The Dissertation Committee for Steven Andrew Maddox Certifies

that this is the approved version of the following Dissertation:

Using Direct Observation Data to Explore the Impact of a Second-Grade Science Program on the Quality of Scientific Investigations and Science Discourse

Committee:

Christian T. Doabler, Supervisor

Sarah R. Powell

Jessica R. Toste

Gregory J. Roberts

Using Direct Observation Data to Explore the Impact of a Second-Grade Science Program on the Quality of Scientific Investigations and Science Discourse

by

Steven Andrew Maddox

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2022

Dedication

This dissertation is dedicated to my mother, Patricia Ann Schaefer Maddox, for the support and love that she has shown me not only throughout my doctoral program but throughout my life. Mom, thank you for everything that you do for me every day. You are my best friend, and I love you.

Acknowledgements

I would like to thank Dr. Christian Doabler for all of the support that he has provided me throughout my doctoral program. Additionally, I would like to thank Dr. Greg Roberts as well as SMARTER Consulting for the support provided in the data analysis portion of this dissertation. My thanks also go out to Drs. Jessica Toste and Sarah Powell for serving on my dissertation committee. Finally, I would like to recognize the agencies that made this dissertation possible – the National Science Foundation, The Meadows Center for Preventing Educational Risk, The University of Texas at Austin's Department of Special Education, the SAGA Lab, as well as The University of Virginia.

Abstract

Using Direct Observation Data to Explore the Impact of a Second-Grade Science Program on the Quality of Scientific Investigations and Science Discourse

Steven Andrew Maddox, Ph.D.

The University of Texas at Austin, 2022

Supervisor: Christian T. Doabler

In 2018, the National Science and Technology Council (2018) highlighted the importance of ensuring all students, including those with disabilities, reach proficiency in the disciplines of science, technology, engineering, and mathematics (STEM). This call is apropos as it comes at a time when a considerable number of U.S. students struggle to become STEM literate. Data from the 2019 NAEP fourth-grade science measure suggest only 36% of students scored at or above the *Proficiency* level (NCES, 2021). One viable solution for getting all students on track for success in science is to design effective early elementary science programs. In a recent study, Doabler et al. (2021) tested the efficacy of Sci2, a second-grade, Tier 1 program focused on the disciplinary core ideas, crosscutting concepts, and scientific practices associated with Earth's Systems in the Next Generation Science Standards (NGSS Lead States, 2013). Findings indicated Sci2 students significantly outperformed their peers in control classrooms on three of the four science outcome measures. Given these results, there is need to understand why the Sci2 program demonstrated

promise to improve science achievement. Therefore, the current study sought to unpack the effects of Sci2 by examining direct observation data collected in the 18 participating classrooms. This study used mean observation item scores and independent samples *t*-tests to address four research questions: (1) To what extent do pre-randomly assigned teachers differ at baseline in terms of the quality of scientific investigations and science discourse facilitated during science instruction; (2) To what extent do treatment (Sci2) and control teachers differ during the treatment time period in terms of the quality of scientific investigations and science discourse facilitated during science instruction; (3) To what extent do treatment (Sci2) and control teachers differ in terms of their capacity to maintain high-quality scientific investigations and science discourse after the treatment time period; and (4) To what extent does the quality of scientific investigations and science discourse observed in the treatment time period maintain from the end of treatment to the maintenance time period? Implications will also be discussed.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

IMPORTANCE OF STEM LITERACY

Recent projections from the U.S. Bureau of Labor Statistics (2021) estimate that occupations in the science, technology, engineering and mathematics (STEM) fields will be among the fastest growing occupations through at least the year 2030. For example, some of the most in-demand jobs will be those in the healthcare industry (e.g., registered nurses, certified nurse's assistants) due, at least in part, to an aging population that will require such services, among other factors (U.S. Bureau of Labor Statistics, 2021). To further illustrate the projected growth in the STEM fields, it is anticipated that computer-related skills (e.g., Information Technology) will also be of dire need in the future due to the large number of employees continuing to telecommute to work as a result of the COVID-19 pandemic (U.S. Bureau of Labor Statistics, 2021). These rising demands for a STEM-trained workforce spotlight the critical need to ensure *all* students, including students with or at risk for learning disabilities (LD), receive high-quality, effective STEM education as early as the early elementary grades (i.e., kindergarten to second grade).

CURRENT STATUS OF U.S. SCIENCE LEARNING

Despite the importance of providing all students with high-quality, effective STEM education, recent data suggests the field is struggling to meet this goal. For example, on the 2019 administration of the National Assessment of Educational Progress (NAEP) science assessment, only 36% of fourth-grade students scored high enough to be determined *Proficient* (National Center for Educational Statistics [NCES], 2021). This was a decrease from the previous NAEP science administration in 2015 where 38% of fourth-grade students reached a *Proficient* level (NCES, 2021). The NCES (2021) reports similar cross-year NAEP-Science results for fourth-grade students with disabilities. In 2019, 15% of fourth-grade students with disabilities were

considered *Proficient* on the NAEP-Science assessment, compared to 18% in 2015. Collectively, these NAEP data suggest the enacted science curriculum in U.S. schools may be falling short of helping all students gain STEM proficiency.

When comparing the science achievement of U.S. students to their international peers, data suggests that the STEM instruction provided in U.S. classrooms is inferior to many nations. For example, in 2019, 64 countries administered the Trends in Mathematics and Science Study (TIMSS) assessment to fourth-grade students (TIMSS, 2019). Relative to their international fourth-grade peers, students in the U.S. ranked just 10th in overall science achievement. Together, findings from the TIMSS and the NAEP science assessments showcase the urgent need for U.S. classrooms, particularly early elementary classrooms, to deliver high-quality science instruction.

MANIFESTATIONS OF LEARNING DIFFICULTIES IN SCIENCE

While standardized outcome measures like the NAEP and TIMSS provide valuable snapshots into the current science performance of U.S. students, their administrations do not begin until fourth grade. The late timing of these outcome measures in the upper elementary grades may divert needed attention to delivering effective science instruction in early elementary classrooms. Compelling research suggests students' foundational knowledge around science ideas, concepts, and practices in the early elementary grades is critical for later science learning.

Morgan et al. (2016), for example, analyzed the Early Childhood Longitudinal Study, Kindergarten Class of 1998-1999 (ECLS-K) dataset with the primary goal of determining the extent to which achievement gaps in the discipline of science exist for students in kindergarten and how these gaps maintain over time. The data suggested the presence of significant achievement gaps in students' science-related background knowledge depending on students' socio-demographic factors. For example, African-American students entering kindergarten were found

to score 0.62 standard deviations lower in terms of science background knowledge when compared to their White peers. Additionally, Morgan and colleagues (2016) noted this gap in kindergarten was predictive of differences in science achievement through eighth grade. While the findings of Morgan et al. (2016) paint a bleak picture in terms of the long-term projections of students who struggle early in science, recent research has produced promising findings that it is plausible to get early elementary students on track for success in science by immersing them in well-designed science instruction.

To illustrate this, Kim et al. (2021) conducted a cluster randomized controlled trial (RCT) to determine the extent to which the Model of Reading Engagement (MORE) intervention improved the science knowledge of 674 first-grade students from 38 classrooms. Approximately 21% of students were considered English learners and 7% were eligible to receive special education services. Through the use of evidence-based practices aimed at fostering reading engagement (e.g., opportunities for discussion) and science content knowledge (e.g., concept mapping), Kim and colleagues (2021) found that participants who received instruction from the MORE program significantly outscored their peers in the control condition on science-related measures of vocabulary, listening comprehension, and argumentative writing. This finding that content literacy instruction can positively impact younger students' science achievement is important given the scant research available supporting the use of content literacy instruction with students in early elementary grades. As noted by Kim et al. (2021), much of the research conducted to date has focused on the academic achievement (e.g., science, social studies, and reading) of students in the upper grades. Research on science-focused content literacy instruction commonly does not target younger students because they lack the foundational literacy skills necessary to gain content knowledge from reading expository text (Williams et al., 2016). However, integrating

certain reading strategies (e.g., repeated exposure to vocabulary, concept mapping) into content literacy instruction may augment foundational deficits and lead to increased comprehension in more technical content areas (i.e., science; Kim et al., 2021).

In another recent study, Doabler et al. (2021) conducted a cluster RCT, investigating the initial efficacy of the Scientific Explorers (Sci2) program on second-grade students' science knowledge. Participating in the study were 294 second-grade students across 18 classrooms. Of the participating students, 3.5% were considered English learners and 22.2% were eligible to receive special education services. The Sci2 program is a scripted science program that focuses on the disciplinary core ideas, crosscutting concepts, and scientific practices associated with Earth's Systems and identified by the NGSS Lead States (2013). Centered on principles of guided inquiry science instruction (Coyne et al., 2011; Therrien et al., 2017), the Sci2 program uses a systematic, explicit approach to support teachers in facilitating hands-on and simulation-based scientific investigations and science discourse in whole-class settings. Such instruction is geared toward students' understanding of the disciplinary core ideas, crosscutting concepts, and scientific practices associated with Earth's Systems and identified by the NGSS Lead States (2013). Findings suggested that students who received Sci2 instruction in the treatment classrooms significantly outperformed their peers in control classrooms on measures of (a) vocabulary, (b) the NGSS science and engineering practices, and (c) NGSS core disciplinary ideas. Additionally, results indicated the Sci2 program was efficacious for all students, regardless of students' background knowledge in science prior to participating in Sci2 instruction. Findings from these two studies (i.e., Kim et al., 2021; Doabler et al., 2021) suggest well-designed, early elementary science programs are beginning to show promise for improving the science achievement of a diverse range of students.

Early Elementary Science Instructional Materials

Similar to science assessments not being administered until the late elementary grades (e.g., NAEP, TIMSS), there is further evidence that science learning in the early grades is not a priority for many schools or teachers. For example, the What Works Clearinghouse (WWC) provides a total of 462 academic and behavioral programs and interventions for K-12 students (WWC, n.d.). Of these, there are only 14 programs or interventions that focus on science (WWC, n.d.). Furthermore, the only program with potentially positive effects for early elementary students' science learning is Teach for America (TFA; WWC, n.d.), which in and of itself should not necessarily be considered an intervention. Given early elementary teachers' lack of access to high-quality evidence-based science programs and recognizing that early achievement gaps in the discipline of science can persist for many years, there is an urgent need to provide teachers with science programs or interventions that have been found to be effective for all learners.

Purpose of the Current Study

As demonstrated in Kim et al. (2021) and Doabler et al. (2021), a primary aim of science intervention research is to investigate the effects of science programs on student science achievement. However, alone, information on treatment effects is insufficient for ascertaining why science programs successfully increase students' science achievement (or not). Therefore, it is important to explore other data sources that may help the field understand why science programs work, for whom, and under which conditions.

Against that backdrop, this study sought to conduct a secondary analysis of extant observation data collected in Doabler et al. (2021). This work sought to extend the Sci2 program of research by exploring whether and to what extent teachers' instructional practices in science (i.e., use of scientific investigations and facilitation of science discourse) were impacted by the

implementation of the Sci2 program, and whether observed changes were maintained over time. Recent observation research suggests that systematic, explicit programs can improve the quantity and quality of instruction delivered by teachers in core reading (Nelson-Walker et al., 2013) and early mathematics (Doabler et al., 2018) classrooms. Since the Sci2 program is grounded in systematic, explicit instruction and thus provides scripted lessons to support teachers in facilitating high-quality scientific investigations and science discourse, it was hypothesized that the findings reported in Nelson-Walker et al. (2013) and Doabler et al. (2018) would replicate in the current study. That is, a systematically designed and explicitly delivered science program would have the capacity to improve the quality of teaching practices delivered in core second-grade science instruction. More specifically, it was hypothesized that Sci2 teachers would facilitate higher quality scientific investigations and science discourse during their science instruction than control teachers. Additionally, it was anticipated that the program would support treatment teachers in maintaining those effects after the treatment time period and the conclusion of its implementation support.

Research Questions

The current study sought to test these hypotheses by addressing four research questions. Figure 1 provides a visual timestamp for when the research questions were situated within the lifespan of the Sci2 efficacy trial (Doabler et al., 2021).

1. To what extent do pre-randomly assigned teachers differ at baseline in terms of the quality of scientific investigations and science discourse facilitated during science instruction?
2. To what extent do treatment (Sci2) and control teachers differ during the treatment time period in terms of the quality of scientific investigations and science discourse facilitated during science instruction?

3. To what extent do treatment (Sci2) and control teachers differ in terms of their capacity to maintain high-quality scientific investigations and science discourse after the treatment time period?
4. To what extent does the quality of scientific investigations and science discourse observed in the treatment time period maintain from the end of treatment to the maintenance time period?

Chapter 2: Literature Review

The purpose of this chapter is to operationally define the two science teaching practices targeted for investigation in the current study and review the relevant existing literature surrounding them. Specifically, the current study focused on teachers' use of high-quality scientific investigations and facilitation of science discourse opportunities during second-grade science instruction. It was anticipated that providing teachers with a purposefully-designed, second-grade science program along with corresponding professional development (PD) opportunities would improve the quality of their science teaching practices and allow them to maintain the quality of such practices across time. This chapter begins with a review of the extant literature on early elementary science programs (i.e., kindergarten to second grade) and concludes with a discussion of how scientific investigations and science discourse opportunities represent the cornerstone of high-quality science instruction and learning.

Elementary Science Programs

Recognizing the importance of science, it is important that all early elementary students, including students with disabilities, have the opportunity to become literate in science. Part of achieving the goal of science proficiency for *all*, is ensuring that teachers have access to well-designed science programs that are supported by scientific evidence, suggesting they meet the instructional needs of the full range of learners. Unfortunately, there is a current dearth of validated science programs available for use in early elementary classrooms (WWC, n.d.).

To emphasize this dearth of available programs, of the 462 K-12 academic interventions and programs reviewed by the What Works Clearinghouse (WWC), only 14 focus on the discipline of science (WWC, n.d.). Of these 14 science programs, regrettably, only one (i.e., Teach for America) focused on the early elementary grades (i.e., kindergarten to second grade) and had

positive effects. This paucity raises the question of why more validated, early elementary science programs are not available. It is plausible that the timing of high stakes assessments is a contributing factor. In most states, including Texas (Texas Education Agency, n.d.), science is not first assessed until fourth and fifth grades. Regardless, there is an urgent need to design and disseminate strong early elementary science programs. This is especially true given the results of Morgan et al. (2016) – that achievement gaps in science background knowledge can be seen in kindergarteners and may persist for years.

More specifically, early elementary science programs need to emphasize the practices outlined in the NGSS. Recognizing that K-12 students did not have the skills necessary to take on STEM-related jobs after graduation, the NRC outlined a general set of knowledge and skills that students would need to achieve a basic literacy in STEM – namely, knowledge surrounding physical, life, Earth, and space science, as well as skills related to engineering and technology (NRC, 2012). Eventually, this broad knowledge and skill set would be crafted into the standards seen in the NGSS (Osborne & Quinn, 2017). Early elementary students in particular (i.e., kindergarten to second grade) are expected to engage in some of the same activities as their older peers – such as asking questions, creating models, and supporting an argument using evidence, among others – but at a more novice level (NGSS, n.d.).

Inclusion of Students with Disabilities

This need to ensure the availability of strong early elementary science programs is underscored by the increasing rate at which students with disabilities are receiving instruction in general education classrooms. Fifty years ago, approximately 20% of students with disabilities received services in the public school system (U.S. Department of Education, 2010). However, that figure has significantly increased. Recent estimates suggest that nearly 95% of students with

disabilities are educated in the general education classrooms (NCES, 2020). For students with LD, findings suggest approximately 72% receive their instruction in a general education setting (NCES, 2020). Therefore, it is critical that high-quality evidence-based practices focused on students with disabilities (and specifically LD) become incorporated more into general education settings to ensure that *all* students are able to access the content being taught for each subject, including science.

This improvement in services for students with disabilities can be directly tied to a number of acts passed by lawmakers, key among them being the Individuals with Disabilities Education Improvement Act (IDEIA; 2004). The IDEIA (2004), along with similar laws, ensured that students with disabilities were able to receive a "free and public education in the least restrictive environment to the maximum extent possible" (Kart & Kart, 2021, p. 2). As noted by the NCES (2020), for many students with disabilities (especially LD), this means spending a considerable amount of time in general education classrooms. However, for inclusion to be successful, general education teachers – not just special education teachers – need to have a working knowledge of evidence-based strategies for teaching students with disabilities (Hornby, 2015). It is for this reason that a more earnest effort needs to be made to incorporate what are typically considered special education strategies into the Tier 1 setting. This is especially true for the discipline of science, given the low rate at which students with disabilities are considered to be *Proficient* in the subject (NCES, 2021).

Research on Early Elementary Science Instruction

Encouragingly, researchers have begun to embrace the importance of bringing well-designed science instruction (i.e., teaching that incorporates evidence-based strategies) into the early elementary grades. For example, Wells et al. (2015) conducted a multi-state, cluster RCT,

randomly assigning 49 schools to either receive a school garden intervention or a waitlist control condition that delivered typical classroom science instruction. The purpose of the study was to determine whether students' science achievement in terms of knowledge surrounding plants and nutrition increased through the implementation of the school garden intervention. The intervention, which consisted of weekly one-hour sessions, occurred in 151 classrooms with 3,061 students in second- through fifth-grade. Students learned specifically about concepts in nutrition, horticulture, and plant science. Results suggested that students that received the school garden curriculum scored significantly better than their peers in the control condition on a measure of nutrition and plant science knowledge with a mean difference score of 0.47.

More recently, Kim et al. (2021) conducted a cluster RCT to investigate the effects that their MORE (i.e., Model of Reading Engagement) intervention might have on 674 first-grade students' content knowledge in the discipline of science. Specifically, this science-content literacy program utilized ten 60-minute lessons to address students' knowledge around animals that live in the Arctic and how they have adapted to survive in such a harsh environment. Results suggested that, when compared to their peers in the control condition, treatment students who were in classrooms that used the MORE program scored significantly better on measures of science vocabulary (Hedges' $g = 0.30$), listening comprehension (Hedges' $g = 0.40$), argumentative writing (Hedges' $g = 0.24$), as well as reading comprehension (Hedges' $g = 0.11$).

In another cluster RCT, Doabler et al. (2021) investigated the effects of the Sci2 program on 294 second-grade students' science achievement across 18 classrooms. Over the course of two weeks, treatment teachers (i.e., Sci2) delivered 10 lessons that focused on the core disciplinary ideas and crosscutting concepts associated with Earth's Systems that were outlined by the NGSS Lead States (2013). More specifically, lessons were centered around the role that both wind and

water play in shaping Earth's landforms, as well as the impact that these changes can have on both humans and animals. For each lesson, students were presented with requisite background knowledge before taking part in activities where they were asked to "act as" scientists by designing solutions for various scenarios and then discuss their findings. Students in control classrooms received science instruction through the use of both district-developed materials and a science program available on the commercial market. Similar to other studies based in early-elementary classrooms (e.g., Wells et al., 2015; Kim et al., 2021), overall promising effects for the Sci2 program were reported. Compared to their peers in the control condition, students in Sci2 classrooms demonstrated stronger outcomes on three science measures than their peers in control classrooms, with effects ranging from 0.48 to 0.94 (Hedges' *g*).

Encouragingly, results from Wells et al. (2015), Kim et al. (2021), and Doabler et al. (2021) highlight the promise of early elementary science programs to improve student science outcomes. However, questions remain as to why and under what conditions these programs were found to work. One plausible reason is that the act of providing teachers with a well-designed science program and ongoing PD likely supported teachers' use of evidence-based practices during their science instruction. For example, by design, the Sci2 program incorporates daily scientific investigations which thereby are anticipated to spark students' meaningful engagement in science discourse. As such, teachers' use of the Sci2 program and their receipt of the program's corresponding PD supports may have played a contributing role in improving student science achievement.

Therefore, the purpose of the current research was to investigate whether and to what extent the Sci2 program improved the quality of the science instructional practices provided in treatment classrooms and supported the Sci2 teachers in maintaining the quality of such practices after the

conclusion of the initial efficacy trial conducted by Doabler et al. (2021). Specifically, the current study sought to shed light on Sci2 teachers' facilitation of high-quality scientific investigations and science discourse opportunities relative to teachers in control classrooms. The following section briefly summarizes the literature behind these two science teaching practices (i.e., scientific investigations and science discourse) and discusses their relevancy to the current work.

QUALITY OF SCIENCE INSTRUCTION

One of the foundational tenets of being a "professional" scientist is to be able to "understand the core principles and theoretical constructs of their field" and to be able to use that knowledge "to make sense of new information or tackle novel problems" (NRC, 2012, p. 25). In this pursuit, scientists engage in scientific practices that permit observation of natural phenomena, collection of critical data, and the communication and dissemination of their findings. One way to immerse early elementary students in the practices of science is through well-designed scientific investigations where students explore natural phenomena, collect data, and disseminate their findings through meaningful science discourse.

Scientific Investigations

In the current study, scientific investigations are operationally defined as hands-on and/or simulation-based experiments that are conducted with the goal of having students explore, understand, and explain natural phenomena (NRC, 2012). It is important to note that, similar to the work of professional scientists, scientific investigations in early elementary grades should employ an iterative process. In this way, students receive opportunities to extend their understandings developed in prior investigations to future ones (NRC, 2012; Windschitl, 2017).

In the early elementary grades, investigations, whether they be hands-on or simulation-based, should be purposefully designed to offer students the opportunity to develop foundational

knowledge and build fluency with the science and engineering practices identified in the NGSS Lead States (2013). These practices include students' ability to: (a) ask questions in the discipline of science and/or define problems in the discipline of engineering, (b) create and use models, (c) conduct investigations, (d) critically examine data, (e) think computationally and/or mathematically, (f) generate answers for science questions and/or develop solutions for engineering problems, (g) engage in science discourse, and (h) disseminate findings (NRC, 2012). As such, students need to have opportunities to conduct investigations where they deeply engage in these NGSS practices.

To illustrate student engagement, in a quasi-experimental study, Samarapungavan et al., (2008) explored the extent to which kindergarten students investigated features of monarch butterflies. One-hundred students from six classrooms either received science instruction using the Scientific Literacy Project (i.e., treatment) or no science instruction at all (i.e., comparison). Lessons in the treatment condition fostered opportunities for students to learn about entomologists and the tools and practices that these professionals use in the field. Specifically, students received opportunities ask questions, collect data, and engage in science discourse. Following completion of the intervention, Samarapungavan and colleagues (2008) administered a researcher-developed outcome measure called the Science Learning Assessment to students in both conditions. Using an ANOVA, they found that students in the treatment condition significantly outperformed their control peers, $F(1, 98) = 44.10$, $p < 0.01$, with mean scores of 16.91 and 12.03, respectively, and a Hedges' g effect size of 1.39 (Social Science Statistics, 2022).

In situations where it may not be possible to observe and investigate natural phenomena in authentic settings due to safety and financial reasons, simulation-based models may be more appropriate (Li & Tsai, 2013; Rosenbaum et al., 2007; Squire & Klopfer, 2007). For example,

students might utilize a technology-based application to manipulate different variables that contribute to avalanches (e.g., amount of snow and wind) and investigate the extent to which these adjustments affect the strength and direction of an avalanche. Regardless of whether students engage in hands-on or simulation-based investigations, if designed and delivered well, both offer valuable experiences for students to work with the scientific practices.

Science Discourse

In the current work, science discourse is operationally defied as a verbal interaction among scientists, where an observation of a targeted phenomenon is initially presented and then that explanation or claim is subsequently scrutinized (NRC, 2012; Schwarz et al., 2017). This back-and-forth discussion is critical as it can either build support for a claim or disconfirm it (Berland et al., 2017). In short, science discourse can be conceptualized as an argumentative vetting process that engages the research community.

While science discourse is considered a cornerstone of science, it is not typically facilitated in today's classrooms (Berland et al., 2017). This shortfall can be attributed to several factors. In a recent review, Bae et al. (2021) identified different factors that may inhibit high-quality science discourse in classrooms. At the individual level, for example, students may not be able to engage in science discourse due to limitations with the English language and the difficult nature of science vocabulary. Research also suggests that science instruction may fail to embrace the background knowledge that students bring to the classroom, which can limit students' opportunities to critique or co-construct their peers' ideas around science concepts. In light of these shortcomings and the important role science discourse plays in student science learning, some researchers have begun to investigate how to best promote and measure science discourse opportunities in science classrooms.

In a recent study, for example, Borko et al. (2021) sought to investigate the extent to which the Practicum Academy to Improve Science Education (PRACTISE) PD program was able to improve four fifth-grade teachers' use of science discourse in their classrooms. The PD program provided participating teachers with both "conceptual" and "practical" instructional strategies intended to help their students remain engaged in lessons and participate in science discourse (Borko et al., 2021, p. 6). During the course of the study, the participating teachers video recorded eight science lessons that were subsequently analyzed. Findings suggested that participating teachers not only provided their students with more science discourse opportunities, but richer opportunities as well.

Additionally, others researchers have, in part, investigated which instructional practices can be viewed as critical to the support of science discourse. For example, Fishman and colleagues (2017) have taken steps in measuring science discourse by developing the Science Discourse Instrument (SDI). During the SDI's development, Fishman et al. (2017) used previous research to identify three teacher practices (i.e., *ask*, *press*, and *link*) and three student practices (i.e., *explain/claim*, *co-construct*, and *critique*) that, collectively, constitute what might be considered high-quality science discourse. For teachers, the goal is to encourage students to expand on their claims and understand how ideas can build off of one another (Fishman et al., 2017). The student-centered components measure how well students are able to articulate a well thought out claim that is supported by evidence while, at the same time, connecting their ideas to those of others and being able to scrutinize ideas with potential flaws.

Two studies (i.e., Fishman et al., 2017; Osborne et al., 2019) have applied the SDI to investigate the effects of PD programs on teachers' ability to improve their science instruction by incorporating these components of discourse. In the earlier study, Fishman et al. (2017) found that

the PD program, which included a week-long learning opportunity during the summer and follow-up sessions during the school year, was able to significantly improve teachers' use of high-quality science discourse. For example, teachers were able to gradually ask more open-ended questions which, in turn, provided students with more opportunities to expand on their answers and to support those answers with evidence. In the more recent study of the SDI, Osborne et al. (2019) found similar results. Interestingly, both studies (i.e., Fishman et al., 2017; Osborne et al., 2019) observed an absence of students critiquing the responses of their peers. The authors conjectured that teachers' ability to elicit opportunities for students to question the answers provided by their peers may require additional training. Finally, the interrater reliability for the SDI was reported as $\geq$ .70, which represents moderate agreement between observers when coding components of science discourse.

CONCLUSION

In sum, given the importance of all students being afforded the opportunity to become literate in science, and the practices that encompass scientists' work in particular, the purpose of the current research was to investigate the impact of the Sci2 program on teachers' use of instructional practices during second-grade science instruction. Specifically, the current study focused on teachers' ability to engage students in both well-designed scientific investigations and meaningful science discourse.

Chapter 3: Method

To investigate the extent to which the Sci2 program was able to impact teachers' ability to incorporate high-quality scientific investigations and science discourse into their science instruction, this study was guided by four research questions:

1. To what extent do pre-randomly assigned teachers differ at baseline in terms of the quality of scientific investigations and science discourse facilitated during science instruction?
2. To what extent do treatment (Sci2) and control teachers differ during the treatment time period in terms of the quality of scientific investigations and science discourse facilitated during science instruction?
3. To what extent do treatment (Sci2) and control teachers differ in terms of their capacity to maintain high-quality scientific investigations and science discourse after the treatment time period?
4. To what extent does the quality of scientific investigations and science discourse observed in the treatment time period maintain from the end of treatment to the maintenance time period?

RESEARCH DESIGN

The current study conducted a secondary analysis of direct observation data collected during a recent initial efficacy trial conducted by Doabler et al. (2021). The original study took place during the 2019-2020 school year and investigated the treatment effects of a second-grade science program (Sci2) on the science outcomes of 294 second-grade students. In total, 18 second-grade classrooms participated in the study. Employing an RCT design, Doabler et al. (2021) randomly assigned classrooms within schools to treatment or control conditions. Treatment classrooms (n = 9) implemented the Sci2 program, whereas classrooms assigned to the control

condition (n = 9) employed district-approved, business-as-usual science instruction. Trained research staff conducted four direct observations of each participating classroom (i.e., Rounds 1-4). Using the direct observation data, the current study explored the effects of the Sci2 program and its corresponding PD sessions on the quality of second-grade science instruction. In this study, quality of science instruction was operationally defined as the extent to which second-grade teachers facilitated high-quality scientific investigations and science discourse opportunities.

PARTICIPANTS AND SETTING

The original study (Doabler et al., 2021) took place in a suburban school district in Texas with a student enrollment of 50,204, including 3,709 second-grade students. Demographic data provided by the school district indicated that 38% of the students identified as White, 31% Hispanic, 18% Asian, 9% African American, 4% Two or More Races, <1% Native American, and <1% Pacific Islander. Additionally, 28% of the students received free or reduced-price lunch, 11% were considered English learners, and 10% of students received special education services.

A total of 18 second-grade classrooms distributed across three schools participated in the original study (Doabler et al., 2021). Of the 18 classrooms, nine were randomly assigned to treatment condition and nine to control. Treatment classrooms implemented the Sci2 program and control classrooms used district-approved science instructional materials. All 18 classroom teachers were certified in the state of Texas and provided science instruction five days per week. On average, the teachers had 10.92 years of overall teaching experience (SD = 2.50) with a mean of 7.31 years of experience in second-grade classrooms (SD = 4.39). The majority of participating classrooms delivered academic instruction in English with the exception of one classroom that took part in the district's Dual Language Immersion/Two-Way biliteracy program. However, for

the purposes of the original study, this teacher was given permission to provide science instruction in English.

EXPERIMENTAL CONDITIONS

The current study analyzed observation data collected across the two experimental conditions employed in the original Sci2 study (Doabler et al., 2021). The first served as the treatment condition, where nine teachers were randomly assigned to implement the recently developed Sci2 program. The second condition represented the control group. Teachers randomly assigned to the nine control classrooms continued to implement business-as-usual science materials and instructional practices for the duration of the study. Details on the two conditions are provided in the sections below.

Treatment Condition

The Sci2 program was designed through an iterative process wherein prototype components were first developed and then subsequently tested in classrooms from Texas and Virginia (Doabler et al., 2021). Researchers then utilized data obtained from these initial field tests to revise the program and assemble for initial efficacy testing. The purpose of these initial user tests was to determine the feasibility and usability of the components.

Sci2 focuses specifically on Earth's systems, with particular attention being paid to how wind and water change Earth's landscapes and how these changes impact both humans and animals. Specifically, the program targets the causes and effects of the concepts of weathering, erosion, and deposition. For example, during one of the program's simulation-based investigations students learn how erosion in mountainous regions can lead to an increased amount of flooding which, in turn, can impact the infrastructure of people living at the base of mountains.

During the original study (Doabler et al., 2021), nine classrooms implemented the Sci2 program. Across a two-week time period, each treatment teacher taught 10 lessons that lasted approximately 30 minutes. Each Sci2 lesson is comprised of four major components – (a) Spark Student Inquiry, (b) Vocabulary, (c) Read-Aloud, and (d) Investigations. During the Spark Student Inquiry component, instruction seeks to activate students' prior knowledge and provide students the necessary context to conduct their scientific investigation. For example, when focusing on weathering, students were tasked with making observations about whether the presence of moss, which was used to model grass, would affect the amount of weathering and discussing their observations with their peers.

In the Vocabulary component of Sci2, students were taught key vocabulary terms (e.g., deposition, erosion, and weathering) deemed critical to the focus of the lesson through the use of extended and direct vocabulary instructional strategies (Coyne et al., 2010). Teachers supported students' development of critical science vocabulary and directly taught key terms by providing age-appropriate definitions, examples of how to use the terms in sentences, as well as the use of graphic organizers (Doabler et al., 2021). Additionally, teachers extended their vocabulary instruction by providing students with further opportunities to engage with these science terms in subsequent lesson components of Sci2 (i.e., Read-Aloud, and Investigations). The Vocabulary component is considered an important piece of the Sci2 program because of the crucial role vocabulary plays in science discourse (Fishman et al., 2017; NASEM, 2018; Osborne et al., 2016).

The Read-Aloud portion of the lessons consisted of teachers using "big books" (i.e., oversized story books) to emphasize the vocabulary that students had just learned as well as other concepts directly related to that day's lesson (Doabler et al., 2021). Teachers used guiding questions as a scaffold to bridge the gap between students' current level of knowledge and the new

material they were learning. Student engagement was encouraged through the use of various cartoon characters that were consistent throughout the big books.

Finally, in the Investigations component of Sci2, students were provided the opportunity to engage with either hands-on or simulation-based activities to cement their learning of the new content being presented. For example, during one lesson students used sugar cubes and moss to represent how rocks go through the process of weathering when falling down mountains. Students began by collecting data in the form of tracing the sugar cube before weathering, then shaking the cube in a plastic container to represent the process of weathering, and finally collecting data again to see how the sugar cube changed. At the end of the Investigations portion of the lessons, each lesson is closed by having students engage in scientific argumentation of their findings in either verbal or written forms.

Professional Development

Prior to the start of teachers implementing the Sci2 program, treatment teachers participated in two 3-hr PD sessions (Doabler et al., 2021). Each session was conducted by research staff and focused on building teachers' content and pedagogical knowledge for teaching early science instruction focused on Earth science. Specifically, the purpose of the PD sessions was to instruct teachers on concepts surrounding Earth's systems, instructional practices in science that are considered evidence-based, as well as how to implement the Sci2 program with fidelity. During these sessions, teachers were also provided with opportunities to practice various components of the Sci2 program and received feedback from research staff on how well these program components were being implemented. Additionally, throughout the duration of the study the treatment teachers also received ongoing coaching sessions (at least three) wherein they received feedback on their fidelity of implementation and the quality of their instruction.

Control Condition

In the control classrooms, science instruction was delivered using a combination of materials developed by the school district and a core science program available on the commercial market. The commercially available core science program used was STEMscopes (Accelerate Learning, 2017), a program focusing on the inquiry-based 5E model of instruction; a model which itself consists of five phases – engagement, exploration, explanation, elaboration, and evaluation (Bybee et al., 2006). While STEMscopes appears to be widely used, at this time, no data are available through the WWC (n.d.) on the efficacy of STEMscopes in improving students' science achievement.

For each control classroom, research staff conducted four observations of science instruction (Doabler et al., 2021). On average, each science lesson lasted for 37.8 minutes (*SD* = 7.6) and an average of 18.1 students (SD = 1.4) were present for each lesson. Approximately two-thirds (i.e., 67%) of the science lessons were developed by teachers. The most common instructional formats were whole group and independent, and teachers primarily used a combination of educational technology and science books to deliver their instruction. With both the educational technology and science books, teachers were more likely to take the lead and provided students with few (if any) opportunities to engage with the mode of instruction.

OBSERVATION MEASURES

The current study investigated observation data collected through two observation tools. Each tool required a moderate level of inference among observers to conduct the direct observations. As defined by Gersten et al. (2005), moderate-inference observation tools allow observers "to make an informed judgment about the quality or nature of predetermined instructional events without the strict coding structure of a low-inference measure" (p. 198). The

first observation tool, called the Instructional Quality of Scientific Investigations (IQSI; Doabler et al., 2020), employed in the current study focused on scientific investigations, which is operationally defined as hands-on or simulation-based experiments that are conducted with the goal of having students explore, understand, and explain natural phenomena (NRC, 2012). The second observation tool, called the Science Discourse Instrument (SDI; Fishman et al., 2017; Osborne et al., 2019) targeted science discourse, which is defined as a verbal interaction among students, where an observation of a targeted phenomenon is initially presented and then that explanation or claim is subsequently supported and scrutinized (NRC, 2012; Schwarz et al., 2017). Trained observers in the original Sci2 (Doabler et al., 2021) study used the ISQI and SDI observation tools to document the quality of scientific investigations and science discourse facilitated in treatment and control classrooms (for examples of the ISQI and SDI, see Appendix). Data collected from the IQSI and SDI observation tools served as dependent variables for the current study.

Instructional Quality of Scientific Investigations (Doabler et al., 2020)

The IQSI is a researcher-developed, moderate-inference tool designed to assess the quality of teachers' scientific investigations during science instruction. For this measure, trained observers rated teachers on a scale of 1 (i.e., low) to 5 (i.e., high) for how well they incorporated six practices into their instruction. These practices included: (a) demonstrations of scientific content and practices, (b) planning out investigations, (c) carrying out investigations, (d) models, (e) scaffolding learning opportunities, and (f) independent practice opportunities. If teachers did not include a scientific investigation during their instruction the observers marked each practice as NI (i.e., no investigation).

Science Discourse Instrument (Fishman et al., 2017)

The Science Discourse Instrument (SDI; Fishman et al., 2017) is a moderate-inference instrument that requires observers to rate how well teachers facilitated opportunities for science discourse during science instruction. Specifically, the SDI targets four discourse components at the teacher level (i.e., *ask*, *press*, *link*, and *feedback*) and three components at the student level (i.e., *explain/claim*, *co-construct*, and *critique*). Within the teacher practices, the *ask* component focuses on how well teachers include open-ended questions in their instruction. Compared to short-answer questions that seek to assess students' knowledge of facts (Cazden, 2001; Edwards & Mercer, 1987; Lemke, 1990), open-ended questions lend themselves to the facilitation of more robust discussion given the possibility of multiple answers that could be valid (Fishman et al., 2017). These multiple answers, in turn, result in argumentation amongst students to be able to arrive at an agreed upon "correct" answer (Fishman et al., 2017).

Relatedly, the *press* component documents how well teachers use guiding and follow-up questions to encourage their students to expand on their answers (Fishman et al., 2017). By doing this, teachers further student argumentation by encouraging student collaboration to generate the best answer as well as the use of evidence to support answers (Chin, 2006). Also at the teacher-level, the SDI measures how well teachers are able to *link* students' answers to those provided by their peers (Fishman et al., 2017). The goal here being for teachers to promote a common understanding of a targeted concept, practice, or phenomenon. This is achieved by the students' discussion eventually dispelling faulty arguments and further developing those that are sound. Finally, the *feedback* component – added to the SDI measure by Doabler et al. (2021) – measured how well teachers provided academic feedback to students that was either affirmative or corrective and specific to their answers.

At the student level, the SDI documents the quality of three student discourse components. The first component, *explain/claim*, measures how well students are able to elaborate on their answers and provide evidentiary support of those responses. Researchers (e.g., Webb, 1989; Chi, 2009; Franke et al., 2015) have found that the ability to expand on one's answers, correlates with higher student achievement in the content being learned (e.g., science). In a similar vein, the *co-construct* component documents students' ability to either build off of an answer they heard or ask their peers to elaborate on their answers when an idea was unclear. The final student component, *critique*, focuses on how well students are able to critically analyze their peers' ideas. This is an important skill within science discourse as it signifies conceptual understanding.

In the original Sci2 study (Doabler et al., 2021), trained observers employed the SDI measure to rate the quality of these seven teacher- and student-level components during core science instruction. Each observed lesson contained two separate coding periods (~15 min each). For each coding period, observers rated the quality of the seven components using a five-point rating scale, with a rating of 0 representing no science discussion and rating of 4 representing consistent, high-quality discussion.

OBSERVATION PROCEDURES

The original efficacy trial conducted by Doabler et al. (2021) was segmented into four observation rounds that spanned 13 calendar weeks (i.e., December 2019 to March 2020). The first round was considered a baseline phase where all teachers implemented their typical classroom science instruction. This initial observation round took place between one and four weeks prior to the delivery of Sci2 PD workshop. During rounds two and three, intervention teachers implemented the Sci2 program, while control teachers utilized a combination of instructional materials provided by their school district as well as STEMscopes, a commercially-available

science program (Accelerate Learning, 2017). The second and third rounds took place across three weeks. The fourth round occurred one to four weeks after conclusion of the Sci2 program implementation (Doabler et al., 2021). This final observation round represented a return to baseline and thus all teachers implemented their typical classroom science instruction.

Research staff observed each participating classroom one time during each of observation round, for a total of four observations per treatment and control teacher across the study time period. Before research staff were able to independently conduct observations of either treatment or control teachers' instruction, Doabler et al. (2021) provided several training sessions that focused on direct observation procedures and the multi-faceted observation system employed in the original Sci2 study. In all, observers conducted a total of 72 observations. Research staff included the project's Principal Investigator, an Assistant Professor of Special Education for The University of Texas at Austin. Also included was the Project Coordinator with experience totaling more than 20 years in education, a former teacher with seven years' experience, two doctoral students, as well as a recent graduate with experience in game development.

Inter-Observer Reliability

Of the 72 observations conducted in the original Sci2 efficacy trial (Doabler et al., 2021), 26 observations (36%) included two observers (i.e., a primary and secondary observer) who simultaneously evaluated inter-observer reliability. The current study calculated two forms of inter-observer reliability with both being an item-for-item analysis, wherein the primary observer's score was compared directly to the secondary observer's score for each observation item. First, the study calculated exact matches or agreements (i.e., if primary and secondary observers provided the same rating for a given item). Second, because obtaining exact agreement on moderate-high inference observation tools can be difficult even when making informed judgments (Gersten et al.,

2005; Valentine & Cooper, 2003), the current study also examined whether there were one-point discrepancies between primary and secondary observers' ratings. To calculate both forms of inter-observer reliability, the number of matching scores was divided by the number of matching scores plus non-matching scores and then multiplied by 100. In terms of exact agreement, inter-observer reliabilities for scientific investigations (i.e., the ISQI measure) and science discourse (i.e., the SDI measure) were 58.33% and 68.13%, respectively. Regarding the one-point discrepancies, estimates for scientific investigations (i.e., the ISQI measure) and science discourse (i.e., the SDI measure) were 74.36% and 94.51%, respectively.

DATA ANALYSIS

Four research questions guided the current study. Note that the term quality of science instruction represents the extent to which second-grade teachers facilitated high-quality scientific investigations and science discourse during core science instruction. With each of the 18 participating second-grade classrooms (treatment = 9; control = 9), trained research staff conducted four direct observations of each classroom (i.e., Rounds 1-4). To address the first three research questions – investigating group differences between treatment and control teachers at each time period (i.e., baseline, treatment, and maintenance) – this study conducted independent samples *t*-tests. Visual analysis was applied to address the fourth research question – specifically investigating the extent to which the quality of scientific investigations and science discourse was able to maintain from the end of the treatment time period to the maintenance time period. Given the systematic, explicit nature of the Sci2 program and its corresponding PD, it was anticipated that treatment teachers would deliver higher quality teaching practices (i.e., scientific investigations and science discourse) than their control peers during the treatment time period and after the study concluded (i.e., during the maintenance time period).

Chapter 4: Results

Tables 1 provides the means, standard deviations, and sample sizes for both scientific investigations and science discourse for each observation time period by condition. On a scale of 1-5, observed means for scientific investigations ranged from 1.39 to 2.94 (*SDs* = 0.29 to 0.60). On a 0-4 scale, observed means for science discourse ranged from 1.41 to 1.60 (*SDs* = 0.18 to 0.60).

It is important to note that while Doabler et al. (2021) conducted four observation rounds during the original Sci2 study, the current study calculated a mean of scores for rounds two and three together since both treatment rounds included Sci2 implementation. These combined rounds are referred hereafter as the "treatment time period." As a result, the current study examined observation data from three time periods (i.e., baseline, treatment, maintenance).

Research Question 1: Baseline Differences Between Conditions

The first research question focused on differences between pre-randomly assigned teachers in terms of the quality of the scientific investigations and science discourse documented at baseline. Table 2 provides results from the independent samples *t*-test. Given the study's sample size ($n = 18$), the Welch's correction for the independent samples *t*-test was employed even though the equality of variances assumption was not violated. For this first research question there were no significant differences found between pre-randomly assigned conditions for scientific investigations ($p = 0.637$) or science discourse ($p = 0.298$). Teachers from both conditions were similar on quality of the scientific investigations and science discourse delivered in science instruction, suggesting equivalence prior to the onset of treatment.

Research Question 2: Treatment Time Period Differences Between Conditions

Table 2 provides results related to whether there were differences between conditions during the treatment time period. While the equality of variances assumption was not violated for either teaching practice (i.e., investigations or discourse), the Welch's correction was still utilized because of the small sample size included in the study. During the treatment time period, there was a significant difference between the Sci2 and control teachers in terms of quality of scientific investigations conducted ($p < 0.001$) with a reported effect size (Hedges' g; Borenstein et al., 2009) of 2.52. Conversely, treatment (Sci2) and control teachers were not significantly different with regards to the quality of science discourse opportunities facilitated during science instruction ($p = 0.318$).

Research Question 3: Maintenance Time Period Differences Between Conditions

Table 2 presents the differences between Sci2 and control teachers during the maintenance time period. As with the previous research questions, the equality of variances assumption was not violated for either teaching practice; however, the Welch's correction was still utilized. No significant differences emerged between the conditions with respect to the quality of science discourse. However, analyses revealed a trend-level effect indicating treatment (Sci2) teachers delivered higher quality scientific investigations than control teachers at the maintenance time period ($p = 0.071$, Hedges' $g = 0.88$).

Research question 4: Treatment to Maintenance Differences Within Participants

For the fourth research question, the current study employed a visual analysis to explore whether both treatment and control teachers maintained the quality of scientific investigations and science discourse observed in the treatment time period from the end of treatment to the maintenance time point. Figure 2 shows the Hedges' g values for both teaching practices (i.e.,

scientific investigations and science discourse) at each time point. While recognizing that there was not a significant difference between the treatment and control groups at maintenance, a visual inspection of Figure 2 shows that treatment teachers were able to maintain high-quality teaching practices acquired during the treatment time period with respect to scientific investigations. Furthermore, a Hedges' *g* of -0.22 for scientific investigations at baseline and then an effect size of 0.88 at the maintenance time point suggests that treatment teachers maintained some level of the instructional practices implemented during the treatment time period.

Chapter 5: Discussion

The purpose of this study was to explore observation data collected during an RCT focused on testing the initial efficacy of Sci2, a second-grade whole-class science program (Doabler et al., 2021). While the original efficacy trial produced encouraging findings related to Sci2's promise to increase student science achievement, its primary dependent variable was student science outcomes (Doabler er al., 2021). However, no curriculum teaches itself (Ball et al., 2005). Therefore, to further build the knowledge base of effective science instruction for all learners, including those at risk for learning difficulties, it seemed important to unpack whether the treatment effects of Sci2 spilled over into improving the quality of science teaching practices employed in core science instruction. Specifically, the current study sought to explore whether the Sci2 program impacted the quality of two science teaching practices: (a) scientific investigations, and (b) science discourse.

Moreover, prior research has found that explicit and systematic programs can improve the quantity and quality of instruction delivered in core reading (Nelson-Walker et al., 2013) and early mathematics (Doabler et al., 2018) classrooms. Because the Sci2 program is grounded in a systematic, explicit instructional framework, it seemed reasonable to predict the same principle found in Nelson-Walker et al. (2013) and Doabler et al. (2018) would hold true with the Sci2 program. The current study sought to test this hypothesis by addressing four research questions. Results for each question along with implications related to practice and research are discussed below. Finally, this section presents limitations of the current research.

Research Question 1: Baseline Differences Between Conditions

The first research question explored whether there were baseline differences between the pre-randomly assigned treatment and control teachers regarding the quality of scientific

investigations and science discourse facilitated during their science instruction. Given teachers were expected to use the district-adopted science curriculum as part of their core science instruction, it was anticipated that conditions would not differ during the first round of observations (i.e., baseline). Consistent with this hypothesis, baseline results revealed no statistical differences regarding the quality of scientific investigations (p = 0.637) or science discourse (p = 0.298) facilitated in pre-randomly assigned treatment and control classrooms. This finding of baseline equivalence was important as it strengthened the case that any increase in the quality of scientific investigations and science discourse in treatment classrooms found during the other observation time periods was likely attributed to the implementation of Sci2.

RESEARCH QUESTION 2: TREATMENT TIME PERIOD DIFFERENCES BETWEEN CONDITIONS

The second research question investigated whether there were differences between Sci2 and control teachers regarding the quality of scientific investigations and science discourse facilitated during the treatment time period. Findings suggested a significant difference between the conditions, favoring treatment teachers, for the quality of scientific investigations ($p < 0.001$). The effect for scientific investigations was large (g = 2.52), suggesting the promise of the Sci2 program to improve teachers' capacity to facilitate engaging scientific investigations for second-grade students.

Considering that there was such a large effect for scientific investigations (i.e., g = 2.52), there are a number of possibilities as to why this was found. First, it could be that since science instruction is not usually a focus during the early elementary grades, the teachers might not have received much support in the area of science prior to the intervention. Therefore, receiving the number of resources that they did could have facilitated the growth observed amongst the Sci2 teachers. Another possibility is that the teachers were receiving resources, but they were not as

high-quality as those received during Sci2. In other words, the materials that teachers had been using may not have been as conducive to incorporating scientific investigations as the Sci2 materials were. Consequently, what this means for educators and school districts is that they should have a more critical eye regarding the materials they are using in the discipline of science, and whether these materials serve the ultimate goal of helping students become proficient in science.

The data further suggested a non-significant difference regarding the quality of science discourse ($p = 0.318$). One plausible reason for this finding is that the study conducted by Doabler et al. (2021) began more than six months into the school year. It could be that implementing a new program, even one that was grounded in systematic, explicit instruction, midway into a school year is far too late to change teachers' capacity to manage productive science discourse in whole-class settings. Research suggests that managing group-level discourse in a successful manner is one of most difficult instructional challenges faced by teachers (NRC, 2001). In a similar vein, it could be that the lesson scripting offered in the Sci2 program was insufficient to adequately support teachers with facilitating science discourse in second-grade classrooms. Additionally, it could be that teachers need more PD in order to effectively promote meaningful science discourse in whole-class settings. Finally, it is plausible that Doabler et al. (2021) employed an observation protocol that was ill-suited for early elementary science classrooms. The SDI observation tool (Fishman et al., 2017; Osborne et al., 2019) was originally designed for and validated in third- through fifth-grade classrooms.

RESEARCH QUESTION 3: MAINTENANCE TIME PERIOD DIFFERENCES BETWEEN CONDITIONS

The third research question explored whether Sci2 and control teachers differed in terms the quality of scientific investigations and science discourse opportunities provided at the maintenance time period. Doabler et al. (2021) collected maintenance data approximately one to

four weeks following the end of the treatment time period. Findings indicated no significant differences for the quality of science discourse. Analyses did reveal, however, a trend-level effect for the quality of scientific investigations (p = 0.071; g = 0.88), suggesting treatment teachers appeared to transfer the features of the Sci2 program's investigations into their business-as-usual science instruction. Considering the p-value of 0.071, it might be that if Doabler et al. (2021) had a larger sample size, it could have fallen below the 0.05 threshold to be considered significant. However, this finding is preliminary and therefore further research is warranted to determine if teachers can sustain the practices prescribed in systematic, explicit science programs.

RESEARCH QUESTION 4: TREATMENT TO MAINTENANCE DIFFERENCES WITHIN PARTICIPANTS

The fourth research question explored the extent to which the quality of scientific investigations and science discourse observed in the treatment time period maintained from the end of treatment to the maintenance time period. Figure 2 provides a visual of how the effect sizes (i.e., Hedges' g) changed over time for each instructional practice (i.e., scientific investigations and science discourse). While recognizing that there was not a significant difference between groups for either instructional practice at the maintenance time point, visual inspection of Figure 2 provides preliminary evidence that teachers were able to maintain some instructional components related to scientific investigations after Sci2 had ceased. Specifically, the Hedges' g score at maintenance (i.e., 0.88) was positive and significantly higher than that found at baseline (i.e., -0.22).

While the effect size decreased between the treatment and maintenance time periods, an effect of 0.88 at maintenance for scientific investigations can still be considered large. It could be that with a larger sample size and more data points the effect size at maintenance could be closer to that observed during treatment time period. Another plausible explanation for the decrease in

effect size from treatment to maintenance could be that teachers began a new science unit after the treatment time period and found it difficult to generalize the practices that they learned in Sci2 to that new unit. Additionally, for units other than Earth Science, it will take teachers time to plan for these new lessons. Therefore, if the Sci2 research staff had waited for a longer period of time after the end of the intervention time period to conduct their maintenance observations, they might have seen an effect size closer to that observed during treatment.

IMPLICATIONS FOR PRACTICE

One implication for practice of this study is that features of high-quality scientific investigations, and possibly science discourse, can be explicitly taught to teachers. For example, the IQSI (Doabler et al., 2020) measured teachers' use of demonstrations (i.e., modeling) and scaffolded learning opportunities (i.e., differentiating instruction for students that struggle). Both of these practices are teachable and ones that could easily be addressed during PD. Additionally, the SDI observation tool (Fishman et al., 2017) captured whether teachers asked open-ended questions and provided feedback to students. Posing open-ended questions and providing academic feedback are principles of instruction that can be explicitly taught to and acquired by teachers. While significant results were not found for science discourse during this study, future studies may lead to different findings.

Another implication is the importance of providing teachers with well-designed evidence-based science programs. As demonstrated in the current study, providing such programs may help teachers facilitate high-quality scientific investigations during their science instruction. Specifically, findings suggest the Sci2 program served as a roadmap for delivering effective science instruction in early elementary classrooms. Therefore, it is highly recommended that

school districts consider making purposefully-designed science interventions more readily available for their teachers.

Implications for Research

In terms of future research, one implication is the need for replication. To date, only one study has investigated the effects of the Sci2 program. Specifically, the initial study involving the Sci2 program was conducted in second-grade classrooms located in central Texas. To further establish the program's initial efficacy and investigate its capacity to facilitate high-quality scientific investigations and science discourse, additional research is necessary. Therefore, a recommendation is for researchers to conduct a conceptual replication study of the Sci2 program with a larger sample of classrooms from a different geographical region.

A second research implication is to continue to test the impact of explicit and systematic science programs in early elementary students. While some research has investigated the efficacy of such science programs for early elementary students (e.g., Wells et al., 2015; Kim et al., 2021; Doabler et al., 2021), a dearth of research continues around these types of programs (WWC, n.d.). Because supporting students' understanding of science is essential, continued research in this area is warranted. For example, researchers might consider investigating how to best support early elementary students' use of science discourse and to what extent that looks different than supporting students in the upper grades.

Limitations

While there was an earnest effort to conduct a rigorous investigation of an extant dataset, inevitably limitations were encountered. One such limitation was the current study's statistical power. The analysis included a total of 72 direct observations. To obtain a more robust estimate of instructional quality in the discipline of science, researchers may need to observe classrooms

more than four times throughout the intervention time period. In the original Sci2 study, for example, it would have been ideal if Doabler et al. (2021) could have observed each participating teacher more than two times during the treatment period and collected additional maintenance data beyond the conclusion of the Sci2 program's implementation. However, limited financial resources and the educational disruptions caused by the onset of COVID-19 precluded additional classroom observations.

Another potential limitation was the application of the SDI observation tool (Fishman et al., 2017) in early elementary science classrooms. Fishman et al. (2017) previously validated the SDI in third-, fourth-, and fifth-grade classrooms. Therefore, Doabler and colleagues (2021) were likely one of the first research teams to employ the SDI in the context of second-grade classrooms. While the Sci2 research staff maintained the integrity of the SDI measure, it is plausible that refinements were required to best cater to the science instruction that occurs in early elementary classrooms. Consequently, it could be that some observation items had lower ratings because the SDI protocol is better suited for the upper elementary grades.

In a similar vein, the IQSI (Doabler et al., 2020) was a researcher-developed measure and was not validated prior its use in the Sci2 pilot study. Consequently, it is plausible that the IQSI did not measure what it purported to measure (i.e., the quality of scientific investigations). More research is warranted to determine whether the IQSI reliably and validly documents the quality of scientific investigations conducted in second-grade classrooms.

Additionally, this study may have been limited by the fact that observations were conducted in real time. These types of observations require observers to make in vivo inferences. Future research should consider video recordings to permit repeated reviews of science instruction.

This approach would likely mitigate the influence observer inference plays when conducting real-time observations.

Finally, the current study used observation tools that relied on moderate levels of observer inference. While moderate-inference observation tools allow observers to make informed judgments when capturing targeted behaviors, research suggests obtaining high levels of exact agreement with such measures can be difficult (e.g., Gersten et al., 2005). The current study experienced similar challenges. It may be that more in-depth observer training is required to reliably use the ISQI and SDI measures in real-time observations. Also, had Doabler et al. (2021) audio recorded or videotaped the science instruction, it might have been possible to obtain stronger observer agreement on the quality of the targeted behaviors.

CONCLUSION

Recognizing that systematic and explicit intervention programs have been able to impact teachers' instructional quality in both core reading (e.g., Nelson-Walker et al., 2013) and early mathematics classrooms (e.g., Doabler et al., 2018), the current study sought to determine whether the same type of impact could be observed in early science instruction. In other words, could an early science program that utilizes systematic and explicit instruction improve the quality of science instruction? This is a crucial question that comes at a time when science instruction is not often prioritized in the early grades. If, as a nation, we would like to promote science proficiency for all, arguably more has to be done to address the science achievement gap that can begin as early as the kindergarten year.

Overall, findings from this study suggested that implementation of the Sci2 program improved the quality of science instruction delivered in second-grade classrooms. Specifically, relative to control teachers, teachers in Sci2 classrooms were able to incorporate higher quality

scientific investigations into their core science instruction during the treatment time period. Results related to science discourse did not follow a similar suit. However, from an educational research perspective, these findings can be viewed in a positive light. Above all, the non-significant findings regarding the quality of science discourse highlight the need to replicate the original Sci2 study (Doabler et al., 2021) with a larger sample size, additional observation data points, and refined observation protocols that are better catered to the science instruction that occurs in early elementary classrooms.

Table 1

Descriptive Statistics by Teaching Practice

	Control			Treatment		
	M	*SD*	Sample	*M*	*SD*	Sample
Scientific Investigations						
Baseline	1.500	0.520	9	1.389	0.456	9
Treatment	1.676	0.327	9	2.944	0.595	9
Maintenance	1.463	0.286	9	1.778	0.391	9
Science Discourse						
Baseline	1.603	0.322	9	1.468	0.189	9
Treatment	1.480	0.182	9	1.599	0.292	9
Maintenance	1.413	0.313	9	1.587	0.598	9

Table 2

Group Differences by Time Period

	t	df	*p*	Hedges' *g*
Baseline				
Scientific Investigations	-0.482	15.732	0.637	-0.22
Science Discourse	-1.084	12.951	0.298	-0.49
Treatment				
Scientific Investigations	5.606	12.418	< 0.001	2.52
Science Discourse	1.037	13.404	0.318	0.47
Maintenance				
Scientific Investigations	1.95	14.658	0.071	0.88
Science Discourse	0.776	12.071	0.453	0.35

Figure 1

Research Question Placement Within Sci2 Timeline

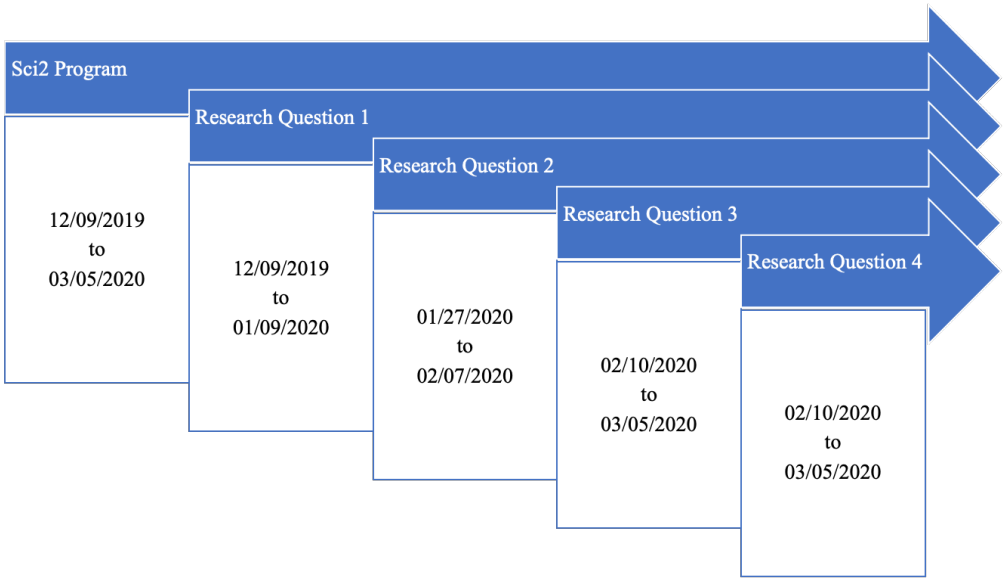
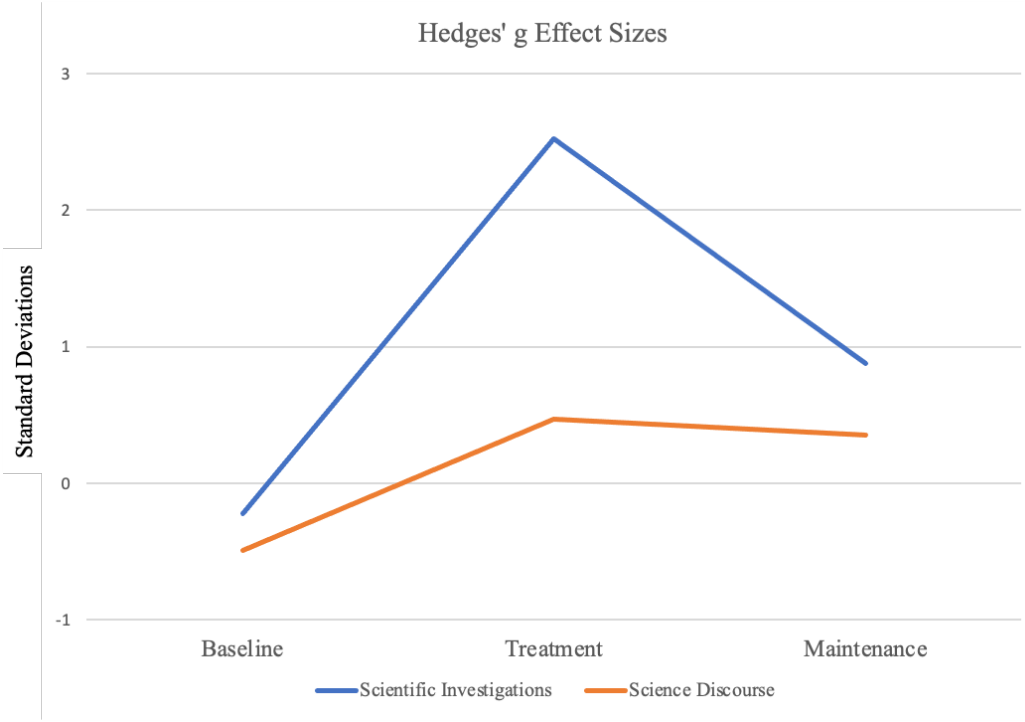


Figure 2

Hedges' g Differences by Time Period



Appendix

Sci2 Direct Observation System

Sci2 Observation Instrument – 2019/2020 (Year-3: PILOT)

Date: ☐☐/☐☐/☐☐	School:_____ Tchr ID:_____ # Students _____	Treatment ____ Control ____ Sci2 Lesson #: _____
Projected Lesson Time:_____min Start Time ☐☐.☐☐ End Time ☐☐.☐☐	Observer Initials _____ IOA: ____ IOA Initials: _____	Observation Rd (circle one): 1 2 3 4 5 6

SCIENTIFIC CONTENT & INSTRUCTION

1. Name of the ***science program/curriculum*** used during instruction? _____ ____ Unrecognizable_____
   - The materials used during instruction appeared teacher developed? circle: Yes_____ No_____ Blend_____
   - Were Sci2 materials used during instruction (*Sci2 Lesson Cards, Big books, Student Journals*)? circle: Yes No

2. ***Instructional format*** in which science instruction occurred: (check all that apply)
   - Whole-group Small-group Partners Independent
   - Primary format used_____

3. Was the instruction structured around a ***phenomenon*** (i.e., observable event)? Yes____ No____
   - If yes, identify the phenomenon? _____ (e.g., habitats, pollination, erosion, H_2O on Earth)

4. Was a ***crosscutting concept(s)*** addressed? 1. Patterns____ 2. Cause & effect____ 3. Scale, proportion, & quantity____ 4. Systems & system models____ 5. Energy & matter____ 6. Structure & function____ 7. Stability & change_____
   - *Note*. Concept(s) does not have to be explicitly stated by the teacher or displayed on board, doc cam, etc…

5. Did the lesson include ***vocabulary*** instruction? Yes____ No ____
   - Type of vocab instruction (check one)? Direct____ Implicit ____ Review ____

Direct = T & S state defs & use in sentences; *Implicit* = T states terms after the fact; *Review* = T asks S meaning of prior terms

6. Did the lesson incorporate ***education technology***? Yes____ No____
   If yes, was the technology (a) teacher-led only, (b) used individually by students, or (c) both? Circle one
   If yes, check only one below:
   - Technology allowed for *direct* interaction with science content (students use apps, search websites)____
   - Technology permitted *indirect* interaction with science content (PPTs, videos, projected & read online books)____

7. Did the lesson incorporate aspects of ***literacy*** related to science? Yes____ No____
   - If yes, how was literacy incorporated? (check one): Teacher read aloud:___ Independent:___ Partner:___ Other:___

8. Describe the lesson's primary science-based activity? New Activity____ Review or Continuation of Prior Activity____
   - Did the activity directly relate to the phenomena (#3) and/or crosscutting concepts (#4)? Yes____ No____
   - Was the relation between the activity and the phenomena/crosscutting concepts made explicit to students Y N
   - Did the activity include or focus on a testable question? Yes____ (investigation) No____
     - Describe the question? _____
     - Did students have an opportunity to pose and/or practice ***asking the question***? Yes No
   - Did students ***carry out*** a hands-on experiment by ***developing*** or ***using scientific models***? Yes____ No____
   - Did students have any part in ***planning*** the activity? Yes Tchr/St Teacher-led only No
   - Did students have opportunities to ***explain*** or ***discuss findings*** from the activity? Yes____ No___
     - If students constructed explanations, did they ***engage in data-driven scientific argumentation***? Yes____ No___

   Describe the activity _____
   _____

   *Tchr/St = blend of teacher & student

9. Other content areas addressed in the *science* lesson:
   - Mathematics? Yes____ No____
     - If, yes, did students engage in ***mathematics and computational thinking***? Yes____ No____
   - Did students ***analyze and interpret data***? Yes____ No____
   - Writing? Yes, students write explanations using full sentences_____ Yes, but one word responses only ____ No____

Sci2 Direct Observation System

FEATURES OF SCIENTIFIC DISCOURSE (SDI)	
10. **Discourse Purpose** (*share background knowledge, review prior content, share observations, interpret data, build explanations, design investigations, compare/evaluate claims, non-science discussion*) – Circle the primary purpose	
11. **Discourse Structures** (*whole class, small group, pairs, individuals*) – Check all that apply and circle the primary structure	
12. **Participation in Whole Class Science Discussions** (On average, how many student contributions to each discussion/question) Check all observed & circle primary format *No evidence of class discussion* *1 or 2 students* *3-5 students* *6-10 students* *>10 students*	

SCIENCE DISCOURSE INSTRUMENT (SDI)	Rating	No Evid	Low		Med	High
Teacher Discursive Forms						
13. **ASK** (Nature of teachers' questioning) – *Do teachers' ask open-ended questions (how and why) intended to elicit diverse student responses?*	A:	0	1	2	3	4
	B:	0	1	2	3	4
14. **PRESS** (Teacher press) – *Do teachers press students to support their contributions with evidence and/or reasoning?*	A:	0	1	2	3	4
	B:	0	1	2	3	4
15. **LINK** (Teacher linking contributions) – *Does the teacher connect students' ideas and positions in a way that helps build and develop the discussions?*	A:	0	1	2	3	4
	B:	0	1	2	3	4
16. **FEEDBACK*** (Teachers academic feedback) – *Does the teacher offer both affirmative and corrective academic feedback that is specific to student responses?*	A:	0	1	2	3	4
	B:	0	1	2	3	4
Student Discursive Forms						
17. **EXPLAIN/CLAIM** (Nature of students' responses) – *Do students offer explanations or claims/conjectures supported by evidence?*	A:	0	1	2	3	4
	B:	0	1	2	3	4
18. **COCONSTRUCT** (Student co-constructing) – *Do students' contributions link to and build on each other to co-construct understanding?*	A:	0	1	2	3	4
	B:	0	1	2	3	4
19. **CRITIQUE** (Student critique) – *Do students' offer critiques of the contributions of other students or the teacher ("I agree…" or "I disagree…" opportunities)*	A:	0	1	2	3	4
	B:	0	1	2	3	4

SDI Rating Scale

4 = There are **consistent, high quality** examples of the target practice.
3 = There are **occasional, high quality** examples of the target practice.
2 = There was **one high quality example** (rarely) OR **multiple examples of lower quality**.
1 = There is **no evidence** of the target practice.
0 = There is **no science discussion** happening in the segment.

SCIENTIFIC INVESTIGATIONS (Activities)	Rating	Low		Medium		High
20. **Demonstrations** of scientific content and practices (e.g., teacher models)	NI	1	2	3	4	5
21. **Plan out the investigation (**students contributed to planning of investigation)	NI	1	2	3	4	5
22. **Carry out the investigation** (students learned to define the investigated features)	NI	1	2	3	4	5
23. **Models** (hands-on materials used by the teacher and students in the investigation)	NI	1	2	3	4	5
24. **Scaffolded learning opportunities** (differentiation for struggling learners; think time)	NI	1	2	3	4	5
25. **Independent practice opportunities** (students conduct investigation on their own)	NI	1	2	3	4	5

OVERALL QUALITY OF SCIENCE INSTRUCTION	Low		Medium		High
26. **Overall student interest in the science lesson** (lesson appeared interesting to students)	1	2	3	4	5
27. **Overall discourse opportunities** (discussion was rich and involved majority of class)	1	2	3	4	5
28. **Overall teaching of key science vocabulary**	1	2	3	4	5
29. **Overall teaching for scientific understanding**	1	2	3	4	5

References

Accelerate Learning. (2017). *STEMscopes PreK-12 Learning.*

Bae, C. L., Mills, D. C., Zhang, F., Sealy, M., Cabrera, L., & Sea, M. (2021). A systematic review of science discourse in K-12 urban classrooms in the United States: Accounting for individual, collective, and contextual factors. *Review of Educational Research, 91*(6), 831-877. https://doi.org/10.3102/00346543211042415

Ball, D. L., Hill, H. C., & Bas, H. (2005). Knowing mathematics for teaching: Who knows mathematics well enough to teach third grade, and how can we decide? *American Educator, 29*(1), 14-17, 20-22, 43-46. https://hdl.handle.net/2027.42/65072

Berland, L. K., McNeill, K. L., Pelletier, P., & Krajcik, J. (2017). Engaging in argument from evidence. In C. V. Schwarz, C. Passmore, & B. J. Reiser (Eds.), *Helping students make sense of the world using next generation science and engineering practices* (pp. 229-258). National Science Teachers Association. https://doi.org/10.2505/9781938946042

Borenstein, M., Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 221-235). Russell Sage Foundation.

Borko, H., Zaccarelli, F. G., Reigh, E., & Osborne, J. (2021). Teacher facilitation of elementary science discourse after a professional development initiative. *The Elementary School Journal, 121*(4). https://doi.org/10.1086/714082

Bybee, R. W., Taylor, J. A., Gardner, A., Van Scotter, P. Powell, J. C., Westbrook, A., & Landes, N. (2006). *The BSCS 5E instructional model: Origins and effectiveness.* https://media.bscs.org/bscsmw/5es/bscs_5e_full_report.pdf

Cazden, C. B. (2001). *Classroom discourse* (2nd ed.). Heinemann.

Chi, M. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1,* 73-105. https://doi.org/10.1111/j.1756-8765.2008.01005.x

Chin, C. (2006, November). *Teacher questioning in science classrooms: What approaches stimulate productive thinking?* Presentation at the International Science Education Conference, Singapore.

Coyne, M. D., Kame'enui, E. J., & Carnine, D. (2011). *Effective teaching strategies that accommodate diverse learners* (4th ed.). Pearson Education.

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

Doabler, C. T., Nelson, N. J., Kennedy, P. C., Stoolmiller, M., Fien, H., Clarke, B., Gearin, B., Smolkowski, K., & Baker, S. K. (2018). Investigating the longitudinal effects of a core mathematics program on evidence-based teaching practices in mathematics. *Learning Disability Quarterly, 41*(3), 144-158. https://www.doi.org/10.1177/0731948718756040

Doabler, C. T., Therrien, W. J., Longhi, M. A., Roberts, G., Hess, K. E., Maddox, S. A., Uy, J., Lovette, G. E., Fall, A. M., Kimmel, G. L., Benson, S., VanUitert, V. J., Wilson, S. E., Powell, S. R., Sampson, V., & Toprac, P. (2021). Efficacy of a second-grade science program: Increasing science outcomes for all students. *Remedial and Special Education, 42*(3), 140-154. https://www.doi.org/10.1177/0741932521989091

Doabler, C. T., Therrien, W. J., Powell, S., & Sampson, V. (2020). *Instructional Quality of Scientific Investigations* [unpublished observation measure]. Meadows Center for Preventing Educational Risk, The University of Texas at Austin.

Edwards, D., & Mercer, N. (1987). *Common knowledge: The development of understanding in the classroom.* Methuen.

Fishman, E. J., Borko, H., Osborne, J., Gomez, F., Rafanelli, S., Reigh, E., Tseng, A., Million, S., & Berson, E. (2017). A practice-based professional development program to support scientific argumentation from evidence in the elementary classroom. *Journal of Science Teacher Education, 28*(3), 222-249. https://www.doi.org/10.1080/1046560X.2017.1302727

Franke, M. L., Turrou, A. C., Webb, N. M., Ing, M., Wong, J., Shin, N., & Fernandez, C. (2015). Student engagement with others' mathematical ideas. *The Elementary School Journal, 116*(1), 126-148. https://doi.org/10.1086/683174

Gersten, R. M., Baker, S. K., Haager, D., & Graves, A. W. (2005). Exploring the role of teacher quality in predicting reading outcomes for first-grade English learners: An observational study. *Remedial and Special Education, 26*(4), 197-206. https://doi.org/10.1177/07419325050260040201

Hornby, G. (2015). Inclusive special education: Development of a new theory for the education of children with special educational needs and disabilities. *British Journal of Special Education, 42*(3), 234-256. https://doi.org/10.1111/1467-8578.12101

Individuals with Disabilities Education Improvement Act, 20 U.S.C. § 300.8 (c)(4)(2004).

Kart, A., & Kart, M. (2021). Academic and social effects of inclusion on students without disabilities: A review of the literature. *Education Sciences, 11*(16), 1-13. https://doi.org/10.3390/educsci11010016

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

Lemke, J. (1990). *Talking science: Language, learning and values*. Ablex. https://doi.org/10.5860/choice.28-5211

Li, M. C., & Tsai, C. C. (2013). Game-based learning in science education: A review of relevant research. *Journal of Science Education and Technology, 22*, 877-898. https://doi.org/10.1007/s10956-013-9436-x

Morgan, P. L., Farkas, G., Hillemeier, M. M., & Maczuga, S. (2016). Science achievement gaps begin very early, persist, and are largely explained by modifiable factors. *Educational Researcher,* 1-18. https://doi.org/10.3102/0013189x16633182

National Academies of Sciences, Engineering, and Medicine. (2018). *English learners in STEM subjects: Transforming classrooms, schools, and lives.* The National Academies Press. https://www.doi.org/10.17226/25182

National Center for Education Statistics. (2021, December 17). *NAEP Report Card: Science.* Institute of Education Sciences, U.S. Department of Education. https://www.nationsreportcard.gov/science/nation/achievement/?grade=12

National Center for Education Statistics (2020, November 3). *Students with Disabilities, Inclusion of.* https://nces.ed.gov/fastfacts/display.asp?id=59

National Research Council. (2001). *Adding it up: Helping children learn mathematics.* Mathematics Learning Study Committee.

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas.* Committee on a Conceptual Framework for New K-12 Science Education Standards. Board on Science Education, Division of Behavioral and Social Sciences and Education. The National Academies Press. https://doi.org/10.17226/13165

National Science & Technology Council, Committee on STEM Education. (2018). *Chartering a course for success: America's strategy for STEM education.* https://doi.org/10.1119/1.5088484

Nelson-Walker, N. J., Fien, H., Kosty, D. B., Smolkowski, K., Smith, J. L. M., & Baker, S. K. (2013). Evaluating the effects of a systematic intervention on first-grade teachers' explicit reading instruction. *Learning Disability Quarterly, 36*(4), 215-230. https://www.doi.org/10.1177/0731948712472186

Next Generation Science Standards: For States, By States. (n.d.). *Read the standards*. https://nextgenscience.org/search-standards?keys=&tid%5B%5D=98&tid_1%5B%5D=114

Next Generation Science Standards Lead States. (2013). *Next generation science standards: For states, by states.* The National Academies Press. https://doi.org/10.17226/18290

Osborne, J. F., Borko, H., Fishman, E., Zaccarelli, F. G., Berson, E., Busch, K. C., Reigh, E., & Tseng, A. (2019). Impacts of a practice-based professional development program on elementary teachers' facilitation of and student engagement with scientific

argumentation. *American Educational Research Journal, 56*(4), 1067-1112. https://doi.org/10.3102/0002831218812059

Osborne, J. F., Henderson, B., MacPherson, A., Szu, E., Wild, A., & Yao, S. Y. (2016). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching, 53,* 821-846. https://doi.org/10.1002/tea.21316

Osborne, J., & Quinn, H. (2017). The framework, the NGSS, and the practices of science. In C. V. Schwarz, C. Passmore, & B. J. Reiser (Eds.), *Helping students make sense of the world using next generation science and engineering practices* (pp. 23-31). National Science Teachers Association. https://doi.org/10.2505/9781938946042

Rosenbaum, E., Klopfer, E., & Perry, J. (2007). On location learning: Authentic applied science with networked augmented realities. *Journal of Science Education Technology, 16*(1), 31-45. https://doi.org/10.1007/s10956-006-9036-0

Samarapungavan, A., Mantzicopoulos, P., & Patrick, H. (2008). Learning science through inquiry in kindergarten. *Science Education,* 868-908. https://doi.org/10.1002/sce.20275

Schwarz, C. V., Passmore, C., & Reiser, B. J. (2017). Moving beyond "knowing about" science to making sense of the world. In C. V., Schwarz, C. Passmore, & B. J. Reiser (Eds.), *Helping students make sense of the world using next generation science and engineering practices* (pp. 3-22). National Science Teachers Association. https://doi.org/10.2505/9781938946042

Social Science Statistics. (2022, October 9). *Effect size calculator for t-test.* https://socscistatistics.com/effectsize/default3.aspx

Squire, K., & Klopfer, E. (2007). Augmented reality simulations on handheld computers. *Journal of Learning Sciences, 16*(3), 371-413. https://doi.org/10.1080/10508400701413435

Texas Education Agency (n.d.). *STAAR Resources*. https://www.tea.texas.gov/student-assessment/testing/staar/staar-resources

Therrien, W. J., Benson, S. K., Hughes, C. A., & Morris, J. R. (2017). Explicit instruction and next generation science standards aligned classrooms: A fit or a split? *Learning Disabilities Research and Practice, 32*(3), 149-154. https://doi.org/10.1111/ldrp.12137

TIMSS 2019 U.S. Highlights Web Report (NCES 2021-021). U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics. Available at https://www.nces.ed.gov/timss/results19/index.asp.

U.S. Bureau of Labor Statistics. (2021). *Employment Projections: 2020-2030 Summary.* U.S. Department of Labor. https://www.bls.gov/news.release/ecopro.nr0.htm

U.S. Department of Education, Office of Special Education and Rehabilitative Services. (2010). *Thirty-Five Years of Progress in Educating Children with Disabilities Through IDEA*. Office of Special Education and Rehabilitative Services.

Valentine, J. C., & Cooper, H. (2003). *What Works Clearinghouse study design and implementation assessment device (Version 1.0)*. U.S. Department of Education.

Webb, N. M. (1989). Peer interaction and learning in small groups. *International Journal of Education Research, 13,* 21-39. https://doi.org/10.1037/0022-0663.74.5.642

Wells, N. M., Myers, B. M., Todd, L. E., Barale, K., Gaolach, B., Ferenz, G., Aitken, M., Henderson, C. R., Tse, C., Pattison, K. O., Taylor, C., Connerly, L., Carson, J. B., Gensemer, A. Z., Franz, N. K., & Falk, E. (2015). The effects of school gardens on

children's science knowledge: A randomized controlled trial of low-income elementary schools. *International Journal of Science Education, 37*(17), 2858-2878. https://doi.org/10.1080/09500693.2015.1112048

What Works Clearinghouse. (n.d.). *Find what works based on the evidence.* Institute of Education Sciences, U.S. Department of Education. https://www.ies.ed.gov/ncee/wwc/FWW/Results?filters=,Science

Williams, J. P., Kao, J. C., Pao, L. S., Ordynans, J. G., Atkins, J. G., Cheng, R., & DeBonis, D. (2016). Close analysis of texts with structure (CATS): An intervention to teach reading comprehension to at-risk second graders. *Journal of Educational Psychology, 108*, 1061-1077. https://doi.org/10.1037/edu/0000117

Windschitl, M. (2017). Planning and carrying out investigations. In C. V. Schwarz, C. Passmore, & B. J. Reiser (Eds.), *Helping students make sense of the world using next generation science and engineering practices* (pp. 135-158). National Science Teachers Association. https://doi.org/10.2505/9781938946042