Routledge
Taylor & Francis Group

Journal of Research on Educational Effectiveness

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uree20

# The Efficacy of Two Models of Professional Development Mediated by Fidelity on Fourth Grade Student Reading Outcomes†

**Elizabeth Swanson, Alicia A. Stewart, Elizabeth A. Stevens, Nancy K. Scammacca, Philip Capin, Bethany H. Bhat, Greg Roberts & Sharon Vaughn**

**To cite this article:** Elizabeth Swanson, Alicia A. Stewart, Elizabeth A. Stevens, Nancy K. Scammacca, Philip Capin, Bethany H. Bhat, Greg Roberts & Sharon Vaughn (2023): The Efficacy of Two Models of Professional Development Mediated by Fidelity on Fourth Grade Student Reading Outcomes†, Journal of Research on Educational Effectiveness, DOI: 10.1080/19345747.2023.2181897

**To link to this article:** https://doi.org/10.1080/19345747.2023.2181897

Published online: 20 Mar 2023.

JOURNAL OF RESEARCH ON EDUCATIONAL EFFECTIVENESS
https://doi.org/10.1080/19345747.2023.2181897



# The Efficacy of Two Models of Professional Development Mediated by Fidelity on Fourth Grade Student Reading Outcomes†

Elizabeth Swanson[a]*, Alicia A. Stewart[b]*, Elizabeth A. Stevens[c], Nancy K. Scammacca[a], Philip Capin[a], Bethany H. Bhat[a], Greg Roberts[a], and Sharon Vaughn[a]

[a]The Meadows Center for Preventing Educational Risk, The University of Texas at Austin, Austin, Texas, USA; [b]Department of Special Education and Interventions, Central Connecticut State University, New Britain, Connecticut, USA; [c]Department of Learning Sciences, Georgia State University, Atlanta, Georgia, USA

**ABSTRACT**

This study addressed the effects of Strategies for Teaching Reading, Information, and Vocabulary Effectively (STRIVE), a distributed professional development (PD) model designed to help teachers implement reading comprehension and vocabulary practices in fourth grade social studies classes. Schools ($n = 81$ schools, $n = 235$ teachers, $n = 4{,}757$ students) were randomly assigned to one of three conditions: researcher-supported PD, school-supported PD, or business as usual (typical instruction). Findings revealed significant effects for both treatment conditions when compared to the business-as-usual condition for content knowledge ($g = 0.51$–$0.55$), vocabulary learning ($g = 0.49$) and reading comprehension in content ($g = 0.16$–$0.26$). Statistically significant effects were not observed for the Gates MacGinitie Reading Comprehension ($g = 0.04$–$0.06$), however, the effect size for the Gates MacGinitie Vocabulary test was statistically significant for the school-supported PD group ($g = 0.03$–$0.07$). Findings establish the efficacy of the STRIVE PD model on student reading outcomes and supports the efficacy of using more sustainable methods of PD that feature school supported follow up PD. Fidelity did not mediate any outcomes.

**ARTICLE HISTORY**
Received 11 January 2021
Revised 26 January 2023
Accepted 7 February 2023

**KEYWORDS**
Professional development; elementary school; literacy

## Investigating the Role of a Distributed Model of Professional Development on Fourth Grade Student Reading Outcomes

Reading in the social studies provides broad reaching benefit to students. In fact, elementary school students who receive an additional 30 minutes per day of social studies instruction and text reading outperform their peers on general reading outcomes (Tyner

CONTACT Elizabeth Swanson easwanson@austin.utexas.edu The Meadows Center for Preventing Educational Risk, The University of Texas at Austin, 1912 Speedway D4900, Austin, TX 78712, USA.
*Both these authors have contributed equally to this manuscript.
†Reports of the findings from subsections of the fully powered sample were presented at a national conference and on the Meadows Center for Preventing Educational Risk website.

& Kabourek, 2020). Syntheses and meta-analyses investigating the impact of reading instruction using informational text report moderate to large effects on reading and content knowledge outcomes (Gajria et al., 2007; Swanson et al., 2014). A recent series of studies investigating the efficacy of a set of vocabulary and reading comprehension instructional practices embedded within social studies for middle-grade students further support these meta-analytic findings. Vaughn et al. (Vaughn, Swanson, et al., 2013; Vaughn et al., 2015, 2017, 2019) reported that a 6-week dose of lessons that included explicit vocabulary instruction coupled with text-based discussion and team based learning impacted students' content knowledge and informational text comprehension,

## The Impact of Professional Development (PD) on Student Outcomes

Learning content from text reading is important, yet teachers of older students report feeling underprepared to teach their students how to learn from informational text (Anders et al., 2000; Ness, 2011). Providing a one-time PD to a large group of teachers is common practice. However, it is rarely sufficient to influence sustained change in classroom practice (Darling-Hammond et al., 2017). For change in practice to occur that is potent enough to influence student outcomes, teachers require follow up opportunities to engage around a targeted set of practices (Opfer et al., 2011). These follow up opportunities are associated with positive effects on teachers' knowledge, skills, and practices (Garet et al., 2001, 2008) as well as student literacy outcomes (Klingner et al., 2004). A meta-analysis of 17 PD studies (Basma & Savage, 2018) revealed that the overall effect size across studies was $g = 0.23$ when PD focused on literacy practices in kindergarten to fifth grade classrooms. The nature of PD represented in 15 of the studies was described as "typical" or "traditional" PD that took the form of workshops or summer institutes delivered face to face on a designated PD day with no students present. The two studies where teachers were reflective of their practice during follow up PD opportunities produced the largest effects on student reading outcomes (Amendum, 2014; Fine & Kossack, 2002), establishing a gap in the literature exploring the efficacy of PD that includes ongoing support opportunities for teachers. In addition, prior studies provide relatively little information about the effect of PD on upper elementary students' reading comprehension. Among 4th and 5th graders, reading comprehension effects ranged from 0.07 to 0.49 (Duffy et al., 1986; Klingner et al., 2004; Porche et al., 2012). No study provided information about the effect of PD on students' vocabulary outcomes and few studies provided fidelity of implementation. None of the studies examined the role of fidelity in influencing student outcomes.

## The STRIVE PD Model

STRIVE PD includes an initial workshop and two follow-up teacher study team meetings that provide teachers (1) multiple opportunities to practice using the lesson materials with peers and students, (2) opportunities to reflect, problem solve and extend learning, and (3) support for implementing the practices with high levels of fidelity. In addition, STRIVE PD is distributed over three units of study. The workshop occurs prior to Unit 1 and teacher study team meetings occur prior to Unit 2 and Unit 3.

This approach distributes knowledge and practice allowing teachers to manage a limited number of lesson materials containing a feasible number of instructional practices.

### What Teachers Learn during STRIVE PD

STRIVE PD provides instruction to teachers on a set of vocabulary and reading comprehension practices grounded in the construction-integration model of comprehension (Kintsch, 1988). According to the model, comprehension occurs in two stages. In the first stage—construction—readers approach the text and encounter new information. In the second stage, the reader integrates the new information with prior information to create a mental representation of the text. As they read, this mental representation is updated, producing comprehension and content learning. According to Kintsch's (1988) model and more recent research (Ahmed et al., 2016; O'Reilly et al., 2019), prior knowledge, including vocabulary knowledge, is a major driving factor in text comprehending and learning.

The instructional practices taught during STRIVE PD include evidence based knowledge building, vocabulary, and comprehension practices. Before reading, teachers introduce the lesson and tie the day's topic to prior learning. They also spend time teaching a selected set of high-utility vocabulary words that are essential to understanding the content presented in the text. During reading, students read text, engage in class discussion and use a main idea strategy called Get the Gist (Klingner et al., 2012). After reading, students return to the vocabulary words to increase understanding aided by what students learned during text reading. In an earlier study of STRIVE PD (Simmons et al., 2010), 48 teachers learned either the reading comprehension practices, the vocabulary practices, or engaged in business as usual (BAU) fourth-grade social studies instruction. Students in both treatment groups outperformed their BAU peers on a social studies content measure and students in all groups performed equally well on a standardized measure of reading comprehension. Students in the vocabulary treatment group also outperformed their BAU peers on social studies vocabulary measure. Findings from this study revealed the need to combine the vocabulary and reading comprehension practices into one set of lessons. It also provided indication that STRIVE PD can be effective in influencing student outcomes. In the current study, we investigate not only efficacy of STRIVE PD on a range of student outcomes but also *why* STRIVE PD works.

### Examining Why PD Works

One way to investigate why PD works is to consider the role of hypothesized mediators (Roberts et al., 2018). Researchers have long hypothesized that PD produces a change in classroom practice that subsequently effects student outcomes (Clarke & Hollingsworth, 2002; Garet et al., 2001). Empirically examining the causal-mediated effect of fidelity on student outcomes remains a necessary examination in the field of education and can serve to add depth to results from intervention research. For example, within a randomized controlled trial (RCT), group differences on outcome measures can be attributed to the intervention. Mediating variables that are hypothesized to change prior to changes in the outcome measures benefit from random assignment as well allowing us to assume

that changes to the mediator—in this case fidelity—are also attributed to the intervention. If the pathway from PD to student outcomes is mediated through fidelity, then we can conclude that the PD increased student outcomes in part by improving fidelity (i.e., adherence to the independent variable; Roberts et al., 2018).

Prior research examining the mediating role of fidelity in student outcomes indicates that fidelity matters. When teachers implemented practices they learned during PD with higher levels of fidelity, students in preschool (Hamre et al., 2010), early elementary school (Unlu et al., 2016), the middle grades (Roberts et al., 2017; Vaughn, Swanson et al., 2013; Vaughn et al., 2015), and high school (Cantrell et al., 2013) performed better on reading outcomes. In one study (Vaughn, Roberts et al., 2013), researchers examined the impact of Collaborative Strategic Reading (CSR; the Get the Gist strategy is part of CSR and STRIVE) on middle school student outcomes and reported that implementation fidelity mediated the effect of group assignment on reading comprehension outcomes. Teachers who implemented CSR with higher fidelity made a greater impact on reading comprehension outcomes.

In prior examinations of STRIVE PD (Hairrell et al., 2011; Simmons et al., 2010), teachers implemented the practices they learned with high levels of fidelity and reported the practices as feasible, appropriate, and important. This provides us with some indication that STRIVE PD influenced classroom practice. However, the causal relation between STRIVE PD, fidelity, and student outcomes has not yet been examined empirically.

### *Examining PD Conditions that Are Led by School Personnel*

In prior studies when researchers delivered PD, teachers reported a shallow understanding about the instructional practices (Daniel & Lemons, 2018; Datnow, 2002) and a perceived lack of ownership of the practices (Coburn, 2003). Others claim that while U.S. schools spend an average of $18,000 per teacher per year—mostly to bring outside PD deliverers into the school—teachers rarely take up the practices (TNTP, 2015). Intuition may suggest that when PD is localized (i.e., delivered by school personnel or teachers' peers), teachers are more likely to take up the practices and student outcomes are influenced. However, we could locate no empirical studies that examined the phenomenon.

One focus of the current study is to examine the effect of PD led by school leaders and teachers on student outcomes. For this reason, we include two treatment conditions. One is a researcher supported PD (RPD) condition where researchers led the initial PD session and all teacher study team meetings. The second treatment condition is school-supported PD (SPD) where researchers provided the initial PD session and school-based leaders provided the teacher study teams (i.e., shifting ownership to teachers). Both of these treatments are compared to a BAU comparison condition.

### *Purpose of the Study*

Findings prior studies examining the efficacy of PD (Basma & Savage; 2018; Garet et al., 2016; Yoon et al., 2007) combined with studies that have examined the role of fidelity (e.g. Unlu et al., 2016; Vaughn, Roberts et al., 2013) make evident that (1) most PD

research has examined the effects of traditional PD models with few studies examining PD models that provide follow up opportunities to teachers in the form of booster PD sessions; (2) most PD studies have been conducted with teachers of kindergarten through 3rd graders; (3) few PD studies examine vocabulary or reading comprehension outcomes; (4) no PD study has examined the efficacy of teacher-led PD, and (5) we could find no PD study examining a chain of effects from PD to fidelity (i.e., treatment adherence) to student outcomes. Examining the efficacy of a PD model that features teacher collaboration may provide essential information about how to best provide teachers the skills required to teach text comprehension and content. Therefore, the following research questions guided our work:

1. What are the effects of the STRIVE PD supported by researchers (i.e. RPD) or school personnel (i.e. SPD) compared to BAU on the content knowledge, vocabulary, and reading comprehension outcomes of fourth grade students?
2. To what extent does fidelity mediate the main effect of STRIVE PD on fourth grade student reading outcomes?

We hypothesized that student content knowledge and vocabulary outcomes would differ in favor of treatment groups when compared to students in the BAU group. Findings from prior studies of the STRIVE PD model (Simmons et al., 2010) and similar sets of instructional practices (Vaughn et al., 2015, 2017, 2019) indicate that group performance did not differ to a statistically significant level. Therefore, while we measured comprehension as a distal outcome, we did not anticipate an effect. Because evidence suggests fidelity can mediate both content knowledge (Vaughn et al., 2015) and reading comprehension (Hairrell et al., 2011; Vaughn, Roberts et al., 2013), we hypothesized that fidelity would at least partially mediate the main effect of PD on content knowledge, vocabulary, and reading comprehension. If teachers can deliver the instructional practices and students perform at least equally well across the SPD and RPD conditions, findings from this study can provide the field with evidence that PD models aligned with uptake and sustainability theory is efficacious, setting the stage to further examine sustainability over longer periods of time.

## Method

### *Research Design*

We conducted an RCT with school-level randomization (reviewed and approved by The University of Texas Institutional Review Board, approval number 2015-03-0100). The fully powered sample was recruited across the span of three annual cohorts: Cohort 1, in year 1, consisted of 38 schools. Cohort 2, in year two, consisted of 30 schools, and Cohort 3, in year 3, consisted of 13 schools. Schools in each cohort were randomly assigned to one of three conditions: RPD ($n = 26$), SPD ($n = 28$) or BAU ($n = 27$).

## Participants and Setting

### Schools

Principals and 4th grade teachers in 81 schools across six school districts in the southwestern United States participated in the study. Two schools that were randomized to participate in Cohort 3 did not complete the study. One was a school assigned to the BAU condition declined to participate after a new principal was appointed. No data were collected at this school. A second school, assigned to the RPD condition, sent teachers to the initial PD session and allowed student data to be collected at pretest and after the first unit of content. The school withdrew from the study to focus on improving state assessment performance. This produced an overall attrition of 2.5%, well within the limits of attrition deemed tolerable by the What Works Clearinghouse (What Works Clearinghouse, 2020). We determined not to include data from these schools in the analyses because it likely was not missing at random and represented such a small proportion of the entire dataset (Kang, 2013). This produced a final sample of 79 schools.

### Teachers

All 235 fourth-grade teachers (210 female and 25 male; Table 1) from the 79 schools consented to participate in the study; teachers received compensation for attending PD meetings and completing project tasks outside of their regularly scheduled workday.

**Table 1.** Teacher and school coordinator demographics.

| | School supported PD | Researcher supported PD | BAU comparison |
|---|---|---|---|
| Teachers | | | |
| Number | $n = 80$ | $n = 77$ | $n = 78$ |
| Gender | | | |
| Male | 9 | 11 | 5 |
| Female | 71 | 66 | 73 |
| Years teaching | | | |
| *M* | 10.90 | 9.68 | 10.20 |
| Certification | | | |
| Elementary | 78 | 76 | 73 |
| Secondary | 13 | 14 | 9 |
| Special Education | 8 | 2 | 5 |
| ESL | 22 | 12 | 14 |
| Bilingual | 14 | 18 | 23 |
| Degree | | | |
| Bachelors | 54 | 49 | 45 |
| Master's | 25 | 27 | 29 |
| Coordinators | | | |
| Number | $n = 29$ | $n = 25$ | $n = 25$ |
| Gender | | | |
| Male | 0 | 0 | 1 |
| Female | 29 | 25 | 24 |
| Years teaching | | | |
| *M* | 13.90 | 13.60 | 12.20 |
| Job titles | | | |
| Teacher | 15 | 7 | 13 |
| Assistant Principal | 0 | 1 | 1 |
| Principal | 0 | 1 | 0 |
| Implement specialist | 0 | 1 | 0 |
| Instructional coach | 14 | 15 | 11 |

*Note*. ESL: English as a second language; BAU: business as usual condition; One teacher in the researcher supported condition, one teacher in the school supported condition, and four teachers in the BAU condition did not report degree status.

There were no statistically significant differences between teachers in the three conditions in average years of teaching experience ($F(1, 34) = 0.583$, $p = 0.446$) or the percentage of teachers who held master's degrees ($\chi^2(2) = 1.607$, $p = .448$).

### Coordinators

We asked the principal at each school to identify a coordinator who served as the liaison between the campus and the research team. Coordinators in schools assigned to the SPD condition also led the teacher study team meetings at their schools (Table 1).

### Students

Parents of 4,757 students in grade four provided consent for their children to participate in data collection efforts (77% return rate). Student demographics for the full sample are detailed in Table 2. Most students were of Hispanic or White ethnicity. Across groups, more than 50% of students qualified to receive free or reduced lunch. About 10% of students were identified with a disability. Most students' parents reported speaking English in the home.

## STRIVE Intervention Procedures

### STRIVE PD

The PD provided to teachers in the SPD and RPD groups was distributed over an 18-week time period. Teachers in both treatment conditions participated in one PD session

**Table 2.** Student demographics.

| | School supported | | Researcher supported | | BAU | |
|---|---|---|---|---|---|---|
| | *N* | % | *n* | % | *n* | % |
| Gender | | | | | | |
| Male | 962 | 49.6 | 787 | 46.5 | 654 | 47.7 |
| Female | 900 | 46.4 | 816 | 48.3 | 653 | 47.6 |
| Ethnicity | | | | | | |
| White | 444 | 22.9 | 321 | 19.0 | 204 | 14.9 |
| African American | 83 | 4.3 | 73 | 4.3 | 61 | 4.4 |
| Hispanic | 1281 | 66.1 | 1190 | 70.4 | 1019 | 74.3 |
| Asian | 26 | 1.3 | 6 | 0.4 | 34 | 0.7 |
| Native American or Pacific Islander | 1 | 0.1 | 4 | 0.2 | 1 | 0.1 |
| Two or More | 42 | 2.2 | 22 | 1.3 | 16 | 1.2 |
| Economic disadvantage | | | | | | |
| Free or Reduced Lunch | 882 | 54.3 | 793 | 53.8 | 761 | 62.3 |
| None | 611 | 37.6 | 507 | 34.4 | 311 | 25.5 |
| Participates in special education | 186 | 9.6 | 183 | 10.8 | 125 | 9.1 |
| Home language | | | | | | |
| Arabic | 0 | 0.0 | 0 | 0.0 | 1 | 0.1 |
| English | 1426 | 73.6 | 1275 | 75.4 | 880 | 64.1 |
| German | 1 | 0.1 | 1 | 0.1 | 0 | 0.0 |
| Spanish | 324 | 16.7 | 321 | 19.0 | 279 | 20.3 |
| Tamil | 1 | 0.1 | 0 | 0.0 | 0 | 0.0 |
| Thai | 0 | 0.0 | 1 | 0.1 | 0 | 0.0 |
| Vietnamese | 4 | 0.2 | 3 | 0.2 | 0 | 0.0 |
| Korean | 0 | 0.0 | 2 | 0.1 | 0 | 0.0 |
| Other | 2 | 0.2 | 4 | 0.2 | 0 | 0.0 |
| Limited English Proficient | 321 | 16.6 | 318 | 18.6 | 261 | 19.0 |

*Note*: Parents could choose more than one home language or could leave the item blank.

prior to the school year in preparation for delivering Unit 1. Teacher study team meetings led by either researchers (RPD) or school leaders (SPD) were held prior to Units 2 and 3. Units were designed so that teachers and students were introduced to the vocabulary and comprehension practices incrementally to build knowledge and use over time. Unit 1 focused on explicit vocabulary instruction, text reading with discussion, and get the gist. Unit 2 added the summary writing practice. Unit 3 added a context clue strategy.

The initial 6-hour workshop led by researchers included an overview of the instructional practices, modeling and practice opportunities with Unit 1 practices (i.e. explicit vocabulary instruction, text reading with discussion, get the gist), and the importance of treatment adherence to support student outcomes. Researchers engaged in explicit instruction to show teachers how to use the instructional practices: (1) Purpose for learning: Researchers stated the purpose for each practice and connected them to evidence documented by research; (2) Modeling: Researchers modeled instructional practices for participants and showed brief videos of the instruction. After each video, teachers engaged in discussions about how they plan to implement practices with their students; (3) Guided Practice: Teachers worked in pairs to rehearse the instructional practices. Researchers provided specific feedback to teachers. Presenters also encouraged teachers to take notes about how they planned to implement each practice in their classrooms; (4) Prepare for Independent Practice: Teachers shared ideas with one another regarding how they intended to implement STRIVE lessons with their students. After attending the initial, day-long training, teachers implemented Unit 1 lessons in their classrooms.

Teachers participated in two teacher study team meetings to prepare for Units 2 and 3. Prior to Unit 2, teachers learned how to teach students the summary writing practice and prior to Unit 3, teachers learned about a context clue strategy. These meetings lasted approximately two hours and were led by members of the research team (RPD) or a school coordinator (SPD). To promote extensive collaboration, meetings included an average of four participants and were held after school. Researchers and school-based coordinators followed the same meeting agenda, consisting of three components: (1) Reflect on instruction in the prior unit, (2) Introduce new practices, and (3) Set one major goal for the coming unit. To prepare school coordinators to lead teacher study team meetings, they attended the initial PD session. In addition, three weeks prior to each study team meeting, they received a PowerPoint presentation and handouts. After reviewing the materials on their own, they participated in a 30-minute phone call with researchers to ask questions about the content.

### *Vocabulary and Reading Comprehension Instructional Practices*

All schools in all conditions utilized the same fourth-grade progressive state standards and the same state-developed timeline for delivering social studies. Therefore, the only difference between treatment and comparison groups was inclusion in STRIVE PD where teachers learned to use a set of vocabulary and reading comprehension instructional practices within social studies. Teachers implemented the instructional practices across three, six-week units of study. Each instructional unit consisted of 12, 45-minute social studies lessons (36 lessons total). Teachers delivered two lessons per week.

*Before Reading Practices.* Teachers were provided PD on how to implement two instructional practices before text reading: (1) building and activating background knowledge: Teachers were provided illustrations of the text content and asked to lead discussions about the illustrations prompting students to make connections between prior knowledge and new content. (2) explicit vocabulary instruction: Each lesson included two, high-utility vocabulary words important to social studies understanding. Teachers guided students in completing the first half of a semantic map. They introduced the word using a student-friendly definition, led a discussion guided by a visual representation of the word, and provided examples of the word in the appropriate context. In Unit 3, teachers taught students a context clue strategy to derive the meaning of words rather than providing them with a student-friendly definition.

*During Reading Practices.* Teachers were taught to use the following practices during reading: (1) Text reading with questioning: Teachers led a text-based discussion framed by various question types to encourage literal and inferential thinking. (2) Get the Gist strategy: Gist statements are main idea statements that support content comprehension (Klingner et al., 2012). The strategy included two steps: (a) Who or what is this about? and (b) What is the most important idea about the "who" or "what?" Teachers modeled the strategy and then guided students to compose brief main idea statements after each major section of reading.

*After Reading Practices.* Participating teachers were asked to use after text reading practices designed to develop a deeper understanding of vocabulary and social studies content: (1) explicit vocabulary instruction: Students returned to the semantic maps. Word associations provided a list of four words; students chose two related to the target word. Students wrote a sentence using the word to demonstrate understanding. A turn-and-talk activity provided students an opportunity to apply their understanding of words in a way that connected to their own lives (e.g., If you could go on an *expedition*, where would you go and why?). A word building activity required students to add prefixes or suffixes to the target word to create new words. (2) summary writing: Teachers showed students how to use gist statements from sections of the text to write a summary of the entire passage.

## Fidelity

### *Strive PD Fidelity*

All teachers attended the initial PD session and teacher study team meetings completing a PD fidelity form upon completion of each meeting. Teachers rated the extent to which they felt prepared to teach each instructional component and the PD quality (e.g. the lesson components were well-described). Items in both sections were rated on a scale of 1 to 4 (1 = strongly disagree to 4 = strongly agree). Table 3 shows that teachers rated high levels of readiness to teach each instructional component and rated the PD quality as high.

**Table 3.** Professional development fidelity.

| | Initial PD | | Teacher Study Team 1 | | Teacher Study Team 2 | |
|---|---|---|---|---|---|---|
| Preparedness to teach each component* | Researcher supported *M* (*SD*) | School supported *M* (*SD*) | Researcher supported *M* (*SD*) | School supported *M* (*SD*) | Researcher supported *M* (*SD*) | School supported *M* (*SD*) |
| Before reading | | | | | | |
| Background knowledge | 3.74 (.47) | 3.84 (.40) | n/a | n/a | n/a | n/a |
| Explicit vocabulary instruction | 3.86 (.35) | 3.87 (.37) | n/a | n/a | n/a | n/a |
| During reading | | | | | | |
| Ask and answer questions | 3.75 (.49) | 3.81 (.42) | n/a | n/a | n/a | n/a |
| "Get the gist" main idea statements | 3.71 (.51) | 3.68 (.55) | n/a | n/a | n/a | n/a |
| "Get the gist" routine in collaborative learning pairs | n/a | n/a | 3.64 (.61) | 3.56 (.55) | n/a | n/a |
| After reading | | | | | | |
| "Gist to summary" | n/a | n/a | 3.66 (.53) | 3.48 (.60) | n/a | n/a |
| Comprehension purpose question | 3.75 (.51) | 3.81 (.40) | n/a | n/a | n/a | n/a |
| Vocabulary maps in collaborative learning pairs | 3.70 (.54) | 3.87 (.34) | 3.72 (.51) | 3.78 (.45) | n/a | n/a |
| Overall quality | 3.96 (.21) | 3.90 (.30) | 3.77 (.45) | 3.82 (.39) | 3.88 (.33) | 3.72 (.57) |

*Note.* Overall quality was rated on a 4-point scale; n/a: not applicable because the instructional practice was not introduced.
*Teachers rated how prepared they felt to teach each practice after the professional development sessions, with 1 being not prepared and 4 being prepared.

## *STRIVE Implementation Fidelity*

The fidelity code sheet contained two sections: instructional adherence and instructional quality. The code sheet was adapted from a previous study (e.g., Simmons et al., 2010) that used the same fidelity data collection methods (i.e., audio recordings detailed below) to align with STRIVE practices. In the instructional adherence section, STRIVE instructional practices and were coded on a four-point, Likert-type rating scale ranging from 1 (low alignment with intended method) to 4 (high alignment with intended method). If a component was not expected during a lesson or there was not enough time for the teacher to implement a specific component (i.e., fire drill or other school-related interruption), coders scored the item as "not applicable." If a component was expected but did not occur, coders assigned a score of 0. Instructional quality was coded on a 4-point scale and focused on teacher's general instructional performance using seven items that addressed lesson pacing, the use of feedback, frequency of practice opportunities, teacher preparedness, clarity of questions, explicit instruction, and enthusiasm.

*Fidelity Coding Procedure.* The research team asked treatment teachers to audio-record all STRIVE lessons. Teachers assigned to the BAU condition recorded three weeks of BAU social studies instruction (i.e. one week per six-week period).

For both treatment condition's (i.e., RPD and SPD) audios were selected as follows: (1) Two audios per school were randomly selected and coded; (2) Because it was important for mediation analysis to have a code for each STRIVE practice, if an audio was missing a code for a STRIVE practice, we randomly chose an additional audio. We repeated this step until each school had two codes for every STRIVE practice. In the

BAU condition, (1) teachers were randomly assigned to one of the three, six-week periods, and (2) one audio per teacher was randomly selected for coding. Five BAU teachers submitted no audio recordings. A total of 207 lessons (RPD = 64 lessons; SPD = 70 lessons; BAU = 73 lessons) were selected across conditions for coding.

The principal investigator conducted a 4-hour training for seven members of the research team that included learning the fidelity form codebook and practice assigning codes using audio recordings. Using the gold standard method for establishing inter-rater agreement (Gwet, 2001), the principal investigator assigned one audiotape for the coders to independently score for inter-rater agreement. The mean inter-rater agreement with the gold standard was 92%. Coding took place over the course of one month. To maintain agreement, one-third of audio recordings were double coded. Agreement between coders was above 90% for all double-coded audio recordings.

*Fidelity Results.* We used fidelity data to identify the extent to which instruction in the treatment conditions aligned with the instructional practices as intended (Table 4). Recorded STRIVE lessons averaged 44 minutes (SD = 22) in length. Scores for the RPD group ranged from 2.75 (building background knowledge) to 3.68 (explicit vocabulary instruction before reading). Scores for the SPD group ranged from 2.42 (building background knowledge) to 3.61 (explicit vocabulary instruction before reading). Fidelity ratings for the RPD and the SPD conditions were mid-high to high for most practices, indicating that teachers implemented the instructional practices as intended.

We assessed whether the STRIVE instructional practices were observed in the BAU condition (Table 4). Recorded BAU social studies lessons averaged 34 minutes (SD = 19) in length. The median recorded lesson lasted 26 minutes aligning with results from a recent study reporting that elementary school students receive about 28 minutes of social studies instruction per day (Tyner & Kabourek, 2020). Building background knowledge ($M = 2.28$), explicit vocabulary instruction ($M = 1.88$), and asking questions during reading ($M = 2.77$) were observed in more than 50% of lessons, with alignment to the treatment instructional practices in the low to medium high range. Lesson closure

**Table 4.** Instructional fidelity.

| | Researcher supported | | School supported | | BAU | |
|---|---|---|---|---|---|---|
| STRIVE components | *M* (SD) | Times observed | *M* (SD) | Times observed | *M* (SD) | Times observed |
| Before reading | | | | | | |
| Background knowledge | 2.75 (.95) | 55 | 2.42 (1.18) | 60 | 2.28 (.90) | 57 |
| Explicit vocabulary instruction | 3.68 (.71) | 59 | 3.61 (.73) | 62 | 1.88 (.97) | 40 |
| During reading | | | | | | |
| Pose comprehension purpose question | 3.14 (1.31) | 59 | 2.45 (1.48) | 62 | 2.82 (1.17) | 11 |
| Questions during text Reading | 3.25 (.97) | 60 | 3.28 (1.0) | 67 | 2.77 (1.01) | 53 |
| Discuss comprehension purpose question | 2.84 (1.32) | 37 | 2.87 (1.32) | 39 | 2.33 (1.53) | 3 |
| Gist statements | 3.22 (.84) | 46 | 3.00 (1.03) | 54 | 2.14 (.90) | 7 |
| After reading | | | | | | |
| Explicit vocabulary instruction | 3.26 (.94) | 42 | 3.44 (.91) | 39 | 1.33 (.58) | 3 |
| Summaries | 3.04 (1.07) | 23 | 2.89 (1.19) | 27 | 1.25 (.50) | 4 |
| Lesson closure | 3.00 (1.11) | 30 | 2.77 (1.22) | 30 | 1.92 (1.02) | 24 |

*Note.* Implementation of components was rated on a 4-point scale (1-lowest and 4 highest).

was observed in almost one-third of BAU lessons ($M = 1.92$). Explicit vocabulary instruction after reading, discussing the comprehension purpose question, gist statements, and summaries were observed in fewer than 10% of BAU observations. Most practices were either misaligned as evidenced by low alignment scores (e.g. vocabulary instruction) or not present in the BAU.

## Measures

We adopted a measurement plan that reflects a balanced approach (Clemens & Fuchs, 2021; Gersten et al., 2005) that includes a series of far-transfer (i.e., distal), mid-transfer, and proximal measures were used to examine STRIVE PD efficacy. Including multiple vocabulary and reading comprehension measures allows us to examine, "… student outcomes sensitive to the performance change the intervention is intended to bring about… [and] student outcomes not strictly aligned with the intervention (Institute of Education Sciences & National Science Foundation 2013, p. 22)." For a summary of our balanced approach to measurement, see Figure 1. Aligned with WWC standards (2020), all measures demonstrate face validity and reliability, no measure is over-aligned with the intervention, and administration of all measures was identical across all conditions.

## Far-Transfer Measures

The Gates MacGinitie Reading Comprehension (GMRC) and Vocabulary (GMRV) Subtests were administered to students by professional test administrators trained by a Ph.D. level researcher within two weeks prior to (using form S) and two weeks immediately following (using form T) STRIVE implementation. In one district that contained 42 schools, the Board of Trustees passed an initiative to reduce the amount of time students in the district were engaged in testing activities resulting in excluding students from the GMRC subtest.

| Test Features | | | | | |
|---|---|---|---|---|---|
| **Level of Alignment** | **Test Name** | **Construct Measured** | **Content** | **Response Mode** | **Timing of Administration** |
| Most aligned | Unit Test of Vocabulary | Vocabulary | Vocabulary taught during STRIVE lessons | Matching | At the end of each unit |
| | Unit Test of Content Knowledge | Content Knowledge | Content taught during STRIVE lessons | Multiple choice | At the end of each unit |
| | Content Reading Comprehension | Reading Comprehension | Social studies text on topics not covered in STRIVE lessons | Multiple choice | Post test |
| | Gates MacGinitie Vocabulary | Vocabulary | Vocabulary varying in type and difficulty | Multiple choice | Pretest and post test |
| Least Aligned | Gates MacGinitie Reading Comprehension | Reading Comprehension | Passages of various genres and text structures | Multiple choice | Pretest and post test |

**Figure 1.** Balanced approach to measuring content knowledge, vocabulary and reading comprehension.

GMRC Subtest (4th edition; MacGinitie et al., 2000). The GMRC is a group administered, 35-minute timed assessment of reading comprehension. Internal consistency reliability for the reading comprehension subtest ranges from 0.91 to 0.93, and alternative form reliability is 0.80 to 0.87.

GMRV Subtest (4th edition; MacGinitie et al., 2000). The GMRV is a group administered, 45-item, 20-minute timed assessment of vocabulary knowledge. The Kuder Richardson 20 reliability for this measure is 0.90–092.

### Mid-Transfer Measure

*Content Reading Comprehension.* The research team developed a content reading comprehension measure to test students' understanding of expository texts containing social studies content that was not yet taught to any student in any condition. Because it covered content (1) listed in the state-developed standards and (2) not yet taught according to the state-developed scope and sequence, this measure is considered a mid-transfer measure. Students read five passages and responded to six multiple-choice items following each passage. Passages ranged from 197 to 233 words and conformed to a fourth-grade reading level with a Lexile range of 700–900 L. Scores could range from 0 to 30. Internal consistency reliability for the assessment was 0.89, exceeding WWC (2020) standards.

### Proximal Measures

Two curriculum-based measures tested students' content and vocabulary knowledge at the end of each 6-week unit. These measures were not overly aligned with the intervention (What Works Clearinghouse, 2020). Items on the content knowledge measure covered content that students in all conditions learned as part of the state-developed standards and scope and sequence. Items on the vocabulary knowledge measure were taken from the state adopted textbook and standards that were used across all conditions.

Teachers administered the unit tests of content and vocabulary knowledge to their students. Researchers developed a test protocol that included scripted administration instructions. Teachers were trained on test administration during the initial PD workshop. The research team delivered and retrieved copies of the tests during the week of administration. Because these measures were used to assess curriculum mastery and many students in the sample were novice English speakers, Spanish versions of the three curriculum-based measures were made available to teachers in all three conditions. None of the BAU teachers opted to use the Spanish version. As a result, the students in the SPD and RPD conditions who responded to the Spanish version of the content measures were dropped before analyses of outcomes were conducted.

*Unit Tests of Content Knowledge.* Participating students responded to a content knowledge assessment at the end of each unit (Unit 1 = 13 items; Unit 2 = 15 items; Unit 3 = 19 items; Possible score 0–47). Each item included a brief sentence stem followed by four answer choices. We utilized data from Cohort I students to evaluate the item properties of the content tests. Items determined to be too difficult for students (based on a

criterion of 40% or fewer students responding correctly) were revised. Cohort 2 and Cohort 3 participants responded to this revised assessment. To allow for analysis of scores on the content measures with the three cohorts combined, total scores for each student were computed based on the common items administered to students in both cohorts. Item-total correlations were lower than desired, ranging from 0.33 to 0.41. Given these low item-total correlations, internal consistency ranged from 0.77 to 0.84 across the three content tests, exceeding the What Works Clearinghouse (2020) standards for reliability.

*Unit Tests of Vocabulary Knowledge.* Using Espin et al. (2001) procedures, researchers developed three vocabulary-matching tests of knowledge of terms taught during each of the three units. The tests consisted of 24 items for Unit 1 (score range = 0 to 24), 16 items for Unit 2 (score range = 0 to 16), and 19 items for Unit 3 (score range = 0 to 19). For each assessment, students matched each word with a brief definition. Students completed the measure in approximately 20 minutes. We utilized data from students included in cohort 1 to evaluate the item properties of the unit test of vocabulary knowledge. None of the items on the vocabulary measure were determined to be too difficult for students (based on a criterion of 40% or fewer students responding correctly). Within our sample, item-total correlations ranged from 0.54 to 0.58. Internal consistency reliabilities ranged from 0.89 to 0.93, exceeding What Works Clearinghouse (2020) standards. We acknowledge that the reliability estimates may be inflated because matching tests use a common set of response choices.

## Data Analysis

The efficacy of the STRIVE PD model on the comprehension, vocabulary, and social studies content knowledge of fourth-grade students was evaluated by estimating a series of multilevel models (MLMs). We estimated main effects as average treatment effects (ATE) using *MPlus* Version 8.4 (Muthén & Muthén, 1998–2019) with full-information maximum likelihood (FIML) estimation. Multilevel models account for dependencies in nested data structures by estimating residual components (i.e., random effects, errors, etc.) at each level and by partitioning total variance into its level-specific component parts, yielding unbiased parameter estimates. In the MLMs, students were nested within teachers and teachers were nested within schools. Schools were randomized to one of three conditions at level 3 of the model. Pairwise contrasts were modeled using dummy codes (0 = BAU). We fit preliminary null models to evaluate patterns of clustering; final models were specified accordingly. In total, 10 MLMs were analyzed to determine the main effects of STRIVE PD on student outcomes. To account for the number of hypothesis tests conducted, we implemented the Benjamini–Hochberg (B–H) procedure for controlling the false discovery rate (FDR; Benjamini & Hochberg, 1995). The ten models analyzed were treated as one family under the B–H procedure. We rank-ordered the $p$ values from the ten effect estimates in the five MLMs to determine if any exceeded the adjusted $p$ value criterion of 0.035 for the FDR, which was calculated using Kornbrot's (Kornbrot, 2021) spreadsheet.

Because schools placed limitations on testing time and we did not expect pretest differences on content knowledge, vocabulary or reading, these measures were administered at posttest only. Therefore, the analytical model for these outcomes is represented by Equation (1),

$$Unit\ Test\ Outcome_{ijk} = \gamma_{000} + \gamma_{001}SPD_k + \gamma_{002}RPD_k + e_{ijk} + r_{0jk} + u_{00k} \quad (1)$$

where $Unit\ Test\ Outcome_{ijk}$ is the post-test score for student $i$ in teacher $j$ in school $k$ where $SPD_k$ and $RPD_k$ represent assignment to the SPD and RPD treatment conditions, and $e_{ijk}$, $r_{0jk}$, $u_{00k}$ represent random effects at the student, teacher, and school levels, respectively. Pretests were administered for the GMRC and GMRV, and the models for these measures included the pretest score as a covariate. This analytical model is represented by Equation (2),

$$\begin{aligned} \text{GMRT Outcome}_{ijk} = {} & \gamma_{000} + \gamma_{100}(\text{Pretest}_{ijk}) + \gamma_{010}(\text{Pretest}_{jk}) + \gamma_{001}(\text{Pretest}_k) \\ & + \gamma_{002}(\text{SPD})_k + \gamma_{003}(\text{RPD})_k + e_{ijk} + r_{0jk} + u_{00k} \end{aligned} \quad (2)$$

Equation (2) extends Equation (1) by adding pretest covariates with the student level pretest centered on the teacher mean, the teacher-level parameter centered around the school mean, and the school-level value centered around the grand mean. Including pretest scores as a covariate optimizes statistical power (Venter et al., 2002), and centering covariates aids in the interpretation of the model parameters (Hox, 2002). As before, random effects at different levels of the model are represented by the latter three terms in the equation.

To determine if fidelity mediated the effect of assignment to condition on student outcomes, we conducted mediation analyses using multi-level structural equation models (ML-SEM; Preacher et al., 2010). As a first step in investigating fidelity as a mediator of treatment effects, we sought to determine if the nine procedural fidelity items measured a single latent construct that could be represented in the ML-SEM as a factor score. To that end, we conducted a confirmatory factor analysis (CFA) on the fidelity data for teachers in the RPD and SPD conditions using the lavaan 0.6-6 package (Rosseel, 2012) in R (R Core Team, 2019). Given that not all of the nine items were present in each fidelity observation, we used FIML estimation to address missing data. In the process of refining the factor model, the item "Discuss comprehension purpose question related to passage main idea" was dropped to improve fit. This item was not observed often in the fidelity recordings. Additionally, the residuals for the several of the items were allowed to covary to enhance fit and the variance for one item was fixed to 0 because it was not significantly different from 0. The chi-squared statistic for the model fit was statistically significant ($\chi^2(28) = 130.87$, $p < .001$), which is a common finding with large sample sizes. Other fit indices indicated excellent fit (CFI = 1.0; RMSEA <0.001). We also tested for measurement invariance across the RPD and SPD teachers and found that the factor model was invariant groups at the scalar level.

Given these results, we calculated factor scores for STRIVE and BAU teachers and used these as the measure of fidelity in the mediation ML-SEMs. We took this approach rather than including the CFA as a measurement model component in the ML-SEM in order to have a meaningful fidelity score for both STRIVE and BAU teachers. As described above, the BAU teacher data contained few instances where STRIVE components were observed and rated. As a result, item scores varied little across the BAU

condition and modeling the BAU teacher data in the CFA was problematic. Therefore, using factor scores for all teachers (computed based on the CFA of the data from the teachers in the two STRIVE PD conditions) allowed us to represent fidelity for each teacher in a meaningful way in the ML-SEM.

Mediation analyses were conducted only for outcomes where significant main effects were found. In the mediation models, we included paths for direct and indirect effects of treatment assignment on student outcomes. The nested structure of the data was modeled in the same way as in the analyses for main effects. The fidelity factor score for each teacher and an aggregation of the fidelity factor scores within each school was added to the appropriate nested model where we had found a significant main effect for the SPD vs. BAU or the RPD vs. BAU contrasts in the main effect models. To find the indirect effect we multiplied the mediator at level 3 (commonly referred to as the *a* path), regressed on the treatment effect, by the outcome regressed on the mediator at level 3 (commonly referred to as the *b* path). For the GMRT Vocabulary, we also included the centered pretest variable in the model.

Five effect sizes can be estimated in a three-level cluster-randomized model with allocation and treatment at level 3 (see Hedges, 2011). These differ primarily in the standard deviation used to "normalize" group differences. As a result, they differ in their interpretation and usefulness. We report two of these five effects: school-level effect sizes and student-level effect sizes. School-level effect sizes are meaningful because level-3 units were randomized and the STRIVE PD treatments were delivered at the school level. Differences were standardized on the school-level variance. A school-level effect can be interpreted as the standardized mean difference between two similar schools, where one implemented the treatment and the other implemented business as usual. We also report student-level effect sizes using the pooled within-total variance to standardize differences and the Hedges (2007, 2011) procedures for adjusting student-level effect sizes for clustering (specifically, we used Equations (17) and (18) in Hedges, 2011). In the models where the pretest was included as a covariate, we used covariate-adjusted means to calculate the effect size and then adjusted it for clustering. The student-level effect size represents the difference between the average student in the average teacher's classroom in the average school implementing a STRIVE PD treatment compared to a very similar average student in an average classroom in an average BAU school. School-level effect sizes are much larger than those at the student level, in our case and generally, because the majority of the total variance in school settings exists at lower levels of the model (e.g., students). The two types of effect sizes complement one another. However, they do not replicate each other, and they should not be conflated when interpreting our results.

## Results

In the results, both school-level and student-level effect sizes adjusted for the cluster-randomized design are presented. Descriptive statistics for all outcome measures and student-level effect sizes are presented in Table 5. Because analytical models differed for the measures administered at pre and posttest and those administered at posttest only,

**Table 5.** Student-level descriptive statistics by condition.

| | School supported PD | | | Researcher supported PD | | | BAU Comparison | | |
|---|---|---|---|---|---|---|---|---|---|
| | *M* | *SD* | *n* | *M* | *SD* | *n* | *M* | *SD* | *n* |
| Unit Content knowledge | 29.27 | 11.43 | 1822 | 28.49 | 11.22 | 1572 | 19.91 | 8.53 | 1363 |
| Unit Content vocabulary | 38.40 | 16.66 | 1822 | 38.05 | 16.75 | 1572 | 25.38 | 14.35 | 1363 |
| Content reading comprehension | 8.01 | 4.56 | 1822 | 7.28 | 4.53 | 1572 | 6.23 | 4.36 | 1363 |
| Gates Vocabulary Pretest (ESS) | 458.20 | 41.45 | 1710 | 456.65 | 43.11 | 1476 | 458.28 | 38.07 | 1289 |
| Gates Vocabulary Posttest (ESS) | 477.16 | 44.43 | 1634 | 475.11 | 44.69 | 1426 | 472.70 | 41.30 | 1210 |
| Gates Reading Comp Pretest (ESS)* | 468.24 | 43.92 | 1068 | 472.00 | 43.43 | 734 | 472.79 | 42.68 | 593 |
| Gates Reading Comp Posttest (ESS)* | 483.29 | 43.87 | 996 | 489.42 | 43.34 | 709 | 486.18 | 39.07 | 581 |

*Note*. PD: professional development; BAU: business as usual; ESS: Extended Scale Scores.
*One of the six school districts did not allow the research team to administer Gates MacGinitie Reading Comprehension due to district-imposed caps on assessment time.

model parameters are contained in tables organized by model (Table 6 for the posttest only measures and Tables 7 and 8 for the pretest and posttest measures).

## Missing Data

Across all the unit content measures, no consented student had missing scores because teachers administered these measures and provided make-up opportunities for students who were absent on the day of testing. For the GMRC and GMRV, which university-affiliated assessment coordinators administered, complete data for both the pretest and posttest GMRV were available for 85% of consented students; 95% completed the pretest and 91% completed the posttest. In the school districts where the GMRC was administered, 86% of students had complete pretest and posttest data; 95% completed the pretest and 91% completed the post-test. FIML estimation, which allows for participants with missing data to be included in the analysis without data imputation, was used to estimate the multilevel models. FIML can be used in this way if data are missing completely at random or missing at random. Analysis of missing data patterns indicated that the likely explanation for the missing data is that students were absent when the GMRC and GMRV measures were administered or had transferred out of the school during the year. As a result, we concluded that the data met the assumptions of missing at random.

## Baseline Equivalence

To check for baseline equivalence, we examined the data for pretest differences between students in SPD and RPD and BAU schools on the GMRV and GMRC subtests. The differences in scores on both subtests were not statistically significant. Effect sizes at pretest on the GMRV subtests were small, −0.01 ($SE = 0.17$) for the comparison of SPD and BAU and −0.04 ($SE = 0.17$) for RPD and BAU and met the WWC criteria for baseline equivalence (WWC, 2020). For the students who provided responses to the GMRC, the effect sizes at pretest were somewhat larger, −0.22 ($SE = 0.25$) for the comparison of SPD and BAU and −0.08 ($SE = 0.24$) for RPD and BAU, but were within the WWC standard of 0.05 and 0.25 for baseline equivalence when the pretest is included as a covariate in the analysis of posttest scores on the GMRC subtest. Additionally, prior to combining data from the cohorts we determined that there were no statistically

**Table 6.** Unit tests model results.

| | | Coeff. | *SE* | *p* | Intercept (SE) | Level 2 variance | Level 2 ICC | Level 3 variance | Level 3 ICC | Residual | School-level g (SE) | Student-level g (SE)[b] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Unit content knowledge | SPD | 8.90 | 1.32 | <.001[a] | 19.59 (0.68) | 7.23 | 0.06 | 19.26 | 0.17 | 86.18 | 1.30 (0.21) | 0.55 (0.09) |
| | RPD | 8.57 | 1.20 | <.001[a] | | | | | | | 1.42 (0.22) | 0.51 (0.08) |
| Unit content vocabulary | SPD | 12.35 | 1.99 | <.001[a] | 24.94 (1.15) | 17.07 | 0.07 | 43.21 | 0.17 | 197.24 | 1.19 (0.21) | 0.49 (0.09) |
| | RPD | 12.54 | 1.88 | <.001[a] | | | | | | | 1.32 (0.22) | 0.49 (0.08) |
| Content reading comprehension | SPD | 1.74 | 0.52 | .001[a] | 6.06 (0.36) | .81 | 0.04 | 2.95 | 0.15 | 16.34 | 0.65 (0.20) | 0.26 (0.08) |
| | RPD | 1.14 | 0.53 | .032 | | | | | | | 0.43 (0.20) | 0.16 (0.08) |

[a]Difference between treatment condition and BAU was statistically significant at $p<.016$. $N_{Students} = 4{,}757$; $N_{Teachers} = 222$; $N_{Schools} = 79$.
[b]Student-level g(SE)s were corrected for clustering at the school level using the method described in Hedges (2007).

**Table 7.** Gates-MacGinitie Reading Test – vocabulary model results.

| | Vocabulary |
|---|---|
| Pretest | |
| Level 1 (CWC) | 0.78 (0.02) |
| Level 2 (CWC) | 0.94 (0.06) |
| Level 3 (GMC) | 1.09 (0.03) |
| Coeff. (*SE*) | |
| SPD | 5.03 (1.52) |
| RPD | 2.26 (1.81) |
| *p* | |
| SPD | 0.001 |
| RPD | 0.21 |
| Intercept (*SE*) | 469.75 (0.98) |
| Level 2 variance | 35.18 |
| Level 2 ICC | 0.06 |
| Level 3 variance | 13.91 |
| Level 3 ICC | 0.02 |
| Residual | 556.70 |
| *N* students | 4027 |
| *N* teachers | 220 |
| *N* schools | 79 |
| School-level *g* (*SE*)[a] | |
| SPD | 0.50 (0.20) |
| RPD | 0.21 (0.20) |
| Student-level *g* (*SE*)[1] | |
| SPD | 0.07 (0.03) |
| RPD | 0.03 (0.03) |

*Note*: CWC: centered within clusters; GMC: grand mean centered; SE: standard error; SPD: school supported professional development; RPD: researcher supported professional development. [a]Student-level g(SE)s were corrected for clustering at the school level using the method described in Hedges (2007).

**Table 8.** Gates-MacGinitie Reading Test – reading comprehension model results.

| | Reading comprehension |
|---|---|
| Pretest | |
| Level 1 ($CWC_{Students}$) | 0.70 (0.02) |
| Level 1 ($CWC_{Teachers}$) | 0.65 (0.10) |
| Level 2 (GMC) | 0.89 (0.07) |
| Coeff. (*SE*) | |
| SPD | 3.39 (3.23) |
| RPD | 4.92 (3.35) |
| *p* | |
| SPD | 0.30 |
| RPD | 0.14 |
| Intercept (*SE*) | 483.07 (2.59) |
| Level 2 variance | 48.42 |
| Level 2 ICC | 0.06 |
| Residual | 719.03 |
| *N* students | 2167 |
| *N* teachers | 84 |
| *N* schools | 37 |
| School-level *g* (*SE*)[a] | |
| SPD | 0.13 (0.45) |
| RPD | 0.24 (0.49) |
| Student-level *g* (*SE*)[a] | |
| SPD | 0.04 (0.12) |
| RPD | 0.06 (0.12) |

*Note*: CWC: centered within clusters; GMC: grand mean centered; SE: standard error; SPD: school supported professional development; RPD: researcher supported professional development.
[a]Student-level g(SE)s were corrected for clustering at the school level using the method described in Hedges (2007).

significant or meaningful differences between cohorts on pretest or post-test measures. We worked in the same region of Texas and in some of the same districts in each cohort to minimize demographic differences across years. Thus, we did not include cohort effects in our analytical models.

### Content Knowledge

Students in the SPD condition ($\gamma_{001}$ = 8.90, $SE$ = 1.32, $p$ <.001) and the RPD condition ($\gamma_{002}$ = 8.57, $SE$ = 1.20, $p$ <.001) scored significantly higher on the unit test of content knowledge than BAU. The school-level effect size for the difference between the RPD condition and the BAU condition was $g$ = 1.42 ($SE$ = 0.22; 95% CI [0.98, 1.86]); the student-level effect size was $g$ = 0.55 ($SE$ = 0.09; 95% CI [0.38, 0.73]). For the comparison of the SPD condition and BAU, the school-level effect size was $g$ = 1.30 ($SE$ = 0.21; 95% CI [0.88, 1.72]); the student-level effect size was $g$ = 0.51 ($SE$ = 0.08; 95% CI [0.35, 0.67]).

### Vocabulary

Students in the RPD condition ($\gamma_{001}$ = 12.54, $SE$ = 1.88, $p$ <.001) and the SPD condition ($\gamma_{002}$ = 12.35, $SE$ = 1.99, $p$ <.001) scored significantly higher on the unit test of content vocabulary than students in the BAU. The school-level effect size for the comparison of RPD and BAU was $g$ = 1.32 ($SE$ = 0.22, 95% CI [0.89, 1.75]); the student-level effect size was $g$ = 0.49 ($SE$ = 0.08, 95% CI [0.33, 0.65]). The school-level effect size for the comparison of school-provided PD and BAU was similar in magnitude, $g$ = 1.19 ($SE$ = 0.21, 95% CI [0.78, 1.61]) as was the student-level effect size of $g$ = 0.49 ($SE$ = 0.09, 95% CI [0.32, 0.66]).

Scores on the GMRV did differ significantly between schools in the SPD condition and those in the BAU condition ($\gamma_{002}$ = 5.03, $SE$ = 1.52, $p$ = .001). The school-level effect size was $g$ = 0.50, $SE$ = 0.20, 95% CI [0.11, 0.88]) and the student-level effect size was $g$ = 0.07, $SE$ = 0.03, 95% CI [0.02, 0.12]). However, students in the RPD condition did not score significantly higher than those in BAU ($\gamma_{001}$ = 2.26, $SE$ = 1.81, $p$ = .21; $g$ [school level] = 0.21, $SE$ = 0.20, 95% CI [−0.18, 0.61]; $g$ [student level] = 0.03, $SE$ = 0.03, 95% CI [−0.03, 0.09]).

### Reading Comprehension

Scores on content reading comprehension showed significant differences between students in SPD and those in BAU schools ($\gamma_{002}$ = 1.74, $SE$ = 0.52, $p$ = .001, $g$ [school level] = 0.65, $SE$ = 0.20, 95% CI [0.26, 1.04]; $g$ [student level] = 0.26, $SE$ = 0.08, 95% CI [0.11, 0.42]). Students in schools in the RPD also scored significantly higher than those in comparison schools ($\gamma_{001}$ = 1.14, $SE$ = 0.53, $p$ =.032, $g$ [school level] = 0.43, $SE$ = 0.20, 95% CI [0.03, 0.82]; $g$ [student level] = 0.16, $SE$ = 0.08, 95% CI [0.01, 0.32]). No statistically significant differences were found between either STRIVE condition and BAU for the GMRC. Because the analytical model for the GMRC involved a subset of schools and pretest means could not be included at the teacher level, the school-level

effect sizes were estimated at level 2 of the model. For the comparison of the SPD and BAU conditions, the school-level effect size was $g = 0.13$ ($SE = 0.45$, 95% CI [−0.75, 1.02]); the student-level effect size was $g = 0.04$ ($SE = 0.12$, 95% CI [-0.20, 0.27]). The school-level effect size for the comparison of the RPD and BAU was $g = 0.24$ ($SE = 0.49$, 95% CI [−0.18, 0.30]); the student-level effect size was $g = 0.06$ ($SE = 0.12$, 95% CI [−0.17, 0.29]).

### Mediation Analyses

We estimated models in which factor scores for fidelity mediated the effect of assignment to condition on student outcomes when statistically significant main effects were found (i.e., all of the unit content tests and the GMRV for the SPD condition compared to BAU instruction). Both direct effects and indirect effects were included in these models. Statistically significant path coefficients for indirect effects would indicate that fidelity mediates the relationship between assignment to condition and outcome. Direct effects remained statistically significant, as would be expected. However, in all models the coefficients for the indirect effect were found to not differ significantly from 0. See Table 9 for the model parameters for the unit content tests and Table 10 for the model parameters for the GMRV.

## Discussion

This RCT examined the efficacy of STRIVE PD—a distributed PD model featuring a set of vocabulary and reading comprehension instructional practices—on student content learning and reading outcomes. The PD model features ongoing PD that highlights collaboration among fourth grade teachers toward common instructional and

**Table 9.** Unit tests mediation model results.

| | Content knowledge | | Content vocabulary | | Content reading comprehension | |
|---|---|---|---|---|---|---|
| | Coeff. (SE) | *p* | Coeff. (SE) | *p* | Coeff. (SE) | *p* |
| $\text{Intercept}_{\text{Outcome}}$ | 18.34 (2.55) | <.001 | 22.84 (3.89) | <.001 | 6.00 (1.11) | <.001 |
| $\text{Intercept}_{\bar{Fidelity}}$ | −1.25 (0.04) | <.001 | −1.25 (0.04) | <.001 | −1.25 (0.04) | <.001 |
| $\text{Fidelity}_{L2}$→Outcome | 0.58 (0.89) | .52 | −0.70 (1.45) | .63 | −0.09 (0.46) | .85 |
| SPD →*Fidelity* ($a_1$) | 1.33 (0.06) | <.001 | 1.33 (0.06) | <.001 | 1.33 (0.06) | <.001 |
| RPD →*Fidelity* ($a_2$) | 1.22 (0.07) | <.001 | 1.22 (0.07) | <.001 | 1.21(0.07) | <.001 |
| *idelity*→Outcome (b) | −1.61 (2.23) | .47 | −1.05 (3.51) | .77 | 0.03 (0.93) | .98 |
| SPD→Outcome ($c_1$) | 10.65 (2.91) | <.001 | 15.08 (4.30) | <.001 | 1.97 (1.20) | .10 |
| RPD→Outcome ($c_2$) | 10.15 (3.07) | .001 | 14.92 (4.68) | .001 | 1.26 (1.33) | .35 |
| Indirect Effect ($a \times b$) | | | | | | |
| INDB_SPD ($a_1 \times b$) | −1.96 (2.68) | .47 | −1.27 (4.25) | .77 | 0.03 (1.13) | .98 |
| INDB_RPD ($a_2 \times b$) | −2.15 (2.98) | .47 | −1.40 (4.68) | .77 | 0.04 (1.24) | .98 |
| Level 2 variance | 7.45 | | 17.21 | | 0.83 | |
| Level 2 ICC | 0.07 | | 0.07 | | 0.04 | |
| Level 3 variance | 17.43 | | 39.59 | | 2.85 | |
| Level 3 ICC | 0.16 | | 0.16 | | 0.14 | |
| Residual | 85.90 | | 196.56 | | 16.22 | |
| *N* students | 4555 | | 4555 | | 4555 | |
| *N* teachers | 208 | | 208 | | 208 | |
| *N* schools | 77 | | 77 | | 77 | |

*Note*. *SE*: standard error; SPD: school supported professional development; RPD: researcher supported professional development.

**Table 10.** Gates-MacGinitie Reading Test mediation model results.

| | Vocabulary | |
|---|---|---|
| | Coeff. (SE) | *p* |
| $Intercept_{Outcome}$ | 468.78 (3.82) | <.001 |
| $Intercept_{Fidelity}$ | −1.25 (0.04) | <.001 |
| Level 1 (CWC) | 0.81 (0.02) | <.001 |
| Level 2 (CWC) | 0.96 (0.09) | <.001 |
| Level 3 (GMC) | 1.10 (0.05) | <.001 |
| $Fidelity_{L2}$→Outcome | 1.35 (2.12) | .52 |
| SPD →*Fidelity* (a) | 1.22 (0.07) | <.001 |
| *Fidelity*→Outcome (b) | −2.08 (3.63) | .57 |
| SPD→Outcome (c) | 5.51 (4.38) | .21 |
| Indirect Effect ($a \times b$) | | |
| INDB_SPD ($a \times b$) | −2.52 (4.40) | .57 |
| Level 2 variance | 25.60 | |
| Level 2 ICC | 0.04 | |
| Level 3 Variance | 9.16 | |
| Level 3 ICC | 0.02 | |
| Residual | 546.22 | |
| N Students | 2623 | |
| N Teachers | 142 | |
| N Schools | 53 | |

*Note*. CWC: centered within clusters; GMC: grand mean centered; *SE*: standard error; SPD: school supported professional development.

implementations. We also examined the mediating role of fidelity in order to understand the causal chain from PD to fidelity to student outcomes.

Results from the study indicated statistically significant differences for students on measures of content knowledge ($g = 0.51$, 0.55), content vocabulary ($g = 0.49$), and a distal measure of vocabulary ($g = 0.03$, 0.07) when their teacher participated in STRIVE PD. Statistically significant effects were also detected in favor of the SPD group on a mid-transfer measure of content reading comprehension ($g = 0.26$). Positive effects were detected in favor of the treatment groups on a distal measure of reading comprehension ($g = 0.04$, 0.06), but the differences were not statistically significant. These effects are particularly meaningful because all students in all conditions received the same social studies curriculum and were taught the same concepts using the same timeline. The only difference between the treatment and BAU groups were the use of the vocabulary and reading comprehension instructional practices teachers learned during STRIVE PD.

Findings compare favorably to those in a meta-analysis (Basma & Savage, 2018) on professional development reporting a significant effect of 0.22 across a range of reading outcomes. Effects also compare favorably to the magnitude of the effects on content knowledge outcomes (ES = 0.17–0.40) and content reading comprehension (ES = 0.20–0.29) reported by Vaughn et al. (Vaughn, Swanson et al., 2013; Vaughn et al., 2017, 2019) who investigated a similar set of instructional practices among middle-grade students. Reading comprehension findings generally align with recent PD study results at the upper elementary level that detected no statistically significant group differences (Porche et al., 2012; Van Keer & Verhaeghe, 2005). The collaborative and distributed nature of STRIVE PD may contribute to STRIVE PD's favorable outcomes relative to prior studies. In general, studies that use more collaborative PD methods reported larger effects than studies using a more traditional one-stop PD model (e.g. Amendum, 2014; Fine & Kossack, 2002; Klingner et al., 2004). Findings provide additional evidence that

the use of teacher study teams can influence student reading outcomes in a meaningful way.

### PD Featuring Teacher Ownership

We also examined the efficacy of differing PD conditions that purposefully shift ownership to teachers (Coburn, 2003). We did this by testing PD led by researchers (i.e., RPD), PD led by school leaders (i.e., SPD), and BAU. Findings indicate that students whose teachers participated in the RPD and the SPD conditions outperformed the BAU condition on measures of content knowledge and content vocabulary. In addition, students whose teachers participated in the SPD condition outperformed their BAU peers on measures of content reading comprehension and general vocabulary. This provides indication that PD models delivered through a more sustainable school-supported model are at least as effective as a researcher-supported model on some outcomes (e.g., content knowledge and vocabulary) and potentially more effective on others (e.g., content reading comprehension and general vocabulary).

Our study was not designed to identify exactly what features of the SPD condition contributed to the differences in student outcomes. However, we can hypothesize potential explanations for findings that may be examined in future studies. For example, one explanation may be related to facilitators to uptake of new practices. A top cited facilitator for high implementers in prior studies (e.g., Klingner et al., 2003) is support from university personnel to establish strong knowledge of the practices. This study provides causal indication that combining an initial PD session delivered by researchers with school support (i.e. the SPD condition) leads to positive student outcomes across a range of measures. Another explanation may be related to social networks created during teacher study teams (e.g. Coburn et al., 2012). Teachers' interactions with one another provide ongoing access to knowledge, feedback and social support that help teachers understand new approaches. Social networks also facilitate the persistence of practices as teachers navigate shifting school demands (Hargreaves & Goodson, 2006). Finally, prior work theorizes that transfer of ownership (i.e., SPD condition) may influence uptake and sustainability (Coburn, 2003). We could not locate any studies investigating the effect of PD models designed to transfer ownership to key stakeholders on student outcomes. In this way, the current study makes an important contribution to the literature base. These findings provide evidence that practices theorized to increase sustainability are *efficacious*, setting the stage for future research examining the *sustainability* of the practices over time.

### The Role of Fidelity

Our second research question focused on the mediating role of fidelity on student outcomes. Based on results from prior examinations (e.g., Vaughn, Roberts et al., 2013; Vaughn et al., 2015), we hypothesized that fidelity would at least partially mediate the path from PD to student outcomes. Results from this study do not confirm the hypothesis. While this finding does not align with evidence from several prior studies (Cantrell et al., 2013; Roberts et al., 2017; Unlu et al., 2016; Vaughn, Roberts et al., 2013; Vaughn

et al., 2015), it does align with others that did not detect evidence of a mediating role of fidelity on reading outcomes (e.g. Boardman et al., 2016; Neugebauer, 2016). In the current study, we defined fidelity as adherence to the STRIVE instructional practices as designed. In Van Dijk et al.'s (2019) recent review of studies that investigated the role of adherence in reading outcomes, 41% of the studies showed no relation, 29% were "unclear," 24% were positive and 3% showed a negative relation. Like others who reported similar findings (e.g. Neugebauer, 2016), we do not conclude that fidelity is an unimportant ingredient for influencing student reading outcomes. Instead, we suggest that a possible explanation of current findings may be related to the potential resilience of the instructional practice. In other words, perhaps it doesn't take much fidelity to the practices for students to do well, i.e., even a "little bit" of STRIVE accelerates students' performance on content knowledge, vocabulary, and content reading comprehension.

Alternatively, issues related to the measurement and modeling of fidelity may explain the outcome. The fidelity measure used in the current study aligns with fidelity measures from prior high-quality intervention research (e.g. Hairrell et al., 2011; Simmons et al., 2010; Vaughn, Roberts et al., 2013; Vaughn et al., 2015) some of which included examinations of the mediating role of fidelity (e.g. Vaughn, Roberts et al., 2013; Vaughn et al., 2015). Even with this alignment, there may be elements of fidelity specific to the STRIVE PD and lessons that *did* mediate findings and were not captured by the fidelity measure. Our measure focused on one commonly used feature of fidelity—adherence—that in other studies was shown to be a good indicator of fidelity (e.g., Fogarty et al., 2014; van Dijk et al., 2019). However, a quantitative measure of differentiation or dosage may have contributed to mediation results (van Dijk et al., 2019). Other examples may be how much social studies content was covered or the quality with which components were implemented. Another issue facing analysis in this, and prior studies is the reliability and validity of fidelity measures (Capin et al., 2018; Swanson et al., 2013). Measurement error in the mediating variable exacts substantial consequences by introducing noise into the fidelity data that obscures the ability to detect an effect. Although it is not included in power calculations, measurement error attenuates power by adding error variance to analytical models. For this reason, developing reliable and valid fidelity measures and procedures are paramount.

## Practice

Findings from this examination of STRIVE PD's efficacy have several implications for practice. Few PD studies examine the influence on student outcomes and when they do, relatively small effects are reported (Basma & Savage, 2018). This study is the first, school-level, fully powered study examining not only student literacy outcomes but also how student outcomes differ when ownership of PD activities is transferred to teachers. Most PD delivered in schools is done so through one-shot PD sessions. With mounting evidence that ongoing teacher support is necessary for change in classroom practice (Amendum, 2014; Fine & Kossack, 2002; Opfer et al., 2011), findings from this study add some nuance to prior recommendations. First, the content of the PD sessions should include activities that support an understanding of how to use lessons, procedures, and materials so the principles taught in PD can be readily

manifested in classroom teaching. Second, highly collaborative PD distributed over time was beneficial for teachers as they learn and implement new practices. Finally, collaboration can effectively be led by school leaders and teachers through the use of teacher study teams.

### *Limitations and Future Research*

District policy in one participating district did not permit administration of the general reading comprehension measure. Therefore, the sample used to examine the efficacy of STRIVE PD on general reading comprehension was under-powered resulting in possibly obscuring the true effect. There are challenges to examining fidelity as a mediator including ongoing issues with assuring reliability and validity of measurement that the field continues to grapple with solving. In particular, future research would benefit from a broader set of PD fidelity data. Additionally, findings from this study are limited to the target sample of 4th grade teachers and students from those classes. Questions remain about the efficacy of STRIVE in other settings, with other teachers and with other students (i.e., rural school settings). Findings also provide evidence that when the ownership of follow-up PD is shifted to teachers, student level effects are at least equally effective as compared with researcher-led follow up PD. This supports the promise of the theory that a shift in ownership leads to sustainability (Coburn, 2003). Future work can better examine the role of a shift in ownership of PD in sustainability.

## Conclusion

Findings contribute to knowledge of the efficacy of PD supporting a collaborative, distributed PD model based on sociocultural learning theory as effective in impacting fourth grade reading outcomes. Also, the study is among the first to examine the effect of PD on students' content knowledge, vocabulary, content reading comprehension, *and* general reading comprehension. Findings support PD for teachers that includes lessons, procedures, and materials within collaborative PD sessions distributed over time.

## Open Research Statements

This manuscript was not required to disclose open research practices, as it was initially submitted prior to JREE mandating open research statements in April 2022.

## Funding

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through [Grant R305A150407] to The University of Texas at Austin. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

## ORCID

Elizabeth Swanson http://orcid.org/0000-0002-2716-4078
Alicia A. Stewart http://orcid.org/0000-0001-6770-5046
Elizabeth A. Stevens http://orcid.org/0000-0002-8412-1111
Nancy K. Scammacca http://orcid.org/0000-0002-7484-5976
Philip Capin http://orcid.org/0000-0003-4955-9879
Bethany H. Bhat http://orcid.org/0000-0002-7330-8204
Greg Roberts http://orcid.org/0000-0001-6333-7442
Sharon Vaughn http://orcid.org/0000-0001-8305-5549

## Data Availability Statement

The data that support the findings of this study are available from the corresponding author, Elizabeth Swanson, upon reasonable request and the establishment of a data sharing agreement between the requestor and The University of Texas at Austin. The contents of this manuscript along with analysis and reporting on the fully powered sample has never been presented or published elsewhere.

## 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

Amendum, S. (2014). Embedded professional development and classroom-based early reading intervention: Early diagnostic reading intervention through coaching. *Reading & Writing Quarterly*, *30*(4), 348–377. https://doi.org/10.1080/10573569.2013.819181

Anders, P. L., Hoffman, J. V., & Duffy, G. G. (2000). Teaching teachers to teach reading: Paradigm shifts, persistent problems, and challenges. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. III, pp. 719–742). Lawrence Erlbaum Associates.

Basma, B., & Savage, R. (2018). Teacher professional development and student literacy growth: A systematic review and meta-analysis. *Educational Psychology Review*, *30*(2), 457–481. https://doi.org/10.1007/s10648-017-9416-4

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* (1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x/

Boardman, A. G., Buckley, P., Vaughn, S., Roberts, G., Scornavacco, K., & Klingner, J. K. (2016). Relationship between implementation of collaborative strategic reading and student outcomes for adolescents with disabilities. *Journal of Learning Disabilities*, *49*(6), 644–657. https://doi.org/10.1177/0022219416640784

Cantrell, S. C., Almasi, J. F., Carter, J. C., & Rintamaa, M. (2013). Reading intervention in middle and high schools: Implementation fidelity, teacher efficacy, and student achievement. *Reading Psychology*, *34*(1), 26–58. https://doi.org/10/1080/02702711.2011.577695

Capin, P., Walker, M. A., Vaughn, S., & Wanzek, J. (2018). Examining how treatment fidelity is supported, measured, and reported in K–3 reading intervention research. *Educational Psychology Review*, *30*, 1–35. https://doi.org/10.1007/s10648-017-9429-z

Clarke, D., & Hollingsworth, H. (2002). Elaborating a model of teacher professional growth. *Teaching and Teacher Education*, *18*(8), 947–967. https://doi.org/10.1016/S0742-051X(02)00053-7

Clemens, N. H., & Fuchs, D. (2021). Commercially developed tests of reading comprehension: Gold standard or fool's gold? *Reading Research Quarterly*, *0*(0), 1–13. https://doi.org/10.1002/rrq.415

Coburn, C. (2003). Rethinking scale: Moving beyond numbers to deep and lasting change. *Educational Researcher*, *32*(6), 3–12. https://doi.org/10.3102/0013189X032006003

Coburn, C. E., Russell, J. L., Kaufman, J. H., & Stein, M. K. (2012). Supporting sustainability: Teachers advice networks and ambitious instructional reform. *American Journal of Education*, *119*(1), 137–182. https://doi.org/10.1086/667699

Daniel, J., & Lemons, C. (2018). Teacher perspectives on intervention sustainability: implications for school leadership. *School Leadership & Management*, *38*(5), 518–538. https://doi.org/10.1080/13632434.2018.1439465

Darling-Hammond, L., Hyler, M. E., & Gardner, M. (2017). *Effective teacher professional development*. Learning Policy Institute. https://learningpolicyinstitute.org/product/teacher-prof-dev.

Datnow, A. (2002). Can we transplant educational reform, and does it last? *Journal of Educational Change*, *3*(3/4), 215–239. https://doi.org/10.1023/A:1021221627854

Duffy, G. G., Roehler, L. R., Meloth, M. S., Vavrus, L. G., Book, C., Putnam, J., & Wesselman, R. (1986). The relationship between explicit verbal explanations during reading skill instruction and student awareness and achievement: A study of reading teacher effects. *Reading Research Quarterly*, *21*(3), 237–252. https://doi.org/10.2307/747707

Espin, C. A., Busch, T. W., Shin, J., & Kruschwitz, R. (2001). Curriculum-based measurement in the content areas: Validity of vocabulary-matching as an indicator of performance in social studies. *Learning Disabilities Research & Practice*, *16*(3), 142–151. https://doi.org/10.1111/0938-8982.00015

Fine, J. C., & Kossack, S. W. (2002). The effect of using rubric-embedded cognitive coaching strategies to initiate learning conversations. *Journal of Reading Education*, *27*(2), 31–37.

Fogarty, M., Oslund, E., Simmons, D., Davis, J., Simmons, L., Anderson, L., Clemens, N., & Roberts, G. (2014). Examining the effectiveness of a multi component reading comprehension intervention in middle schools: A focus on treatment fidelity. *Educational Psychology Review*, *26*(3), 425–449. https://doi.org/10.1007/s10648-014-9270-6

Gajria, M., Jitendra, A. K., Sood, S., & Sacks, G. (2007). Improving comprehension of expository text in students with LD: A research synthesis. *Journal of Learning Disabilities*, *40*(3), 210–225. https://doi.org/10.1177/00222194070400030301

Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., Uekawa, K., Falk, A., Bloom, H., Doolittle, F., Zhu, P., Sztejnberg, L., & Silverback, M. (2008). *The impact of two professional development interventions on early reading instruction and achievement*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Garet, M. S., Heppen, J. B., Walters, K., Smith, T. M., & Yang, R. (2016). *Does content-focused teacher professional development work? Findings from three institute of education sciences studies*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Garet, M., Porter, A., Desimone, L., Birman, B., & Yoon, K. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal*, *38*(4), 915–945. https://doi.org/10.3102/00028312038004915

Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children*, *71*(2), 149–164. https://doi.org/10.1177/001440290507100202

Gwet, K. (2001). *Handbook of inter-rater reliability: How to estimate the level of agreement between two or multiple raters*. Advanced Analytics, LLC.

Hairrell, A., Rupley, W. H., Edmonds, M., Larsen, R., Simmons, D., Willson, V., Byrns, G., & Vaughn, S. (2011). Examining the impact of teacher quality on fourth-grade students' comprehension and content-area achievement. *Reading & Writing Quarterly*, *27*(3), 239–260. https://doi.org/10/1080/10573569.2011.560486

Hamre, B. K., Justice, L. M., Pianta, R. C., Kilday, C., Sweeney, B., Downer, J. T., & Leach, A. (2010). Implementation fidelity of MyTeachingPartner literacy and language activities: Association with preschoolers' language and literacy growth. *Early Childhood Research Quarterly*, *25*(3), 329–347. https://doi.org/10.1016/j.ecresq.2009.07.002

Hargreaves, A., & Goodson, I. (2006). Educational change over time? The sustainability and non-sustainability of three decades of secondary school change and continuity. *Educational Administration Quarterly*, *42*(1), 3–41. https://doi.org/10.1177/0013161X05277975

Hedges, L. V. (2007). Correcting a significance test for clustering. *Journal of Educational and Behavioral Statistics*, *32*(2), 151–179. https://doi.org/10.3102/1076998606298040

Hedges, L. V. (2011). Effect sizes in three-level cluster-randomized experiments. *Journal of Educational and Behavioral Statistics*, *36*(3), 346–380. https://doi.org/10.3102/1076998610376617

Hox, J. (2002). *Multilevel analysis techniques and applications/Joop Hox*. Lawrence Erlbaum Associates.

Institute of Education Sciences & National Science Foundation. (2013*). Common guidelines for education research and development*. https://ies.ed.gov/pdf/Commo nGuid elines.pdf.

Kang, B. (2013). Fast determinantal point process sampling with application to clustering. In *Advances in Neural Information Processing Systems, 26*.

Kintsch, W. (1988). The use of knowledge in discourse processing: A construction-integration model. *Psychological Review*, *95*(2), 163–182. https://doi.org/10.1037/0033-295X.95.2.163

Klingner, J. K., Ahwee, S., Pilonieta, P., & Menendez, R. (2003). Barriers and facilitators in scaling up research-based practices. *Exceptional Children*, *69*(4), 411–429. https://doi.org/10.1177/001440290306900402

Klingner, J. K., Vaughn, S., Arguelles, M. E., Tejero Hughes, M., & Ahwee Leftwich, S. (2004). Collaborative strategic reading: "real-world" lessons from classroom teachers. *Remedial and Special Education*, *25*(5), 291–302. https://doi.org/10.1177/07419325040250050301

Klingner, J., Vaughn, S., Boardman, A., & Swanson, E. (2012). *Now we get it! Boosting comprehension with collaborative strategic reading*. Jossey-Bass.

Kornbrot, D. (2021, June 15). *False discovery Benjamini-Hochberg*. https://dianakornbrot.wordpress.com/excel-statistics-spreadsheets/false-discovery-benjamini-hochberg/.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., Dreyer, L. G., & Hughes, K. E. (2000). *Gates-MacGinitie reading tests fourth edition*. Hougton Mifflin Harcourt.

Muthén, L. K., & Muthén, B. O. (1998–2019). *Mplus user's guide*. Author.

Ness, M. (2011). Teachers' use of and attitudes toward informational text in K–5 classrooms. *Reading Psychology*, *32*(1), 28–53. https://doi.org/10.1080/02702710903241322

Neugebauer, S. (2016). Stable or situated understandings of adolescent reading engagement across readers and raters. *The Journal of Educational Research*, *109*(4), 391–404. https://doi.org/10.1080/00220671.2014.968914

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

Opfer, V. D., Pedder, D. G., & Lavicza, Z. (2011). The role of teachers' orientation to learning in professional development and change: A national study of teachers in England. *Teaching and Teacher Education*. *27*(2), 443–453. https://doi.org/10.1016/j.tate.2010.09.014

Porche, M., Pallante, D., & Snow, C. (2012). Professional development for reading achievement. results from the collaborative language and literacy instruction project (CLLIP). *The Elementary School Journal*, *112*(4), 649–671. https://doi.org/10.1086/665008

Preacher, K., Zyphur, M., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, *15*(3), 209–233. https://doi.org/10.1037/a0020141

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Roberts, G., Lewis, N. S., Fall, A.-M., & Vaughn, S. (2017). Implementation fidelity: Examples from the reading for understanding initiative. In G. Roberts, S. Vaughn, S. N. Beretvas, &

Wong, V. (Eds.), *Treatment fidelity in studies of educational intervention* (pp. 61–79). Routledge.

Roberts, G., Scammacca, N., & Roberts, G. J. (2018). Causal mediation in educational intervention studies. *Behavioral Disorders*, *43*(4), 457–465. https://doi.org/10.1177/0198742917749560

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. http://www.jstatsoft.org/v48/i02/.

Simmons, D., Hairrell, A., Edmonds, M., Vaughn, S., Larsen, R., Willson, V., Rupley, W., & Byrns, G. (2010). A comparison of multiple-strategy methods: Effects on fourth-grade students' general and content-specific reading comprehension and vocabulary development. *Journal of Research on Educational Effectiveness*, *3*(2), 121–156. https://doi.org/10.1080/19345741003596890

Swanson, E., Hairrell, A., Kent, S., Ciullo, S., Wanzek, J. A., & Vaughn, S. (2014). A synthesis and meta-analysis of reading interventions using social studies content for students with learning disabilities. *Journal of Learning Disabilities*, *47*(2), 178–195. https://doi.org/10.1177/0022219412451131

Swanson, E., Wanzek, J., Haring, C., Ciullo, S., & McCulley, L. (2013). Intervention fidelity in special and general education research journals. *The Journal of Special Education*, *47*(1), 3–13. https://doi.org/10.1177/0022466911419516

TNTP. (2015). *The mirage: Confronting the hard truth about our quest for teacher development*. Authors.

Tyner, A., & Kabourek, S. (2020). *Social studies instruction and reading comprehension: Evidence from the early childhood longitudinal study*. Thomas B. Fordham Institute.

Unlu, F., Bozzi, L., Layzer, C., Smith, A., Price, C., & Hurtig, R. (2016). Linking implementation fidelity to impacts in an RCT. In G. Roberts, S. Vaughn, S. N. Beretvas, & V. Wong (Eds.), *Treatment fidelity in studies of educational intervention* (pp. 100–129). Routledge.

van Dijk, W., Lane, H., & Gage, N. A. (2019, October 7). The relation between implementation fidelity and students' reading outcomes: A systematic review of the literature. *Preprint EdArXiv*. https://doi.org/10.35542/osf.io/vhrp5

Van Keer, H., & Verhaeghe, J. P. (2005). Comparing two teacher development programs for innovating reading comprehension instruction with regard to teachers' experiences and student outcomes. *Teaching and Teacher Education*, *21*(5), 543–562. https://doi.org/10.1016/j.tate.2005.03.002

Vaughn, S., Fall, A.-M., Roberts, G., Wanzek, J., Swanson, E., & Martinez, L. R. (2019). Class Percentage of students with reading difficulties on content knowledge and comprehension. *Journal of Learning Disabilities*, *52*(2), 120–134. https://doi.org/10.1177/0022219418775117

Vaughn, S., Martinez, L. R., Wanzek, J., Roberts, G., Swanson, E., & Fall, A.-M. (2017). Improving content knowledge and comprehension for English language learners: Findings from a randomized control trial. *Journal of Educational Psychology*, *109*(1), 22–34. https://doi.org/10.1037/edu0000069.supp

Vaughn, S., Roberts, G., Swanson, E. A., Wanzek, J., Fall, A.-M., & Stillman-Spisak, S. J. (2015). Improving middle-school students' knowledge and comprehension in social studies: A replication. *Educational Psychology Review*, *27*(1), 31–50. https://doi.org/10.1007/s10648-014-9274-2

Vaughn, S., Roberts, G., Klingner, J. K., Swanson, E. A., Boardman, A., Stillman-Spisak, S. J., Mohammed, S. S., & Leroux, A. J. (2013). Collaborative strategic reading: Findings from experienced implementers. *Journal of Research on Educational Effectiveness*, *6*(2), 137–163. https://doi.org/10.1080/19345747.2012.741661

Vaughn, S., Swanson, E., Roberts, G., Wanzek, J., Stillman-Spisak, S., Solis, M., & Simmons, D. (2013). Improving reading comprehension and social studies knowledge in middle school. *Reading Research Quarterly*, *48*(1), 77–93. https://doi.org/10.1002/rrq.039

Venter, A., Maxwell, S., & Bolig, E. (2002). Power in randomized group comparisons: the value of adding a single intermediate time point to a traditional pretest-posttest design. *Psychological Methods*, *7*(2), 194–209. https://doi.org/10.1037//1082-989X.7.2.194

What Works Clearinghouse. (2020). *What Works Clearinghouse standards handbook, version 4.1*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Shapley, K. L. (2007). *Reviewing the evidence on how teacher professional development affects student achievement. Issues & Answers. REL 2007-No. 033*. Regional Educational Laboratory Southwest.