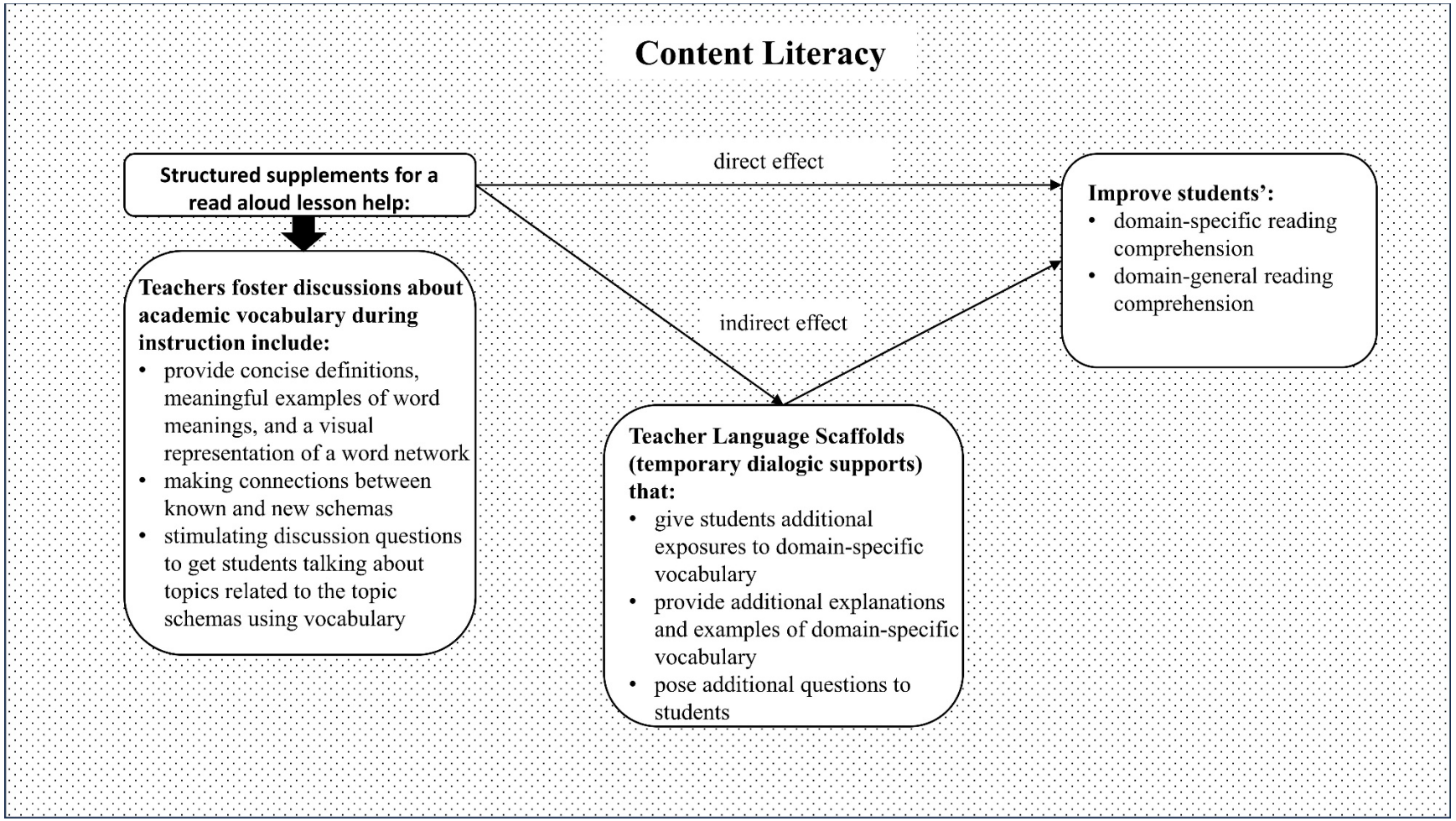


Figure 2. Intervention General and Topic Schemas

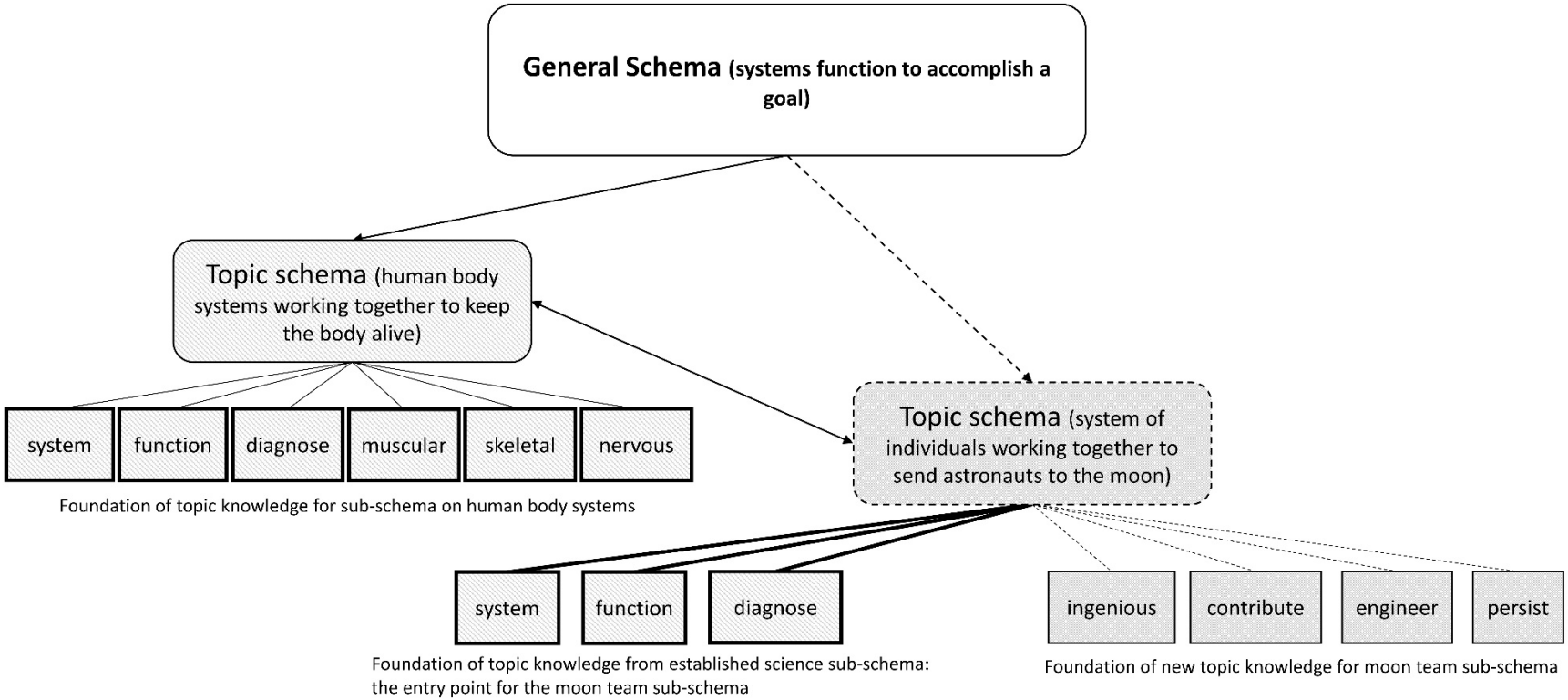


Figure 3. Consort Diagram for Sample Flow (Teachers & Students)

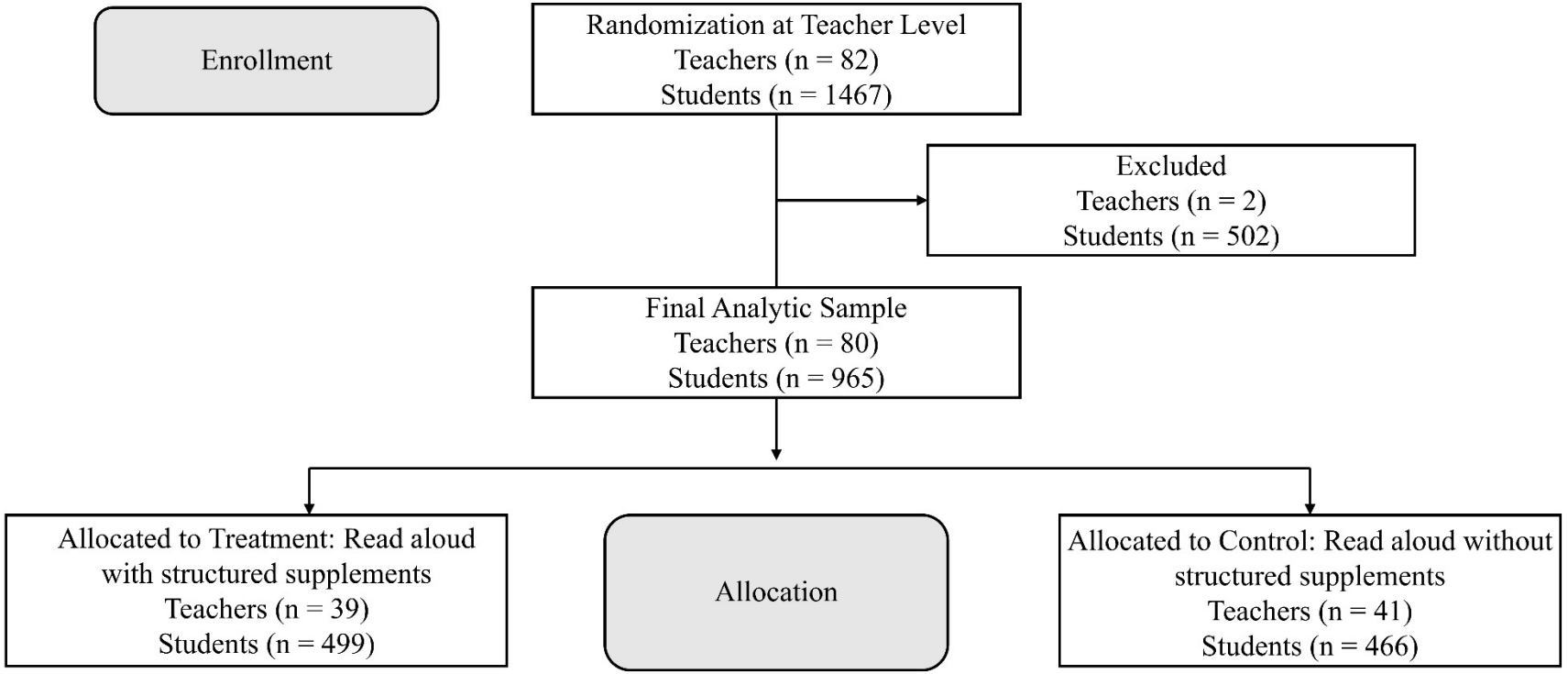


Figure 4: Schema Mapping Activity Using "Madlibs"

function muscles system nervous skeleton contribute

Human body

My human body is a _system_. The human body system has many parts. These parts function together to make the body work to keep us alive and healthy. One part of the human body system is your _muscles_. Your heart muscle has an important _function_. It moves on its own with each heartbeat. Another part of your body system is your _skeleton_. Your bones' function is to help support your body! Finally, the human body system needs the _nervous_ system to tell your body what to do. All these different parts of your human body system _contribute_ to helping your body function.

Moon team ***astronauts ingenious system function engineer contributed***

The moon team was a _system_. The moon team system had many people. These people functioned together to send astronauts to the moon and back. One part of the moon team system was an _ingenious_ math expert like Katherine Johnson. She had an important _function_. She figured out the path to the moon and back. Another part of the moon team system was an ingenious _engineer_ like Chuck Lowry. The engineer's function was to figure out how to slow down Apollo 11 when it fell to Earth! Finally, the moon team system needed _astronauts_ like Neil Armstrong to fly to the moon. All these different people of the moon team system _contributed_ to helping the moon team make it to the moon and back.

How is the moon team system similar to the human body system? (Pick the word that completes the sentence) (1 min)

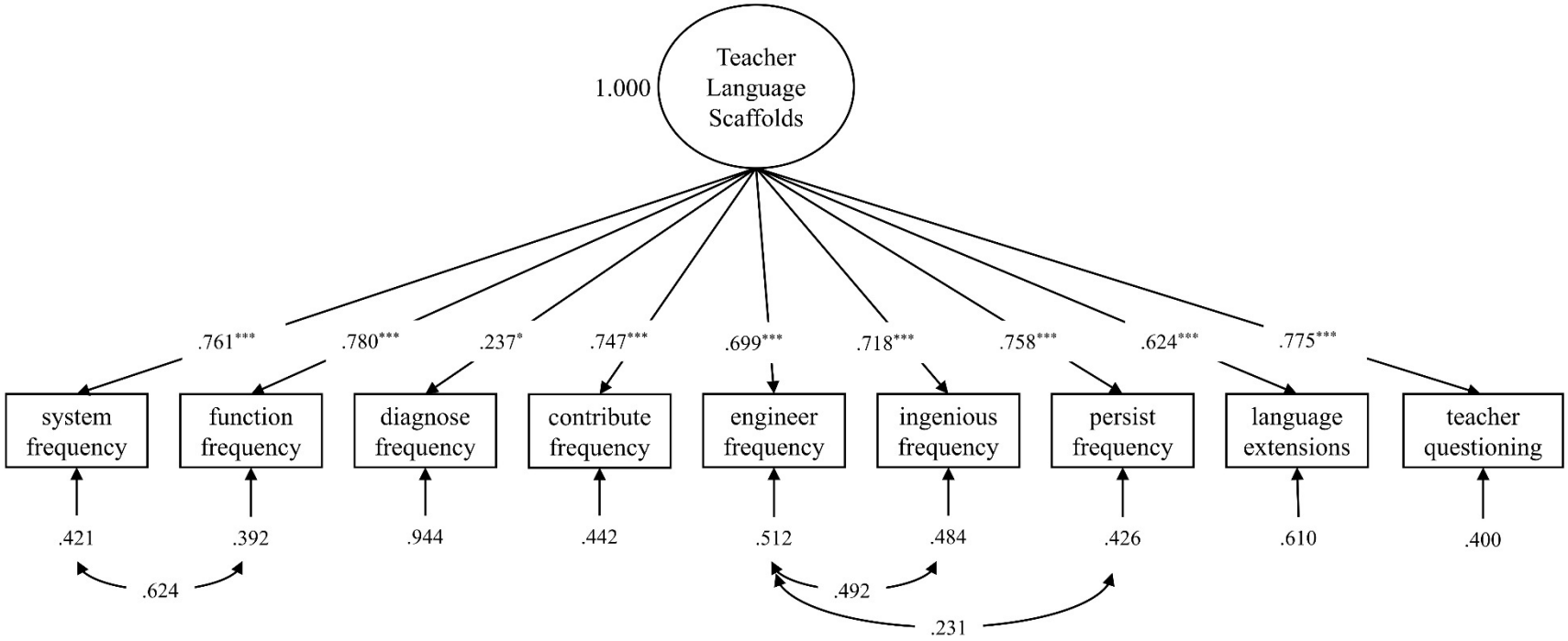




persist engineer contribute ingenious

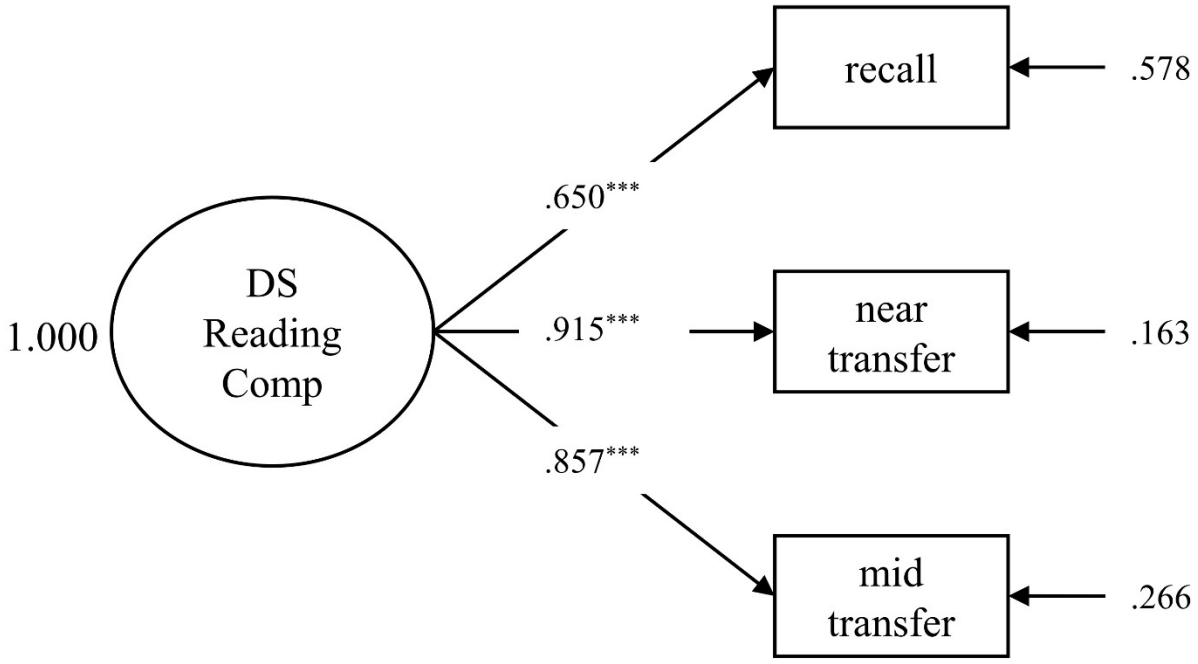
Both the human body system and the moon team system have many parts that _contribute_ to helping it function.

Figure 5: Teacher Language Scaffolds CFA



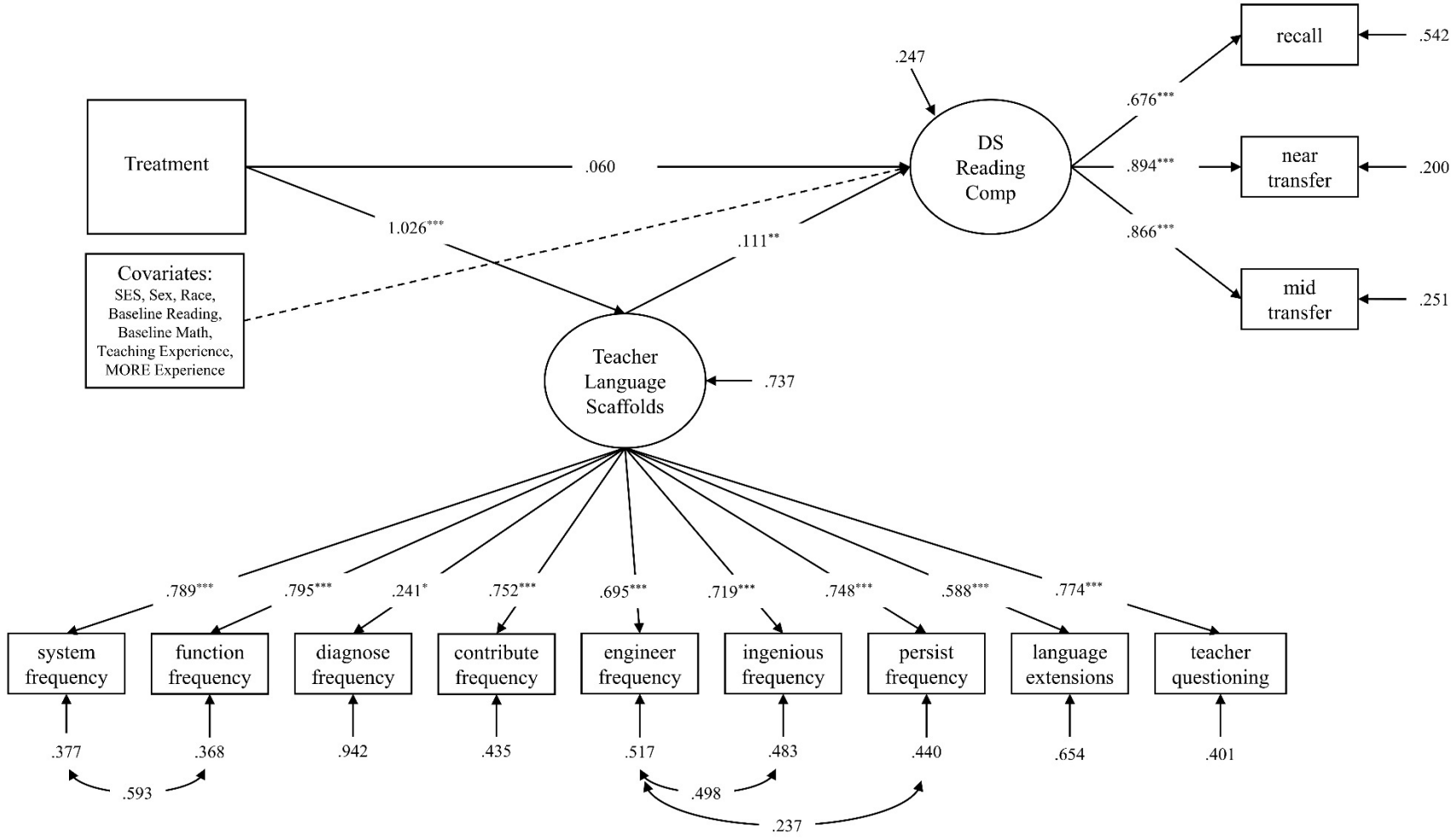
Note. All coefficients are standardized. Significant coefficients are starred and *p* values taken from unstandardized values. *** $p < .001$. * $p < .05$.

Figure 6: Domain-Specific Reading Comprehension CFA



Note. All coefficients are standardized. Significant coefficients are starred and *p* values taken from unstandardized values. *** $p < .001$.

Figure 7: SEM Exploring the Mediation Role of Teacher Language Scaffolds on Students' Domain-Specific Reading Comprehension



* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 1. Structured Supplements for a Read Aloud Lesson

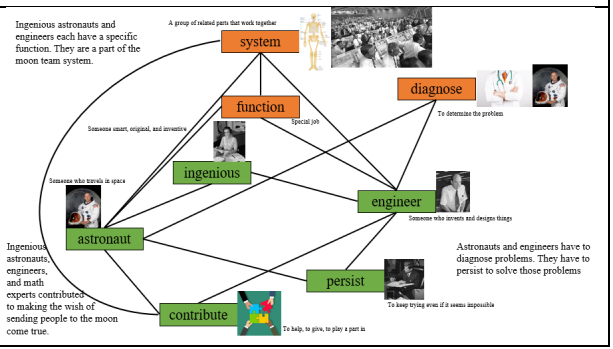
Structured Supplements	**Example**
Activate existing schema	Class completes an anchor chart reviewing the three human body systems and their functions ***Human Body Systems*** **Part** | **Function** muscular system skeletal system nervous system
Introduce new domain-specific vocabulary with explanation of word's meaning and examples	*Let's say the word ingenious together.* *Ingenious.* *An ingenious person is someone who is original or inventive and smart. Before there were cars, people used to have to walk or ride horses everywhere. But ingenious inventors and engineers created cars and buses, making it so much easier for us to travel to different places – like from home to school. These people were ingenious because they were inventive and smart! They created something new.*
Review topic/vocabulary knowledge	Partners answer questions with specific inclusion criteria for what students answers should contain (i.e., target words) ***How was Chuck Lowry an ingenious engineer? (1 min)*** **Must Have**: Answer includes the word ingenious or engineer. **Amazing**: Answer also includes ingenious, engineer, contribute, function, or system.
Map new topic schema onto different and already established schema	Partner discussion and "madlibs" schema mapping activity.	See Figure 4
Concept mapping activity connecting words/concepts	Concept mapping activity	

Table 2. Fidelity of Implementation

Variable	Overall *M* (*SD*)	Treatment *M* (*SD*)	Control *M* (*SD*)	Difference
Lesson Adherence (0-8)	6.76 (1.63)	6.49 (1.89)	7.03 (1.28)	-.46(.33)
Lesson Adherence (%)	84.5 (20.33)	81.08 (23.68)	87.83 (16.05)	-5.81(4.1)
Lesson Time	34.77 (14.42)	44.54 (12.42)	25.26 (8.80)	19.18***(2.2)

Note. M = Mean; *SD* = Standard Deviation. The difference was estimated from a regression model that controls for fixed effects of randomization blocks (schools). Standard errors are in parentheses in the difference column.

$^{***}p < .001$

Table 3. Descriptive Statistics for the Final Grade 3 Sample

Characteristics	*N*	Overall	Treatment	Control
White	965	25%	24%	27%
Black	965	32%	34%	29%
Hispanic	965	31%	29%	33%
Asian	965	9%	9%	9%
Other	965	3%	4%	2%
Male	965	48%	48%	48%
Multilingual	965	25%	23%	26%
High SES	965	29%	27%	30%
Med SES	965	35%	35%	36%
Low SES	965	35%	37%	34%
Baseline MAP Reading, *M* (*SD*)	959	192.32 (18.34)	192.67 (18.33)	191.95 (18.36)
Baseline MAP Math, *M* (*SD*)	955	194.18 (14.87)	194.45 (14.92)	193.9 (14.84)
Recall, *M* (*SD*)	942	3.46 (1.46)	3.58 (1.43)	3.32 (1.47)
Near Transfer, *M* (*SD*)	934	8.28 (3.5)	8.63 (3.43)	7.89 (3.54)
Mid Transfer, *M* (*SD*)	910	6.87 (3.21)	7.14 (3.15)	6.57 (3.25)
Content Comprehension, *M* (*SD*)	965	14.49 (6.83)	15.3 (6.59)	13.63 (6.97)
EOG, *M* (*SD*)	964	538.29 (10.11)	538.46 (10.15)	538.12 (10.07)

Note. MAP = Measure of Academic Progress. EOG = End-of-Grade assessment. SES = socioeconomic status. Other = students who identify as multicultural or Native American. *M* = mean; *SD* = Standard Deviation. Treatment students = 499; Control students = 466.

Table 4. Correlation Matrix of Pre- and Post-Assessment Measures

Variable	1	2	3	4	5	6
1. MAP Reading Pre-test	--					
2. MAP Math Pre-test	0.80	--				
3. Recall	0.58	0.56	--			
4. Near Transfer	0.73	0.68	0.57	--		
5. Mid Transfer	0.69	0.64	0.55	0.76	--	
6. Content Comprehension	0.68	0.63	0.54	0.94	0.93	--
7. EOG	0.86	0.76	0.58	0.73	0.69	0.69

Note. All correlations statistically significant at the .05 level. MAP = Measure of Academic Progress; EOG = End-of-Grade assessment.

Table 5. Descriptive Statistics for Teacher-Level Variables and Teacher Language Scaffolds

Variable	N	Overall *M* (*SD*)	Treatment *M* (*SD*)	Control *M* (*SD*)	Raw *M* Difference	Cohens *d*
System Frequency	75	13.76 (14.18)	21.59 (15.6)	6.13 (6.66)	15.46	1.3
Function Frequency	75	4.75 (5.61)	7.38 (6.66)	2.18 (2.47)	5.20	1.04
Diagnose Frequency	75	.27 (.72)	.27 (.51)	.26 (.89)	.01	.01
Contribute Frequency	75	2.69 (4.41)	4.51 (5.41)	.92 (1.98)	3.59	.89
Engineer Frequency	75	1.79 (2.97)	2.73 (3.61)	.87 (1.8)	1.86	.65
Ingenious Frequency	75	4.57 (6.53)	7.38 (7.09)	1.84 (4.57)	5.54	.93
Persist Frequency	75	3.65 (4.74)	5.11 (5.66)	2.24 (3.09)	2.87	.63
Language Extensions	75	5.8 (6.49)	6.14 (6.38)	5.47 (6.67)	.67	.1
Teacher Questioning	75	65.65 (37.56)	80.05 (36.93)	51.63 (32.94)	28.42	.81
Years Teaching	80	10.48 (8.88)	9.70 (9.37)	11.21 (8.44)	-1.51	-.17
Past Experience with Intervention	80	.9 (.94)	.86 (.92)	.93 (.96)	-.07	-.07

Note. M = Mean; *SD* = Standard Deviation. Past Experience with intervention is on a 6-point scale; each point indicated the number of times a teacher has taught one of the intervention units. Treatment = 39 teachers; Control = 41 teachers

Table 6. Correlation Matrix for Teacher-Level Variables

Variable	1	2	3	4	5	6	7	8	9	10
1. System Frequency	--									
2. Function Frequency	0.83	--								
3. Diagnose Frequency	0.14	0.09	--							
4. Contribute Frequency	0.60	0.62	0.07	--						
5. Engineer Frequency	0.55	0.54	0.04	0.56	--					
6. Ingenious Frequency	0.53	0.55	0.06	0.60	0.71	--				
7. Persist Frequency	0.54	0.54	0.13	0.48	0.61	0.48	--			
8. Language Extensions	0.35	0.44	0.32	0.45	0.55	0.50	0.58	--		
9. Teacher Questioning	0.60	0.61	0.13	0.49	0.45	0.46	0.61	0.50	--	
10. Years Teaching	-0.22	-0.20	-0.16	-0.13	-0.11	-0.15	-0.08	-0.14	-0.09	--
11. Past Experience with Intervention	-0.08	0.04	-0.06	0.00	-0.20	-0.19	-0.16	-0.11	0.16	0.08

Note. All correlations above .3 are significant at the .05 level.

Table 7. Results of Multiple Imputation Hierarchical Linear Models Predicting Treatment Effects of Structured Supplements on Measures of Reading Comprehension

Variable	Recall	Near Transfer	Mid Transfer	Content Comprehension	EOG
			Fixed effects		
Intercept, γ_{00}	-.15 (.2)	-.32† (.18)	-.5* (.21)	-.42* (.2)	0 (.07)
Treatment Effect, γ_{01}	.17*(.07)	.17**(.06)	.18*(.07)	.18*(.07)	.01(.02)
			Variance Components		
Level 1, ε_{ijk}	.53	.34	.34	.28	.09
Level 2, u_{0jk}	.04	.04	.07	.06	0
N	942	934	910	934	964

Note. Point estimates derived from hierarchical models including the treatment indicator, school-fixed effects, student demographics, and reading and math pretest scores, with teacher-random intercepts. Content Comprehension is a composite of near- and mid-transfer assessments. EOG = End-of-Grade assessment.

† $p < 0.10$. * $p < 0.05$. ** $p < 0.01$.

Table 8. Sensitivity Check: Results of Regression Models Predicting Treatment Effects of Structured Supplements on Measures of Reading Comprehension Using Aggregated Classroom-Level Data

	Recall	Near Transfer	Mid Transfer	Content Comprehension	EOG
Intercept	1.57 (3.43)	-3.96 (2.55)	-5.11 (3.18)	-4.21 (2.67)	-1.31 (1.08)
Treatment	.23*(.1)	.25***(.07)	.21*(.09)	.24**(.08)	.04(.03)
R^2	.80	.90	.84	.89	.95
N	79	80	80	80	80

Note. Point estimates derived from OLS regression models using classroom-level aggregated data including treatment indicator, school-fixed effects, student demographics, and reading and math pretest scores. Content Comprehension is a composite of near- and mid-transfer assessments. EOG = End-of-Grade assessment.

* $p < 0.05$. ** $p < 0.01$. *** $p < 0.001$.

Table 9. Annotations of Selected Transcripts of Treatment/Control Teachers from Clive Elementary School

Teacher	Quote	Analysis
Ms. Brighton (Control)	"The route and what's a route? The way that they travel, right."	Ms. Brighton uses the word *ingenious* 11 times, but never stops to provide a child-friendly definition of *ingenious* or to discuss the word's meaning in different contexts. Adhering closely to the lesson script, she reads aloud: "how did the team face these challenging obstacles…the moon team needed ingenious members…we needed ingenious and smart mathematicians." Instead, Ms. Brighton gives a quick explanation of the word *route*. While a quick clarifying explanation is fine, it is important to note that *route* is a word mentioned in passing four times whereas *ingenious* is a word that appears 11 times. Yet, Ms. Brighton never explained its meaning outside of the brief explanation that was a part of the text mentioned above. In general, Ms. Brighton largely follows the same pattern of discourse, adhering to the script but rarely stopping to define words or discuss word meanings.
Ms. Jones (Treatment)	"Now the word ingenious itself can be applied to almost all of these people we talked about today. It simply means – it has that word genius. If you are a genius, does anybody know what that word means?" [inaudible student responses] "If somebody says, 'Oh hey, you're a genius,' what are they actually saying to you?" [inaudible student response] "Yeah, you're smart. So if you are	Unlike Ms. Brighton, Ms. Jones had the structured supplement to introduce the form and meaning of the word *ingenious*. And still she continued to discuss the word's meaning beyond the structured supplement. She asked her students what *ingenious* meant and then provided additional information about the word by leveraging a familiar word nestled within *ingenious – genius*, a word that

	a genius person – person is a noun – ingenious is an adjective."	some students seemed to know as evidenced by them shouting out the word's meaning. Ms. Jones draws an important distinction here: *genius* is a noun and *ingenious* is an adjective that describes a genius. This was all unscripted, but the structured supplement likely highlighted that this was a word to examine closely.

Note. School and teacher names are all pseudonyms.

Supplementary Appendices

S-A.1. Baseline Characteristics and Balance Checks at Time of Randomization

	Overall *M* (*SD*)	Treatment *M* (*SD*)	Control *M* (*SD*)	Difference
White (%)	0.216	0.214	0.218	-0.030†
	(0.412)	(0.411)	(0.413)	(0.018)
Black (%)	0.354	0.361	0.347	0.018
	(0.478)	(0.481)	(0.476)	(0.022)
Hispanic (%)	0.320	0.301	0.338	-0.019
	(0.467)	(0.459)	(0.473)	(0.021)
Asian (%)	0.080	0.082	0.077	0.010
	(0.271)	(0.275)	(0.267)	(0.015)
Other (%)	0.031	0.042	0.020	0.021*
	(0.172)	(0.200)	(0.140)	(0.009)
Male (%)	0.516	0.519	0.513	0.009
	(0.500)	(0.500)	(0.500)	(0.026)
Limited English Proficiency	0.255	0.244	0.266	-0.004
	(0.436)	(0.430)	(0.442)	(0.020)
High SES	0.249	0.253	0.244	-0.001
	(0.432)	(0.435)	(0.430)	(0.016)
Med SES	0.361	0.361	0.362	-0.004
	(0.481)	(0.481)	(0.481)	(0.022)
Low SES	0.387	0.384	0.389	0.009
	(0.487)	(0.487)	(0.488)	(0.020)
Standardized Winter MAP Reading RIT Fall 202122	189.983	190.791	189.213	1.070
	(18.603)	(18.450)	(18.727)	(1.074)
Standardized Winter MAP Math RIT Fall 202122	192.048	192.703	191.424	1.094
	(15.472)	(15.167)	(15.742)	(0.887)

Note. MAP = Measure of Academic Progress. EOG = End-of-Grade assessment. SES = socioeconomic status. Other = students who identify as multicultural or Native American. M = mean; SD = Standard Deviation. Total students = 1,467; Treatment students = 718; Control students = 749. Difference was calculated from HLM models with school fixed effects and teacher random effects.

† $p < 0.10$. * $p < 0.05$.

S-A.2. Teacher Baseline Characteristics and Balance Checks at Time of Randomization

	Overall	Treatment	Control	Difference
Past MORE Experience	0.853	0.765	0.941	-0.119
	(0.919)	(0.890)	(0.952)	(0.301)
Board Certified (%)	0.250	0.265	0.235	0.046
	(0.436)	(0.448)	(0.431)	(0.100)
Years of Classroom Experience	11.868	11.441	12.294	-1.020
	(7.651)	(7.688)	(7.705)	(2.721)
N	68	34	34	68

Note. Post MORE Experience is on a scale of 0-6 and each value represent the number of MORE units taught. Board Certified refers to Nation Board Certified Teacher.

S-A.3. Baseline Characteristics and Balance Checks of Final Analytic Sample

	Full Sample	Treatment M(SD)	Control M(SD)	Difference
White (%)	0.250	0.240	0.270	-0.047*
	(0.430)	(0.430)	(0.440)	(0.024)
Black (%)	0.320	0.340	0.290	0.041
	(0.470)	(0.470)	(0.460)	(0.026)
Hispanic (%)	0.310	0.290	0.330	-0.031
	(0.460)	(0.450)	(0.470)	(0.025)
Asian (%)	0.090	0.090	0.090	0.007
	(0.290)	(0.280)	(0.290)	(0.019)
Other (%)	0.030	0.040	0.020	0.029*
	(0.170)	(0.210)	(0.130)	(0.011)
Male (%)	0.480	0.480	0.480	0.001
	(0.500)	(0.500)	(0.500)	(0.033)
Limited English Proficiency	0.250	0.230	0.260	-0.013
	(0.430)	(0.420)	(0.440)	(0.025)
High SES	0.290	0.270	0.300	-0.020
	(0.450)	(0.450)	(0.460)	(0.022)
Med SES	0.350	0.350	0.360	0.002
	(0.480)	(0.480)	(0.480)	(0.027)
Low SES	0.350	0.370	0.340	0.023
	(0.480)	(0.480)	(0.470)	(0.024)
Standardized Winter MAP Reading RIT Fall 202122	192.320	192.670	191.950	0.525
	(18.340)	(18.330)	(18.360)	(1.313)
Standardized Winter MAP Math RIT Fall 202122	194.180	194.450	193.900	0.518
	(14.870)	(14.920)	(14.840)	(1.033)

Note. MAP = Measure of Academic Progress. EOG = End-of-Grade assessment. SES = socioeconomic status. Other = students who identify as multicultural or Native American. M = mean; SD = Standard Deviation. Total students = 965; Treatment students = 499; Control students = 466. Difference was calculated from HLM models with school fixed effects and teacher random effects.

* $p < 0.05$.

S-A.4. Teacher Baseline Characteristics and Balance Checks of Final Analytic Samples

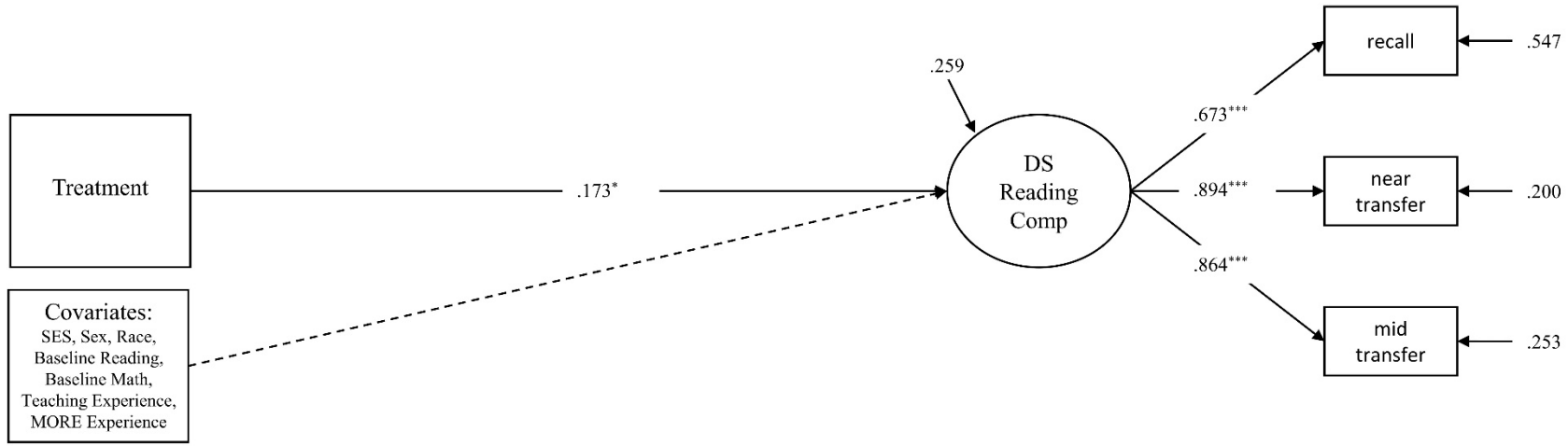
	Overall	Treatment	Control	Difference
Past MORE Experience	0.806	0.765	0.848	-0.046
	(0.839)	(0.890)	(0.795)	(0.302)
Board Certified (%)	0.254	0.265	0.242	0.032
	(0.438)	(0.448)	(0.435)	(0.103)
Years of Classroom Experience	11.985	11.441	12.545	-1.310
	(7.647)	(7.688)	(7.682)	(2.841)
N	67	34	33	68

Note. Post MORE Experience is on a scale of 0-6 and each value represent the number of MORE units taught. Board Certified refers to Nation Board Certified Teacher.

S-A.5. Benjamini-Hochberg Procedure to Test for False Discoveries

	Reported p-value (p_x)	X = p-value rank	M = # of p-values	New Critical p-value (p_x' =0.05x/5)	Finding p-value < New Critical p-value?	Statistical Significance after BH Correction?	Largest p-value rank less than or equal to p_x'
Content Comprehension	0.008	1	5	0.01	TRUE	TRUE	0.04
Near transfer	0.00600	2	5	0.02	TRUE	TRUE	0.04
Recall	0.01200	3	5	0.03	TRUE	TRUE	0.04
Near transfer	0.01800	4	5	0.04	TRUE	TRUE	0.04
EOG	0.69100	5	5	0.05	FALSE	FALSE	0.04

S-A.6. Sensitivity check for results of SEM predicting the treatment effect of structured supplements on measure of reading comprehension



Note. * $p < .05$. *** $p < .001$.