Final Report of the i3 Impact Study of Making Sense of SCIENCE

2016-17 THROUGH 2017-18

November 18, 2020

Andrew P. Jaciw, Co-principal Investigator

Thanh Nguyen, Co-principal Investigator

Li Lin

Jenna L. Zacamy

Connie Kwong

Sze-Shun Lau

Empirical Education Inc.

ACKNOWLEDGEMENTS

First and foremost, we are grateful to the students, teachers, and staff in participating schools and districts in California and Wisconsin for their assistance and cooperation in conducting this research. Without them, this research would have not been possible. We would like to express our gratitude to our esteemed Technical Working Group (Anne Chamberlain, Angela DeBarger, Heather Hill, Ellen Kisker, James Pellegrino, Rich Shavelson, Guillermo Solano-Flores, Steve Schneider, Jessaca Spybrook, and Fatih Unlu) for their expert guidance based on state-of-the art knowledge and invaluable insights from the field and research. Additionally, we thank Anne Wolf, our National Evaluation of Investing in Innovations technical assistance liaison, for her assistance during the entire research process, as well as for her comments on early drafts of the report, although any errors are our own. We are indebted to Frank Jenkins and Ruth Childs for providing their psychometric expertise, which was critical in informing our decisions about key outcome measures.

We would like to thank our research partner Heller Research Associates (HRA), who conducted the implementation study and the scale-up study as part of this evaluation, for being a steady and perceptive thought partner throughout the course of the grant. Last but certainly not least, we thank the Making Sense of SCIENCE program team for the opportunity to collaborate on this project. The program team's enduring collaborative spirit and rich understanding of the context of science education have greatly benefited this work.

This work has been supported by the U.S. Department of Education's Investing in Innovation program, through Award Number U411B140026. The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education. As the independent evaluator, Empirical Education Inc. was provided with independence in reporting the results.

ABOUT EMPIRICAL EDUCATION INC.

Empirical Education Inc. is a Silicon Valley-based research company that provides tools and services to help K-12 school systems make evidence-based decisions about the effectiveness of their programs, policies, and personnel. The company brings its expertise in research, data analysis, engineering, and project management to customers that include the U.S. Department of Education, educational publishers, foundations, leading research organizations, and state and local education agencies.

Reference this report: Jaciw, A.P., Nguyen, T., Lin, L., Zacamy, J., Kwong, C., Lau, S. (2020). *Final Report of the i3 Impact Study of Making Sense of SCIENCE, 2016-17 through 2017-18.* (Empirical Education Rep. No. Empirical_MSS-7030-FR1-2020-O.1) San Mateo, CA: Empirical Education, Inc. Retrievable from https://empiricaleducation.com/mss

Table of Contents

Chapter 1. Introduction

BACKGROUND

In a 2010 report offering strategies for improving K–12 STEM education, the President's Council of Advisors on Science and Technology (PCAST) asserted the paramount importance of STEM education in the advancement of the United States in many critical areas, including but not limited to health, energy, environment, and national security. More specifically, PCAST highlighted STEM education's diverse role in society, ranging from creating individuals who will earn livable wages and make informed choices as citizens, to producing a workforce flexible and capable enough to compete in the new industries of the 21st century, to promoting a society that can continue to make new discoveries about ourselves and the universe.

But the report also offered a stark depiction of the current reality of STEM education.

> "Schools often lack teachers who know how to teach science and mathematics effectively, and who know and love their subject well enough to inspire their students. Teachers lack adequate support, including appropriate professional development as well as interesting and intriguing curricula. School systems lack tools for assessing progress and rewarding success. The Nation lacks clear, shared standards for science and math that would help all actors in the system set and achieve goals. As a result, too many American students conclude early in their education that STEM subjects are boring, too difficult, or unwelcoming, leaving them ill-prepared to meet the challenges that will face their generation, their country, and the world" (PCAST, 2010).

These anecdotal shortcomings culminated in, and are corroborated by, reports of students' science achievement on various assessments. The results from the 2015 National Assessment of Educational Progress (NAEP)—the largest nationally representative assessment of America's students in a number of subjects, including science—revealed that only 38% of fourth graders, 34% of eighth graders, and 22% of twelfth graders achieved a level of proficient or higher. We should, however, acknowledge that there have been relatively small successes, such as an improvement (by 4 points) since 2009 in grades 4 and 8 (no significant difference in grade 12), and a trend toward narrowing the achievement gap, with Black and Hispanic students making greater gains than White students (NAEP, 2019). International assessments, such as the Trends in International Mathematics and Science Study (TIMSS), show that the average scores of America's students have remained flat since 1995, and the United States slipped in rank, from third in 1995 to fifth in 2015, among the 17 education systems that participated in the 1995 and 2015 grade 4 TIMSS assessments (National Science Board, 2018).

It was against this backdrop that emerged the re-envisioning of K-12 science education. The vision of what students should know and be able to do in science was first laid out in the National Research Council's "Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas" in 2012 (NRC, 2012). Authors of the Framework pointed out that the impetus for the project grew not just from the recognition that there is much room for improvement in science education, but also the desire to take advantage of an opportunity that presented itself at the time: a large number of states were in the process of adopting the Common Core State Standards (CCSS) in mathematics and English Language Arts (ELA). In contrast with existing standards, which often emphasized content knowledge, the Framework emphasized three-dimensional learning: Cross-cutting Concepts (CCCs), Science and Engineering Practices (SEPs), and Disciplinary Core Ideas (DCIs). In 2013, the Next Generation Science Standards (NGSS) were released, specifying the targets for student learning that were based on the vision set forth in the Framework (NRC, 2013). The release of the

Guide to Implementing the NGSS followed in 2015 (NRC, 2015). Among the recommendations were ones that drew attention to the importance of teacher professional learning, along with building leadership capacity at the school, district, and regional levels. The guide reiterated the finding from numerous previous research that while teachers' knowledge of the content is necessary, it is insufficient. Teachers' mastery of the content must also be accompanied by their ability to translate their own knowledge of science into effective lessons for students (Duschl et al., 2007; Heller et al., 2012).

But here again exists a gap between our aspirations for science education and the realities on the ground. A recent (2018) national survey of science teachers conducted by Horizon research showed that very few elementary science teachers have college or graduate degrees in science, engineering, or science education, with fewer than half having had at least one college course in chemistry, environmental science, or physics, and close to none in engineering. In regard to feeling prepared about science content knowledge, only a quarter of elementary school teachers report feeling very prepared to teach life science, one-fifth feel very prepared to teach Earth science, and a mere 13 feel very prepared to teach physical science. In regard to being prepared pedagogically, only 23% feel very prepared to develop students' conceptual understanding, less than 33% in monitoring and assessing students' understanding, and fewer than 25% in anticipating areas where students might have difficulties (Banilower et al., 2018).

The literature suggests that one approach to narrowing these gaps is through providing teachers with high quality professional learning, which rests on research-based principles that professional learning should be "intensive, ongoing, and connected to practice," "focus on student learning and address the teaching of specific content," "align with school priorities and goals," and "build strong working relationships among teachers" (Darling-Hammond et al., 2009). Despite what is known in the field about effective professional learning, few professional learning programs focus on improving teachers' specialized pedagogical knowledge and skills for teaching science and mathematics. Instead, most focus solely on content or on classroom management (Sztajn et al., 2012).

While teachers are considered to be one key driver of student achievement (Duschl et al., 2007; Hill et al., 2005), they operate within a set of systems at multiple levels. As such, the Guide to Implementing the NGSS also called for teacher professional learning being a sustained component of a comprehensive support system at the school, district, and state levels, with the principal's understanding of and support for instructional change being a primary driving force for sustained implementation. Also important is science education leaders' willingness and capacity to leverage networks and partners (e.g. researchers, higher education institutions, and science technology centers) in facilitating high-quality professional learning and providing ongoing support (NRC, 2015).

Concurrent with these discussions in how to implement the NGSS was the equally essential question of how to create assessments that can measure what and how well students have learned. Prior to NGSS, science assessments were not designed to measure the three-dimensional learning that is at the center of the new standards. Developing new NGSS-aligned assessments to inform classroom instruction and to monitor science learning at a broad scale was going to be a considerable challenge (NRC, 2014).

It was in 2014–15, in the midst of the evolutions in the landscape of science teaching and assessment, that Making Sense of SCIENCE received a validation grant from the Investing in Innovation (i3) grant from the U.S. Department of Education. Developed by WestEd, Making Sense of SCIENCE is a model for teacher professional learning that aims to raise students' science achievement and teachers' science content knowledge. Under the five-year i3 grant, WestEd partnered with Empirical Education Inc. (Empirical) to conduct an impact evaluation using a school-randomized control trial and with Heller Research Associates (HRA) to conduct implementation and scale-up studies of the model. What follows is the

report on the impact evaluation in this context, with connections to the implementation and scale-up studies where relevant.

In this chapter, we continue with an overview of the Making Sense of SCIENCE model and impact evaluation, including the confirmatory and exploratory research questions that guide our inquiry. Chapter 2 outlines our methods, including the experimental study design, samples, and data collection. Chapter 3 includes a summary of the results of fidelity of implementation of Making Sense of SCIENCE. Chapter 4 presents findings on teacher content knowledge and pedagogical content knowledge, and Chapter 5 presents findings on teacher attitudes and beliefs, opportunity to learn, and school culture. Chapters 6–9 report findings related to student outcomes, including: student achievement in Earth and space science and physical science (Chapter 6); student achievement on the state assessments in ELA, math, and science (Chapter 7); student communication of science ideas in writing (Chapter 8); and non-academic student outcomes (Chapter 9). We discuss the significance and implication of the findings and offer conclusions in Chapter 10.

OVERVIEW OF MAKING SENSE OF SCIENCE

Making Sense of SCIENCE is a model for teacher professional learning aimed at raising students' science achievement, particularly Earth and space science and physical science, through improving science instruction. The professional learning model focuses on the critical connections between science understanding, classroom practice, and literacy integration in ways that support the implementation of NGSS and CCSS.

Logic Model

The Making Sense of SCIENCE theory of action is based on the premise that professional learning, when situated in an environment of collaborative inquiry and supported by school leadership, has a cascade of effects on teachers' content and pedagogical content knowledge, teachers' attitudes and beliefs, school climate, and students' opportunities to learn. These effects produce improvements in student science achievement, as well as other non-academic outcomes, such as enjoyment of science, self-efficacy and agency in science learning, and aspirations for future use of science in adulthood and careers.

The Making Sense of SCIENCE model comprises six components. Components 1, 2, and 3 are related to leadership capacity building for three groups: site coordinators, Leadership Cadre (LC) members, and administrators. Components 4, 5, and 6 are related to teacher professional learning. Implementation of the six components takes place over the course of a two-year period.

Component 1 focuses on **professional learning for site coordinators**, one in each of the two states in which the study was conducted. The site coordinators' professional learning includes mentoring, coaching, individual assistance, and provision of support materials as site coordinators support LC members for the duration of the study, particularly at the summer courses and school-year Professional Learning Communities (PLC) meetings. Support materials available to the site coordinators included relevant articles and resources related to professional learning and implementing NGSS. Site coordinators also received resources to help them support facilitators (e.g., observation logs, rubrics), to guide their work with the LC members (e.g., sample agendas, emails), and to reach out to school/district administrators (e.g., handouts, report summaries). The expected outcomes for site coordinators are increased ability to build LC capacity and facilitate teacher professional learning and PLC meetings, and greater skill in providing technical assistance to upper administrators.

Component 2 focuses on **LC professional learning.** This component is designed around building the capacity for members of the LC, which includes teacher leaders, district staff, and regional partners (e.g., from universities, museums).

Making Sense of SCIENCE provides professional learning to LC members through: 1) LC workshops that prepare all LC members to support participating teachers and the implementation of Making Sense of SCIENCE, 2) a Teacher Course Facilitation Academy that prepares a subset of LC members to facilitate the Summer Teacher Course, and 3) the PLC Facilitation Academy that prepares a subset of LC members to facilitate the school year follow-up PLC meetings. The expected outcomes for this component are LC members' greater skill in facilitating teacher learning and supporting collaboration, as well as increased capacity to provide ongoing support for NGSS implementation.[1]

Component 3 focuses on **administrator professional learning.** Making Sense of SCIENCE provides professional learning to school administrators through workshops, once per year in each of the two years, to prepare them for supporting teachers and science teaching. The expected outcomes for administrators are an increased understanding of required shifts for NGSS instruction and greater understanding of effective teaching practices and professional learning.

Components 4, 5, and 6 are all related to **teacher professional learning.** Teacher professional learning consists of summer courses (35 hours each year for two years) and PLC meetings throughout both school years (12 hours each year for two years). More specifically, these three components focus on the delivery, process, and structure of the summer courses (component 4), teacher attendance at summer courses (component 5), and teacher attendance at the school-year PLC meetings (component 6). The expected outcomes for teachers are improved content knowledge (Hill et al., 2005; Kanter & Konstantopoulos, 2010) and pedagogical content knowledge, as guided by the definition of pedagogical content knowledge in the Refined Consensus Model (Carlson et al., 2019). Additional expected outcomes for teachers include a shift toward NGSS-aligned instructional practices and a shift in attitudes and beliefs, such as greater confidence in science teaching (Murphy et al., 2007), stronger belief that students are capable learners (National Research Council, 1996), and greater value placed on reflective practices (National Research Council, 1996).

As depicted in the logic model graphic (Figure 1), the impacts from the site coordinators, LC professional learning, administrator professional learning, and teacher professional learning would in turn cascade into impacts on teachers, schools, and classrooms. Posited impacts on teachers are discussed above. Posited impacts on school climate include administrators prioritizing and being involved in teacher professional and science teaching (Casey et al., 2012); greater availability of science resources and supplies; greater support for teacher collaboration (Iveland et al., 2017); greater trust and collaboration among teachers and between teachers and administrators (Briscoe & Peters, 1997; Brahier & Schäffner, 2004; Hallam et al., 2015; Urick et al. , 2018; Graham, 2007); and improved school culture conducive for learning (Bryk, 2010).

These changes for teachers and schools are posited to lead to changes in students' opportunity to learn (OTL) in the classroom: more time spent on science instruction; greater integration of science and literacy (Cervetti et al., 2012); and increased opportunities to engage in phenomena-based exploration (Achieve et al., 2016), scientific argumentation (NRC, 2013), and sense-making of hands-on investigations (McNeill et al., 2015). Students would also have more opportunities to engage in multi-dimensional learning (i.e. the integration of DCIs, SEPs, and CCCs).

The impact of Making Sense of SCIENCE on leaders, teachers, schools, and students' OTL would ultimately lead to students' higher achievement in science and literacy, as well as changes in their attitudes and beliefs, such as greater enjoyment of science, interest in science-related careers, and sense of self-efficacy in their learning (Tytler & Osborne, 2012; Bandura et al., 2001; Cavagnetto et al., 2020; Ainley & Ainley, 2011).

[1] Outcomes for the Leadership Cadre are addressed in the corresponding implementation report for this evaluation by Heller Research Associates (Wong et al., 2020).

As indicated by the arrows in the logic model, the causal pathways of these intermediate outcomes are by no means linear and one-directional. They have complex relationships with feedback loops, wherein the effects of x on y could trigger further effects of y on x, thus amplifying the final impacts.[2]

[2] The logic model terminologies and definitions provided by WestEd are presented in Appendix A.

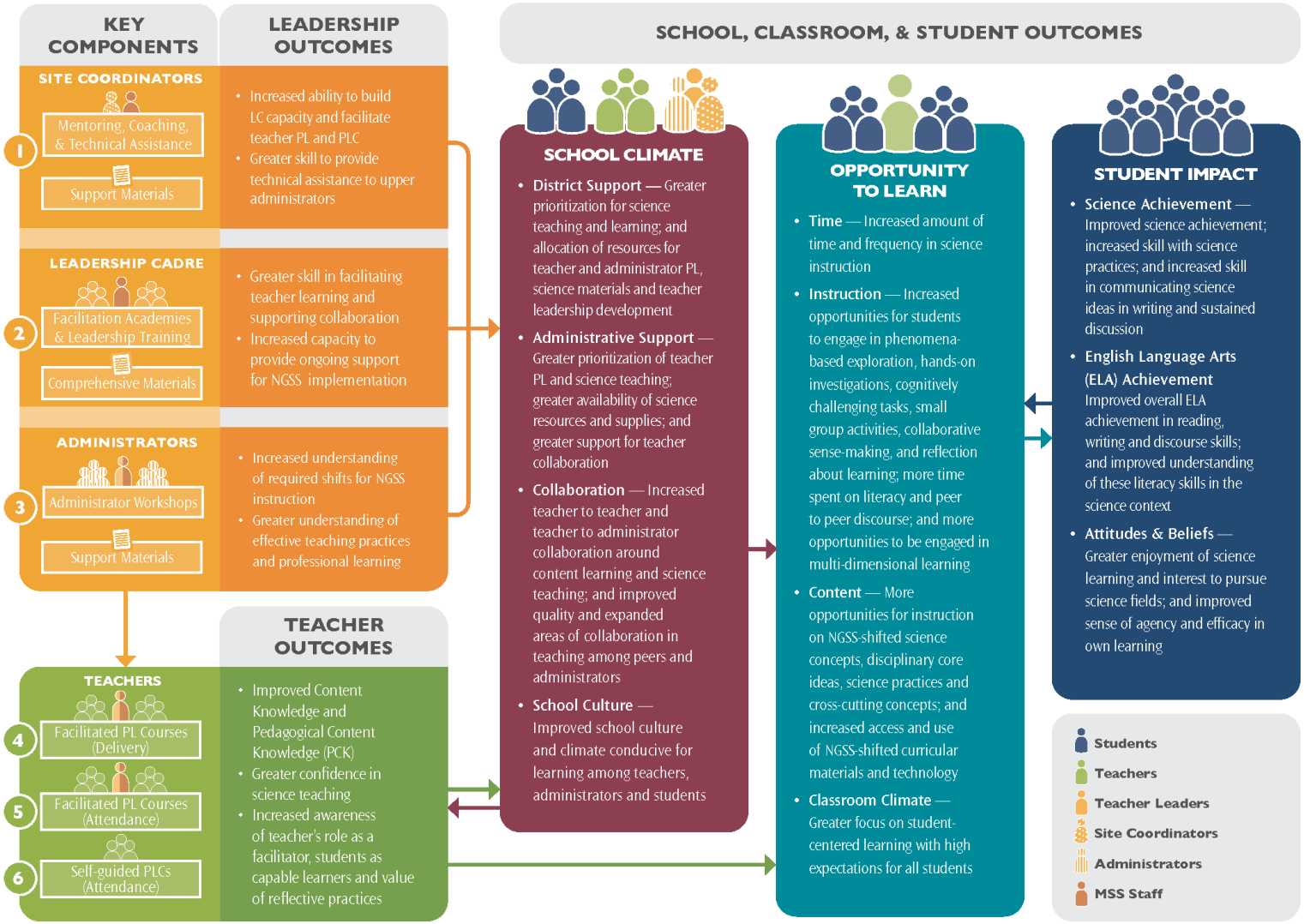


FIGURE 1. THE MAKING SENSE OF SCIENCE LOGIC MODEL

PL = professional learning; PLCs = Professional Learning Communities; NGSS = Next Generation Science Standards

Previous Research on Making Sense of SCIENCE

Making Sense of SCIENCE has participated in many rigorous experimental and quasi-experimental evaluations in the past two decades. We discuss here findings from three of the most recent studies, two of which met What Works Clearinghouse (WWC) group design standards without reservations (under review standards 3.0). One RCT conducted across 6 states and 49 school districts, with over 260 elementary teachers and 7,000 students found positive effects on teacher science content knowledge (ES ranging from 1.81 to 1.93 standard deviations, each significant at the $\alpha = .001$ level) and on student science content knowledge (ES ranging from 0.37 to 0.60 standard deviations, $p < .001$) (Heller et al., 2012). Another RCT conducted with over 130 middle school teachers from six sites with approximately 6,000 students found positive effects on teacher science content knowledge (ES = 0.38, $p < .01$) and on teacher confidence in teaching *Force and Motion* (ES = 0.49, $p < .01$). Before adjusting for multiple comparisons, students in the treatment group outperformed their counterparts in the control group by an effect size of 0.11 standard deviations ($p = .04$) for the full sample and by 0.31 standard deviations ($p = .04$) for the subset of English Language Learners. The effects on students for both samples were no longer significant after adjusting for multiple comparisons (Heller, 2012). The third RCT, which has not been reviewed by WWC, was conducted with middle school teachers from 62 schools across 11 districts. The study found that students of treatment teachers outperformed students of control teachers on a state standardized test by 0.17 standard deviations ($p = .09$), equivalent to nearly 6 months of additional learning based on a 9-month school year (Heller et al., 2017).

OVERVIEW OF THE IMPACT STUDY

This study evaluates the Making Sense of SCIENCE teacher professional learning model on elementary school teachers teaching science in fourth and fifth grades. The study is a two-year school randomized control trial (RCT) conducted in 66 elementary schools (60 randomization units)[3] in seven school districts across California and Wisconsin in the 2016–17 ("Year 1") and 2017–18 ("Year 2") school years.

For this study, program developers aimed to recruit a diverse group of districts that served high percentages of high-need students (defined as students who are low income or are English learners). Program developers also wanted to work in states that had either adopted or were highly likely to adopt the NGSS, or had state science standards that were based on the Framework for K–12 Science Education. California was chosen because the state adopted the NGSS in 2013, and the districts that were interested in participating in the study ranged in size and served the target student population. Wisconsin was chosen because during the recruitment period, the state was going through the process of adopting the NGSS, and the program developers had a committed Wisconsin district partner that served a high percentage of low-income students, but was different from the student population in California. In addition, program developers had previously worked in Wisconsin and had a highly qualified state coordinator, along with good relationships with the science leaders in the Wisconsin Department of Public Instruction.

The Making Sense of SCIENCE professional learning courses used in this intervention were Dynamic Earth in 2016–17 and Planet Earth in 2017–18. Note that these were teacher professional learning courses, not student curriculum. One

[3] There were 66 participating schools but only 60 units of randomization because there were 12 small schools that were combined into "dyads," each comprising two schools, and randomized as a single unit. Dyads were formed to accommodate small schools that did not have enough eligible teachers to participate in the study. In these cases, we allowed schools to partner up and implement MSS as if they were one school. Henceforth, units of randomization (54 schools and 6 dyads) will be referred to as "schools" for the purposes of the study.

characteristic of the Making Sense of SCIENCE approach is that the professional learning often takes on science topics for which teachers are the least prepared to teach. For elementary teachers, this includes Earth and space science and physical science topics (see Banilower et al., 2018). The courses Dynamic Earth and Planet Earth were developed as part of the grant, based on the Making Sense of SCIENCE approach to professional learning, to:

1. align with the NGSS content (e.g., DCIs, SEPs, CCCs);
2. help teachers understand the Earth and space science and physical science disciplinary core ideas necessary for teaching fourth and fifth grade students; and
3. give teachers an adult-level, first-hand experience with learning in the three-dimensional manner called for by the NGSS.

The study is guided by the following confirmatory and exploratory research questions.

Confirmatory Research Questions

1. What is the impact of Making Sense of SCIENCE, after two years of implementation, on student science achievement in Earth and space science and physical science among fourth- and fifth-grade students in intervention schools, compared to fourth- and fifth-grade students in control schools receiving the business-as-usual science instruction?
2. What is the impact of Making Sense of SCIENCE among fourth- and fifth-grade students *in the lowest third of pretest achievement*, after two years of implementation, on science achievement in Earth and space science and physical science in intervention schools, compared to students in control schools receiving the business-as-usual science instruction?
3. What is the impact of Making Sense of SCIENCE on teachers' science content knowledge in Earth and space science and physical science compared to teachers in the business-as-usual control schools, after two full years of implementation in schools?

Exploratory Research Questions

The study also examines a series of exploratory research questions, which can be loosely classified by their purpose into three sets. The first set aims to examine impacts on additional measures of student achievement including an assessment of "communicating science ideas in writing," state tests in ELA and math in fourth and fifth grades, and a state science assessment. The state science assessment was limited to fourth grade in Wisconsin. Data for the California state science assessment were not available during the study years. The second set of exploratory research questions aims to evaluate possible impacts on precursor variables that could potentially mediate impacts on teacher content knowledge and student achievement. This includes self-reported measures: teachers' instructional practices, OTL, and school climate; and self-reported measures of students' non-academic outcomes (including enjoyment of science, agency in learning science, self-efficacy in learning science, cognitive demand of science instruction, quality of and activities in their science class, and aspirations about future use of science in their adulthood and career). The third set of research questions aims to examine average and differential impacts across more specific samples and conditions than for the sample-wide analysis that was used to assess the confirmatory impacts. This includes assessing average impacts on student science achievement for subsamples of teachers and students with longer exposure to the program and stronger implementation, as well as by grade, state, and district. We also include in this category the evaluation of impact on student science achievement using (1) alternative scaling of the student science achievement assessment, (2) a subtest based on items that were more-highly

discriminating of ability, and (3) separate scores for Earth and space science and physical science sub-strands. We also assess whether impact on student and teacher outcomes differs by whether or not the teacher is also a teacher leader.

Chapter 2. Study Methods

EXPERIMENTAL DESIGN

Recruitment

Recruitment Process

The recruitment and informed consent process took place in 2015–16, a year prior to the start of the evaluation. In the fall, WestEd recruited schools and teachers from seven districts in Wisconsin and California to participate in the evaluation.[4] In the winter of 2015–16, WestEd, Empirical, and HRA held a series of online and in-person meetings with teachers and administrators to introduce them to the Making Sense of SCIENCE model and evaluation and to walk them through the informed consent process. When interested teachers and administrators were unable to attend these meetings, researchers sent them the recorded webinars with the same information and offered them the opportunity to ask questions before consenting. The research team emphasized to study participants the importance of remaining in the study to minimize attrition, cautioned against risks of contamination, and informed participants on steps to take should either occur.[5]

Determining Eligibility

The research team relied on a number of criteria to determine the eligibility of schools and teachers. Eligible schools had to belong to school districts that serve low-income students, as defined by the percentage of students who are eligible for the Free and Reduced-Price Lunch (FRPL) program. Schools also had to have at least three eligible teachers who were willing to participate in the evaluation. Schools that were interested in participating but did not meet the minimum threshold of three committed teachers at the school were allowed to partner with another school in their district to participate in the study as one singular unit. The two schools had to be sufficiently near each other geographically so that teachers could collaborate across schools. Teachers and administrators from the two schools had to actively express understanding that the implementation model requires teachers to collaborate across schools and to function as a unit for the duration of the study.

Teachers were eligible if they were expected to teach science to at least one class that had students in the fourth, fifth, or a combination of the two grades in both Years 1 and 2 of the study. Teachers who taught self-contained special education classes with all students taking an alternative state test and teachers who taught sheltered English Language Learner classes were not eligible to participate in the study.

---

[4] Empirical submitted to each district a research application that included an overview of the study, the data collection plan, sample instruments, and draft consent forms. Upon receiving the district's approval of the research application, the CEO of Empirical and the superintendent of each district (or assigned designee) signed a district agreement detailing what data the district agreed to provide and the roles and responsibilities of each party in supporting the evaluation. Each district also assigned a point of contact to streamline communication between Empirical and the various district departments.

[5] Whenever we were notified that a teacher left the school, we followed up with the site coordinator or the teacher leader to ask whether the teacher moved to another teaching position in one of the participating districts. There were four known cases of teachers transitioning from a treatment to a control school, and in all four, we contacted the teacher to remind them about the risk of contamination to the study should knowledge or materials acquired through MSS be shared with colleagues at the control school. All teachers acknowledged receipt of the message.

Site Description

The study was conducted in California and Wisconsin across seven school districts: two small suburban districts and five urban districts (one small, one midsize, and three large). The districts ranged from approximately 10–30% in regard to English learner students and 10–20% in regard to students with individualized education plans. With respect to race/ethnicity, districts ranged from 2–45% Black students and 2–66% Hispanic students (NCES, 2014). In regard to economic status, the percentages of families with income below the poverty level ranged from 7–34%, with annual median household income ranging from approximately 40,000 to 70,000 USD (NCES, 2020).

Randomization

We randomized schools to either *Making Sense of SCIENCE* or business-as-usual ("control").[6] We conducted randomization at the school level to not disrupt school-level implementation of science programs in their respective conditions and to avoid contamination that could possibly occur if randomization had been conducted within schools. Making Sense of SCIENCE professional learning encourages collaboration of teachers within schools through formal (e.g., through PLCs) and informal channels (e.g., sharing or discussing lesson plans among individual teachers). This type of collaboration would put the study at high risk for contamination if randomization was conducted at units lower than the school (e.g., teacher- or class-level).

We randomized schools in the winter of the 2015–16 school year. Our power analysis in the study plan assumed 60 units randomized with 54 retained, assuming a Minimum Detectable Effect Size for impact on student science achievement of .22. We achieved 60 school during recruitment. There was a total of six dyads[7] (12 schools): four dyads (8 schools) in the *Making Sense of SCIENCE* group and 2 dyads (4 schools) in the control group. Our analysis of impact on student science achievement in Spring 2018 included 55 schools from among those randomly assigned.

We randomized schools within matched pairs within districts. The primary criteria used to establish matched pairs were school-level state standardized achievement scores in mathematics for grades 4 and 5 (from 2014–15 in California and 2013–14 in Wisconsin) and pre-randomization school-average teacher performance on an assessment of science content knowledge ("teacher pretest"). We also used the average amount of time allocated for science instruction, as reported by teachers on the baseline survey, and an index of school climate based on 22 Likert scale items from the teacher baseline survey to inform block formation. The process of identifying matched pairs involved calculating Euclidean distances between schools within districts using standardize measures of student achievement and teacher pretest scores as the base dimensions. We also took into consideration similarities in teacher-reported measures of school climate and time spent on science.[8]

---

[6] The business-as-usual group received delayed treatment starting in late spring of Year 2 (2017–18), after all data collection activities for the study were complete.

[7] See Chapter 1, page 7 for a definition of dyads.

[8] After matching pairs of schools based on their proximity in terms of Euclidean distances, in several cases, we overrode the results to make sure that schools with certain characteristics were included in the same pair. This strategy was used to ensure that schools that share a specific characteristic do not, by chance, all end up in the same condition (i.e., since two schools in a given pair end up assigned to opposite conditions, putting schools with a shared characteristics in the same pair ensures that at least some schools with that characteristic end up in each condition). This was done with factors that would be undesirable to be imbalanced between, or completely confounded with, conditions. For example, we made sure that schools with programs that focused heavily on science were placed into a pair, so that not all such schools would end up in only one condition by chance. If this happened, it would complicate the

The Counterfactual

Critical to gauging the contrast between treatment and control is the understanding of the context and the counterfactual, particularly in regard to NGSS adoption, participation in professional learning for science instruction, and science instructional resources used by participating teachers.

NGSS Adoption

Given that Making Sense of SCIENCE intends to support districts and schools in their transition to NGSS-aligned instructions, it is important to understand where participating states were on the trajectory of NGSS adoption and implementation. The expectation was that the control group in a state that was far along in implementing NGSS-aligned content and instruction, compared to the control group in a state that was early in its implementation of NGSS, would likely have a weaker contrast with the *Making Sense of SCIENCE* group in regard to district and school supports for NGSS-aligned instructions.

As mentioned above, NGSS was released in April 2013. California adopted *the NGSS for California Public Schools* soon after in September 2013. In November 2014, the State Board of Education (SBE) approved a statewide plan to implement the new standards. In November 2016, the SBE adopted a new science curriculum framework to provide guidance to educators, parents, and publishers, to support implementing California content standards. Organized by theme and grade-level, the framework provides a vision of science instruction in a classroom and examples for teachers to use as a starting point (d'Alessio, 2018).

In contrast to California's statewide model of adoption and implementation of NGSS, Wisconsin is a local-control state in regard to standards. Wisconsin allowed districts to use the Wisconsin Standards for Science (based on the NGSS), locally determined standards, or the NGSS themselves. Districts in the state varied in regard to how far along they were in implementing NGSS. To illustrate, certain districts made available instructional guides during the study years while others did not release an NGSS-aligned curriculum until after the study was completed.

Instructional Materials

We asked teachers to select the top three instructional resources that they used or planned to use for science instruction. Figure 2 and Figure 3 below present the types of instructional materials used by participating teachers in California and Wisconsin, disaggregated by randomization status.

---

interpretation of the difference in outcomes between *Making Sense of SCIENCE* and control schools. As a final step, we ran sensitivity checks on several possible pairing schemes by randomizing according to each scheme 1,000 times. We assessed which scheme produced the greatest balance, on average, on critical characteristics. We selected the approach that produced the best balance on average, and then conducted one official randomization using that scheme. Randomization resulted in balance between conditions on important baseline factors. For variables used to define blocks, we observed the following differences between conditions following randomization: Standardized Effect Size (ES) of -0.13 (p = .385) for school-average math pretest, ES of 0.07 (p = .611) for school average ELA pretest, ES of -0.10 (p = .634) for teacher baseline content knowledge score (based on the MOSART assessment), with 78.1% of schools being Title 1 in control and 88.2% being in Title 1 in treatment (p = .121 associated with the difference in means), and with 75.4% and 79.6% of students being eligible for Free or Reduced Price lunch in control and treatment schools respectively, (p = .360 associated with the difference in means). (All effect sizes reported here use the pooled standard deviation of outcomes reported at the school level in the denominator.)

In California, notably, nearly half of teachers reported using or planning to use Mystery Science in both *Making Sense of SCIENCE* and control groups. For all resources, the difference in the proportion of *Making Sense of SCIENCE* teachers and control teachers selecting the resource was less than 10 percentage points. A Chi-square test of the proportions of resources selected by teachers across the two groups suggested no differences between the two groups: X^2 (5, n = 103) = 2.16, p = .827 (Figure 2).

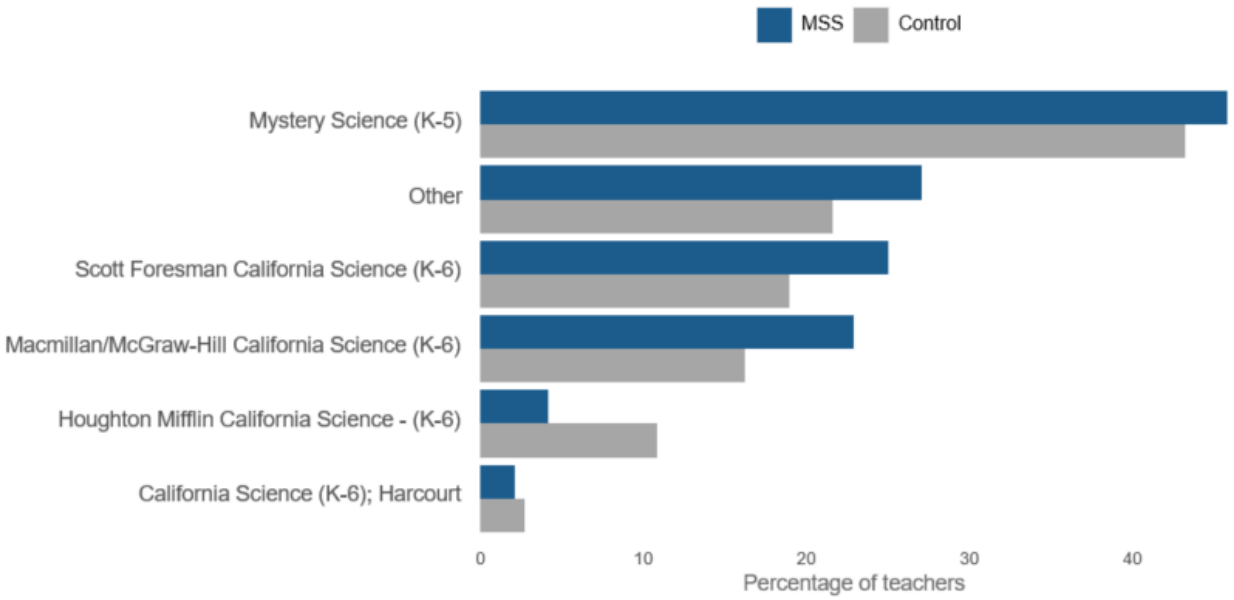


FIGURE 2. DISTRIBUTION OF INSTRUCTIONAL MATERIALS USED BY STUDY PARTICIPANTS IN CALIFORNIA, BY RANDOMIZATION STATUS

Note. Sample consists of 85 teachers in California: 48 *Making Sense of SCIENCE* + 37 control. Because each teacher could select more than one instructional resource, the percentages do not sum to 100%. *Full Option Science System (FOSS) (K-5); Delta Education* was a seventh option, but none of the teachers in the sample selected it.

Figure 3 displays the instructional resources used by study participants in Wisconsin. The top three instructional resources that are most frequently selected by teachers are Science A-Z, Discovery Education Science Elementary, and BrainPop, each selected by approximately 70% of *Making Sense of SCIENCE* teachers. The differences in proportions of teachers selecting the resource were less than 10 percentage points for all resources except for two (Discovery Education Science Elementary and Science A-Z). Notably, only 35% of control teachers, compared to 73% of *Making Sense of SCIENCE* teachers, reported using Science A-Z. A Chi-square test of the proportion of instructional resources across the two conditions, however, shows no statistical significance: X^2 (5, n = 159) = 3.32, p = .651.

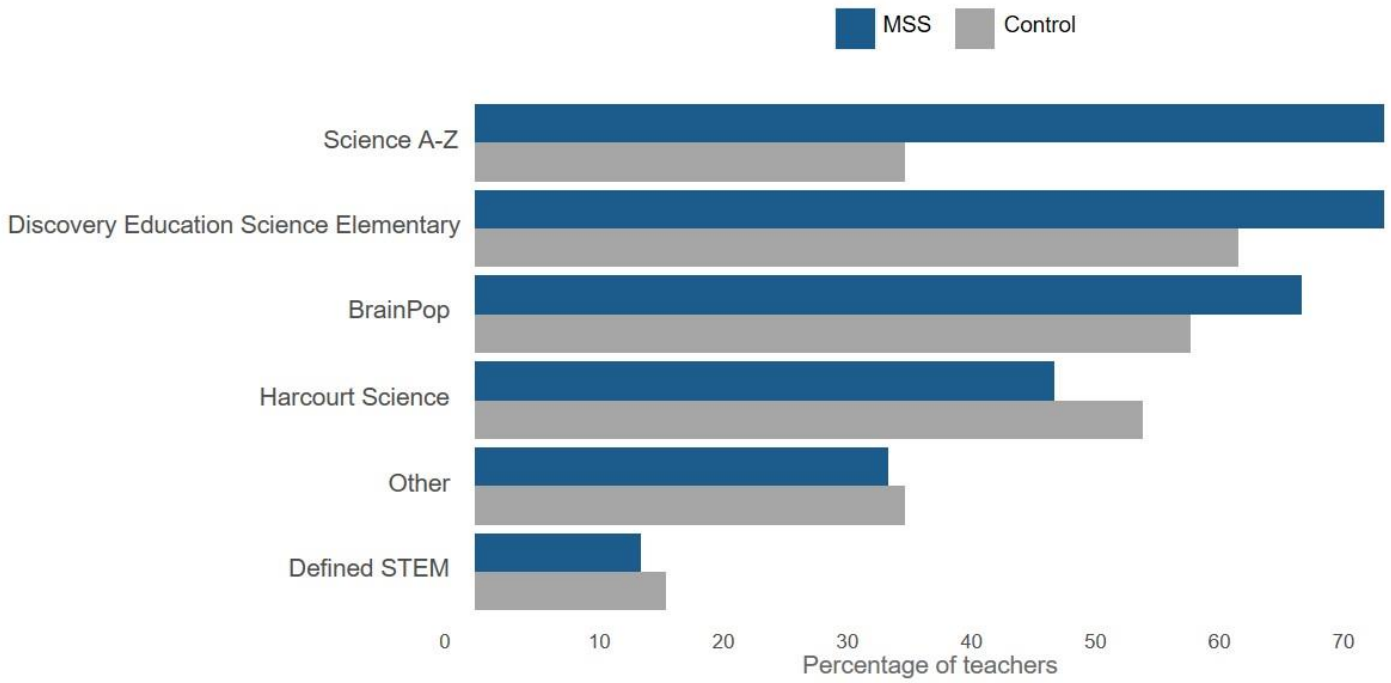


FIGURE 3. DISTRIBUTION OF INSTRUCTIONAL MATERIALS USED BY STUDY PARTICIPANTS IN WISCONSIN, BY RANDOMIZATION STATUS

Note. Sample consists of 56 teachers in Wisconsin: 30 *Making Sense of SCIENCE* (MSS) + 26 control.

Professional Learning

Given that the two major components of the Making Sense of SCIENCE professional learning models were the summer institute and the school-year PLCs, we examined the professional learning that the *Making Sense of SCIENCE* and control groups received during the study years to understand the treatment-control contrast. In the fall of Year 2 (2017–18), we asked teachers how much science professional learning[9] they received between June and October 2017. As part of tracking and monitoring fidelity of implementation (discussed in Chapter 3), we also collected teachers' attendance at Making Sense of SCIENCE summer courses and school-year PLCs. We asked teachers in the *Making Sense of SCIENCE* group to exclude any Making Sense of SCIENCE PLCs and Making Sense of SCIENCE summer courses. In Figure 4, we compare responses from the two groups. We also report the amount of professional learning that *Making Sense of SCIENCE* teachers received through the Making Sense of SCIENCE summer courses, as measured by their attendance at these courses.

While the majority of control teachers (79%) reported not having received any professional learning during this time period, the majority of *Making Sense of SCIENCE* teachers (90%) reported receiving 9–40 hours (2–5 days). Notably, *Making*

[9] Examples of professional learning that we asked teachers to consider included district professional learning, summer institutes, museum workshops, online courses, and conference sessions.

Sense of SCIENCE teachers also reported receiving additional professional learning beyond what was provided by Making Sense of SCIENCE.[10]

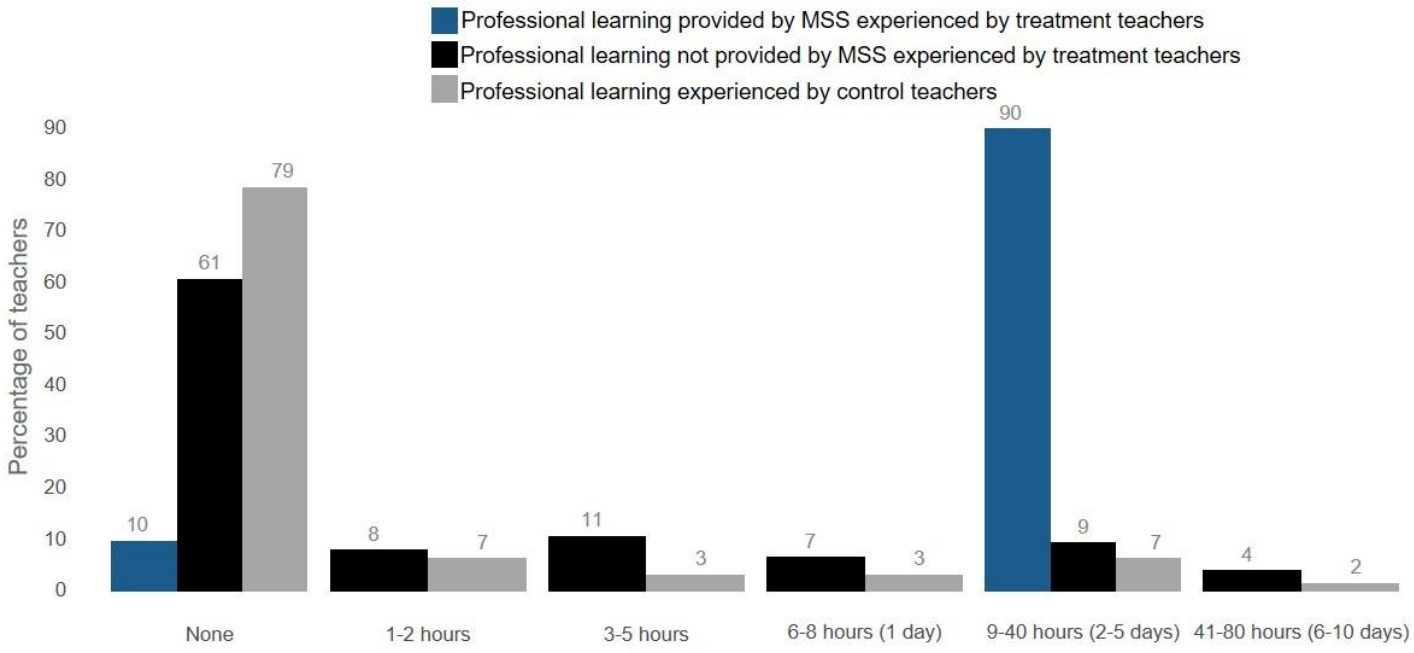


FIGURE 4. TIME TEACHERS SPENT IN SUMMER AND FALL SCIENCE PROFESSIONAL LEARNING (NOT PROVIDED BY MAKING SENSE OF SCIENCE)

Note. Sample consists of 142 teachers: 81 *Making Sense of SCIENCE* (MSS) + 61 control. There was a seventh option of "more than 10 days," but no teacher in this sample selected that option.

A similar trend was observed for participation in PLC meetings. Seventy-two percent of control teachers, compared to four percent of *Making Sense of SCIENCE* teachers, reported not having participated in any science-related PLC meetings between February 16 and April 25, 2018. At the other extreme, 8% of control teachers, compared to 49% of *Making Sense of SCIENCE* teachers, reported spending more than four hours in science-related PLC meetings during the same time period. Note that in Figure 5, we did not report the time spent at PLC meetings that were not provided by Making Sense of SCIENCE (as we did with the summer courses in Figure 4) because we didn't ask *Making Sense of SCIENCE* teachers this question.

[10] There is a possibility that this is an indication of recall bias. Having participated in Making Sense of SCIENCE professional learning, it might be challenging for *Making Sense of SCIENCE* teachers to discern between Making Sense of SCIENCE and non-Making Sense of SCIENCE professional learning.

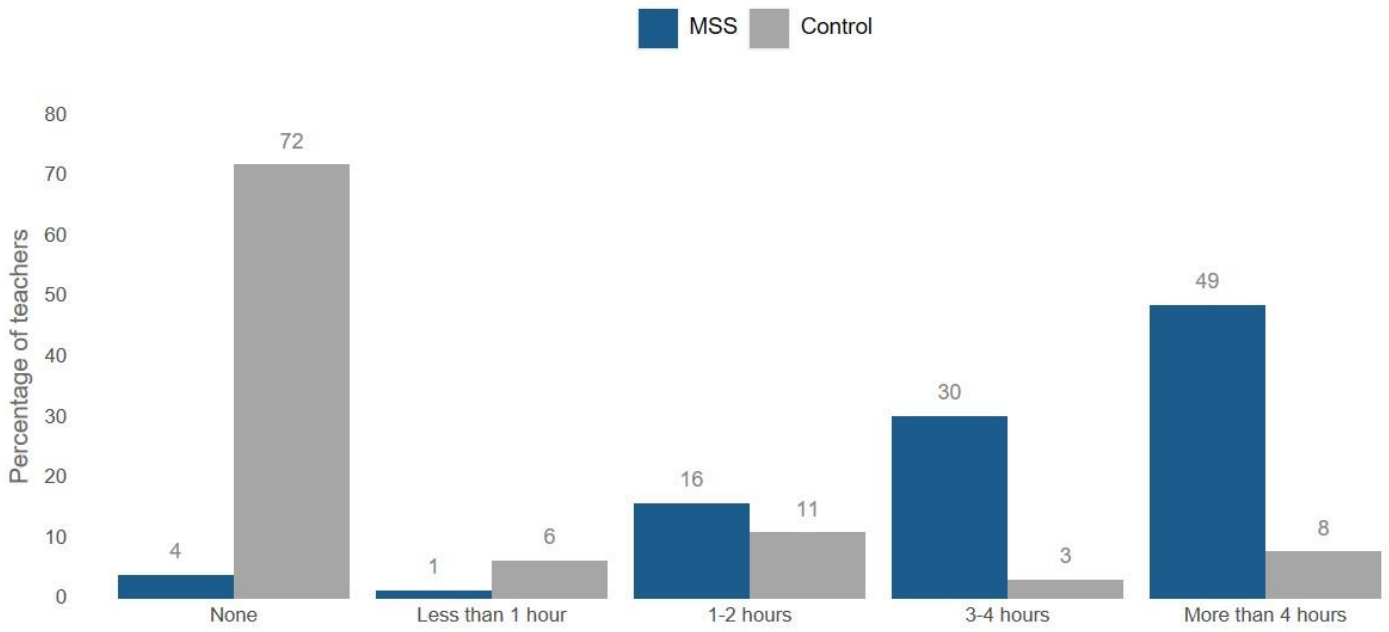


FIGURE 5. TIME IN PROFESSIONAL LEARNING COMMUNITY MEETINGS

Note. Sample consists of 140 teachers: 76 *Making Sense of SCIENCE* (MSS)+ 64 control. Responses are from the spring Year 2 survey asking teachers to about the total time spent at PLC meetings between February 16 and April 25, 2018. We asked *Making Sense of SCIENCE* teachers about the total time they spent in Making Sense of SCIENCE PLC meetings. We asked control teachers about the total time spent in science-related PLC meetings.

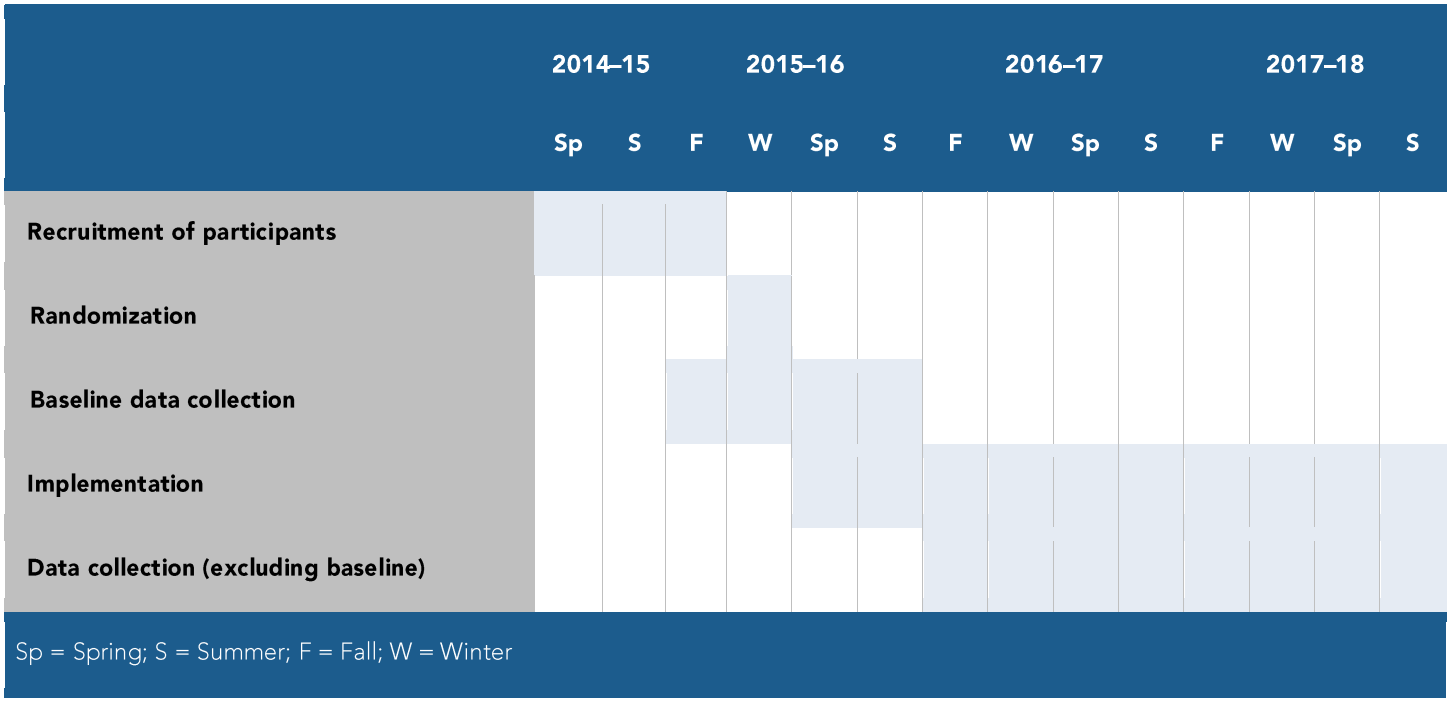
In summary, in regard to NGSS adoption and implementation, California was one of the earliest states to adopt the standards. The state released a curriculum framework at the beginning of the study. Wisconsin, a local-control state, was more varied in its adoption and implementation, with districts allowed to choose to use NGSS, the state standards, or locally determined standards. In California, the types of instructional resources used appeared to be similar across the *Making Sense of SCIENCE* and control groups. In Wisconsin, they also appeared to be similar, except for a disproportionate number of *Making Sense of SCIENCE* teachers, compared to control teachers, reported using Science A-Z. Regarding professional learning, *Making Sense of SCIENCE* teachers reported receiving more science professional learning and spending more time in PLC meetings than control teachers.

SCHEDULE OF MAJOR MILESTONES

The impact study is part of a five-year grant that began in January 2015 and ended in September 2020, with two six-month no-cost extensions to finalize reporting. The first two years of the grant (2014–15 and 2015–16) were allocated to study design, recruitment (of districts, schools, and teachers), and randomization. Implementation occurred during two years: 2016–17 and 2017–18. The remainder of the time was allocated to data analysis, reporting, and dissemination. The impact study formally ends with the submission of this final report to WestEd and NEi3. Table 1 presents a timeline of the major activities of the study during study design, implementation, and data collection for the impact study. Information about

the parallel efforts by HRA in conducting the implementation and scale-up studies under this grant are presented in a separate report (Wong et al., 2020).

TABLE 1. TIMELINE OF MAJOR ACTIVITIES IN THE STUDY



MEASURES

The impact evaluation collected a rich set of assessment and survey data from teachers, students, and school districts. Table 2 presents the full set of measures, their source, data collection timeline, and reliability statistics.

TABLE 2. MEASURES

Measure	Adapted from	2014–15	2015–16	2016–17	2017–18	Reliability
Teacher content knowledge and pedagogical content knowledge						
Science Content Knowledge	Evaluator-developed: 46 selected response items adapted from MOSART		Rolling Pre-RA			.91
	Evaluator-developed: 32 selected response items adapted from MOSART			Pilot in spring	spring	.78
Pedagogical Content Knowledge	Evaluator-developed: Items developed by HRA in prior research studies about Making Sense of SCIENCE (e.g. Heller et al., 2010; Daehler et al., 2015)			Pilot in spring	spring	.77 (interrater reliability)

TABLE 2. MEASURES

Measure	Adapted from	2014–15	2015–16	2016–17	2017–18	Reliability
Teacher attitudes and beliefs, opportunity to learn, school climate	Measures were constructed by researchers based on survey items that were developed by evaluators or adapted from the National Survey of Science & Math Education and Surveys of Enacted Curriculum.		Rolling Pre-RA	fall, winter, spring	fall, winter, spring	
Teacher attitudes and beliefs	8 measures					.61 to .97
Opportunity to learn - Time on science	1 outcome					NA
Opportunity to learn – instruction	4 outcomes					.69 to .89
Opportunity to learn – content	10 outcomes					.74 to .92
School climate	7 outcomes					.68 to .91
Demographics; Education & Teaching Background	Evaluator-developed survey items		Rolling Pre-RA			NA
Student achievement measures						
Science achievement	Evaluator-developed assessment: 30 fourth grade and 29 fifth-grade selected response items			Pilot in spring	spring	.69 (4TH grade form) .56 (5th grade form)
Communication of science ideas in writing	Evaluator-developed assessment 8 constructed response items			Pilot in spring	spring	80.7 – 96.4 (interrater reliability)
Math	State assessments	spring	spring	spring	spring	WI: .92 (gr 4 & 5)[b] CA: .82 (gr 4 & 5) (marginal reliability[a])

TABLE 2. MEASURES

Measure	Adapted from	2014–15	2015–16	2016–17	2017–18	Reliability
English language arts	State assessments	spring	spring	spring	spring	WI: .90 (gr 4 & 5)[b] CA: .88 (gr 4 & 5) (marginal reliability[a])
Science	State assessments	spring	spring	spring	spring	WI: .88 (gr 4)[b] CA: NA
Student non-academic measures						
Survey scale 1 - Aspirations	Adapted from Friday Institute for Educational Innovation Elementary School STEM - Student Survey			Pilot in spring	spring	.584
Survey scale 2 – Quality of Science Class (Learning Environment / Classroom Management)	Adapted from Colorado Education Initiative Student Perception Survey			Pilot in spring	spring	.812
Survey scale 3 – Self-Efficacy	Adapted from TIMSS 2015 Student Questionnaire			Pilot in spring	spring	.762
Survey scale 4 – Activities in Science Classrooms	Adapted from TIMSS 2007 Student Questionnaire			Pilot in spring	spring	.602
Survey scale 5 – Quality of Science Class (Science Instruction)	Adapted from TIMSS 2015 Student Questionnaire			Pilot in spring	spring	.629
Survey scale 6 – Agency in Learning	Evaluator-developed			Pilot in spring	spring	.109
Survey scale 7 – Cognitive Demand of Science Class	Evaluator-developed			Pilot in spring	spring	.576
Survey scale 8 – Enjoyment of Science	TIMSS 2015 Questionnaire			Pilot in spring	spring	.865

TABLE 2. MEASURES

Measure	Adapted from	2014–15	2015–16	2016–17	2017–18	Reliability
Student demographics (from district or school administrative records)		N/A	N/A	N/A	N/A	N/A

Note. Reliability is Cronbach's alpha unless otherwise noted.

MOSART = Misconception-Oriented Standards-Based Assessment Resources for Teachers; RA = randomization; HRA = Heller Research Associates; CA = California; WI = Wisconsin

See Appendix B for details on the process for constructing each of the outcomes, including items used, aggregation method, and Cronbach's alpha.

[a] Description of "marginal reliability" is provided by the California Department of Education Assessment Development & Administration Division (2019).

[b] Reliability statistics are provided by the Wisconsin Department of Public Instruction (2018).

Teacher Baseline Content Knowledge Assessment

Teachers completed a science content knowledge assessment ("teacher pretest"), which measured their baseline knowledge of Earth science and physical science, upon joining the study. The pretest was developed by WestEd using items adapted from the Misconception-Oriented Standards-Based Assessment Resources for Teachers (MOSART) test inventory. MOSART consists of multiple-choice items that are linked to the K–12 physical science, K–12 Earth science, and K-8 life science content in the NRC National Science Education Standards, as well as to misconceptions related to science concepts as documented in research literature. MOSART "probe[s] for any conceptual shift(s) as a result of professional learning activities, course work, or other intervention" (MOSART, 2011). WestEd staff selected 46 items from the MOSART test inventory and piloted the test with 15 teachers. Results from the pilot yielded an internal consistency reliability (Cronbach's alpha) of .91.

We administered the pretest to teachers at the end of each informed consent and data collection meeting. Teachers submitted completed assessments directly to research team members. We then entered teachers' answer choices using a double data entry process by two research team members.[11] The resulting file, which had one set of pretest responses for each teacher, was then scored and warehoused by an Empirical warehouse engineer.

Teacher Content Knowledge Assessment

The TCK assessment was an evaluator-developed instrument consisting of 32 selected-response items (29 were retained in scoring) and administered as one form. The assessment included items taken or adapted from the MOSART test and the New York State Education Department's Regents High School Examination. The items used in Year 2 (2017–18) were

[11] The data entered were compared using a reconciliation tool to identify any discrepancies, which could have occurred if the data entry staff made a manual error or if the two data entry staff members had different opinions about certain responses that are not clearly marked. In case of a discrepancy, those who entered the data discussed and agreed on a resolution.

piloted during Year 1 (2016–17) of the study. We selected items for use on the basis of their degree of difficulty, high point-biserial correlations, and alignment with NGSS DCIs. The Cronbach's alpha value for retained items for this scale was .78. More-detailed, item-specific information is provided in Appendix C Teacher Content Knowledge Assessment: Descriptive Statistics and Item-Level Information.

Teacher Pedagogical Content Knowledge

The PCK instrument used in this study, adapted from PCK items developed and used by HRA in prior research studies about Making Sense of SCIENCE (e.g. Heller et al., 2010; Daehler, Wong, & Heller, 2015), was a cluster of prompts centered around one of the constructed-response assessment tasks presented to fourth and fifth-grade students in this study. The instrument aimed to assess three areas of teachers' abilities and knowledge relating to weather and erosion:

1. ability to interpret student work;
2. knowledge of typical student difficulties;
3. knowledge of effective instructional strategies for supporting fourth- and fifth-grade students in making observations, providing evidence, and constructing scientific explanations; and
4. the explicitness of teachers' pedagogical reasoning.

The instrument asked teachers to evaluate a hypothetical student's response to a question about erosion that asked students to respond to three prompts: (1) make a claim, (2) provide evidence, and (3) explain reasoning. Teachers were asked to (a) state the strengths and weaknesses in the student's response, (b) state specific difficulties students may have in responding to the item, separately from the specific difficulties exhibited by the hypothetical student, and (c) describe activities to support the student. The interrater reliability (that is, percent agreement between scorers) for the PCK items was 76.7%.

Teacher Surveys

The teacher baseline survey was administered to inform randomized blocks (matched pairs or triplets), to track participants administratively, to establish baseline equivalence for the analysis sample, to use as covariates in the impact analysis to increase the precision of the estimates, and to serve as potential moderators of impact.

The purposes of the post-randomization surveys administered during the study years were to track administrative information about teachers' teaching positions (e.g. grades and subjects taught), to measure teachers' responses on intermediate outcomes of interest, to understand the supports and barriers to implementation, and to provide information about the treatment-control contrast.

The development of the teacher surveys was a collective effort between Empirical and HRA, in close consultation with WestEd. Empirical and HRA researched and compiled a set of items from the 2012 National Survey of Science and Math Education (Weis, 2013), Surveys of Enacted Curriculum, and from each respective team's item banks. The items covered a range of topic areas, such as the following.

- **Science instruction** – the amount of time allocated to science instruction, the level of priority given to particular topic areas, and the barriers to teaching and learning science
- **Attitudes and beliefs** – the level of influence and confidence with respect to teaching science
- **Teaching philosophies and pedagogical techniques** – beliefs about students; instructional practices

- **NGSS** – familiarity and level of comfort with the NGSS, and the extent to which they believe NGSS aligns with their own teaching philosophies
- **School climate** – school climate and the dynamics among administrators and teachers at their school; collaboration among peers
- **Professional learning** – previous professional learning experiences
- **Education and teaching background** – number of college courses in science; years of teaching experience (on baseline survey only)
- **Demographic information** – race/ethnicity, gender (on baseline survey only)

There were approximately 40–65 questions on each survey. Using responses from these surveys, we finalized the 30 intermediate outcomes related to teacher attitudes and beliefs, student OTL, and school climate. Cronbach's alphas for the resulting composites ranged from .61 to .97 (see Table B1 in Appendix B for details, including Cronbach's alpha for each outcome).

Empirical administered all teacher surveys for the impact study through the online survey platform.

Assessment of Student Science Achievement[12]

To assess students' science achievement, we used an evaluator-developed assessment covering Earth and space science, physical science, and life science. We created separate assessments for fourth grade and for fifth grade using items that were appropriate for students in the particular grade. There were 10 "inquiry items" that were suitable for both grades and were included in both assessments. We selected items from several sources (e.g., MOSART and NAEP) to address general NGSS-aligned specifications (information about student test forms and basic item statistics are provided in Appendix D). Students received the assessment through an online platform that included voiceover functionality such that students could click on the questions to hear them read aloud. Students had approximately one hour to complete the assessment and the student survey (described further below).

Communication of Science Ideas in Writing

To assess students' communication of science ideas in writing, we used a pool of eight constructed response (CR) items compiled by HRA. Among the eight items, six were drawn from NAEP and two were developed by HRA to address necessary specifications not covered by the NAEP items. Four of eight items were appropriate for and administered in fourth and fifth grades. The remaining four items were administered in fifth grade only. Information about the student test forms are provided in table D1 of Appendix D. The interrater reliability ranged from 80.7% to 96.3%, with median value 92.7% (Wong et al., 2020). The full details concerning test development and scoring are provided in the companion report to this one by HRA (Wong et al., 2020).

Student Survey

The student survey was administered along with the science assessment. The survey aimed to measure two types of outcomes. The first type included outcomes related to students' opportunity to learn, such as quality of science classroom

[12] A description of how we went about selecting a student science assessment, the development of the assessment, and the approach used to scale the posttest scores are described in Appendix D.

in regard to classroom management and science instruction, activities in science, and the level of cognitive demand for those tasks. The second type included distal, non-achievement outcomes, such as students' sense of agency and self-efficacy in science learning, aspirations for future science learning in their adulthood, and application of science in their careers. The student survey consisted of six scales adapted from the Friday Institute for Educational Innovation, TIMSS 2015 Questionnaire, and the Colorado Education Initiative. Modifications include the addition or removal of items, and modifications to the response scales. We also created two survey scales to measure cognitive demand and agency in learning. Cronbach's alpha values ranged from .11 to .87 with a median value of .62 (scales are provided in Appendix E).

District/School Data Requests

From all participating school districts, we collected the following data for first through fifth graders in participating schools: class rosters, student demographic data, and state assessment data for the 2014–15 through 2017–18 school years. For state assessment data, we requested math, ELA, and science scores for students in all tested grades. Additionally, we requested third-grade math and ELA assessment data ("pretest scores") for all students with a posttest in spring 2017–18.

For students in Wisconsin, the assessment data are based on the Wisconsin Forward Exams: math and ELA for grades 3–8 and science for grade 4.[13] The assessments, which were first rolled out in 2015–16, are administered online and are based on the Wisconsin Academic Standards. In Year 2 of the study (2017–18), when final outcome data would be collected, Wisconsin was still administering the Wisconsin Forward Exam with science items aligned to Wisconsin's Model Academic Standards for Science and enhanced by the NGSS. It was not until spring 2019 that Wisconsin administered the new science tests aligned with the Wisconsin Standards for Science, which is based on NGSS (Wisconsin DPI, n.d.a).

For students in California, the ELA and math assessment data are based on the Smarter Balanced Assessment Consortium (SBAC) assessments from first through fifth grade.[14] The SBAC assessments, which use computer-based tests and performance tasks, are based on the CCSS. The science assessment is administered to students in grades 5 and 8, and once in high school. However, since 2013, California has developed the California Science Test (CAST) that is aligned with the NGSS for California Public Schools. No science test scores are available for California for 2016–17 and 2017–18 because the state was in the process of piloting and field testing the CAST assessment and has not made student scores from these two years available.

Other Data Collected

The study team also surveyed administrators (see Appendix F for a description). The key findings based on the administrator survey responses are reported in HRA's implementation report (Wong et al., 2020). In addition, the study team conducted a pilot of classroom video recordings in Year 1 (2016–17) to estimate parental consent response rates and determine the feasibility of scheduling for the full sample of schools. Due to the low consent response rate, particularly in districts that require active parental consent to collect video recordings of classrooms, we decided to not proceed with video recordings for the full sample in Year 2. Instead, in Year 2 (2017–18), we collected audio recordings of science lessons for a subset of teachers. We also asked teachers to provide accompanying lesson artifacts, such as lesson plans and student works, and to audio record themselves responding to a set of interview questions about the lessons. Again, due to the low consent response and completion rates, these data were not analyzed for the impact study. They were, however,

[13] In Wisconsin, students are tested in math and ELA in grades 3-8, science in grades 4 and 8, and social studies in grades 4, 8, and 10.

[14] In California, students are tested in math and ELA in grades 3–8 and 11, and science in grades 5, 8, and once in high school.

analyzed and included in the implementation study by HRA (Wong et al., 2020). We provide additional details about the video and audio data collection efforts in Appendix G.

FORMATION OF THE STUDY SAMPLE

School sample. At baseline, in the winter of 2015–16, we randomly assigned 60 schools to either *Making Sense of SCIENCE* or control (see Table 3 for all sample sizes).

Teacher sample (RCT confirmatory analyses). Within these schools, 269 teachers were randomly assigned to conditions by virtue of their schools being randomly assigned. This initial roster of teachers responded to baseline surveys and took a baseline science content knowledge assessment prior to random assignment. This group of teachers is referred to as the "Present at Randomization" (PAR) sample.

From among this group of PARs, we randomly sampled 183 teachers to be included in data collection activities. Random sampling was used as a cost-cutting measure. We refer to this probability sample as the "Baseline Representative Sample" (BRS). We follow members of BRS through the completion of the study, allowing us to assess levels of total attrition, differential attrition, and potential for bias from the randomized sample becoming compromised over time. Our analysis of confirmatory impacts on teacher content knowledge included 118 of the BRS teachers.[15] Notably, only 88 of these teachers remained teaching in study-eligible grade levels by spring Year 2 (2017–18). We engaged the other 30 to take the teacher content knowledge assessment to obtain a larger sample of the originally randomized groups.

Teacher sample (exploratory analyses). Given that more BRS teachers left the study than we had expected by the second year,[16] we engaged in additional recruitment efforts of teachers. This provided a larger sample of 147 teachers for assessing impacts on teacher attitudes and beliefs, opportunity to learn, and school climate as measured by teacher surveys.

Student sample (confirmatory analyses). The student sample included all students who were in classes of study-participating teachers in spring of Year 2 (2017–2018) and had a science achievement outcome on the evaluator-developed assessment.[17] In spring of Year 2, there were 147 study teachers, including those who were present at randomization and those who joined after randomization. In the 147 study teachers' classes, there was a total of 2,140 students, who comprised the sample (or from which we drew students in the lowest third of incoming achievement) for confirmatory analyses of impact on students. Because the student sample includes those who may have joined study schools and who

[15] Attrition is sufficiently low that according to NEi3 and What Works Clearinghouse evidence standards, the potential for bias is low, and the result can potentially meet standards without reservations.

[16] Among the 183 BRS teachers, 84 were no longer in the study by the fall of Year 2 (2017–18). Thirty-six teachers (43%) had left the participating school, 34 teachers (40%) were no longer teaching fourth or fifth grade or were no longer teaching science, and 14 teachers (17%) were no longer in the study for other reasons, including retirement, family situation, a teacher strike at the school, non-responsiveness, or unknown.

[17] We considered identifying the sample of students who were in grades 2 and 3 in study schools prior to random assignment in winter 2015–16 because this group of students would have been randomized to *Making Sense of SCIENCE* or control by virtue of their membership in study schools when schools were randomly assigned. However, the fact that students would be non-randomly placed into the classes of participating teachers well after random assignment (in the 2016–17 and 2017–18 school years) would introduce a potential for bias that would compromise the initial random assignment. Because it would not be possible to maintain the randomization of students, we opted for the larger sample, including all students in study teachers' classes in the 2017–18 school year.

were placed in study teachers' classes after random assignment, we compared baseline achievement of students in both conditions to demonstrate their equivalence.

Student sample (exploratory analysis). As part of our exploratory analyses, we identified two additional student samples. The first consisted of students of a subset of the 147 teachers who are BRS teachers: 1,415 students of 96 BRS teachers. The second consisted of students who were in fifth grade in Year 2 (2017–18) and were also in a study teacher's classroom in Year 1 (2016–17) (n = 340). We recognize that some students could have been joiners into study teachers' classes, into study schools, or both. To demonstrate sample equivalence across conditions, we compared baseline achievement of students for each sample.

TABLE 3. SAMPLE SIZES AT SEVERAL LEVELS OF THE STUDY DESIGN

	Condition	PARS	BRS	Impact on TCK (Confirmatory, Research Question 1) (Mixed Sample)	Impact on TCK (Confirmatory, Research Question 1) (Retained in *Study* Sample)	Impact on student science achievement (Confirmatory, Research Question 2)	Impact on student science achievement (Reduced Sample: BRS teachers only)	Impact on student science achievement (Students in classes of study teachers in 16/17 and 17/18; in 5th grade in17/18)	Impact on student science achievement in the lowest third of incoming achievement (Confirmatory, Research Question 3)
Schools	*MSS*	30	30	27	25	29	28	18	29
	Control	30	30	27	22	26	23	13	26
	Total	60	60	54	47	55	51	31	55
Teachers	*MSS*	136	93	60	45	81	48	22	81
	Control	133	90	58	43	66	48	18	62
	Total	269	183	118	88	147	96	40	143
Students	*MSS*	N/A			N/A	1138	719	173	405
	Control					1002	696	167	310
	Total					2,140	1,415	340	715

Note. N/A = Not applicable. MSS references the group of students who received the Making Sense of SCIENCE program. TCK is Teacher content knowledge. PAR is Present at Randomization. BRS is Baseline Representative Sample.

96 BRS teachers were included in the analysis of impact on student science achievement, and 88 BRS teachers for the analysis of impact on TCK. The difference of eight is mostly due to several teachers administering the student science assessment but not taking the TCK posttest.

ANALYSIS AND REPORTING

Analysis of Impact on Teachers

Hierarchical Linear Models

We used 3-level Hierarchical Linear Models (HLMs) (teacher, school, randomized block levels) with individual teacher scores regressed against baseline covariates, a dummy variable indicating treatment assignment at the school level (*Making Sense of SCIENCE* school = 1, control school = 0), and random effects at the teacher and school levels. Block (pair) effects were modeled as fixed. The teacher content knowledge pretest was included in every analysis. For confirmatory analyses, we estimated impacts using a series of additional models as sensitivity tests, including with several approaches to estimation, with alternative scaling of the posttest, and with random block effects.

Approach to Handling Missing Data

As noted above, we limited the analysis to teachers who were among the sample of BRS teachers and who were available for posttest data collection. We removed teachers missing either their posttest, their pretest, or both. Dummy variable imputation was used with covariates other than the pretest (Puma et al., 2009).

Calculating Attrition

We assessed overall and differential attrition at the school and teacher levels, reflecting the school assignment design. Overall attrition was calculated as the number of randomized units (schools and teachers) that were missing outcome data and were therefore not included in the impact analysis sample. Attrition was also calculated separately for each condition, and the differential attrition rate was the difference in the rates between the two groups. To not double-count attrition, teacher attrition was calculated among non-attritting schools.

Analysis of Impact on Students

Hierarchical Linear Models

We used 3-level HLMs (student, school, randomized block levels), with individual student scores regressed against baseline covariates, a dummy variable indicating treatment assignment at the school level (*Making Sense of SCIENCE* = 1, control = 0), and random effects at student and school levels. Block (pair) effects were modeled as fixed. Student math and ELA pretests were included in every main impact analysis. For confirmatory analyses, we estimated impacts using a series of additional models as sensitivity tests, including with alternative scaling of the posttest, with random block effects, with a reduced-items test, by using Multiple Imputation with missing values for all covariates including the pretest, using Maximum Likelihood (ML) instead of Restricted Maximum Likelihood (REML) estimation, and other approaches.

Calculating Attrition

NEi3 has indicated that because study rosters for students were formed in fall of Year 1 (2016–17), after schools were randomized (winter 2015–16), that all students in the study are likely to be considered post-randomization "joiners." This precludes existence of a true baseline sample of students that can be considered the starting point for analysis of student attrition. Therefore, we do not measure attrition for the analysis of impacts on students. At the school level, from among

the 60 schools randomly assigned, we obtained student outcomes from 29 out of 30 *Making Sense of SCIENCE* schools and 26 out of 30 control schools.

Approach to Handling Missing Data

For the confirmatory analysis of impact on science achievement, we listwise deleted any students who were missing either the posttest score or pretest scores (either the ELA or math pretest scores). Dummy variable imputation was used with covariates other than the pretest (Puma et al., 2009). As part of a sensitivity analysis, we also performed multiple imputation analysis to include students who were missing values of one or both pretests.

Assessing Baseline Equivalence

Baseline equivalence on the pretests was assessed by regressing each pretest score (ELA and math) against a dummy variable indicating treatment status. The models also included school and student random effects and fixed pair effects to have the same error structure as the benchmark model used to estimate impact.

Chapter 3. Fidelity of Implementation of Making Sense Of SCIENCE

Implementation fidelity is "the extent to which an enacted program is consistent with the intended program" (Century et al., 2010). Understanding implementation fidelity sheds light on the impact results. To illustrate, if an evaluation of a program yields no impact, without having insight into fidelity of implementation, we would not be able to discern whether the no impact finding is potentially due to poor implementation. Conversely, if an evaluation has a positive impact, without measuring fidelity of implementation, we would be left with the question of whether stronger implementation would have resulted in even bigger impacts (Carroll et al., 2007). Additionally, measuring fidelity allows for a greater understanding of program implementation more broadly and informs future implementation. For example, identifying components or elements that were difficult to implement as intended would likely guide modifications to implementation plans in the future.

With these objectives in mind, and as a requirement of the NEi3, we calculated fidelity of implementation (FOI) scores for each of the six key components of the Making Sense of SCIENCE professional learning model, as outlined in the logic model in Chapter 1, separately for each of the two study years (Table 4).[18] Table 4 includes FOI results for both calendar years (2016 and 2017).

We also measured FOI pooled across the two years, even though it is not a requirement of NEi3 to do so, for Components 5 and 6 (related to Teacher Professional Learning) in order to factor in the importance of continuity of professional learning for teachers throughout the course of the study (Table 5). We calculated FOI across the two years for two samples of teachers, which we describe below.

The tables are organized by the program components and their respective indicators. All components have one indicator, except for Component 2 on Leadership Cadre Professional Learning and Component 5 on Teacher Professional Learning, each of which consists of three indicators. All indicators include the operational definition, scores and the threshold for levels of implementation at the unit level, scores and the threshold for levels of implementation at the sample level, and whether or not the indicator was met for the specified time period. For components that have more than one indicator—thus requiring aggregation to the component level—and for components that require aggregating to the school and sample level, a gray row below the component displays the thresholds for the school and sample levels, and indicators of whether or not fidelity was met at the sample level for the component.

When FOI was calculated separately for the two years, fidelity was met for all components in each of the two calendar years. For FOI calculated across the two years, fidelity was met for one of two samples for the teacher summer courses, and fidelity was not met for either sample for school year PLCs.

This chapter focuses on presenting the results of the FOI matrix. The full implementation study, was conducted and reported by Heller Research Associates (Wong et al., 2020).

---

[18] For Leadership Professional Learning (Components 1–3), we measure fidelity of implementation by calendar year in order to align fidelity of implementation measures with the timing of the professional learning. Fidelity of implementation for 2016 consisted of professional learning events that took place in the spring of 2015–16, and the following summer and fall of the 2016–17 school year. Similarly, fidelity of implementation for 2017 consists of professional learning events that took place in the spring of 2016–17, and the following summer and fall of the 2017–18. For Teacher Professional Learning (Components 4–6), we measured fidelity of implementation in school years.

FIDELITY MATRIX FOR CALENDAR YEARS 2016 AND 2017

Table 4 displays the six components, their corresponding indicators, and results of fidelity of implementation.

TABLE 4. FIDELITY MATRIX FOR MAKING SENSE OF SCIENCE FOR CALENDAR YEARS 2016 AND 2017

Indicator	Operational definition	Scores for levels of implementation at the unit level	Scores for levels of implementation at the sample level	Fidelity in 2016	Fidelity in 2017
Component 1: Site Coordinators					
Indicator 1: Site Coordinators PL	PL attendance	0: Site coordinator attends less than 85 percent of all events 1: Site coordinator attends at least 85 percent of all events Threshold = 1	0: Fewer than two of the site coordinators meet the unit level threshold 1: Both site coordinators meet the unit level threshold Threshold = 1	**Met at the indicator and component levels** **Score = 1** Site coordinators attended 88% (WI) and 89% (CA) of events	**Met at the indicator and component levels** **Score = 1** Site coordinators attended 100% of events
Component 2: Leadership Cadre					
Indicator 1: Facilitation Academy (FA) for Summer Teacher Course – Delivery	Total number participants in the 5-day Teacher Course Facilitation Academy	0: <8 OR >25 participants for one or more days 1: 8 to 25 participants every day. Threshold = 1	Same as unit level Threshold = 1	**Met at the indicator level** **Score = 1** The FA had 20 to 21 participants every day.	**Met at the indicator level** **Score = 1** The WI FA had 12 to 16 participants every day. The CA FA had 9 participants every day.
Indicator 2. FA for Summer Teacher Course – Attendance	Number of half-days Teacher Course Facilitator attends PL to be trained to facilitate teacher courses - 10 half-days offered	0: attended <8 half days 1: attended at least 8 half days Threshold = 1	0: <100% of the Teacher Course Facilitators attend at least 8 half days 1: 100% Teacher Course Facilitators attend at least 8 half days Threshold = 1	**Met at the indicator level** **Score = 1** CA and WI had 9 and 11 facilitators, respectively. 100% of facilitators attended at least 8 half-days.	**Met at the indicator level** **Score = 1** CA and WI had 9 and 8 facilitators, respectively. 100% of facilitators attended at least 8 half-days.

TABLE 4. FIDELITY MATRIX FOR MAKING SENSE OF SCIENCE FOR CALENDAR YEARS 2016 AND 2017

Indicator	Operational definition	Scores for levels of implementation at the unit level	Scores for levels of implementation at the sample level	Fidelity in 2016	Fidelity in 2017
Indicator 3. FA for PLCs - Attendance	Whether PLC Facilitator attends PL to be trained to facilitate school-based PLCs for teachers - 1 day offered	0: did not attend 1: attended Threshold = 1	0: <60% of treatment schools have a PLC Facilitator attend FA for PLCs 1: ≥ 60% of treatment schools have a PLC Facilitator attend FA for PLCs Threshold = 1	**Met at the indicator level** **Score = 1** 29 of 30 (96.7%) treatment units had at least 1 LC member attend.	**Met at the indicator level** **Score = 1** 23 of 30 (76.67%) treatment units had at least 1 LC member attend.
Criteria for implementing Component 2 with fidelity			Sum of indicator scores Range : 0 - 3 Threshold : 3	**Met at the component level** **Score = 3**	**Met at the component level Score = 3**
Component 3: Administrators					
Indicator 1: Administrator workshop	Whether school administrator attends PL on supporting teachers & science teaching in schools – one day offered	0: did not attend 1: attended Threshold = 1	0: <60% of treatment schools have an administrator attend the administrator workshop 1: ≥60% of treatment schools have an administrator attend the administrator workshop Threshold = 1	**Met at the indicator and component levels** **Score = 1** 25 of 30 (83.3%) treatment units had at least 1 administrator attend.	**Met at the indicator and component levels** **Score = 1** 20 of 30 (66.67%) treatment units had at least 1 administrator attend.

TABLE 4. FIDELITY MATRIX FOR MAKING SENSE OF SCIENCE FOR CALENDAR YEARS 2016 AND 2017

Indicator	Operational definition	Scores for levels of implementation at the unit level	Scores for levels of implementation at the sample level	Fidelity in 2016	Fidelity in 2017
Component 4: Teacher Summer Course Delivery					
Indicator 1: Summer Teacher Course - Delivery	Teacher attends Teacher Course section with average of 9 participants per day and no fewer than 6 on any one day	0: No 1: Yes (teacher attends course section with average of 9 participants per day and no fewer than 6 on any one day) Threshold = 1	School-level 0: <50% teachers in school who met fidelity at the unit-level 1: ≥50% of teachers in school who met fidelity at the unit-level Threshold = 1 Sample-level 0: <75% of schools with a score of 1 1: ≥75% of schools with a score of 1 Threshold =1	**Met at the indicator level** **Score = 1** All teachers who were eligible to receive a score under indicator 1 met the threshold because all course sections had between 15 and 25 participants. Therefore, fidelity was also met for school- and sample-levels.	**Met at the indicator level** **Score =1** All teachers who were eligible to receive a score under Indicator 1 met this threshold because all course sections had between 10 and 22 participants. Therefore, fidelity was also met for school- and sample-levels.

TABLE 4. FIDELITY MATRIX FOR MAKING SENSE OF SCIENCE FOR CALENDAR YEARS 2016 AND 2017

Indicator	Operational definition	Scores for levels of implementation at the unit level	Scores for levels of implementation at the sample level	Fidelity in 2016	Fidelity in 2017
Indicator 2: Summer Teacher Course - Structure	Teacher attends course section with the following structure (activities and time allocation): Science whole group – 80% of allotted time each day Science small group – 80% of allotted time each day Teaching small & large group combined – 60% of allotted time for week Literacy small & large group combined – 60% of allotted time for week	Fidelity determined by summing points for each dimensions of the course structure: 1 point each day (up to 5 points total) when Science Investigation small group time is ≥ 80% of allocated time 1 points each day (up to 5 points total) when Science Investigation whole group time is ≥ 80% of allocated time 2 points when the combined Teaching Investigation time is ≥ 60% of allocated time for week 2 points when the combined Literacy Investigation time is ≥ 60% of allocated time for week 0: <12 2 : 12 or more Threshold = 2	Same as indicator 1	**Met at the indicator level** **Score = 1** All teachers who were eligible to receive a score under indicator 1 met this threshold because all course sections received at least 12 points. Therefore, fidelity was also met for school- and sample-levels.	**Met at the indicator level** **Score = 1** All teachers who were eligible to receive a score under Indicator 1 met this threshold because all course sections received at least 13 points. Therefore, fidelity was also met for school- and sample-levels.

TABLE 4. FIDELITY MATRIX FOR MAKING SENSE OF SCIENCE FOR CALENDAR YEARS 2016 AND 2017

Indicator	Operational definition	Scores for levels of implementation at the unit level	Scores for levels of implementation at the sample level	Fidelity in 2016	Fidelity in 2017
Indicator 3: Summer Teacher Course - Process	Percentage of qualities of the course section that teacher indicates having experienced, from a list of potential qualities for a culture of collaboration and respect	0: Selects <60% of qualities from list 1: Selects ≥60% of qualities from list Threshold = 1	Same as indicator 1	**Met at the indicator level** **Score = 1** All teachers who were eligible to receive a score under indicator 1 met this threshold because all course sessions met the threshold. Therefore, fidelity was also met for school- and sample-levels.	**Met at the indicator level** **Score = 1** All teachers who were eligible to receive a score under Indicator 1 met this threshold because all course sessions met the threshold. Therefore, fidelity was also met for school- and sample-levels.
Criteria for implementing Component 4 with fidelity			Sum of indicator scores Range : 0 - 3 Threshold : 3	**Met at the component level** **Score = 3**	**Met at the component level** **Score = 3**
Component 5: Teacher Summer Course Attendance					
Summer Teacher Course: **Teacher Attendance**	Number half-days teacher attends Summer Teacher Course — 10 half-days offered Dynamic Earth - Summer 2016 Planet Earth – Summer 2017	Teacher-level score 0: <7 half-days attended 1 : ≥7 half-days attended Threshold = 1	School-level score 0: <50% teachers in school meet the threshold 1: ≥50% of teachers in school meet the threshold Threshold = 1 Sample-level threshold 0: <75% of schools with a score of 1 1: ≥75% of schools with a score of 1 Threshold = 1	**Met at the indicator and component level** **Score = 1** Of 125 teachers [a] who were eligible to receive a score, 118 (94.4%) attended at least 7 half-days. 29 of 30 (96.7%) schools had score of 1	**Met at the indicator and component level** **Score = 1** Of 114 [b] teachers who were eligible to receive a score, 100 (87.7%) attended at least 7 half-days. 29 of 30 (96.7%) schools had score of 1

TABLE 4. FIDELITY MATRIX FOR MAKING SENSE OF SCIENCE FOR CALENDAR YEARS 2016 AND 2017

Indicator	Operational definition	Scores for levels of implementation at the unit level	Scores for levels of implementation at the sample level	Fidelity in 2016	Fidelity in 2017
Component 6: School Year PLC					

School Year PLC: Attendance	Total hours teacher attends PLC meetings; 6 2-hour meetings offered each year (total: 12 hours)	0: <6 hours (or 3 PLCs)
		1: at least 6 hours (or 3 PLCs)
			School-level
			If school has more than 3 teachers, then
			0: <50% teachers in school meet the threshold
			1: ≥50% of teachers in school meet the threshold
			If school has 3 or fewer teachers, then
			0: fewer than 2 teachers in school meet the threshold
			1: two or more teachers in school meet the threshold
			
			Sample-level threshold
			0: <75% of schools with a score of 1
			1: ≥75% of schools with a score of 1
			Threshold = 1
				Met at the indicator and component levels Score = 1
				
				121 of 125 (96.8%) of teachers who were eligible to receive a score for this indicator met the threshold.
				30 of 30 (100%) schools had score of 1
					Met at the indicator and component levels Score = 1
					
					103 out of 114 (90.35%) of teachers who were eligible to receive a score for this indicator met the threshold.
					25 of 30 (83.3%) schools had score of 1

Note. PL = professional learning; PLC = Professional Learning Community

[a] This sample of 125 teachers consists of teachers who were in the study early enough in the summer prior to Year 1 (2016–17) to attend the first summer course and still in the study in the fall of Year 1 (2016–17) to participate in the school-year PLC meetings.

[b] This sample of 114 teachers comprises of teachers who were in the study early enough in the summer prior to Year 2 (2017–18) to attend the first summer course and still in the study in the fall of Year 2 (2017–18) to participate in the school-year PLC meetings.

FIDELITY MATRIX ACROSS THE TWO SCHOOL YEARS FOR COMPONENTS 5 AND 6

Table 5 presents FOI across the two school years (2016–17 and 2017–18) for Components 5 (Teacher Summer Course: Attendance) and 6 (PLC meetings: Attendance). WestEd deemed it important to calculate FOI across the two years for these two components because the Making Sense of SCIENCE model theorizes that in order for teachers to have a "sufficient" level of participation, they need to be exposed to the range of content which varied from one year to the next. By assessing FOI separately for each year, we would not be able to determine the proportion of teachers who received professional learning for both years.

We did not measure fidelity of implementation across the two years for Components 1–3 (related to Leadership Professional Learning) based on the rationale that Leadership Cadre members can be successful in facilitating a summer course after attending the 5-day Facilitation Academy for the given course and in leading PLCs with only one year of experience. Similarly, while administrators are more likely to be effective with two years of participation, WestEd hypothesized that participating in the Making Sense of SCIENCE professional learning for one year should allow administrators to adequately support their staff and to make policy changes. Therefore, FOI was based on any representation from the school in each of the two years (that is, meeting the threshold did not require the same administrator to be present in both years). Component 4 (Teacher Professional Learning Delivery, Structure, and Process) was also not included in the FOI calculations across the two years because the quality of professional learning can be independent across the two years.

In FOI calculations across two years for Components 5 and 6, we used two samples. The first sample consisted of all 185 *Making Sense of SCIENCE* teachers, including those who attrited and those who joined the study during the study years.[19] The second sample comprised of 136 Making Sense of SCIENCE teachers who were present at randomization.[20] FOI for sample 1 indicated the extent to which the full sample of *Making Sense of SCIENCE* teachers received the intervention as intended by program developers, which requires attendance at summer courses and PLCs for both study years. FOI for sample 2 provides this information for the sample of baseline teachers. We elected to report FOI for these two samples and not for the various analytic samples because they represent the two broadest groups of participating teachers in the study.[21]

Notably, we observed strong uptake of Making Sense of SCIENCE within each year, among teachers who were in the study early enough in the summer to participate in summer course and were still in the study in the following fall. In Year 1 (2016–17), 94% of teachers, and in Year 2 (2017–18), 88% of teachers met the fidelity threshold for attendance at the summer professional learning institutes; 97% of teachers in Year 1 and 90% of teachers in Year 2 met the fidelity threshold

[19] A total of 329 teachers consented to participate in the study, regardless of when they consented (pre- or post-randomization, at baseline or in Year 1 or Year 2) and whether or not they attrited from the study. Among these 329 teachers, 185 teachers were in *Making Sense of SCIENCE* schools.

[20] At randomization, there were 269 teachers who were randomly assigned to conditions by virtue of their schools being randomly assigned. Among these 269 teachers, 136 were assigned to the *Making Sense of SCIENCE* schools.

[21] Recall that the Making Sense of SCIENCE model emphasizes collaboration among teachers at the schools, particularly at PLC meetings. Given the thresholds determined for the school level, limiting the FOI sample to the subset of teachers included in the analytic samples would have made the school-level FOI results uninterpretable. Additionally, FOI at the teacher level was rolled up to the school level, and the low number of teachers per school in the analytic sample introduced an element of arbitrariness for meeting the fidelity threshold.

for attendance at the PLC meetings. Yet, only 54% of study teachers met the attendance threshold for the summer courses, and 56% of teachers met the attendance threshold for PLC meetings for both years, due to the instability of the sample across the two years. Among the 185 participating teachers who were in *Making Sense of SCIENCE* schools, including those who attrited and those who joined the study during the two years, only 97 teachers (52%) were teaching study-eligible classes when summer professional learning was offered and when classes started in the fall for both study years.

TABLE 5. FIDELITY MATRIX FOR MAKING SENSE OF SCIENCE ACROSS THE TWO SCHOOL YEARS FOR COMPONENTS 5 AND 6

Indicator	Operational definition	Scores for levels of implementation at the unit level	Scores for levels of implementation at the sample level	Fidelity across 2016–17 and 2017–18
Component 5: Teacher Summer Course				
Summer Teacher Course: Teacher Attendance	Number half-days teacher attends Summer Teacher Course — 10 half-days offered for each of the two summer courses	0: <7 half-days attended in either or both of the years. 1: ≥7 half-days attended in both years.		Sample 1: 100 of 185 [a] teachers (54.1 %) met the unit-level threshold Sample 2: 83 of 136 [b] teachers (61.0 %) met the unit-level threshold.
Criteria for implementing Component 5 with fidelity		Threshold for school level: 0: <50% teachers in school meet the threshold 1: ≥50% of teachers in school meet the threshold	Threshold for sample level: 0: <75% of schools with a score of 1 1: ≥75% of schools with a score of 1	Sample 1: 19 of 30 (63.3%) schools met the threshold. Did not meet fidelity. Sample 2: 25 of 30 (83.3%) schools met the threshold. Met fidelity.

TABLE 5. FIDELITY MATRIX FOR MAKING SENSE OF SCIENCE ACROSS THE TWO SCHOOL YEARS FOR COMPONENTS 5 AND 6

Indicator	Operational definition	Scores for levels of implementation at the unit level	Scores for levels of implementation at the sample level	Fidelity across 2016–17 and 2017–18
Component 6: School Year PLC				
School Year PLC: **Attendance**	Total hours teacher attends PLC meetings; 6 2-hour meetings offered each year (total: 12 hours)	0: <total possible 1 (High): total possible hours attended		Sample 1: 103 of 185 teachers (55.7%) met the unit-level threshold Sample 2: 79 of 136 teachers (59.4%) met the unit-level threshold.

Criteria for implementing Component 6 with fidelity		If school has more than 3 teachers, then
		0 <50% teachers in school meet the threshold
		1: ≥50% of teachers in school meet the threshold
		If school has 3 or fewer teachers, then
		0: fewer than 2 teachers in school meet the threshold
		1: two or more teachers in school meet the threshold
			0: <75% of schools with a score of 1
			1: ≥75% of schools with a score of 1
				Sample 1:
				21 of 30 (70.0%) schools met the threshold. Did not meet fidelity.
				Sample 2:
				21 of 30 (70.0%) schools met the threshold. Did not meet fidelity.

Note. PL = professional learning; PLC = Professional Learning Community

[a] There was a total of 329 teachers who consented to participate in the study, regardless of when they consented (pre- or post-randomization, at baseline, or in Year 1 or Year 2) and whether or not they attrited from the study. Among these 329 teachers, 185 teachers were in *Making Sense of SCIENCE* schools.

[b] At randomization, there were 269 teachers who were randomly assigned to conditions by virtue of their schools being randomly assigned. Among these 269 teachers, 136 were assigned to the *Making Sense of SCIENCE* schools.

Chapter 4. Impact on Teacher Content Knowledge and Pedagogical Content Knowledge

INTRODUCTION

A critical outcome of the Making Sense of SCIENCE professional learning model is improvement in teacher content knowledge and pedagogical content knowledge in the multiple science disciplines they are expected to teach. This chapter focuses on the impact of Making Sense of SCIENCE on these two outcomes, and it addresses the following questions.

- **Confirmatory**: What is the impact of Making Sense of SCIENCE after two years of implementation on teacher content knowledge when compared to study participants in control schools receiving the business-as-usual science professional learning?
- **Exploratory:** What is the impact of Making Sense of SCIENCE after two years of implementation on teacher pedagogical content knowledge when compared to study participants in control schools receiving the business-as-usual science professional learning?

This chapter is organized into three sections: teacher content knowledge, teacher pedagogical content knowledge, and a discussion of the key findings on both outcomes. Within each of the first two sections, we 1) describe the measure, the analytic samples (including levels of attrition and baseline equivalence), and our approach to analysis, and 2) present the impact findings for each sample and moderator analyses.

TEACHER CONTENT KNOWLEDGE: METHODS

Measure

As detailed in the methods chapter, the teacher content knowledge assessment comprises items adapted from a variety of sources (MCAS, NECAP, MOSART, and NAEP). There was one form with 32 selected response items.[22] The items were piloted in Year 1 of the study (2016–17) and selected for use in Year 2 (2017–18), on the basis of their degree of difficulty, high point-biserial correlations, and alignment with NGSS disciplinary core ideas. We obtained scale scores through 2-parameter logistic (2-PL) Item Response Theory calibration. The achieved Cronbach's alpha for the assessment was 0.78. (See Appendix C for more information about the teacher content knowledge assessment.)

Sample

Recall that the study started with 269 teachers enrolled in the study at the time of random assignment (described in the methods chapter). Of these, 183 were randomly selected for data collection activities. This group was labelled the Baseline Representative Sample (BRS). For the analysis of impact on teacher content knowledge we focus on the subset of the BRS teachers who completed a posttest in spring 2018. We report results for two samples, both of which we consider to be important.

1. The first ("*Mixed* sample") consists of 118 teachers. This sample represents the largest available sample of teachers out of the 183 BRS group. It includes teachers who were initially randomized and randomly selected for outcomes

[22] Only 29 were included in the analysis as detailed in Appendix C.

data collection, and who also took the posttest. This *Mixed* sample includes 30 teachers (15 treatment, 15 control) who completed the posttest even though they were no longer active in the study at the time the posttest was administered. The 15 treatment teachers all had less than full exposure, with their level of exposure varying depending on when they left the study. Five left the study prior to start of professional learning and therefore had no exposure; others left at some point during the study and had some exposure. [23] Therefore, impact associated with the fuller sample represents the effect of Making Sense of SCIENCE based on active and inactive participants.[24]

2. The second ("*Retained in Study*" sample) (n = 88) is the same as the *Mixed* sample, with one difference. This sample includes only teachers from the *Mixed* sample who were active in the study for the entirety of the implementation period. In other words, this sample differs from the first sample by the exclusion of the 30 teachers who dropped out of the study. Note that "active" implies that they were still participating in the study (i.e. had the opportunity to attend all professional learning) and were teaching eligible grades (i.e. had the opportunity to apply what they learned in professional learning), but not necessarily that they participated in all professional learning activities. Impact associated with the *Retained in Study* sample represents the effects for such teachers.

As program evaluators, we endorse the impact finding of the *Retained in Study* sample; the impact result of the *Mixed* sample does not reflect a typical implementation scenario. For understanding impact under conditions of normal implementation, we are interested in the effects on teachers who currently have the opportunity to attend Making Sense of SCIENCE professional learning (i.e., have normal differences in levels of exposure) and to implement what they have learned thus far from the professional learning in their classrooms. We are less interested in impacts on teachers who were no longer in the study because they had left the school or were no longer teaching a study-eligible grade or subject. Impact on the *Mixed* sample would have included such teachers. [25]

Attrition

We show attrition counts for the *Mixed* sample (Table 6) and for the *Retained in Study* sample (Table 7). With the *Mixed* sample, the study has potential to meet WWC evidence standards without reservations. For the *Retained in Study* sample, the study has potential to meet evidence standards with reservations.

---

[23] Of the other 10 teachers, 4 had left the school, 5 had left a study-eligible grade, and 1 had left a study-eligible subject.

[24] The rationale for evaluating impact on the *Mixed* sample was that we wanted as large a sample as possible for evaluating the impact of Intent To Treat. Limiting attrition also would allow the result to potentially meet WWC Evidence Standards without reservations.

[25] Another alternative to examining impacts on the *Retained in Study* sample of teachers who participated in Making Sense of SCIENCE for the full implementation period would have been to estimate the Complier Average Causal Effect (CACE). (We estimate the impact of Intent To Treat (ITT) using the *Mixed* sample). Conducting a CACE analysis consistent with WWC standards would require us to define compliance as having received *any* of the program. In this study, of the 118 BRS teachers for whom we obtained a posttest, 113 were compliant by this criterion. We did not think the result from a CACE analysis would differ much from the ITT analysis for 118 teachers. Furthermore, we were less interested in estimating impacts for teachers exposed to any amount of Making Sense of SCIENCE (which is what the CACE analysis would estimate) than for teachers exposed to all or most of the program (which we estimated using the *Retained in Study* sample). Therefore, we did not pursue the CACE analysis.

TABLE 6. ATTRITION COUNT FOR THE EVALUATION OF IMPACT ON TEACHER CONTENT KNOWLEDGE (*MIXED* SAMPLE)

	Number of schools randomly assigned	Number of schools in the sample used in analysis of 118 teachers	Attrition of schools	Number of teachers at baseline (in non-attriting schools)	Number of teachers (in non-attriting schools) with posttests)	Attrition of teachers
***MSS* (N)**	30	27	10.0%	84	60	28.6%
Control (N)	30	27	10.0%	81	58	28.4%
Total N	60	54		165	118	
Overall attrition			10.0%			28.5%
Differential attrition			0%			0.2%
Potential for bias			low			low

Note. *MSS* stands for the group that received the Making Sense of SCIENCE professional learning.

TABLE 7. ATTRITION COUNT FOR THE EVALUATION OF IMPACT ON TEACHER CONTENT KNOWLEDGE (*RETAINED IN STUDY* SAMPLE)

	Number of schools randomly assigned	Number of schools in the sample used in analysis of 88 teachers	Attrition of schools	Number of teachers at baseline (in non-attriting schools)	Number of teachers (in non-attriting schools) with posttests	Attrition of teachers
***MSS* (N)**	30	25	16.7%	78	45	42.3%
Control (N)	30	22	26.7%	69	43	37.7%
Total N	60	47		147	88	
Overall attrition			21.6%			40.1%
Differential attrition			10.0%			4.6%
Potential for bias			high			high

Note. *MSS* stands for the group that received the Making Sense of SCIENCE professional learning.

Tests of Baseline Equivalence

We assessed baseline equivalence on the pretest for (a) the baseline sample, (b) the *Mixed* analytic sample, and (c) the *Retained in Study* sample. For the *Mixed* sample, because attrition was low, it was not necessary to establish baseline equivalence for the result to be eligible to satisfy NEi3/WW3 evidence standards without reservations. For the *Retained in Study* sample, because attrition was high, we needed to establish baseline equivalence for the result to meet evidence standards with reservations. Results are displayed in Table 8. We observe that for the *Retained in Study* analytic sample,

baseline equivalence was established, and the impact finding is eligible to meet evidence standards with reservations, provided we adjust for the pretest in the impact model. The impact analysis described below includes this adjustment.

TABLE 8. TESTS OF BASELINE EQUIVALENCE ON PRETESTS BETWEEN MAKING SENSE OF SCIENCE AND CONTROL FOR THE BASELINE REPRESENTATIVE AND ANALYTIC SAMPLES (*MIXED* AND *RETAINED IN STUDY* SAMPLES)

	Baseline	Analytic (*Mixed*)	Analytic (*Retained in Study*)
Pretest			
N (Schools)	60	54	47
N (Teachers)	183	118	88
Point estimate	-0.01	-0.02	-0.02
Standard error	0.01	0.02	0.02
***p* value**	0.509	0.254	0.442
Standardized effect size	-0.07	-0.15	-0.13

Analysis

We evaluated the impact of Intent To Treat on teacher content knowledge after two years of program implementation. We estimated impact using a hierarchical linear model with fixed block (pair) effects and school-level random effects. Covariates included a science content knowledge pretest that was administered before random assignment. (The full HL models are provided in Appendix H; the full list of covariates is shown with the complete impact finding in Appendix I.) We used full ML estimation and report robust standard errors. We addressed missing values for covariates other than the pretest using dummy variable imputation. Cases without a pretest or a posttest were removed.

TEACHER CONTENT KNOWLEDGE: FINDINGS

Overall Impact

In Table 9 and Table 10, we exhibit the main impact findings for the *Mixed* and *Retained in Study* samples. (The full results from the impact models are in Appendix I.) For the *Mixed* sample (n = 118 teachers), we observed a positive but not statistically significant impact, with a standardized effect size of 0.22 (p = .165). For the *Retained in Study* sample (n = 88 teachers), we observed a positive and statistically significant impact, with a standardized effect size of 0.56 (p = .006).

TABLE 9. RESULTS OF IMPACT ANALYSIS FOR THE *MIXED* SAMPLE

	Condition	Means	Standard deviations	No. of schools	No. of teachers	Effect size	*p* value	Change in percentile ranking
Unadjusted effect size[a]	Control	-0.10	0.84	27	58	0.12	.404	4.8%
	MSS	0.03	0.90	27	60			

TABLE 9. RESULTS OF IMPACT ANALYSIS FOR THE *MIXED* SAMPLE

	Condition	Means	Standard deviations	No. of schools	No. of teachers	Effect size	*p* value	Change in percentile ranking
Adjusted effect size[b]	Control	-0.10				0.22	.165	8.7%
	MSS	0.09						

Note. *MSS* defines the group receiving the Making Sense of SCIENCE professional learning. The *p* values are for the corresponding impact estimates in the impact model.

[a] The unadjusted effect size is the impact estimate from a model with pair fixed, and school random effects, and without covariates, divided by the pooled standard deviation of the outcome variable.

[b] The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

TABLE 10. RESULTS OF IMPACT ANALYSIS FOR THE *RETAINED IN STUDY* SAMPLE

	Condition	Means	Standard deviations	No. of schools	No. of teachers	Effect size	*p* value	Change in percentile ranking
Unadjusted effect size[a]	Control	-0.13	0.88	22	43	0.40	.050	15.6%
	MSS	0.21	0.86	25	45			
Adjusted effect size[b]	Control	-0.13				0.56	.006	21.1%
	MSS	0.35						

Note. *MSS* defines the group receiving the Making Sense of SCIENCE professional learning. The *p* values are for the corresponding impact estimates in the impact model.

[a] The unadjusted effect size is the impact estimate from a model with pair fixed, and school random effects, and without covariates, divided by the pooled standard deviation of the outcome variable.

[b] The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

Sensitivity Analyses

We conducted a series of sensitivity analyses for both the *Mixed* and *Retained in Study* samples. The results are summarized in Table 11. The results from the analysis using the benchmark model are repeated at the top of the table to facilitate comparison. All models included school and teacher random effects and pair fixed effects (unless otherwise noted). They are as follows (in the order in Table 11): (a) with posttests calculated using the percent correct metric, (b) with posttests scaled using a 1-Parameter (1-PL) model, (c) a 2-Parameter Logistic (2-PL) model with no covariates, (d) a 2-PL model with the pretest as the only covariate, and (e) the benchmark model with blocks modeled as random instead of as fixed effects.[26] Results from all models support the same conclusions as would be drawn from the benchmark models.

[26] We also considered using Multiple Imputation methods to address missing values of covariates, including the pretest. The method would add value if it allowed a larger sample of teachers by virtue of including ones with a missing pretest. However, none of the teachers in the samples considered here had a missing pretest; therefore, we did not apply this approach.

TABLE 11. RESULTS OF SENSITIVITY ANALYSES OF IMPACTS ON TEACHER CONTENT KNOWLEDGE

	Mixed sample (*n* = 118)	*Retained in Study* sample (*n* = 88)
Benchmark impact model	**Impact estimate: 0.19** **Standard error: 0.13** **Effect size: 0.22** ***p* value: .165**	**Impact estimate: 0.48** **Standard error: 0.16** **Effect size: 0.56** ***p* value: .006**
Percent correct scaling of the posttest [a]	No transformation of the posttest Impact estimate: 0.04 Standard error: 0.03 Effect size: 0.23 *p* value: .154 Taking the square of the posttest Impact estimate: 0.06 Standard error: 0.03 Effect size: 0.27 *p* value: .093	No transformation of the posttest Impact estimate: 0.10 Standard error: 0.03 Effect size: 0.62 *p* value: .002 Taking the square of the posttest Impact estimate: 0.48 Standard error: 0.16 Effect size: 0.55 *p* value: .006
1-PL scaling of the posttest	Impact estimate: 0.21 Standard error: 0.14 Effect size: 0.25 *p* value: .131	Impact estimate: 0.52 Standard error: 0.15 Effect size: 0.60 *p* value: .002
2-PL scaling of the posttest with no covariates (block and school and teacher random effects only)	Impact estimate: 0.11 Standard error: 0.13 Effect size: 0.12 *p* value: .404	Impact estimate: 0.35 Standard error: 0.17 Effect size: 0.40 *p* value: .050
2-PL scaling of the posttest with pretest as the only covariate (2-PL)	Impact estimate: 0.18 Standard error: 0.14 Effect size: 0.21 *p* value: .195	Impact estimate: .39 Standard error: .17 Effect size: .45 *p* value: .034
Benchmark with random block effects	Impact estimate: 0.26 Standard error: 0.18 Effect size: 0.20 *p* value: .148	Impact estimate: 0.48 Standard error: 0.21 Effect size: 0.55 *p* value: .030

[a] When scored as percent correct, posttest distributions were skewed (values of skew are -0.68 for the *Mixed* sample and -0.71 for the *Retained in Study* sample). A square transformation reduced skew (-0.12 for the *Mixed* sample and -0.16 for the *Retained in Study* sample). We analyzed impacts before and after applying the transformation. Posttest distributions based on other approaches to scaling did not exhibit substantial skew.

Moderator Analyses

To evaluate whether attributes of teachers moderated impacts on teacher content knowledge, we expanded the benchmark models that were used to estimate the average impact of Making Sense of SCIENCE on teacher content knowledge to include an interaction term between the moderator of interest and the term indicating random assignment status. The moderating effects were evaluated one at a time.[27]

Mixed Sample

For the *Mixed* sample, we observed no differential impact depending on a teacher's incoming level of content knowledge, with an estimate of -0.553 (p = .535) (interpreted as the change in impact for each unit increase in the pretest). We also found no differential impact depending on years of teaching experience, with an estimate of -0.012 (p = .332) (interpreted as the change in impact for each additional year of teaching). Further, we observed no difference in impact depending on whether a teacher was a teacher leader, with an estimate of 0.034 (p = .892) (interpreted as the added-value impact for a teacher being a teacher leader). These results are also summarized in Table 12.

TABLE 12. MODERATING EFFECTS ON TEACHER CONTENT KNOWLEDGE USING THE *MIXED* SAMPLE

Moderator	Interpretation	No. of teachers	Differential impact	Standard error	Df	*t*	*p* value
Teacher's incoming level of content knowledge	increase in impact for each unit increase in the pretest	118	-0.553	0.823	38	-.63	.535
Years of teaching experience	increase in impact for each additional year teaching	103	-0.012	0.012	27	-.99	.332
Is a Teacher Leader	the added value impact for teachers belonging in the leadership cadre	118	0.034	0.247	37	.14	.892

Note. The pretest is in raw score units with mean 0.60 and standard deviation 0.13. The posttest is in 2-PL scale score units, with mean -0.03 and standard deviation 0.87.

Retained in Study Sample

For the *Retained in Study* sample, we observed no differential impact depending on a teacher's incoming level of content knowledge, with an estimate of 0.167 (p = .914) (interpreted as the increase in impact for each unit increase in the pretest). We also found no differential impact depending on years of teaching experience, with an estimate of -0.012 (p = .378) (interpreted as the increase in impact for each additional year teaching). Further, we observed no difference in impact

[27] For this outcome, we do not evaluate moderating effects simultaneously. We did so in the analysis of impacts on student science achievement where we found a marginal differential impact by ELL status, and the combined analysis was used to assess the robustness of that effect to simultaneously model effects of other moderators.

depending on whether a teacher was a teacher leader, with an estimate of 0.113 (p = .670) (interpreted as the added-value impact for a teacher being a teacher leader). These results are also summarized in Table 13.

TABLE 13. MODERATING EFFECTS ON TEACHER CONTENT KNOWLEDGE USING THE *RETAINED IN STUDY* SAMPLE

Moderator	Interpretation	No. of teachers	Differential impact	Standard error	Df	*t*	*p* value
Teacher's incoming level of content knowledge	increase in impact for each unit increase in the pretest	88	0.167	1.514	15	15	.914
Years of teaching experience	increase in impact for each additional year teaching	78	-0.012	0.013	10	.92	.378
Is a Teacher Leader	the added value impact for teachers belonging in the leadership cadre	88	0.113	0.262	14	.43	.670

Note. The pretest is in raw score units with mean 0.60 and standard deviation 0.14. The posttest is in 2-PL scale score units, with mean 0.04 and standard deviation 0.77.

We also examined if impacts were greater in stronger implementing districts. For the larger sample of teachers (n = 118), based on a Type-3 test of fixed effects, we observed variation in impact across districts (p < .001). Limited to 48 teachers in the three stronger implementing districts, we observed a positive impact on teacher content knowledge of 0.874 scale score units (1.00 standardized effect size, p = .006). For the smaller sample of teachers (n = 88), based on a Type-3 test of fixed effects, we observed variation in impact across districts (p < .001). Limited to 35 teachers in the three stronger implementing districts, we were not able to obtain a stable estimate and we do not report a result.

TEACHER PEDAGOGICAL CONTENT KNOWLEDGE: METHODS

Measure

As our research partner HRA explained in their report (Wong et al., 2020), the Refined Consensus Model (RCM) of pedagogical content knowledge (PCK) (Carlson et al., 2019) posited three levels of PCK:

1) *collective* PCK (cPCK), the professional knowledge held by a group of educators in a field;
2) *personal* PCK (pPCK), the professional knowledge held by an individual teacher; and
3) *enacted* PCK (ePCK), the knowledge and pedagogical reasoning that a teacher uses during the process of planning instruction, teaching, and reflecting on instruction and student outcomes around a particular topic for a particular set of students.

The instrument for teacher pedagogical content knowledge assessed ePCK. More specifically, the instrument assessed the following areas of teachers' abilities and knowledge relating to weather and erosion:

a) ability to interpret student work;

b) knowledge of typical student difficulties;

c) knowledge of effective instructional strategies for supporting fourth- and fifth-grade students in making observations, providing evidence, and constructing scientific explanations; and

d) the explicitness of teachers' pedagogical reasoning.

The instrument used in this study to measure pedagogical content knowledge, adapted from PCK items developed by HRA in prior research studies about Making Sense of SCIENCE (e.g. Heller et al., 2010; Daehler et al., 2015), was a cluster of prompts centered around one of the constructed-response assessment tasks presented to fourth- and fifth-grade students in this study. The item asked teachers to evaluate a hypothetical student's response to a question about erosion that asked students to respond to three prompts: (1) make a claim, (2) provide evidence, and (3) explain reasoning. Teachers were asked to (a) state the strengths and weaknesses in the student's response, (b) state specific difficulties students may have had in responding to the item, separately from the specific difficulties exhibited by the hypothetical student, and (c) describe activities to support the student.

Teachers' responses were rated in the following dimensions.

1. Concept Score: an indicator of teachers' ability to connect instructional activities to specific conceptual goals
2. Explanation score: an indicator of the quality of the explanation including attention to questions of "why" or "how," as well as making claims, providing evidence to support the claim, and explaining how the evidence supports the claim
3. 2-Dimensional Score: an indicator of the extent to which teachers integrated both science concepts and explanation practices (i.e., it is not a sum of the Concept and Explanation score)
4. Holistic scale: an indicator based on the overall assessment of the strength of teachers' PCK as shown in their responses to the three prompts (Wong et al., 2020)

One rater evaluated all responses, with another rater independently rating 25% of responses for a random subset of teachers. For double-rated responses, where there were discrepancies, the raters met to resolve inconsistencies. The interrater reliability (that is, percent agreement between scorers) was 76.7%.

Sample

As with our analysis of impacts on teacher content knowledge, we address impacts on "*Mixed*" and "*Retained in Study*" samples of teachers. There was a close correspondence between the samples used for analysis of impact on teacher content knowledge and teacher pedagogical content knowledge. We provide sample sizes of teachers and randomized clusters in Table 14 and Table 15 below.

Attrition

We show attrition counts for the *Mixed* sample (Table 14) and for the *Retained in Study* sample (Table 15). With the *Mixed* sample, the study is in the WWC category of having a "tolerable threat of bias under both optimistic and cautious

assumption." For the *Retained in Study* sample, the study is in the WWC category of at best "meeting evidence standards with reservations." (For the *Retained in Study* sample, teacher-level attrition is high enough that there is unacceptable threat of bias under cautious assumptions, but tolerable threat of bias under optimistic assumptions; however, loss from study likely resulted from a combination of standard reassignment of teachers to new grade levels as well as from teachers opting out, possibly reflecting endogenous factors influencing continued participation.)

TABLE 14. ATTRITION COUNT FOR THE EVALUATION OF IMPACT ON TEACHER PEDAGOGICAL CONTENT KNOWLEDGE (*MIXED* SAMPLE)

	Count of schools at baseline	Number of schools in the sample used in analysis with 114 teachers	Attrition of schools	Counts of teachers at baseline (in non-attriting schools)	Counts of teachers at baseline with posttests (in nonattriting schools)	Attrition of teachers
***MSS* (*n*)**	30	26	13.3%	81	56	30.86%
Control (*n*)	30	27	10.0%	81	58	28.4%
Total *N*	60	53		162	114	
Overall attrition			11.7%			29.6%
Differential attrition			3.3%			2.5%
Potential for bias			low			low

Note. *MSS* stands for the group of students receiving the Making Sense of SCIENCE professional learning. There is one additional school lost to attrition (no posttests) compared to the analysis of Teacher Content Knowledge. This results in three fewer teachers in the count of teachers at baseline among non-attriting schools.

TABLE 15. ATTRITION COUNT FOR THE EVALUATION OF IMPACT ON TEACHER PEDAGOGICAL CONTENT KNOWLEDGE (*RETAINED IN STUDY* SAMPLE)

	Count of schools at baseline	Number of schools in the sample used in analysis with 87 teachers	Attrition of schools	Counts of teachers at baseline (in non-attriting schools)	Counts of teachers at baseline with posttests (in non-attriting schools)	Attrition of teachers
***MSS* (*n*)**	30	25	16.7%	78	44	43.5%
Control (*n*)	30	22	26.7%	69	43	37.7%
Total *N*	60	47		147	87	
Overall attrition			21.6%			40.8%
Differential attrition			10.0%			5.8%

TABLE 15. ATTRITION COUNT FOR THE EVALUATION OF IMPACT ON TEACHER PEDAGOGICAL CONTENT KNOWLEDGE (*RETAINED IN STUDY* SAMPLE)

	Count of schools at baseline	Number of schools in the sample used in analysis with 87 teachers	Attrition of schools	Counts of teachers at baseline (in non-attriting schools)	Counts of teachers at baseline with posttests (in non-attriting schools)	Attrition of teachers
Potential for bias			high			intermediate

Note. *MSS* stands for the group of students receiving the Making Sense of SCIENCE professional learning.

Tests of Baseline Equivalence

We tested baseline equivalence for (a) the BRS, (b) the *Mixed* sample, and (c) the *Retained in Study* sample. Results are in Table 16.

TABLE 16. TESTS OF BASELINE EQUIVALENCE ON PRETESTS BETWEEN MAKING SENSE OF SCIENCE AND CONTROL FOR THE BASELINE REPRESENTATIVE AND ANALYTIC SAMPLES (*MIXED* AND *RETAINED IN STUDY* SAMPLE)

	Baseline	Analytic (*Mixed*)	Analytic (*Retained in Study*)
Pretest			
***N* (schools)**	60	53	47
***N* (teachers)**	183	114	87
Point estimate	-0.01	-0.02	-0.02
Standard error	0.01	0.02	0.02
***p* value**	.509	.185	.307
Standardized effect size	-0.07	-0.18	-0.16

Analysis

To assess the impact on teacher pedagogical content knowledge, we began with the standard analytic model used in the analysis of teacher content knowledge: a hierarchical linear model with fixed block (pair) effects, school-level random effects, and the same set of covariates. Given that the PCK outcome comprises ordinal responses, we used a logit link function to model the difference between conditions in the cumulative probability of correct response. We collapsed over the top intervals in cases where expected cell counts were less than or equal to 5. We adjusted for clustering of teachers in schools and blocks. We evaluated a model where we included a dummy variable to indicate whether item scores involved both raters. The rater effect was not appreciable and was excluded. The logistic regression models had difficulty converging, especially with pair fixed effects and the same covariates as in the benchmark model for evaluating impacts

on teacher content knowledge. Therefore, we used random effects for matched pairs and the pretest as the only covariate. In addition to the cumulative logistic regression models, we evaluated impacts using multilevel linear regression models.

For the two main samples, we report results for each of the three rating dimensions, and for the holistic rating. For each, we report results on the logistic metric and on the linear scale.

TEACHER PEDAGOGICAL CONTENT KNOWLEDGE: FINDINGS

Overall Impact for the *Mixed* Sample

We display impact results for the *Mixed* sample for the three main dimensions and for the average and holistic scores in Table 17. We observed a statistically significant effect for the holistic score based on a linear model, with a standardized effect size of 0.36 (p = .049).

TABLE 17. IMPACT OF MAKING SENSE OF SCIENCE ON PEDAGOGICAL CONTENT KNOWLEDGE FOR THE *MIXED* SAMPLE

	Linear scale effect size	Cumulative logistic regression effect size
PCK Concept	0.24 (p = .228)	0.31 (p = .212)
PCK 2D	0.28 (p = .101)	0.36 (p = .103)
PCK Explanation	0.23 (p = .157)	0.21 (p = .176)
Holistic	0.36 (p = .049)	0.41 (p = .068)

Note. The effect size based on the linear model is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome distribution. The effect size based on the cumulative logistic model is the Cox Index. The p values are from the associated impact estimates for each model.

Overall Impact for the *Retained in Study* Sample

We display impact results for the *Retained in Study* sample for the three main dimensions and for the average and holistic scores in Table 18. We show the impact estimate from the linear model and the cumulative logistic regression model. Similar to what we observed for the *Mixed* sample, the impact on the holistic score evaluated using the linear model achieves statistical significance, with a standardized effect size of 0.41 (p = .026).

TABLE 18. IMPACT OF MAKING SENSE OF SCIENCE ON PEDAGOGICAL CONTENT KNOWLEDGE FOR THE *RETAINED IN STUDY* SAMPLE

	Linear scale effect size	Cumulative logistic regression effect size
PCK Concept	0.25 (p = .255)	N/A
PCK 2D	0.29 (p = .104)	0.38 (p = .150)
PCK Explanation	0.12 (p = .053)	0.28 (p = .252)
Holistic	0.41 (p = .026)	0.50 (p = .053)

Note. The effect size based on the linear model is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome distribution. The effect size based on the cumulative logistic model is the Cox Index. The p values are from the associated impact estimates for each model. For the PCK Concept dimension, the logistic regression model did not converge to a result.

DISCUSSION

As generalists, elementary school teachers teach across content areas and often do not have the adequate background or content knowledge to teach science. A recent national survey revealed that 77% of elementary school teachers felt "very well prepared" to teach ELA and 73% felt "very well prepared" to teach math, but only 31% felt the same level of preparedness to teach science (Banilower et al., 2018). This discrepancy is likely driven by the focus of policy and resources allocated toward curriculum, professional learning, and testing in ELA and math. Making Sense of SCIENCE responds to the growing need for professional learning to support teachers' preparedness and ability to teach science. Specifically, the teacher professional learning component of Making Sense of SCIENCE is posited to have a direct impact on teacher content knowledge and pedagogical content knowledge, which can play a critical role in improving instructional practices and student achievement (Hill et al., 2005; Kanter & Konstantopoulos, 2010). In this study, Making Sense of SCIENCE was successful in improving teacher content knowledge and pedagogical content knowledge in science. We observed a positive impact of Making Sense of SCIENCE on teacher content knowledge of 0.56 standard deviations (p = .006) for the *Retained in Study* sample of teachers (those who were active in the study for the entirety of the implementation period). This means that a teacher at the median of performance on the test of content knowledge in the treatment condition achieved a score corresponding to the 71st percentile in the performance distribution for teachers in the control condition. The impact was robust across different statistical models. Additionally, exploratory analysis showed there were no differential effects for teachers with different baseline levels of science content knowledge, number of years of teaching experience, or whether the teacher was also a teacher leader.

Exploration of the impact of Making Sense of SCIENCE on pedagogical content knowledge revealed a positive impact of 0.41 standard deviations (p = .026) on scores using the holistic rating, and a positive and marginally significant impact on PCK-explanation scores (ES = 0.12; p = .053) for the *Retained in Study* sample. There was no impact on the PCK-concept scores or on scores based on a multidimensional approach to scoring that incorporated both concepts and explanation (PCK-2D). HRA explains that the holistic rating functions as an overall PCK score, and the score takes into account whether a teacher's written responses exhibited PCK in conceptual understanding or the scientific practice of explanation in any form. In this way, the PCK-explanation and holistic rating may be more sensitive to measuring the impact of the professional learning, which provided more support for developing teachers' ability to foster student engagement in scientific practices, as aligned with NGSS priorities (Wong et al., 2020).

Next, we turn to the impact of Making Sense of SCIENCE on teacher attitudes and beliefs, opportunities to learn, and school climate.

Chapter 5. Impact on Teacher Attitudes and Beliefs, Opportunities to Learn, and School Climate

INTRODUCTION

In the previous chapter, we discussed the impact of the Making Sense of SCIENCE professional learning model on teacher content knowledge and teacher pedagogical content knowledge. Reading from left to right on the Making Sense of SCIENCE logic model, acknowledging that the trajectory may not be linear, we now discuss the next set of outcomes of interest: teacher attitudes and beliefs, opportunity to learn, and school climate. Similar to the previous chapter, this chapter aims to shed light on the "black box" of intermediate outcomes between Making Sense of SCIENCE implementation and student achievement. This chapter differs from the previous in that the intermediate outcomes addressed in this chapter are from the teacher surveys.

In this exploratory analysis, for each intermediate outcome, we address the research question: What is the impact of the Making Sense of SCIENCE professional learning model on the intermediate outcome after two years of implementation of Making Sense of SCIENCE? We also examine whether the impact is different for teacher leaders compared to non-teacher leaders. Through this exploration, we hope to identify areas of the logic model that are supported by the empirical evidence generated in this study, as well as areas that would benefit from further research.

This chapter is organized into three sections. In the first section on methods, we begin by briefly describing the process undertaken by the evaluation and program teams to create the survey composites (henceforth "outcomes") and summarize the resulting list of outcomes. We also summarize the analytic sample and baseline equivalence. We then present the statistical analysis used to address the research. The second section reports the findings. The final section provides a summary of results and key takeaways.

METHODS

Measures

This analysis comprises 30 intermediate outcomes across three domains: teacher attitudes and beliefs, opportunity to learn, and school climate. As described in the background section of this report, however, this study took place during a time of great change in science instruction. The study's planning years (2014–15 and 2015–16) immediately followed the release of the NGSS in April 2013. As was the case with the student and teacher assessments, this shift in science instruction posed immense challenges in identifying survey scales that could measure NGSS-aligned constructs. The research team responded to this challenge by collaborating closely with the program developers to create surveys that consisted of items drawn from multiple sources, including researcher-developed items. Unlike with our confirmatory and preregistered analyses where the main contrasts and outcome domains were established at the start of the study, we felt it important to allow the choice of scales to reflect the refinement of the logic model, in order to provide the program developers with the most up-to-date feedback about impacts on the intermediate outcomes.

After data collection was completed, researchers and program developers collaborated in creating composites to represent the final set of intermediate outcomes of interest, which were continually refined throughout the two years of the study. Scales with Cronbach's alphas below .60 were either discarded or augmented with additional items identified

by the program developers. There were several iterations of this process before evaluators and program developers arrived at 30 constructs measured across the three domains, as presented in Table 19.

As mentioned in the introduction chapter, the study administered a baseline survey and three online surveys to teachers, per year, for two years (2016–17 and 2017–18). Each survey consisted of approximately 40–50 questions, and teachers were requested to set aside 30–45 minutes to complete each survey. This analysis is primarily based on data from surveys administered in spring of the 2017–18 school year, as we were often most interested in the final cumulative impact assessed on the last survey occasion. If an outcome of interest was not measured in the spring 2017–18 survey, we obtained data collected on the winter 2017–18 survey. For a few outcomes, such as time spent on science instruction, we aggregated data from all three surveys from the 2017–18 school year, in order to capture the average for the year. We provide an overview of the outcomes below (see Table B1 in Appendix B for the complete list of the constructs and the details, such as the number of items and Cronbach's alphas, for each construct). Table 19 includes the domains and constructs for intermediate outcomes assessed in this chapter.

Teacher Attitudes and Beliefs

This exploratory analysis includes eight outcomes related to teacher beliefs and attitudes, including teacher confidence in areas such as addressing student performance, instructional practices, and supporting literacy in science; sense of agency and self-efficacy; alignment between their own teaching philosophy and NGSS; and beliefs about students. All survey scales comprise items that are on a 5-point Likert scale. For each outcome, we average over the corresponding items to arrive at the rating for each teacher.

Opportunity to Learn

The program team identified four components of opportunity to learn: time, instruction, content, and classroom climate. This exploratory analysis examines the first three components; data for classroom climate are not used because the achieved scale reliabilities were too low. The *time* component is measured as the total number of hours the teacher taught science during the prior four weeks of instruction. Responses are averaged across the fall, winter, and spring surveys. The *instruction* component includes four outcomes that focus on NGSS-aligned instructional practices. The *content* component includes ten outcomes: three for Disciplinary Core Ideas (DCIs) related to Earth and space science, five for DCIs related to physical science, one for science and engineering practices (SEPs), and one for cross-cutting concepts (CCCs). Teachers are asked to indicate whether they did not teach, touched on, or taught in depth for each topic area. Teacher responses are averaged across the items.

School Climate

This domain focuses on administrator support, teacher-teacher and teacher-administrator relationships, and collaboration. With the exception of the amount of informal teacher collaboration, all outcomes are measured on a 5-point Likert scale. The amount of informal teacher collaboration is measured on a 4-point scale ranging from "none" to "more than 4 hours" during the prior four weeks of instruction.

TABLE 19. DOMAINS AND CONSTRUCTS FOR INTERMEDIATE OUTCOMES

Domain	Construct
Teacher attitudes and beliefs	Philosophically aligned with NGSS Belief in life-long learning Agency in the classroom Belief that students are capable learners Self-efficacy Confidence in addressing student performance expectations Confidence in supporting literacy in science Confidence in science instructional practices
Opportunity to learn [a]	Time • Time spent on science instruction Instruction • Participating in collaborative discourse • Explaining ideas and phenomena • Sense-making of hands-on investigations • Integration of science and literacy Content • Earth and human activity • Earth's place in the universe • Earth's systems • Definition of energy • Conservation of energy transfer • Matter and its interactions • Motion and stability – forces and interactions • Waves • Science and engineering practices (SEPs) • Cross-cutting concepts
School climate	Administrator support involving teachers in science leadership Administrators provide support for teacher collaboration Administrators prioritize support for teacher professional learning activities Peer collaboration valued Trust and respect among peers Trust and respect between teachers and administrators Amount of teacher informal collaboration

[a] A fourth dimension of opportunity to learn was of interest to program developers: classroom climate. However, data for these constructs, such as student-centered learning, had low Cronbach's alphas and thus were excluded from the analysis.

Sample and Baseline Equivalence

The sample of teachers included in this analysis comprises a subset of the 147 teachers (81 *Making Sense of SCIENCE*, 66 control) from 55 schools (29 *Making Sense of SCIENCE*, 26 control) whose students were included in the confirmatory analysis of impact on science achievement (n = 2140) and who completed the fall (n = 142), winter (n = 141), and spring (n = 141) surveys. The sample varies slightly depending on which survey a particular outcome draws upon. This particular sample of teachers is selected for this analysis because it most closely aligns with the sample of students for the main confirmatory analysis.

Among the sample of 147 teachers, there were 37 teachers (21 *Making Sense of SCIENCE*, 16 control) who were also teacher leaders. Teacher leaders included in this sample were those who participated in the study as a classroom teacher and also served in the Leadership Cadre as a teacher leader. Aside from the professional learning afforded to all teachers at *Making Sense of SCIENCE* schools, teacher leaders at *Making Sense of SCIENCE* schools attended an additional 18 hours of Leadership Cadre workshops, which included training on facilitating the school-year PLC meetings. A subset of teacher leaders also attended a 40-hour teacher course facilitation academy on facilitating the summer course institutes. Teacher leaders at control schools served as the point of contact for the school.

Teachers were selected to be teacher leaders prior to school randomization; however, teacher leaders who left the study during the study period were replaced by another teacher leader. Of the 37 teacher leaders included in this analysis, 21 were in *Making Sense of SCIENCE* schools (15 consented prior to randomization), and 16 were in control schools (13 consented prior to randomization).

We tested baseline equivalence on a number of measures—including the teacher content knowledge pretest, teacher baseline level of confidence and perceived level of influence, teacher education and teaching background, and teacher-administrator relationships at the school—for the sample of 147 teachers whose students were included in the confirmatory analysis of impact on science achievement. We found that teachers in control schools reported higher levels of education (ES = -0.40; p = .048). We found the *Making Sense of SCIENCE* and control teachers were equivalent at baseline for other covariates that could be positively associated with outcomes, which we controlled for in impact models: a) content knowledge pretest (ES = -0.21; p = .237), b) confidence in literacy and discourse (ES = 0.01; p = .956), c) perceived level of influence (ES = -0.08; p = .676), and d) school culture between teachers and administrators (ES = -0.35; p = .145).[28]

Statistical Analysis

Main Analysis

The main analysis employs a three-level (teacher, schools and matched pairs) hierarchical linear model that regresses each of the 30 intermediate outcomes on an indicator of assignment status (*Making Sense of SCIENCE* or control) and a series of teacher- and school-level covariates as described in the section on Analysis of Impact of Teachers in Chapter 2.

[28] We used the standard approach to testing baseline equivalence with HL models, by regressing the covariate against the indicator of treatment status and including pair fixed effects and random effects parallel to the impact models. The standardized effect size is the estimated regression coefficient for the treatment variable divided by the pooled standard deviation of the covariate.

Additional Analyses and Sensitivity Checks

We conduct several additional analyses to assess the robustness of the main results. We assess five models where we vary how we model matched pairs and use either Full or Restricted Maximum Likelihood. Results are given in Appendix J.

Moderation Analysis

To assess the extent to which teacher leaders moderate the impact of Making Sense of SCIENCE on intermediate outcomes, we add two terms to the main model described above: a binary term to indicate whether the teacher is also a teacher leader and a treatment-teacher leader interaction term. [29]

For completeness, we report impacts for the full group and subgroups, as well as the differential impact across the subgroups. The presence of a differential effect should be based on whether the test of differential impact is statistically significant, and not whether there is a difference between subgroups in the statistical significance of their individual results.

FINDINGS

Teacher Attitudes and Beliefs

Among the eight constructs related to teacher attitudes and beliefs, for the full sample of teachers, there is a positive and statistically significant impact of Making Sense of SCIENCE on teachers' sense of *Agency in the Classroom* (ES = 0.38, p = .025). The impacts of the remaining outcomes are not statistically significant, though they are all positive, with the exception *Belief That Students Are Capable Learners*. One outcome, *Confidence in Science Instructional Practices,* is marginally statistically significant (ES = 0.26, p = .083).

For the subgroup of teacher leaders, the impact on *Confidence in Addressing Student Performance Expectations* is positive and significant (ES = 0.66, p = .016), and the impact on *Confidence in Science Instructional Practices* is positive and marginally significant (ES = 0.49, p = .058).

We observe a differential impact, favoring teacher leaders for the *Belief That Students Are Capable Learners* outcome (ES = 0.82, p = .022). For *Confidence in Addressing Student Performance Expectations*, we observe a small difference in impact for teacher leaders (ES = 0.60) that is close to reaching statistical significance (p = .058) (Figure 6).

[29] In several cases, teachers chose to be teacher leaders after random assignment. In analyzing average impacts, we do not include teacher-leader status as a covariate because it is not strictly a baseline covariate. Because impacts across teacher-leader categories were not adjusted for teacher-leader status, but subgroup impacts are stratified by teacher-leader status, in a couple of cases, the estimates of overall impact do not lie between estimates of corresponding subgroup impacts.

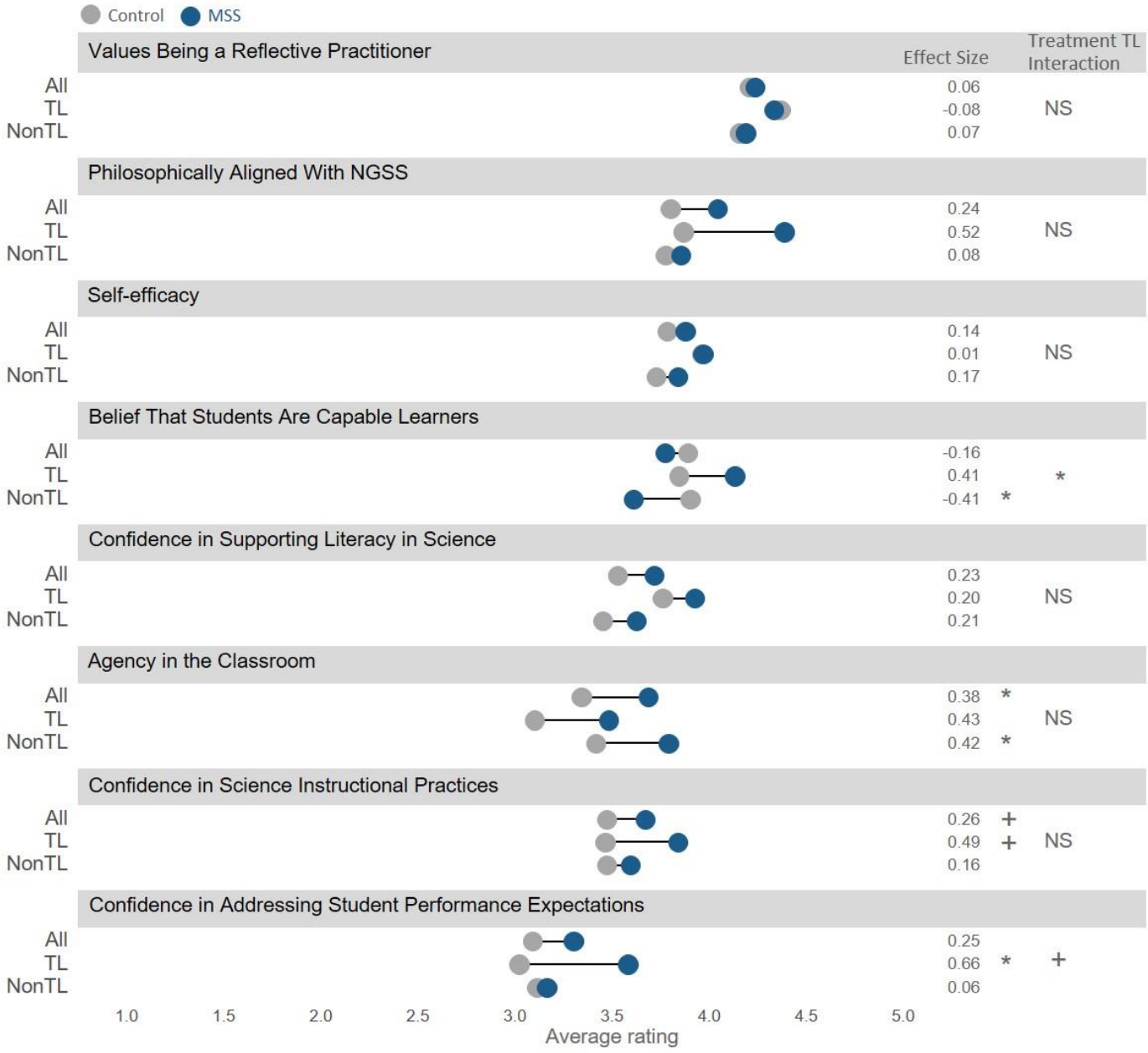


FIGURE 6. IMPACTS ON TEACHER ATTITUDES AND BELIEFS

Note. TL = teacher leaders. MSS = Making Sense of SCIENCE.

The sample consists of 141 teachers (36 teacher leaders and 105 non-teacher leaders) who completed the spring 2017–18 survey.

Gray endpoints represent the raw means for the control group. Blue endpoints represent the adjusted means for the treatment group. The difference between blue and gray endpoints on a line is the regression-adjusted impact estimate for the group in scale score units.

We based *p* values on models that employed transformed outcomes (outcomes were transformed if they have a skew of greater than 0.7). All outcomes were based on a 5-point Likert scale.

For effect size and treatment TL interaction, ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, +$p < 0.1$.

Opportunity to Learn

Opportunity to Learn Derived through Increased Time on Science

We observe a positive impact (ES = 0.40, p = .015) for the *Amount of Time Spent on Science* for the sample of all teachers.[30] For the subsample of teacher leaders, both the *Making Sense of SCIENCE* and control groups report a lower amount of time spent on science than their counterparts in the subsample of teachers who are not teacher leaders. The standardized differential impact for teacher leaders is -0.20, but it is not statistically significant (p = .560) (Figure 7).

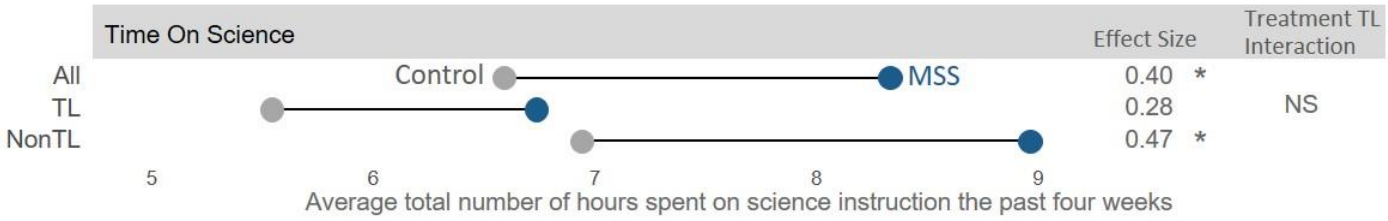


FIGURE 7. IMPACT ON OPPORTUNITY TO LEARN DERIVED THROUGH INCREASED TIME ON SCIENCE

Note. TL = teacher leaders. MSS = Making Sense of SCIENCE.

The sample consists of 141 teachers who completed at least one of three fall, winter, and spring 2017–18 surveys (36 teacher leaders and 105 non-teacher leaders). Sample excludes outliers (three teachers who completed the fall survey, three who completed the winter survey, and seven who completed the spring survey). We defined outliers as having responses greater than the value of the third quartile plus 1.5 times the interquartile range.

Gray endpoints represent the raw means for the control group. Blue endpoints represent the adjusted means for the treatment group.

We based p values on models that employed transformed outcomes (outcomes were transformed if they have a skew of greater than 0.7).

For effect size and treatment TL interaction, ***p < 0.001, **p < 0.01, *p < 0.05, +p < 0.1.

Opportunity to Learn Derived through Teacher Instruction

Among the intermediate outcomes assessed in this exploratory analysis, the impacts on the outcomes belonging to the *Instruction* component of the *Opportunity to Learn* domain offer the greatest promise for demonstrating impact along the path of the logic model. Three of the four outcomes (*Participating in Collaborative Discourse, Sense-Making of Hands-On Investigations,* and *Integration of Science and Literacy)* have positive and significant results (ES of 0.46, 0.40, and 0.49, and p values of .005, .018, and .003, respectively). The remaining outcome, *Explaining Ideas and Phenomena,* also has a positive effect (ES = 0.32) but does not reach statistical significance (p = .064) (Figure 8).

[30] We removed outliers from this analysis.

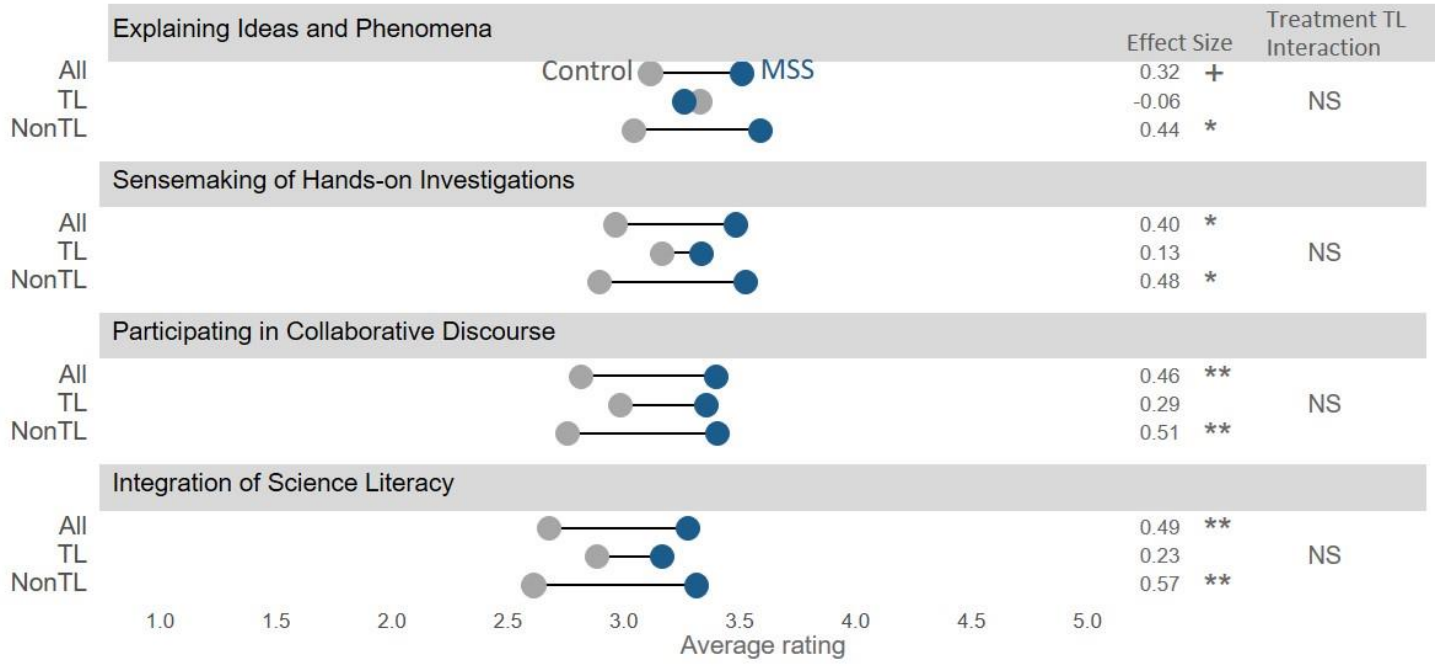


FIGURE 8. IMPACT ON OPPORTUNITY TO LEARN DERIVED THROUGH TEACHER INSTRUCTION

Note. TL = teacher leaders. MSS = Making Sense of SCIENCE.

The sample consists of 142 teachers (36 teacher leaders and 106 non-teacher leaders) who completed the spring 2017–18 survey.

Gray endpoints represent the raw means for the control group. Blue endpoints represent the adjusted means for the treatment group.

We based *p* values on models that employed transformed outcomes (outcomes were transformed if they have a skew of greater than 0.7). All outcomes were based on a 5-point Likert scale.

For effect size and treatment TL interaction, ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, +$p < 0.1$.

Opportunity to Learn Derived through Greater Exposure to NGSS-Aligned Content Areas

Contrary to the trends observed under OTL-instruction, for OTL-content, we observe that while the impact for the full sample of teachers is not statistically significant for any of the outcomes, there are positive and statistically significant impacts observed in the subgroup of teacher leaders for several outcomes: *Definition of Energy* (ES = 0.79, $p = .013$), *Conservation of Energy and Energy Transfer* (ES = 0.90, $p = .002$), and *Science and Engineering Practices* (ES = 0.62, $p = .039$). These three constructs also exhibit a significant or marginally significant differential impact between teachers who are teacher leaders and those who are not, indicating that there is value-added in being teacher leaders for these outcomes. All the positive and statistically significant effects are those within the physical science domain.

We observe no statistically significant impact for any of the Earth and space science outcomes for analyses of the full sample and the subgroups of teacher leaders and non-teacher leaders. In fact, two of three outcomes have negative effect sizes, though they are not statistically significant. However, teachers tended to report having taught Earth and space science in greater depth than physical science and cross-cutting concepts (Figure 9).

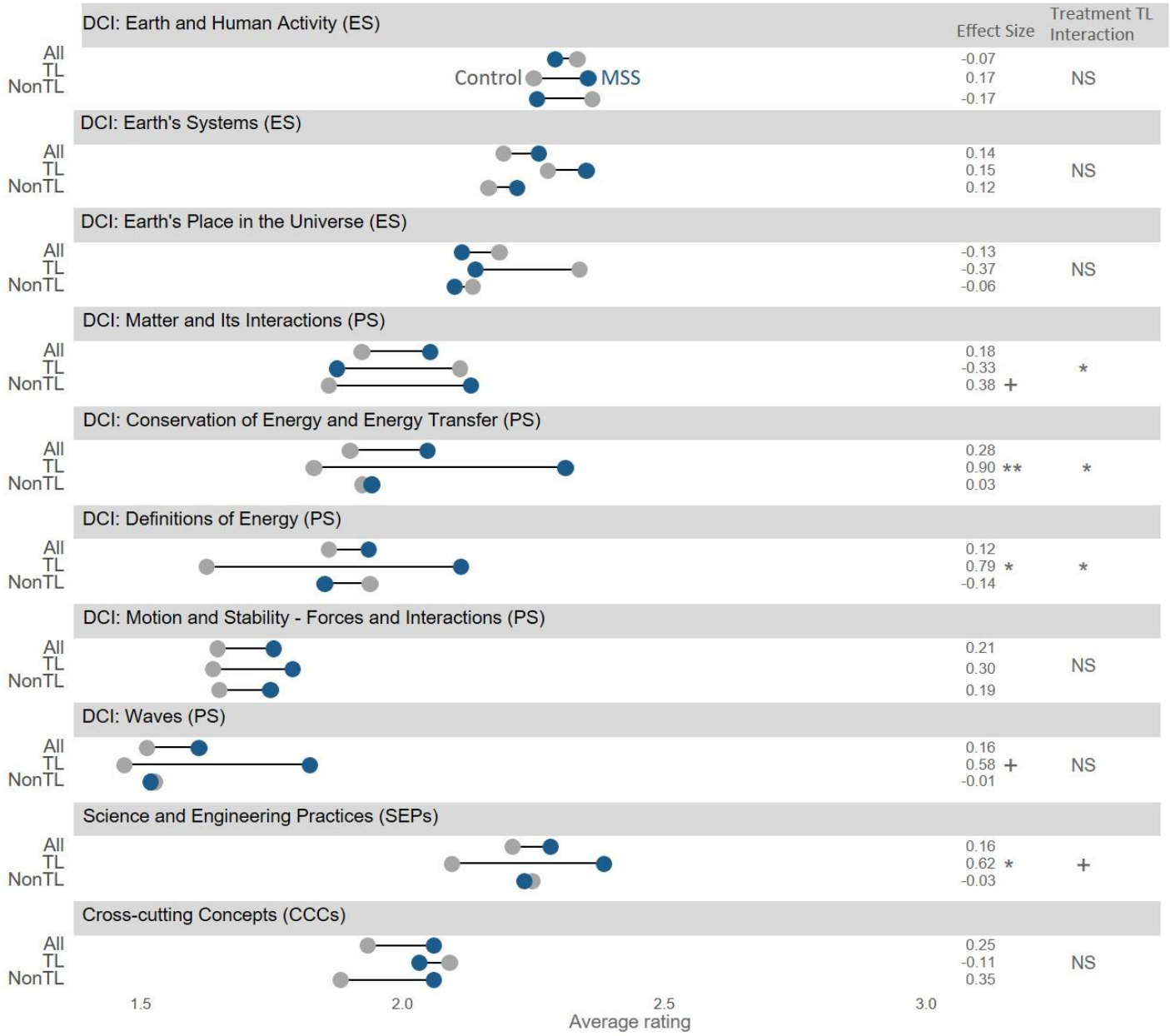


FIGURE 9. IMPACTS ON OPPORTUNITY TO LEARN DERIVED THROUGH GREATER EXPOSURE TO NGSS-ALIGNED CONTENT AREAS

Note. TL = teacher leaders. MSS = Making Sense of SCIENCE. ESS = Earth and space science. PS = physical science.

The sample consists of 141 teachers (36 teacher leaders and 105 non-teacher leaders) who completed the spring 2017–18 survey.

Gray endpoints represent the raw means for the control group. Blue endpoints represent the adjusted means for the treatment group.

We based *p* values on models that employed transformed outcomes (outcomes were transformed if they have a skew of greater than 0.7). All outcomes were based on a 3-point scale (did not teach, touched on, taught in depth).

For effect size and treatment TL interaction, ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, +$p < 0.1$.

School Climate

For the full sample of teachers, we observe a positive and statistically significant impact for *Administrators Providing Support for Teacher Collaboration* (ES = 0.39, p = .025). For the subsample of teacher leaders, we observe a positive and statistically significant impact for *Administrator Support Involving Teachers in Science Leadership* (ES = 1.14, p < .001). Two other constructs related to administrator support—A*dministrators Provide Support for Teacher Collaboration* (ES = 0.55, p = .064) and *Administrators Prioritize Support for Teacher Professional Learning Activities* (ES = 0.63, p = .057)—show positive impacts for the subsample of teacher leaders, but did not reach statistical significance (Figure 10).

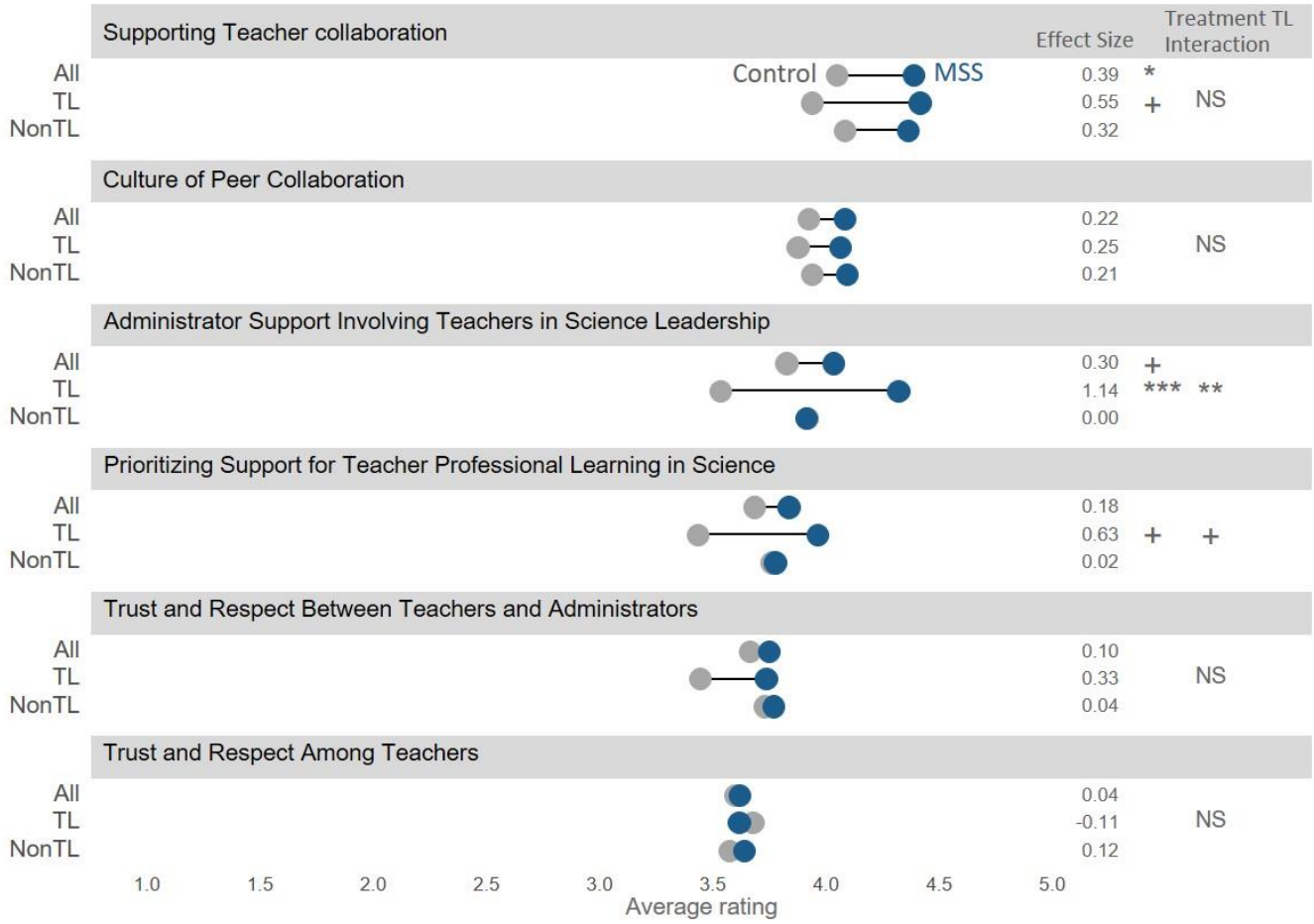


FIGURE 10. IMPACTS ON SCHOOL CLIMATE OUTCOMES

Note. TL = teacher leaders. MSS = Making Sense of SCIENCE.

The sample ranges from 139 to 141 teachers across the six outcomes (36 teacher leaders and 103-105 non-teacher leaders) who completed the spring 2017–18 survey.

Gray endpoints represent the raw means for the control group. Blue endpoints represent the adjusted means for the treatment group.

We based p values on models that employed transformed outcomes (outcomes were transformed if they have a skew of greater than 0.7). We based all outcomes on a 5-point Likert scale.

For effect size and treatment TL interaction, ***p < 0.001, **p < 0.01, *p < 0.05, +p < 0.1.

We also observe a positive impact on the *Amount of Informal Peer Collaboration*, which refers to collaboration that is beyond Making Sense of SCIENCE-directed PLC meetings, for the full sample of teachers (ES = 0.88, $p < .001$) and the subsamples of teacher leaders (ES = 0.86, $p = .003$) and non-teacher leaders (ES = 0.90, $p < .001$), though there is no differential impact for teacher leaders (Figure 11).

FIGURE 11. IMPACTS ON AMOUNT OF INFORMAL PEER COLLABORATION

Note. TL = teacher leaders. MSS = Making Sense of SCIENCE.

The sample consists of 142 teachers (36 teacher leaders and 106 non-teacher leaders) who completed the spring 2017–18 survey.

Gray endpoints represent the raw means for the control group. Blue endpoints represent the adjusted means for the treatment group.

We based *p* values on models that employed transformed outcomes (outcomes were transformed if they have a skew of greater than 0.7). We based all outcomes on a 4-point scale (1 = I did not participate in any informal peer collaboration for science instruction in the past 4 school weeks; 2 = 1-2 hours; 3 = 3 to 4 hours; 4 = more than 4 hours.

For effect size and treatment TL interaction, ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, +$p < 0.1$.

DISCUSSION

Several interpretations emerge from the findings above. First, the results of the intermediate outcomes based on the teacher surveys follow a trend observed for teacher content knowledge and teacher pedagogical content knowledge: effect sizes are primarily positive, though not all are substantively or statistically significant. Moreover, outcomes that are positive and significant (such as teachers' agency in the classroom, level of confidence (for subsample of teacher leaders), time spent on teaching science, instructional practices, the amount of informal peer collaboration, and administrator support for collaboration) are proximal to the summer institutes and the school-year PLC meetings. Important to note is that the outcome *Amount of Informal Peer Collaboration* measures the amount of time teachers spend collaborating with one another *informally*, suggesting that teachers are collaborating above and beyond Making Sense of SCIENCE-directed professional learning activities.

On the contrary, intermediate outcomes that are not statistically significant are relatively further removed from teacher professional learning, such as deeply-ingrained personal beliefs (e.g., believing that students are capable learners or valuing being a reflective practitioner) or outcomes that are more likely to be resistant to change (e.g., trust and respect among teachers and between teachers and administrators).

Analysis of the differential impact for teacher leaders reveals that there is not necessarily a "value-added" on all of these intermediate outcomes for being a teacher leader, though there is some evidence that there is value-added for Making

Sense of SCIENCE teacher leaders in terms of belief that students are capable learners, emphasis on a number of the physical science content areas, and teacher leaders' perception of administrator support for involving teachers in science leadership.

Another interpretation on why impacts are not observed on some of the hypothesized outcomes is that the data are based on self-reported teacher survey data, rather than on more objective measures such as classroom observations, enacted lesson plans, or class artifacts. As such, the data are not immune from shortcomings common to self-reported data, such as social desirability bias and respondents having recall issues. Another interpretation of the findings is related to the shortcomings of the items themselves. For example, the survey items on trust and respect among teachers and between teachers and administrators ask teachers generally about the atmosphere and about the activities of teachers and administrators at the school. In the *Making Sense of SCIENCE* condition, this could mean that teachers are basing their survey responses on reflections about Making Sense of SCIENCE-related activities, as well as non-Making Sense of SCIENCE-related interactions. The items do not ask specifically about the teachers and administrators who are participating in the Making Sense of SCIENCE professional learnings, thus possibly obscuring any positive impact on these outcomes, had the question been explicitly limited to teachers and administrators participating in the study.

Lastly, we hypothesize that the fact that none of the OTL-content outcomes yield positive results for the full sample reflects the intent of the Making Sense of SCIENCE professional learning model. The model focuses on the "how" to teach rather than "what" to teach. While the model supports teachers in their effort to implement NGSS standards, it is not a curriculum and does not provide extensive supplementary instructional material. This hypothesis is corroborated by the fact that for the full sample of teachers, OTL-instruction outcomes are almost all statistically significant, while none of the OTL-content outcomes are.

We note two promising areas for future research exploration. First, given that the data for the intermediate outcomes are self-reported, future research could consider triangulating responses from teacher surveys with those from the administrator and student surveys, as well as with measures typically considered to be more objective measures, such as classroom observations or artifacts from the lessons themselves.[31] Second, given the positive impacts on OTL-instruction but not OTL-content, further research could explore whether having high-quality curricular material is a necessary condition for impact on student achievement.

In the next four chapters, we turn to examining the effectiveness of the Making Sense of SCIENCE professional learning model on student outcomes: Chapter 6 focuses on student science achievement in Earth and space science and physical science (confirmatory research question); Chapter 7 explores student performance on state assessments in ELA, math, and science; Chapter 8 examines students' communication of science ideas in writing; and Chapter 9 explores impacts on students' non-academic outcomes.

[31] As mentioned in Chapter 2 on *Study Methods*, the study did try to collect these more objective measures through video and audio recordings of the classrooms, but the parental consent rates for these data collection efforts were too low for us to include the data in any of the impact analyses.

Chapter 6. Impacts on Student Science Achievement in Earth and Space Science and Physical Science

INTRODUCTION

In this chapter, we present several findings related to the impact of Making Sense of SCIENCE on student science achievement in Earth and space science and physical science. We present findings answering our confirmatory research question on student science achievement, as well as exploration of the impact for more focused samples of students, including subsamples of special interest (by grade and state) and where implementation was considered to be stronger. Our research questions are as follows.

Confirmatory Research Questions

- What is the impact of Making Sense of SCIENCE, after two years of implementation, on student science achievement in Earth and space science and physical science among fourth- and fifth-grade students in intervention schools, compared to fourth- and fifth-grade students in control schools receiving the business-as-usual science instruction?
- What is the impact of Making Sense of SCIENCE among fourth- and fifth-grade students *in the lowest third of ELA achievement*, after two years of implementation, on science achievement in Earth and space science and physical science in intervention schools, compared to students in control schools receiving the business-as-usual science instruction?
- What is the impact of Making Sense of SCIENCE among fourth- and fifth-grade students *in the lowest third of math achievement*, after two years of implementation, on science achievement in Earth and space science and physical science in intervention schools, compared to students in control schools receiving the business-as-usual science instruction?

Exploratory Research Questions Related to Moderators

- Is there a differential impact of Making Sense of SCIENCE, after two years of implementation, based on student ELA and math pretests, state (Wisconsin versus California), grade level (fourth grade versus fifth grade), student English Language Learner (ELL) status, student eligibility for the Free or Reduced-Price Lunch (FRPL) program, or all the moderators combined?

Exploratory Research Questions Related to Science Sub-strand

- What is the impact of Making Sense of SCIENCE, after two years of implementation, on student science achievement, on the *physical science sub-strand*, when compared to study participants in control schools receiving the business-as-usual science professional learning?
- What is the impact of Making Sense of SCIENCE, after two years of implementation, on student science achievement, on the *Earth and space science sub-strand*, when compared to study participants in control schools receiving the business-as-usual science professional learning?

Exploratory Research Questions Related to Focused Sample of Students Based on Exposure to Making Sense of SCIENCE

- *Focused sample 1*: For students of the baseline representative sample (BRS) teachers who participated in the study for both years (2016–17 and 2017–18), what is the impact of Making Sense of SCIENCE on their science achievement, when compared to study participants in control schools receiving the business-as-usual science professional learning?

  We limited this analysis to BRS teachers who were in the study in both years and assessed impacts on their students in spring 2018. This analysis provides an opportunity to evaluate impact among teachers who, in the *Making Sense of SCIENCE* condition, received a greater (the intended) dose of the professional learning.

- *Focused sample 2*: For students enrolled in a *Making Sense of SCIENCE* teacher's class for two years (2016–17 and 2017–18), what is the impact of Making Sense of SCIENCE after two years of implementation on their science achievement, when compared to study participants in control schools receiving the business-as-usual science professional learning?

  We limited this analysis to students who were with study teachers for two years (that is, students in the treatment group were members of a Making Sense of SCIENCE-participating classroom teacher in fourth grade in the first year and with a Making Sense of SCIENCE-participating classroom teacher in fifth grade in the second year).

Exploratory Research Questions Related Impact of Making Sense of SCIENCE by State and Grade Level

- What is the impact of Making Sense of SCIENCE, after two years of implementation, on student science achievement, when compared to study participants in control schools receiving the business-as-usual science professional learning for the following students?
  - students in California
  - students in Wisconsin
  - students in fourth grade
  - students in fifth grade

Exploratory Research Questions Related to Impact of Making Sense of SCIENCE for Strong Implementing Districts

- Is there variation in the impact across all districts?
- Is there a difference in impact between high implementing districts and the rest?
- What is the impact among students from high implementation districts?

We consulted with WestEd in identifying districts where implementation was stronger, on average. WestEd identified three districts, among the seven participating districts across California and Wisconsin, as strong implementers using the following criteria.

1. There was active involvement and support that included at least one strong district leader, plus a coalition of support from other leaders throughout the study.

2. The district valued the program, expressed ongoing interest, and wanted more professional learning.
3. There was active involvement or other support from district administrators.
4. The district developed internal capacity in science (e.g., teacher leaders from the district facilitated summer professional learning).
5. School administrators participated in professional learning.
6. There was continuation of professional learning after the study.

In this chapter, we begin with a description of the methods, including a brief description of the measure, samples for the confirmatory analyses, and approach to analysis. Next, we present the results for each research question. First, we report benchmark results of the confirmatory analysis of impact across fourth and fifth grade, along with sensitivity analyses. Second, we report results of the confirmatory analysis of impact across fourth and fifth grade for students in the lowest third of ELA and math performance, along with sensitivity analyses. Third, for the larger sample, we report differential impacts across categories of students and schools. Fourth, we report impacts on more focused samples, including subsamples of special interest (by grade, and state) and where implementation was considered to be stronger. We conclude with a discussion of the findings.

METHODS

Measure

We used an evaluator-developed assessment with two forms: one consisting of 30 selected-response fourth grade items and another of 29 selected-response fifth grade items covering topics in Earth and space science, physical science, and life science. Both forms included the same 10 "Inquiry" items suitable to both grades. We selected items from several sources (e.g., MOSART and NAEP) to address general NGSS-aligned specifications. (More details about the test construction process are provided in Appendix D.) Students took the assessment online. The online test delivery platform included voiceover functionality such that students could click on the questions to hear it read aloud. Students had approximately one hour to complete both the assessment and the student survey (described further below).

We removed several problematic items (e.g., one item had a very strong distractor response option, and another item was part of a two-part question for which many students missed one of the parts). We also removed three life science items, with this strand not being covered in the professional learning. This resulted in a fourth-grade test with 25 items, and a fifth-grade test with 24 items. Scale scores were calibrated using a 3-PL model. Due to lack of consensus among advisors on the best approach to scaling, we also calibrated scores using 1-PL and 2-PL models and the percent correct metric.

Sample

For the confirmatory analysis of impacts on students, we examined impacts on 2,140 (1138 treatment, 1002 control) students across Grades 4 and 5, for whom we obtained posttest scores on the selected response portion of the spring 2017–18 student assessment of science achievement and who had non-missing ELA and math pretests from third grade. This constitutes a random subset of approximately 75% of students who were in classes of study teachers at the completion of the two-year program implementation period (the remainder were randomly assigned to a test of communication of

science ideas in writing, with results reported in Chapter 8). Of the 2,140 students, 1,303 students were from California, and 837 students were from Wisconsin.[32] Table 20 displays the analytic sample for the 2,140 students.

TABLE 20. ANALYTIC SAMPLE FOR CONFIRMATORY ANALYSIS FOR STUDENTS WITH SCORES ON SELECTED-RESPONSE ITEMS OF THE STUDENT SCIENCE ACHIEVEMENT ASSESSMENT (*N* = 2,140)

	Count of schools in CA	Count of students with posttest in CA	Count of schools in WI	Count of students with posttest in WI
MSS	16	722	13	416
Control	14	581	12	421
Total *N*	30	1303	25	837

Note. *MSS* stands for the group of students receiving the Making Sense of SCIENCE program. The disproportionately greater numbers of students in the MSS condition in California is attributable to one dyad in the treatment group that consisted of two full-sized schools. The dyad was randomly assigned as one unit. (The rationale for using dyads is described earlier in the report.)

Table 21 shows achieved analytic sample sizes at the level of random assignment (school) and the test of equivalence for the ELA and math pretests. We observed that baseline equivalence is established for both ELA pretest and math pretest. The impact analysis can potentially meet NEi3 and WWC evidence standards with reservations given that the impact model adjusts for the effects of the pretests.

[32] Schools were randomized in winter 2015–2016: half to receive the Making Sense of SCIENCE professional learning and half to business-as-usual. Schools and teachers were notified of random assignment status in winter 2015–2016 so preliminary professional learning could begin (i.e., with teachers receiving training in the summer prior to the 2016–17 school year) to support full intervention implementation in fall 2016–2017. Families and students were not notified about the assignment status of their school; therefore, it is unlikely that they would have selected into study schools based on knowledge of the assignment status. What may have plausibly occurred, however, was the placement of students into study teachers' classes within schools, after teachers' participation status was known to school personnel making class assignments for students. This type of sorting may have been repeated again the following year, when student rosters were formed in fall 2016–2017. Impacts on student science achievement reported here are based on outcomes from spring 2017–2018. We asked the treatment group on the administrator survey to what extent participation in Making Sense of SCIENCE was a factor in creating student rosters. Of the 25 responses, 21 administrators said "not at all." The other 4 indicated other factors were considered (e.g., teacher strengths, the capacity of students to work well in groups).

TABLE 21. TESTS OF BASELINE EQUIVALENCE BETWEEN MAKING SENSE OF SCIENCE AND CONTROL ON THE ELA AND MATH PRETESTS

	ELA pretest	Math pretest
***N* (schools)**	55	
***N* (students)**	2,140	
Point estimate	-0.10	-0.05
Standard error	0.09	0.10
***p* value**	.264	.636
Standardized effect size	-0.11	-0.05

Note. Analytic sample for confirmatory analysis for students with scores on selected-response items (n = 2,140)

The sample of students in the lowest third in incoming ELA achievement included 715 students (406 *Making Sense of SCIENCE*, 309 control) in 55 schools. The sample of students in the lowest third in incoming math achievement included 713 students (403 *Making Sense of SCIENCE*, 310 control) also in 55 schools. The counts are given in Table 22.

TABLE 22. ANALYTIC SAMPLE FOR CONFIRMATORY ANALYSIS FOR STUDENTS WITH SCORES ON SELECTED RESPONSE ITEMS (LOWEST THIRD OF PRETEST DISTRIBUTION)

	Lowest third of ELA pretest		Lowest third of math pretest	
	Count of schools	Count of students with posttest	Count of schools	Count of students with posttest
MSS	29	406	29	403
Control	26	309	26	310
Total *N*	55	715	55	713

Note. *MSS* stands for the group of students receiving the Making Sense of SCIENCE professional learning.

We tested baseline equivalence for both ELA pretest and math pretest for students in the lowest third of incoming achievement in both subject areas. Results are in Table 23. We observe that baseline equivalence is established for both the ELA pretest and math pretest. The results of the impact analysis can potentially meet NEi3 and WWC evidence standards with reservations.

TABLE 23. TESTS OF BASELINE EQUIVALENCE ON PRETESTS BETWEEN MAKING SENSE OF SCIENCE AND CONTROL FOR LOWEST THIRD OF PRETESTS

	Lowest third of ELA pretest		Lowest third of math pretest	
	ELA pretest	Math pretest	ELA pretest	Math pretest
***N* (schools)**	55		55	
***N* (students)**	715		713	
Point estimate	-0.05	0.05	-0.01	-0.002
Standard error	0.04	0.07	0.06	0.08
***p* value**	.156	.471	.906	.977
Standardized effect size	-0.12	0.07	-0.01	-0.004

As listed in the research questions, we also limit this sample in several ways to estimate impacts for specific subsamples.

Impact Analysis: Scaling and Impact Models

We evaluated impacts on the spring 2017–18 assessment after two years of Making Sense of SCIENCE implementation. We estimated impacts using a hierarchical linear model with fixed block (pair) effects and school-level random effects. Covariates included z-transformed state assessments scores (in ELA and math) and grade level. Additional covariates were modeled at the student and teacher levels (See Appendix K for HLM specifications, and Appendix L for full results of confirmatory analysis with a listing of covariates.) We used Restricted Maximum Likelihood estimation. We handled missing values for covariates other than the pretest using dummy variable imputation. Cases without a pretest or a posttest were removed. Analysis of impacts on students in the lowest third of incoming achievement on math and ELA pretests used a similar model as the one used to assess impacts on the full sample, with student and school random effects, fixed pair effects, and the same set of covariates. (We provide a brief description of our analysis approach per subgroup prior to presenting those results.)

FINDINGS

Confirmatory Impacts of Making Sense of SCIENCE on Student Science Achievement

Impact

In Table 24, we present the main impact findings for the confirmatory analysis of impact on student science achievement. The means are the averages of the raw scores of the outcomes in each condition. The unadjusted model includes random school and student effects and fixed effects for pairs, but no covariates. The adjusted (benchmark) model includes these effects and the full set of covariates. For the benchmark impact model, we observe a positive but statistically nonsignificant impact, with a standardized effect size of 0.064 (p = .494).

TABLE 24. RESULTS OF CONFIRMATORY IMPACT ANALYSIS FOR THE STUDENT SCIENCE ASSESSMENT OUTCOMES (*N* = 2,140)

	Condition	Means	Standard deviations	No. of clusters	No. of students	Effect size	*p* value	Change in percentile ranking
Unadjusted effect size[a]	Control	0.00	1.00	26	1002	0.004	.967	0.1%
	MSS	-0.07	0.94	29	1138			
Adjusted effect size[b]	Control	0.00				0.064	.494	2.5%
	MSS	0.062						

Note. The *p* values are for the corresponding impact estimates in the benchmark impact model. *MSS* stands for the group of students receiving the Making Sense of SCIENCE professional learning.

[a] The unadjusted effect size is the impact estimate from a model with pair fixed, and school random effects, and without covariates, divided by the pooled standard deviation of the outcome variable.

[b] The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

Sensitivity Analyses[33]

We explored several less parameterized models, starting with a model with no covariates and then progressively including covariates until we arrived at the benchmark model. Standardized effect sizes for program impact ranged between -0.045 and 0.064, and none of the impact estimates reached statistical significance.

Next, we examined several different approaches to estimation, including (1) use of Ordinary Least Squares, (2) HLM but using ML instead of REML, (3) after including a teacher random effect, (4) after removing the pair effect, (5) after modeling a random intercept and treatment effect at the pair level and excluding the school random effect, (6) using Multiple Imputation to address missing values of all covariates, including the pretests, and (7) using a forward selection procedure to limit the number of covariates (treatment status was forced into the model and covariate retained included: indicator of state (California or Wisconsin), Title 1 status, ELA and math pretests, and years teaching; we assessed models with fixed pair effects and no pair effect). Standardized effect sizes for program impact ranged between 0.02 and 0.07, and none of the impact estimates reached statistical significance.

Next, we responded to the fact that there was lack of consensus among our advisors about the best approach to scaling the posttest. We evaluated impact with 24 approaches: 4 (scaling approaches, including percent correct, 1-PL, 2-PL and 3-PL score calibrations) × 3 (models with no covariates, pretests only, and the full set of covariates) × 2 (pair modeled as fixed or random). Standardized effect sizes for program impact ranged between -0.03 and 0.08, and none of the impact estimates reached statistical significance.

Next, we responded to the concern that the test included items that were not discriminating well with respect to the underlying ability scale. A factor analysis confirmed a single main dimension. We limited the fourth- and fifth-grade forms to items with factor loadings of .20 or higher. Limiting the item set increased the Cronbach's alphas from .69 to .72

[33] More detailed results are in Appendix M.

in fourth grade and from .56 to .60 in fifth grade. We conferred with the test developers to see if these items met different specifications or addressed distinct constructs than those with factor loadings below .20. The test developers were not able to identify any differentiating feature of these items. For the reduced set of items, we obtained percent-correct scores and re-examined impacts using the benchmark model and nine other sensitivity analyses described above. Standardized effect sizes ranged between -0.03 and 0.08, and none of the impact estimates reached statistical significance. (The full results of the sensitivity analyses are provided in Appendix M. A discussion about approaches to scaling and evaluation of test dimensionality is included in Appendix D.)

Confirmatory Impacts of Making Sense of SCIENCE on Student Science Achievement for Students in the Lowest Third of Incoming ELA and Math Achievement

Impact

In Table 25 and Table 26, we exhibit the main impact findings for student science achievement outcomes limited to students in the lowest third of incoming ELA achievement and for students in the lowest third of incoming math achievement.

For students in the low range of incoming ELA achievement, we observed a positive and statistically nonsignificant impact, with a standardized effect size of 0.07 (p = .567). For students in the low range of incoming math achievement, we observed a positive and statistically nonsignificant impact, with a standardized effect size of 0.22 (p = .099). (The full results from the impact models are provided in Appendix N.)

TABLE 25. RESULTS OF IMPACT OF MAKING SENSE OF SCIENCE ON SCIENCE ACHIEVEMENT FOR STUDENTS IN THE LOWEST THIRD OF INCOMING ELA ACHIEVEMENT

	Condition	Means	Standard deviations	No. of schools	No. of students	Effect size	p value	Change in percentile ranking
Unadjusted effect size[a]	Control	- 0.69	0.72	26	309	0.13	.180	5.1%
	MSS	- 0.61	0.75	29	406			
Adjusted effect size[b]	Control	- 0.69				0.07	.567	2.9%
	MSS	- 0.63						

Note. *MSS* stands for the group of students receiving the Making Sense of SCIENCE professional learning.

[a] The unadjusted effect size is the impact estimate from a model with pair fixed, and school random effects, and without covariates, divided by the pooled standard deviation of the outcome variable.

[b] The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

TABLE 26. RESULTS OF IMPACT OF MAKING SENSE OF SCIENCE ON SCIENCE ACHIEVEMENT FOR STUDENTS IN THE LOWEST THIRD OF INCOMING MATH ACHIEVEMENT

	Condition	Means	Standard deviations	No. of schools	No. of students	Effect size	*p* value	Change in percentile ranking
Unadjusted effect size[a]	Control	- 0.68	0.72	26	310	0.13	.224	5.0%
	MSS	- 0.60	0.76	29	403			
Adjusted effect size[b]	Control	- 0.68				0.22	.099	8.7%
	MSS	- 0.52						

Note. *MSS* stands for the group of students receiving the Making Sense of SCIENCE professional learning.

[a] The unadjusted effect size is the impact estimate from a model with pair fixed, and school random effects, and without covariates, divided by the pooled standard deviation of the outcome variable.

[b] The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

Sensitivity Analyses

As with the full sample, we explored several less parameterized models, starting with a model with no covariates and then progressively including covariates until we arrived at the benchmark model. For students in the lowest third of incoming ELA achievement, the standardized effect sizes ranged between 0.07 and 0.13, and none of the impact estimates reached statistical significance. For students in the lowest third of incoming math achievement, the standardized effect sizes ranged between 0.08 and 0.22, and none of the impact estimates reached statistical significance.

We also examined several different approaches to estimation (e.g., HLM but using ML instead of REML, and after including a teacher random effect). For students in the lowest third of incoming ELA achievement, the standardized effect sizes ranged between 0.03 and 0.07, and none of the impact estimates reached statistical significance. For students in the lowest third of incoming math achievement, the standardized effect sizes ranged between 0.07 and 0.20, and none of the impact estimates reached statistical significance.

We then reran the benchmark and all sensitivity analyses described above with the reduced-items forms (as described under the full sample analysis). For students in the lowest third in incoming ELA achievement, the standardized effect sizes ranged between 0.04 and 0.12, and none of the impact estimates reached statistical significance. For students in the lowest third of incoming math achievement, the standardized effect sizes ranged between 0.10 and 0.22, and none of the impact estimates reached statistical significance.

More complete results for all of the sensitivity results reported here are presented in Appendix O.

Exploratory Results: Moderator Analyses

For the sample of 2,140 students, we evaluated whether impacts varied by ELA and math pretests, state (WI versus CA), grade (fourth grade versus fifth grade), student English Language Learner (ELL) status, student eligibility for Free or Reduced-Price lunch (FRPL), and all moderators combined. Table 27 shows the results. Note that we do not show estimates of all main effects in the models, limiting them to just the treatment variable and the variable(s) for which we assess the corresponding interaction(s) with treatment.

The pooled standard deviation of the outcome distribution is 0.97 standard deviation units. Therefore, the estimates may be interpreted as approximately representing the change in impact, in standard deviation units of the outcome distribution, associated with a 1-unit increase in each moderator. All moderators are as coded zero or one, with the exception of the pretests, which are z-transformed; therefore, a one-unit increase in either pretest represents an increase of one standard deviation in pretest performance.

We observe that only the differential impact by student ELL status is marginally significant, with a reduction in impact of approximately 0.15 standard deviations (p = .073) associated with being an English Language Learner, based on the model with just the one interaction, and approximately .23 standard deviations (p = .017) for the model that includes all interactions simultaneously. As a reminder, these analyses are exploratory, and we have not performed multiple comparisons adjustments; therefore, we expect some effects to reach statistical significance by chance alone.

TABLE 27. DIFFERENTIAL IMPACTS OF MAKING SENSE OF SCIENCE ON STUDENT SCIENCE ACHIEVEMENT

	Model 1 Moderator is ELA pretest *n* = 2140 J = 55	Model 2 Moderator is Math pretest *n* = 2140 J = 55	Model 3 Moderator is State *n* = 2140 J = 55	Model 4 Moderator is Grade Level *n* = 2140 J = 55	Model 5 Moderator is ELL status *n* = 2109 J = 55	Model 6 Moderator is FRPL status *n* = 1,708 J = 48	Model 7 All moderators included *n* = 1,708 J = 48
Intercept	-0.068 (0.489) *p* = .180	-0.682 (0.489) *p* = .176	-1.085 (0.587) *p* = .078	-0.673 (0.484) *p* = .178	-0.701 (0.494) *p* = .169	-0.471 (0.542) *p* = .395	-0.701 (0.603) *p* = .261
Main Effect of ELA Pretest	0.327 (0.035) *p* < .001						0.321 (0.044) *p* < .001
Main Effect of Math Pretest		0.281 (0.034) *p* < .001					0.291 (0.044) *p* < .001
Main Effect of State (is California)			0.540 (0.453) *p* = .246				0.355 (0.461) *p* = .451
Main Effect of Grade is 4th				0.037 (0.065) *p* = .575			-0.077 (0.082) *p* = .348
Main Effects of Student is ELL					-0.046 (0.065) *p* = .491		0.007 (0.073) *p* = .930
Main Effect of student is FRPL-eligible						0.064 (0.073) *p* = .377	0.066 (0.073) *p* = .346
Treatment	0.059 (0.089) *p* = .520	0.056 (0.089) *p* = .535	0.158 (0.127) *p* = .229	0.128 (0.100) *p* = .214	0.106 (0.093) *p* = .265	0.023 (0.122) *p* = .851	0.118 (0.177) *p* = .515

TABLE 27. DIFFERENTIAL IMPACTS OF MAKING SENSE OF SCIENCE ON STUDENT SCIENCE ACHIEVEMENT

	Model 1 Moderator is ELA pretest	Model 2 Moderator is Math pretest	Model 3 Moderator is State	Model 4 Moderator is Grade Level	Model 5 Moderator is ELL status	Model 6 Moderator is FRPL status	Model 7 All moderators included
Additional impact for each +1 SD in ELA pretest achievement	-0.016 (0.037) $p = .662$						0.001 (0.059) $p = .990$
Additional impact for each +1 SD in Math pretest achievement		-0.028 (0.037) $p = .448$					-0.050 (0.058) $p = .387$
Additional impact (CA – WI)			-0.195 (0.186) $p = .306$				-0.069 (0.191) $p = .721$
Additional impact (4th grade – 5th grade)				-0.111 (0.082) $p = .174$			-0.021 (0.095) $p = .828$
Additional impact (ELL – non ELL)					-0.150 (0.084) $p = .073$		-0.230 (0.096) $p = .017$
Additional impact (FRPL eligible – FRPL non-Eligible)						0.025 (0.097) $p = .795$	0.024 (0.098) $p = .807$
Variance component							
School	.056 $p < .001$	.056 $p < .001$	.055 $p < .001$	.053 $p < .001$	.057 $p < .001$	.046 $p < .001$	.049 $p < .001$
Student	.546	.546	.546	.546	.546	.542	.541

Note. n = student sample size, J = school sample size, ELA = English Language Arts, ELL = English Language Learner, FRPL = Free or Reduced-price lunch

To focus on the main results, we do not report the main effects of other covariates that were used to increase precision, including fixed effects for matched pairs. Several analyses have fewer than 2,140 students and 55 schools indicating missing moderator data for outcomes analyzed. (Cases with missing values for a moderator in a given analysis were listwise deleted; dummy variable imputation was used for missing values of other covariates). Pretests were z-transformed and centered on their grand means with the estimate of the main effect representing the change in posttest corresponding to a one standard deviation increase in the pretest. The interactions with pretest represent the change in impact correspondingly to approximately a one standard deviation increase in the pretest. All other moderators were coded zero or one (0 for Wisconsin, 1 for California; 0 for fifth grade, 1 for fourth grade; 0 for non-English Language Learner, 1 for English Language Learner; 0 for FRPL non-eligible, 1 for FRPL eligible.) The main effects of these variables estimate the average change in achievement in going from the 0 to 1 coded subgroups. The moderating effect represents the added value of the impact of Making Sense of SCIENCE on science achievement in going from the 0 to 1 coded subgroups. Values in parentheses are the standard error for the estimates. The estimate of the treatment effect in a model with interactions should be interpreted as the impact when all moderator values are set to zero. (HLM does not report standard errors for the variance components.)

We also examined differential impacts at the student and school levels after group mean-centering the moderator variables on the school means. We included both the main and interaction effects for both the group mean-centered variables and the school means of those variables. The results tell us whether the impact varies with changes in the school average of the moderator variable, and depending on the individual status of the moderating characteristic. We did this for both pretests and variables indicating ELL and FRPL status.

Results are shown in Table 28. As an example of how to interpret the moderating effects of the variables, consider the two rows "Additional impact for each +1 SD in individual ELA Pretest Achievement" and "Additional impact for each +1 SD in School Average ELA Pretest Achievement". The estimate associated with the former effect is interpreted as a .182 standard deviation increase in impact associated with each one standard deviation increase in the ELA pretest, on average *among students within schools*. The effect is not statistically significant. The estimate associated with the latter effect is interpreted as a .161 standard deviation decrease in impact associated with each one standard deviation increase in the ELA pretest, on average *across schools*. This effect is not statistically significant. The former effect tells us about differential impact among students within schools. The latter effect tells us about differential impact among schools.

Overall, the results are similar to those above, with the only noteworthy interaction being between group mean-centered individual ELL status and treatment.

TABLE 28. DIFFERENTIAL IMPACTS OF MAKING SENSE OF SCIENCE ON STUDENT SCIENCE ACHIEVEMENT (WITH MODERATOR EFFECTS DECOMPOSED INTO WITHIN- AND BETWEEN-SCHOOL COMPONENTS)

	Model 1 Moderator is ELA pretest *n* = 2,140 J = 55	Model 2 Moderator is Math pretest *n* = 2,140 J = 55	Model 3 Moderator is ELL status *n* = 2,109 J = 55	Model 4 Moderator is FRPL eligibility *n* = 1,708 J = 48	Model 5 All moderators included *n* = 1,708 J = 48
Intercept	-0.536 (0.486) *p* = .283	-0.627 (0.506) *p* = .229	-0.803 (0.499) *p* = .122	0.194 (0.554) *p* = .730	0.203 (0.625) *p* = .751
Main Effect of ELA Pretest (GrpMC)	0.326 (0.037) *p* < .001				0.331 (0.045) *p* < .001
Main Effect of Math Pretest (GrpMC)		0.282 (0.035) *p* < .001			0.290 (0.044) *p* < .001
Main Effects of Student is ELL (GrpMC)			-0.034 (0.065) *p* = .602		0.009 (0.080) *p* = .913
Main Effect of FRPL-eligible (GrpMC)				0.077 (0.074) *p* = .301	0.078 (0.074) *p* = .294
Main Effect of School Average ELA Pretest Achievement	0.182 (0.199) *p* = .370				-0.741 (0.451) *p* = .129
Main Effect of School Average Math Pretest Achievement		0.174 (0.195) *p* = .381			0.864 (0.456) *p* = .085
Main Effect of School Average ELL			-0.834 (0.453) *p* = .080		-0.188 (0.478) *p* = .701

TABLE 28. DIFFERENTIAL IMPACTS OF MAKING SENSE OF SCIENCE ON STUDENT SCIENCE ACHIEVEMENT (WITH MODERATOR EFFECTS DECOMPOSED INTO WITHIN- AND BETWEEN-SCHOOL COMPONENTS)

	Model 1 Moderator is ELA pretest	Model 2 Moderator is Math pretest	Model 3 Moderator is ELL status	Model 4 Moderator is FRPL eligibility	Model 5 All moderators included
Main Effect of School Average FRPL-eligible				-1.374 (0.498) p = .013	-1.735 (0.549) p = .009
Main Effect of Treatment	0.028 (0.102) p = .789	0.056 (0.105) p = .601	0.020 (0.144) p = .891	0.115 (0.303) p = .709	-0.217 (0.277) p = .449
Additional impact for each +1 SD in individual ELA Pretest Achievement (GrpMC)	0.182 (0.199) p =.370				-0.010 (0.059) p = .860
Additional impact for each +1 SD in individual Math Pretest Achievement (GrpMC)		-0.028 (0.037) p =.447			-0.046 (0.058) p = .432
Additional impact (ELL – non ELL) (GrpMC)			-0.156 (0.086) p =.069		-0.223 (0.106) p = .041
Additional impact (FRPL eligible – FRPL non-Eligible) (GrpMC)				0.020 (0.100) p =.840	0.019 (0.100) p = .847
Additional impact for each +1 SD in School Average ELA Pretest Achievement	-0.161 (0.229) p = .488				0.365 (0.423) p = .407
Additional impact for each +1 SD in School Average Math Pretest Achievement		-0.061 (0.263) p = .818			-0.417 (0.493) p = .416
Additional impact associated with a change of 0% to 100% ELL students at the school			0.158 (0.431) p = .718		-0.051 (0.418) p = .905
Additional impact associated with a change of 0% to 100% FRPL-eligible students at the school				-0.122 (0.364) p = .742	0.354 (0.359) p = .345
Variance component					
School	.058 p < .001	.061 p < .001	.054 p < .001	.027 p = .001	.019 p < .001
Student	.544	.546	.547	.541	.538

Note. GrpMC = Group Mean Centered, n = student sample size, J = school sample size, ELA = English Language Arts, ELL = English Language Learner, FRPL = Free or Reduced-Price lunch

To focus on the main results, we do not report the main effects of other covariates that were used to increase precision, including fixed effects for matched pairs. Several analyses have fewer than 2,140 students and 55 schools indicating missing moderator data for outcomes analyzed. (Missing values for a moderator in a given analysis was listwise deleted; dummy variable imputation was used for missing values of other covariates.) Values in parentheses are the standard error for the estimates. The estimate of the treatment effect in a model with interactions should be interpreted as the impact when all moderator values are set to zero. (HLM does not report standard errors for the variance components.)

As a supplemental analysis, we also examined whether impacts on student science achievement varied by teacher leader status. Teacher leader status is not strictly a moderator, because for some teachers, the decision to accept the leadership role was made after random assignment. Regardless, we examined if impact varied depending on leadership status. The added-value impact for students of teacher leaders was 0.10 scale score units (p = .398).

Impacts on Specific Subsamples

Impacts on Science Sub-strands

We examined impacts for the (a) Earth and space science and (b) physical science sub-strands. To obtain sub-strand scores, we calibrated scores separately within each grade by sub-strand, then z-transformed resulting scores within grade and by sub-strand, and then combined sub-strand specific z-scores across grades.

The sample is the same as that of the full-sample benchmark impact analysis (n = 2,140). Impacts are shown in Table 29 and Table 30. We observe a positive and statistically nonsignificant impact for both the physical science sub-strand and the Earth and space science sub-strand. The impacts were 0.06 (p = .445) and 0.06 (p = .526) standardized effect sizes, respectively.

TABLE 29. RESULTS OF IMPACT OF MAKING SENSE OF SCIENCE ON THE PHYSICAL SCIENCE SUB-STRAND

	Condition	Means	Standard deviations	No. of schools	No. of students	Effect size	*p* value	Change in percentile ranking
Unadjusted effect size[a]	Control	0.00	1.00	26	1002	0.03	.667	1.3%
	MSS	-0.02	0.97	29	1138			
Adjusted effect size[b]	Control	0.00				0.06	.445	2.4%
	MSS	0.06						

Note. The *p* values are for the corresponding impact estimates in the benchmark impact model. *MSS* stands for the group of students whose teachers received the Making Sense of SCIENCE professional learning.

[a] The unadjusted effect size is the impact estimate from a model with pair fixed, and school random effects, and without covariates, divided by the pooled standard deviation of the outcome variable.

[b] The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

TABLE 30. RESULTS OF IMPACT OF MAKING SENSE OF SCIENCE ON THE EARTH AND SPACE SCIENCE SUB-STRAND

	Condition	Means	Standard deviations	No. of schools	No. of students	Effect size	*p* value	Change in percentile ranking
Unadjusted effect size[a]	Control	0.00	1.00	26	1002	-0.01	.892	-0.5%
	MSS	-0.09	0.95	29	1138			
Adjusted effect size[b]	Control	0.00				0.06	.526	2.4%
	MSS	0.06						

Note. The *p* values are for the corresponding impact estimates in the benchmark impact model. *MSS* stands for the group of students whose teachers received the Making Sense of SCIENCE professional learning.

[a] The unadjusted effect size is the impact estimate from a model with pair fixed, and school random effects, and without covariates, divided by the pooled standard deviation of the outcome variable.

[b] The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

Impacts on Focused Sample 1

Focused sample 1 includes students who were in classes of BRS study teachers who participated in the study in both 2016–17 and 2017–18, and for whom we obtained student posttest scores. The sample included 1,415 students (719 treatment, 696 control) who had both Grade 3 state ELA and math pretests, with 814 students from California and 601 students from Wisconsin.[34] We assessed impact on science achievement using the same approach as with the benchmark analyses; that is, with the same random and fixed effects, covariates, and scaling of the posttest.

In Table 31, we present the main impact findings for focused sample 1. We observe a positive and statistically nonsignificant impact, with a standardized effect size of 0.02 (p = .848).

[34] Sample descriptions and equivalence tests for the focused samples are reported in Appendix P.

TABLE 31. RESULTS OF IMPACT OF MAKING SENSE OF SCIENCE ON STUDENT SCIENCE ASSESSMENT FOR FOCUSED SAMPLE 1

	Condition	Means	Standard deviations	No. of schools	No. of students	Effect size	*p* value	Change in percentile ranking
Unadjusted effect size[a]	Control	0.06	1.03	23	696	-0.04	.752	-1.4%
	MSS	0.02	0.98	28	719			
Adjusted effect size[b]	Control	0.06				0.02	.848	1.0%
	MSS	0.09						

Note. The *p* values are for the corresponding impact estimates in the benchmark impact model. *MSS* stands for the group of students whose teachers received the Making Sense of SCIENCE professional learning.

[a]The unadjusted effect size is the impact estimate from a model with pair fixed, and school random effects, and without covariates, divided by the pooled standard deviation of the outcome variable.

[b]The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

Impacts on Focused Sample 2

Focused sample 2 includes students who were in classes of study teachers both in the 2016–17 school year (when they were in fourth grade) and in the 2017–18 school year (when they were in fifth grade). The sample included 340 students (167 *Making Sense of SCIENCE*, 173 control) who had both grade 3 state ELA and math pretests. Among the 340 students, 178 were from California, and 162 were from Wisconsin. We estimated impacts using a similar impact model as used with the benchmark analysis. The exception was that in many cases, we obtained values from just one school in each pair, therefore we dropped the pair effect altogether.

In Table 32, we present the main impact findings for focused sample 2. We observe a positive and statistically nonsignificant impact, with a standardized effect size of 0.12 (p = .564).

TABLE 32. RESULTS OF IMPACT OF MAKING SENSE OF SCIENCE ON STUDENT SCIENCE ASSESSMENT FOR FOCUSED SAMPLE 2

	Condition	Means	Standard deviations	No. of schools	No. of students	Effect size	*p* value	Change in percentile ranking
Unadjusted effect size[a]	Control	0.05	0.95	13	167	-0.01	.936	-0.6%
	MSS	0.01	0.95	18	173			
Adjusted effect size[b]	Control	0.05				0.12	.564	4.9%
	MSS	0.16						

Note. The *p* values are for the corresponding impact estimates in the benchmark impact model. *MSS* stands for the group of students whose teachers received the Making Sense of SCIENCE professional learning.

[a]The unadjusted effect size is the impact estimate from a model with pair fixed, and school random effects, and without covariates, divided by the pooled standard deviation of the outcome variable.

[b]The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

Impact by State

We examined the impact of Making Sense of SCIENCE on student science achievement by state. We divided the student sample used to assess the overall benchmark impact (n = 2,140) into a California sample with 1,303 students (722 *Making Sense of SCIENCE* and 581 control), and the Wisconsin sample with 837 students (416 *Making Sense of SCIENCE* and 421 control). We applied the benchmark impact model that we used with the full sample, but excluding the "dummy variable" for state.

Table 33 and Table 34 show the main impact findings by state. We observe a negative and statistically nonsignificant impact with a standardized effect size of -0.045 (p = 0.736) among California students, and a positive and statistically nonsignificant impact with a standardized effect size of 0.10 (p = 0.645) among Wisconsin students.

TABLE 33. RESULTS OF IMPACT OF MAKING SENSE OF SCIENCE ON STUDENT SCIENCE ACHIEVEMENT FOR CALIFORNIA

	Condition	Means	Standard deviations	No. of schools	No. of students	Effect size	*p* value	Change in percentile ranking
Unadjusted effect size[a]	Control	-0.03	0.97	14	581	-0.007	.946	-0.3%
	MSS	-0.06	0.95	16	722			
Adjusted effect size[b]	Control	-0.03				-0.045	.736	-1.8%
	MSS	-0.08						

Note. The *p* values are for the corresponding impact estimates in the benchmark impact model. *MSS* stands for the group of students whose teachers received the Making Sense of SCIENCE professional learning.

[a]The unadjusted effect size is the impact estimate from a model with pair fixed, and school random effects, and without covariates, divided by the pooled standard deviation of the outcome variable.

[b] The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

TABLE 34. RESULTS OF IMPACT OF MAKING SENSE OF SCIENCE ON STUDENT SCIENCE ACHIEVEMENT FOR WISCONSIN

	Condition	Means	Standard deviations	No. of schools	No. of students	Effect size	*p* value	Change in percentile ranking
Unadjusted effect size[a]	Control	0.05	1.04	12	421	0.02	.926	0.6%
	MSS	-0.09	0.93	13	416			
Adjusted effect size[b]	Control	0.05				0.10	.645	3.8%
	MSS	0.14						

Note. The *p* values are for the corresponding impact estimates in the benchmark impact model. *MSS* stands for the group of students whose teachers received the Making Sense of SCIENCE professional learning.

[a]The unadjusted effect size is the impact estimate from a model with pair fixed, and school random effects, and without covariates, divided by the pooled standard deviation of the outcome variable.

[b] The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

Impact by Grade

We examined the impact of Making Sense of SCIENCE on student science achievement by grade. We divided the student sample used to assess the overall benchmark impact (n = 2,140) into two samples: Grade 4 (n = 1,220; 611 *Making Sense of SCIENCE* and 609 control) and Grade 5 (n = 920; 527 *Making Sense of SCIENCE* and 393 control). On average, the fifth-grade students experienced greater exposure, as a subset of them will have also experienced Making Sense of SCIENCE in fourth grade. The fourth-grade students will have experienced Making Sense of SCIENCE for the first time in that grade. We applied the benchmark impact model used with the full sample, but excluding the "dummy variable" for grade.

Table 35 and Table 36 show the main impact findings by grade. We observe a positive and statistically non-significant impact both among grade 4 and grade 5 students. The impacts achieved standardized effect sizes of 0.02 (p = .893) and 0.06 (p = .713) among grade 4 and grade 5 students, respectively.

TABLE 35. RESULTS OF IMPACT OF MAKING SENSE OF SCIENCE ON STUDENT SCIENCE ACHIEVEMENT FOR GRADE 4

	Condition	Means	Standard deviations	No. of schools	No. of students	Effect size	*p* value	Change in percentile ranking
Unadjusted effect size[a]	Control	0.00	1.00	24	609	0.004	.970	0.2%
	MSS	-0.08	0.92	27	611			
Adjusted effect size[b]	Control	0.00				0.021	.893	0.8%
	MSS	0.020						

Note. The *p* values are for the corresponding impact estimates in the benchmark impact model. *MSS* stands for the group of students whose teachers received the Making Sense of SCIENCE professional learning.

[a] The unadjusted effect size is the impact estimate from a model with pair fixed, and school random effects, and without covariates, divided by the pooled standard deviation of the outcome variable.

[b] The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

TABLE 36. RESULTS OF IMPACT OF MAKING SENSE OF SCIENCE ON STUDENT SCIENCE ACHIEVEMENT FOR GRADE 5

	Condition	Means	Standard deviations	No. of schools	No. of students	Effect size	*p* value	Change in percentile ranking
Unadjusted effect size[a]	Control	0.00	1.00	21	393	0.02	.857	0.8%
	MSS	-0.06	0.96	28	527			
Adjusted effect size[b]	Control	0.00				0.06	.713	2.2%
	MSS	0.05						

Note. The *p* values are for the corresponding impact estimates in the benchmark impact model. *MSS* stands for the group of students whose teachers received the Making Sense of SCIENCE professional learning.

[a] The unadjusted effect size is the impact estimate from a model with pair fixed, and school random effects, and without covariates, divided by the pooled standard deviation of the outcome variable.

[b] The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

Differential Impacts Across Districts

To maintain the anonymity of districts, we report just the main findings, without reference to names of specific districts or any information that could identify districts. Based on the interaction between district fixed effects and treatment, we observe a statistically significant difference across districts in the impact of the intervention (p = .040 for Type-3 test of fixed effects). We observe a positive, but not statistically significant, difference in intervention impact between strongly implementing districts and the remaining districts, with a standardized effect size of 0.08 (p = .397). Limiting analysis to high implementing districts, we do not observe an impact of the program, with a standardized effect size of -0.03 (p = .735). We also conducted another analysis for which we removed pair fixed effects and school random effects, with the idea being that the goal of looking at impacts for the higher implementing districts is to assess the impact specifically for those districts and schools, and not to generalize to other similar schools within or beyond the study districts. With that analysis, limited to high-implementing districts, we observe a standardized effect size of 0.07 (p = .419).[35]

DISCUSSION

In this chapter, we present the results of the analysis of impact of Making Sense of SCIENCE on student science achievement as measured by the selected-response items of the science assessment. For confirmatory analysis on the full sample of students, we found a small and positive, but statistically nonsignificant, effect size of 0.064 (p = .494). For confirmatory impacts on the subsamples of students who were among the lowest third of incoming ELA and math achievement, we found effect sizes of 0.07 (p = .567) and 0.220 (p = .099), respectively. We note that the impact on the lowest third in incoming math achievement was marginally significant.

For exploratory analyses, we did not observe impacts when employing the focused sample of students of BRS teachers or the sample of students who received full exposure by being in a Making Sense of SCIENCE teacher's classroom in both years. For moderator analyses (to assess whether impacts varied by student pretest performance, state, grade, ELL designations, and FRPL-eligibility), we observed that only the differential impact by ELL status was either marginally significant or significant, depending on the model. The remaining moderators were not significant. In exploring whether impacts varied by whether the district was considered to be a strong-implementing district, we observed a positive, but nonsignificant, differential impact favoring strong-implementing districts. We acknowledge that these analyses were exploratory, and we have not performed multiple comparisons adjustments; therefore, we expect some effects to reach statistical significance by chance alone.

[35] Researchers also began exploring the relationship between fidelity of implementation and impact using a matching approach as discused by Unlu et al. (2013). However, the analysis was fundamentally limited by the availability of the data. We discuss the approach and our attempts to apply it in this study in Appendix Q.

We take a pause here to remind readers about challenges we faced with the measure for student science achievement given the landscape of NGSS-aligned assessments.[36] After extensive efforts to locate an appropriate established assessment failed, evaluators took on the work of developing an assessment knowing full well its challenges and limitations. The team pulled from established sources such as the Massachusetts Comprehensive Assessment System (MCAS) and the National Assessment of Educational Progress (NAEP), recognizing they were based on standards that preceded NGSS. Operationally the assessment turned out to be difficult especially for student with low incoming achievement (see Tables D4 and D5 and discussion in Appendix D).

The experience of researchers and the results of this study have underscored the need for continuing development of NGSS-aligned assessments. As advised by the National Research Council in their guide on developing assessments for NGSS (2014), such assessments would require a reconceptualization of the assessment—including but not limited to integrating multicomponent tasks and performance-based questions, and leveraging the use of matrix-sampling designs to ensure that the NGSS performance expectations are appropriately covered in both depth and breadth.

[36] See Appendix D for a full description of the search for and construction of an NGSS-aligned assessment.

Chapter 7. Exploratory Impacts on Student Achievement on State Assessments

INTRODUCTION

In this section, we report findings about the impact of Making Sense of SCIENCE on student achievement as measured by state assessments in ELA, math, and science. The exploratory research questions are:

- What is the impact of Making Sense of SCIENCE, after two years of implementation, on the student *ELA state assessment* when compared to study participants in control schools receiving the business-as-usual science professional learning?
- What is the impact of Making Sense of SCIENCE, after two years of implementation, on the student *math state assessment* when compared to study participants in control schools receiving the business-as-usual science professional learning?
- What is the impact of Making Sense of SCIENCE, after two years of implementation, on the student *science state assessment* when compared to study participants in control schools receiving the business-as-usual science professional learning?

This chapter includes three sections. The first section on methods provides a brief description of the available spring 2017–18 state assessments in California and Wisconsin. We also include a summary of the analytic sample and our approach to analysis. The second section presents the findings for each assessment. Finally, in the third section, we provide a discussion of those findings.

METHODS

Measure

California administered the Smarter Balanced Summative Assessments (comprehensive, end-of-year assessments for ELA and math) that are aligned with the Common Core State Standards (CCSS) and measure progress toward college and career readiness. The tests are computer adaptive and administered to students in grades 3–8 (California Department of Education, 2019).

Wisconsin administered the Wisconsin Forward Exam, which is designed to measure how well students are doing relative to the Wisconsin Academic Standards. The tests are administered to grades 3–8 in ELA and math and grades 4 and 8 in science (and grades 4, 8, and 10 in social studies) (Wisconsin Department of Public Instruction, n.d.).

The Pearson correlation coefficients between the state science test and the main science assessment used to evaluate impact on student achievement were .68, .68, .69, and .69 for the percent-correct, 1-PL-, 2-PL-, and 3-PL-based scores on the evaluator-designed assessment, respectively.

Sample

ELA and math were tested in Grades 4 and 5 in both states. The ELA sample included 2,108 students (1,128 *Making Sense of SCIENCE*, 980 control) who had ELA outcomes from spring 2017–18. The math sample included 2,108 students (1,128 *Making Sense of SCIENCE*, 980 control) who had math outcomes from spring 2017–18. Science state assessment outcomes were only available for fourth-grade students in Wisconsin. Therefore, the sample included 479 fourth grade students

(251 *Making Sense of SCIENCE*, 228 control). Table 37 and Table 38 include the samples used to evaluate the impacts in the three subject areas.

TABLE 37. ANALYTIC SAMPLE FOR ANALYSIS OF IMPACT OF MAKING SENSE OF SCIENCE ON STUDENT ELA AND MATH STATE ASSESSMENTS

	Count of schools in CA	Count of students with posttest in CA	Count of schools in WI	Count of students with posttest in WI
MSS	16	715	13	413
Control	14	578	11	402
Total N	30	1293	24	815

Note. MSS stands for the group of students whose teachers received the Making Sense of SCIENCE professional learning.

TABLE 38. ANALYTIC SAMPLE FOR ANALYSIS OF IMPACT OF MAKING SENSE OF SCIENCE ON STUDENT SCIENCE STATE ASSESSMENTS

	Count of schools in CA	Count of students with posttest in CA	Count of schools in WI	Count of students with posttest in WI
MSS	0	0	12	251
Control	0	0	10	228
Total N	0	0	22	479

Note. MSS stands for the group of students whose teachers received the Making Sense of SCIENCE professional learning.

We tested baseline equivalence for the ELA and math sample and for the science sample. Results are in Table 39. We observed that for the ELA and math sample, and for the science sample, baseline equivalence is achieved.

TABLE 39. TESTS OF BASELINE EQUIVALENCE ON PRETESTS BETWEEN MAKING SENSE OF SCIENCE AND CONTROL FOR THE ELA, MATH, AND SCIENCE SAMPLES

	ELA and math sample		Science sample	
	ELA pretest	Math pretest	ELA pretest	Math pretest
***N* (schools)**	54		22	
***n* (students)**	2108		479	
Point estimate for difference between conditions	-0.09	-0.04	0.04	0.10
Standard error	0.07	0.07	0.13	0.13
***p* value**	.184	.547	.74	.455
Standardized effect size	-0.09	-0.05	0.05	0.11

Impact Analysis: Scaling and Impact Models

We evaluated impact on state assessment outcomes after two years of program implementation. We studied students' outcomes in three subjects: ELA, math, and science. ELA and math outcomes were accessible in both California and Wisconsin. We rescaled ELA and math outcomes within state and grade to make their scores comparable. Science outcomes were only available in Wisconsin. We estimated impact using a hierarchical linear model similar in form to the benchmark impact model used to assess confirmatory impacts. We handled missing values for covariates (other than the pretest) using dummy variable imputation. We removed cases without a pretest or a posttest.

FINDINGS

In Table 40, Table 41, and Table 42, we show the main impact findings for the ELA, math, and science state assessment outcomes. For ELA, we observe a positive and marginally statistically significant impact, with a standardized effect size of 0.09 (p = .057). For the math state assessment outcomes, we observe a negative but not statistically significant impact, with a standardized effect size of -0.02 (p = .700). For the science state assessment outcome in Wisconsin, we observe a positive but not statistically significant impact, with a standardized effect size of 0.03 (p = .818).

TABLE 40. RESULTS OF IMPACT ANALYSIS FOR THE ELA STATE ASSESSMENT OUTCOMES

	Condition	Means	Standard deviations	No. of schools	No. of students	Effect size	p value	Change in percentile ranking
Unadjusted effect size[a]	Control	-0.00	1.00	25	980	0.00	.948	0.2%
	MSS	-0.11	0.98	29	1128			
Adjusted effect size[b]	Control	-0.00				0.09	.057	3.6%
	MSS	0.09						

Note. *MSS* stands for the group who received the Making Sense of SCIENCE professional learning. The p values are for the corresponding impact estimates in the benchmark impact model.

[a] The unadjusted effect size is the impact estimate from a model with pair fixed, and school random effects, and without covariates, divided by the pooled standard deviation of the outcome variable.

[b] The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

TABLE 41. RESULTS OF IMPACT ANALYSIS FOR THE MATH STATE ASSESSMENT OUTCOMES

	Condition	Means	Standard deviations[a]	No. of schools	No. of students	Effect size	p value	Change in percentile ranking
Unadjusted effect size[a]	Control	-0.00	1.00	25	980	-0.05	.467	-2.0%
	MSS	-0.12	1.02	29	1128			

TABLE 41. RESULTS OF IMPACT ANALYSIS FOR THE MATH STATE ASSESSMENT OUTCOMES

	Condition	Means	Standard deviations[a]	No. of schools	No. of students	Effect size	*p* value	Change in percentile ranking
Adjusted effect size[b]	Control	-0.00				-0.02	.700	-0.8%
	MSS	-0.02						

Note. *MSS* stands for the group who received the Making Sense of SCIENCE professional learning. The *p* values are for the corresponding impact estimates in the benchmark impact model.

[a] The unadjusted effect size is the impact estimate from a model with pair fixed, and school random effects, and without covariates, divided by the pooled standard deviation of the outcome variable.

[b] The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

TABLE 42. RESULTS OF IMPACT ANALYSIS FOR THE SCIENCE STATE ASSESSMENT OUTCOMES

	Condition	Means	Standard deviations[a]	No. of schools	No. of students	Effect size	*p* value	Change in percentile ranking
Unadjusted effect size[a]	Control	391.70	59.70	10	228	0.07	.614	2.7%
	MSS	372.40	49.17	12	251			
Adjusted effect size[b]	Control	391.70				0.03	.818	1.1%
	MSS	393.20						

Note. *MSS* stands for the group who received the Making Sense of SCIENCE professional learning. The *p* values are for the corresponding impact estimates in the benchmark impact model.

[a] The unadjusted effect size is the impact estimate from a model with pair fixed, and school random effects, and without covariates, divided by the pooled standard deviation of the outcome variable.

[b] The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

DISCUSSION

This chapter explores the impact of Making Sense of SCIENCE on the state assessments in ELA, math, and science. We found a marginally significant positive effect of Making Sense of SCIENCE in ELA (ES = 0.09, $p = .057$), but no impact on the math or science state assessments.

While ELA and math were assessed in both states and in both study grades, science was only assessed in Wisconsin in fourth grade during the study years. California was piloting and field-testing a new NGSS-aligned state science assessment, and the scores were not available. This resulted in a reduced sample of students with science assessment data (in only one grade in one state). Additionally, the state assessments focus on general academic knowledge in the three content areas; they measure more distal outcomes of the intervention than the selected response portion of the assessment used to address the confirmatory research question related to student science achievement. We recognize, however, the importance of these high-stakes assessments for state and local practitioners. Given that the focus of the professional

learning on integrating literacy into science instruction includes ways to support classroom practices and strengthen students' abilities to write, read, and discuss about science, the effect of Making Sense of SCIENCE on ELA is promising.

Chapter 8. Impacts on Student Communication of Science Ideas in Writing

INTRODUCTION

In this chapter, we report the results of the following exploratory research question: What is the impact of Making Sense of SCIENCE, after two years of implementation, on students' communication of science ideas in writing when compared to study participants in control schools receiving the business-as-usual science professional learning?

This chapter includes three sections. In the first section on methods, we provide a brief description of the constructed-response items that students responded to in spring 2017–18. We also provide a summary concerning the sample, designs, and analysis used to address the question. In the second section, we report the main impact finding, followed by results from several additional impact models conducted to evaluate the robustness of the estimates. We also summarize the moderator analysis and item-specific impacts. Finally, in the third section, we provide a brief discussion of the results.

METHODS

Measure

As mentioned above, in spring 2017–18, the study administered a student assessment that included selected-response items, constructed-response items, and student survey scales. The analysis in this chapter focuses on the constructed-response items, which were designed to assess student communication of science ideas in writing. There were eight constructed-response items: six items were drawn from NAEP, and the remaining two were developed by HRA to address necessary specifications not covered by the NAEP items. HRA used the rubrics provided on the NAEP website for scoring the NAEP items and iteratively developed rubrics for the other two items (for more detailed information on item and rubric development, see Wong et al., 2020). Four of eight items were appropriate for and administered in fourth and fifth grades. The remaining four items were administered in fifth grade only. The interrater reliability (that is, percent agreement between scorers) ranged from 80.7% to 96.3%, with a median value of 92.7% (Wong et al., 2020). The full details concerning test development and scoring is provided in the companion report to this one by HRA (Wong et al., 2020).

Sample and Baseline Equivalence

The sample for this analysis consisted of 943 students in 60 schools (approximately 25% of the student sample) who were randomly assigned to one of eight assessment forms. Each of the eight forms drew on four of eight constructed-response items. After limiting the sample to students who had at least one valid response and with non-missing ELA and math pretests, we arrived at our analytic sample of 728 students. This sample would allow us to assess baseline equivalence on the ELA and math pretests. We found no difference between the *Making Sense of SCIENCE* group and the control group on either pretest. The final sample size associated with the analysis for each item reflects the number of students—with both pretests—who responded to the item across all forms on which the item appeared. Table 43 includes the achieved analytic sample sizes at the level of random assignment (school) and at the individual student level with tests of baseline equivalence for ELA and math pretests.

TABLE 43. ANALYTIC SAMPLE SIZE AND BASELINE EQUIVALENCE ON PRETESTS BETWEEN MAKING SENSE OF SCIENCE AND CONTROL FOR STUDENT COMMUNICATION OF SCIENCE IDEAS IN WRITING

	ELA pretest	Math pretest
***N* (Schools)**	56	
***n* (Students)**	728 (404 *MSS* + 324 control)	
Point estimate	-0.05	0.09
Standard error	0.07	0.08
***p* value**	.499	.279
Standardized effect size	-0.05	0.10

Note. *MSS* represents the group of students whose teachers received the Making Sense of SCIENCE professional learning.

Scale Scores and Impact Models

HRA, our evaluation partner in this work, conducted the scoring. The process is described in Wong et al. (2020). Empirical received scores in numeric or letter codes. Items with letter codes included some or all of the following response options: Blank, Unsatisfactory/Incorrect, Partial, Essential, Satisfactory, and Complete. Blank and Unsatisfactory/Incorrect became 0. In these analyses, we treated blanks as incorrect, unless they were trailing blanks (i.e., an uninterrupted sequence of missing responses at the end of the test), which we treated as missing. The lowest category of response for an item became 0, and the highest became 1. We then established numeric intermediate values at equal intervals between 0 and 1, depending on the number of response options.

Models for Assessing Impacts on the Average Score Across All Items

After transforming the scores as described above, we calculated a score for each person by averaging their item scores. We estimated the impacts of Making Sense of SCIENCE on communication of science ideas in writing using a hierarchical model with the same covariates and random effects as the one we used to evaluate impacts on science achievement outcomes with selected-response items (the main confirmatory analysis is described in Chapter 6). The exception was that we also included dummy variables to indicate form. Although we achieved balance on forms across conditions ($p = 0.61$), we included form indicators as covariates to obtain additional precision in estimation.

In addition to the benchmark model, we examined the stability of impact estimates by adding specific sets of covariates to progressively "build up" models. We started with a model with no covariates, and then we sequentially added indicators for matched pairs, dummy variables for forms, ELA and math pretests, and finally the full set of covariates.

Moderator Analyses

We also examined the moderating effects of students' ELA pretest scores and students' ELL status on the impact. The rationale was that success on the performance tasks may have required specific levels of literacy skills to understand the

concepts and handle the demands of the task. Therefore, we evaluated the hypothesis that greater impact would be demonstrated for students with stronger literacy skills in English.

Impacts per Item

To provide more-detailed and formative feedback to program developers, we also examined impacts separately for each of the eight constructed-response items. This would give the program developers more-specific information about which science performance tasks Making Sense of SCIENCE has greater impact on. We used two approaches to assessing impacts: (1) an approach that assumes an equal interval scale and using hierarchical linear models, and (2) an approach that models responses on an ordinal scale and uses a cumulative logistic hierarchical (non-linear) regression model. We estimated impacts using a hierarchical model like the one we used in the main cross-item impact analysis described above. The effect size is the linear model's regression-adjusted impact estimate divided by the standard deviation of the outcome variable assessed at the student-level for the control group. Additionally, we calculated the Cox Index using the regression-adjusted impact estimate from the cumulative logistic regression. (We acknowledge the potential for false positives in this analysis from not adjusting results for multiple comparisons).

FINDINGS

Impacts on Communication of Science Ideas in Writing for all Constructed-Response Items Combined

In Table 44, we exhibit the main impact finding. We observe a positive but not statistically significant impact, with a standardized effect size of 0.116 (p = .177).

TABLE 44. RESULTS OF IMPACT OF MAKING SENSE OF SCIENCE ON STUDENT COMMUNICATION OF SCIENCE IDEAS IN WRITING FOR THE FULL SAMPLE (n = 728)

	Condition	Means	Standard deviations[a]	No. of schools	No. of students	Effect size	p value	Change in percentile ranking
Unadjusted effect size[a]	control	0.31	0.23	26	324	0.004	.287	0.3%
	MSS	0.31	0.22	30	404			
Adjusted effect size[b]	control	0.31				0.116	.177	4.6%
	MSS	0.33						

Note. *MSS* stands for the group who received the Making Sense of SCIENCE professional learning.

[a]The unadjusted effect size is the impact estimate from a model with pair fixed, and school random effects, and without covariates, divided by the pooled standard deviation of the outcome variable.

[b]The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

The p values are for the corresponding impact estimates in the benchmark impact model.

Table 45 shows results of additional impact models, where we sequentially added covariates, with the fifth and final model representing the benchmark model, results of which were expanded on in Table 44 above.

TABLE 45. ADDITIONAL MODELS OF IMPACT (*N* = 728)

Fixed Effects	Model 1	Model 2	Model 3	Model 4	Model 5 [a]
Treatment	0.005 (0.026) *p* =.860	0.019 (0.017) *p* =.287	0.018 (0.018) *p* =.326	0.018 (0.014) *p* =.208	0.026 (0.019) *p* =.177
Matched pairs		X	X	X	X
Forms			X	X	X
Pretests (ELA and math)				X	X
Other covariates					X
Standardized effect size	0.02	0.08	0.08	0.08	0.12
Random effects					
School [b]	.005 (0.002) *p* = .002	.0001 (0.0007) *p* = .446	.0005 (0.0008) *p* = .249	0 [c]	0 [c]
Student	.046 (.002) *p* <.001	.046 (.002) *p* <.001	.044 (.002) *p* <.001	.031 (.002) *p* <.001	.028 (.001) *p* <.001

Note. Quantities in parentheses are standard errors.

[a] Model 5 is the benchmark model

[b] The school is the unit randomized.

[c] Zero effect estimate with no *p* value indicates that estimation met boundary condition for this quantity. This often indicates that the quantity is trivially different from zero (Singer & Willett, 2003). We include the school effect in the model given that it is the unit of randomization.

Moderator Analysis

We did not observe a differential impact by level of incoming ELA achievement, with a value-added impact of 0.015 scale score units for each one-unit increase on the pretest (p = .284). We did not observe a differential impact by ELL status, with a value-added impact for non-English Language Proficient students of 0.008 scale score units (p = .839).[37]

[37] Among the sample of 728 students, 522 had information about ELL status and were included in analysis.

Impacts by Item

We also examined impacts on each of the eight constructed-response items individually, with results in Table 46. Items 1 and 2 were developed by HRA; the other items are from NAEP. The main results are based on the linear model, with Cox indices reported in the table note. We observed a positive impact only for Item 2 (Basketball), which was completed by fifth-grade students.

TABLE 46. RESULTS OF ITEM-SPECIFIC ANALYSIS OF THE CONSTRUCTED RESPONSE QUESTIONS

Item	Number of student respondents, overall and by condition	Number of schools, overall and by condition	Grade level of respondents	Standardized effect size
1	449 (239 *MSS*, 210 control)	56 (30 *MSS* vs. 26 control)	4, 5	0.13
2	266 (160 *MSS*, 106 control)	48 (27 *MSS* vs. 21 control)	5	0.42***
3	427 (232 *MSS*, 195 control)	56 (30 *MSS* vs. 26 control)	4, 5	0.05
4	638 (354 *MSS*, 284 control)	55 (29 *MSS* vs. 26 control)	4, 5	0.10
5	260 (153 *MSS*, 107 control)	49 (28 *MSS* vs. 21 control)	5	0.02
6	187(106 *MSS*, 81 control)	49 (28 *MSS* vs. 21 control)	5	0.23
7	195(112 *MSS*, 83 control)	49 (28 *MSS* vs. 21 control)	5	-0.05
8	378 (207 *MSS*, 171 control)	56 (30 *MSS* vs. 26 control)	4, 5	-0.05

Note. *MSS* stands for the group who received the Making Sense of SCIENCE professional learning.

Items were (1) Sandstone, (2) Basketball, (3) Bird Food, (4) Cloudy Days, (5) Pea Seeds, (6) Boiling Water, (7) Ice Cube, and (8) Mineral Scratch. HRA developed items 1 (Sandstone) and 2 (Basketball). The other items were from NAEP.

The Cox Index associated with impact based on cumulative logistic regressions (with the same covariates as used with the linear regression) for the eight items are as follows, respectively: 0.198, 0.501, 0.109, 0.194, 0.244, 1.000, -0.162, -0.049.

None of the items assessed in both grades showed a differential impact across grades: Item 1 ($p = .960$), Item 3 ($p = .397$), Item 4 ($p = .711$), Item 8 ($p = .265$).

* $p \leq .10$, ** $p \leq .05$, *** $p \leq .01$, **** $p \leq .001$

Detailed results of item specific analyses are in Appendix R.

DISCUSSION

In order to assess the impact of Making Sense of SCIENCE on students' communication of science ideas in writing, we administered a set of constructed-response assessment items to a random subset of students. Of the eight items, we selected six from NAEP that were aligned to at least one of the NGSS Science and Engineering Practices. HRA developed the two additional items (Sandstone and Basketball) to represent the multiple dimensions of NGSS (Disciplinary Core Ideas, Science and Engineering Practices, and Crosscutting Concepts). We found a positive, but not statistically significant, impact across the set of items (ES = 0.12, $p = .177$). In the item-specific analysis, we found a large positive impact on the Basketball item (ES = 0.42, $p < .001$). HRA offers several considerations for these findings (Wong et al., 2020). First, they acknowledge that all of the items were difficult for students (five of the six NAEP items were rated "hard," and the HRA-developed items were also challenging). Additionally, they unpack some of the limitations of the scoring of the NAEP items specifically in their ability to capture differences in students' conceptual understanding. Finally, the HRA-

developed items were designed to be NGSS-aligned. HRA concludes that the Basketball item and rubric were particularly sensitive to students' scientific reasoning and that "the type of reasoning elicited in the Basketball item is supported by both strong conceptual knowledge that allows one to make connections between ideas and by strong discourse skills, both of which were emphasized in the Making Sense of SCIENCE professional learning" (Wong et al., 2020).

Chapter 9. Impacts on Student Non-Academic Outcomes

INTRODUCTION

This section reports the impacts of Making Sense of SCIENCE on student non-academic outcomes, as measured by the student survey administered in spring of Year 2 (2017–18). The analysis is exploratory and addresses the following research question: What is the impact of Making Sense of SCIENCE, after two years of implementation, on student non-academic outcomes when compared to study participants in control schools receiving the business-as-usual science professional learning?

As in previous chapters, this chapter includes three sections. The methods section includes a brief description of the survey items, the sample, and our approach to analysis. We then present the findings, followed by a brief discussion.

METHODS

Measure

To assess non-academic outcomes for students, we administered survey scales to students in spring of 2017–18, along with the science assessment. The outcomes were classified into one of two types. The first set of outcomes measured student perception of opportunities to learn in the classroom, such as the types of activities and level of cognitive demand in the science classroom, as well as the quality of the science class in regard to learning environment and science instruction. The second set of outcomes were related to the students' own personal attitudes and beliefs, such as self-efficacy, agency, aspirations, and enjoyment of science.

The survey scales appeared first—before the assessment items—to prevent students' reactions to the assessment from affecting their responses to the survey items. Recall that for the student science assessment, we randomly assigned students to the four test forms in each grade. Each of the four test forms included two survey scales (eight scales in total; two per form). The scales are described in Table 47.

TABLE 47. DESCRIPTION OF STUDENT SURVEY ITEMS

Construct	Form	Content	Description	Cronbach's alpha
Outcomes related to OTL				
Quality of the Science Class (Science Instruction)	3	Teacher use of classroom instruction and learning strategies to promote legitimate science learning experiences	7 items (5-point scale: Strongly Disagree to Strongly Agree)	.63
Quality of Science Class (Learning Environment / Classroom Management)	1	Classroom norms conducive to learning science	12 items (5-point scale: Almost Never to Almost Always)	.81
Cognitive Demand	4	The extent to which students are engaged in thought-provoking science activities	5 items (5-point scale: Almost Never to Almost Always)	.58

Activities in Science Classrooms	2	Partaking in inquiry-based activities during science classes	7 items (5-point scale: Never to Always)	.60
Students' attitudes and beliefs				
Aspirations	1	Expectations for use of science in the future (e.g., in one's job)	4 items (5-point scale: Strongly Disagree to Strongly Agree)	.58
Self-Efficacy	2	Sense of self-efficacy in learning and being adept at science	7 items (5-point scale: Strongly Disagree to Strongly Agree)	.76
Agency in Learning	3	The extent to which students are active participants and agents in their science learning	6 items (5-point scale: Almost Never to Almost Always)	.11
Enjoyment of Science	4	Enjoyment and engagement in science	9 items (5-point scale: Strongly Agree to Strongly Disagree	.87

Sample

For the exploratory analysis of impacts on student non-academic outcomes, we assessed impacts on a sample almost exactly parallel to the sample of students used to evaluate confirmatory impacts on student science achievement (n = 2,140).

Impact Analysis: Scaling and Impact Models

We evaluated impact on student non-academic outcomes after two years of program implementation. Impacts on these outcomes were estimated using the same model that we used to estimate impact on student science achievement. We used a hierarchical linear model with fixed block (pair) effects and school- and student-level random effects. Covariates included state pretests in ELA and math and grade level, along with student-, teacher-, and school-level covariates included in the benchmark model used to assess the impact on students. We handled missing values for covariates (other than the pretest) using dummy variable imputation. We removed cases without a pretest or a survey score.

FINDINGS

Standardized effect sizes across the eight scales range from -0.23 to 0.14 (Table 48), though there is only one statistically significant and negative impact on *Quality of Science Class (Science Instruction)* (p = .030). We do not apply adjustments for multiple comparisons to the results, given the exploratory nature of the analyses. With eight contrasts, there is a high probability that one or more will reach statistical significance by chance alone. Therefore, we caution against over interpreting the meaning of the effect from the one statistically significant result.

TABLE 48. IMPACTS OF MAKING SENSE OF SCIENCE ON STUDENT NON-ACADEMIC OUTCOMES

	n	Point estimate	SE	df	t	*p* value	Standardized effect size
Outcomes related to OTL							
Quality of the Science Class (Science Instruction)	553	-0.12	0.05	24	-2.30	.030	-0.23
Quality of Science Class (Learning Environment / Classroom Management)	534	0.09	0.06	24	1.48	.152	0.14
Cognitive Demand	557	-0.05	0.08	25	-0.06	.575	-0.06
Activities in Science Classrooms	539	0.04	0.05	24	0.88	.388	0.06
Student attitudes and beliefs							
Aspirations	537	-0.04	0.06	24	-0.62	.543	-0.05
Self-Efficacy	558	-0.05	0.05	24	-0.91	.371	-0.07
Agency in Learning	540	-0.04	0.04	24	-0.90	.378	-0.07
Enjoyment of Science[a]	558	-0.02	0.07	25	-0.32	.752	-0.03

[a] The distribution of responses to "Enjoyment of Science" displayed a high level of skew (-1.01 in control and -0.88 in treatment), and we used a square transformation to reduce skew. With the transformation, the impact remained not statistically significant (p = .589).

DISCUSSION

Interestingly, we noticed that the impact results related to opportunities to learn that were reported by students were at odds with our findings for opportunities to learn that were reported by teachers (see Chapter 5), for which we did observe either significant or marginally significant positive impacts. We offer two possible interpretations for the overall lack of impact on student non-academic outcomes. The first—similar to results of impacts on outcomes measured by the teacher surveys—is the limitations with self-reported measures, such as social desirability bias or concern about anonymity of responses. A second possible contributor to the null findings may be that we analyzed the survey scales in item-clusters, as they were administered. That is, we did not conduct additional factor analysis or other item analysis. Future research could include a refinement of the scales, including ensuring their face and construct validity.

Chapter 10: Discussion & Conclusion

SUMMARY OF THE FINDINGS

Findings from this i3 evaluation of Making Sense of SCIENCE provide suggestive evidence of the model's effectiveness in transforming teaching and learning. Among teachers who were active in the study for both years, Making Sense of SCIENCE teachers outperformed the control teachers in teacher content knowledge and in pedagogical content knowledge. Making Sense of SCIENCE teachers also reported spending more time on science instruction and placing greater emphasis on NGSS-aligned instructional practices, such as *Participating in Collaborative Discourse*, *Sense-making of Hands-on Investigations,* and *Integration of Science Literacy.* Additionally, they reported having a greater sense of *Agency in the Classroom* and showed a marginally significant impact on *Confidence in Science Instructional Practices.* They also reported collaborating with each other more often, and notably, beyond the time they already spent in PLC meetings as part of Making Sense of SCIENCE. As for school climate, their self-reports yielded a statistically significant impact on *Administrators' Support of Teacher Collaboration* and a marginally significant impact on *Administrators Involving Teachers in Science Leadership*. These findings are important in that they show improvements in aspects that are linked to greater teacher effectiveness.

Areas where we did not observe impact were those related to long-standing, deeply-ingrained beliefs that we hypothesize are difficult to change or may take longer to observe impact. These include teachers' attitudes and beliefs, such as valuing being a reflective learner or believing that students are capable learners. We also did not observe impacts on outcomes related to student opportunities to learn as measured by exposure to NGSS-aligned content, which we discuss further below.

Other outcomes where we did not see change were those related to trust and respect among teachers and between teachers and administrators; these factors seem to be affected by a myriad of forces beyond professional learning in science for a subgroup of teachers at the school and two days of workshops for administrators. The difficulty in moving the needle on school climate is demonstrated in other studies, even those with a more direct focus or greater emphasis on professional development for administrators (Jacob et al., 2014; Hermann et al., 2019).

In this study, Making Sense of SCIENCE did not show statistically significant impacts on student science achievement, as measured by the researcher-developed assessment. However, we did observe a marginally significant effect on the ELA state assessment, with an effect size of 0.09 (p = .057). This is an encouraging finding given that Making Sense of SCIENCE aims to make connections between science and literacy. Moreover, with the exception of just two results,[38] effect sizes of impacts on students for multiple samples (main sample, focused samples 1 and 2, and samples of the lowest third in incoming ELA and math achievement), and across different measures (communication of science ideas in writing, and state ELA, math, and science assessments) were consistently small, positive effects, albeit not statistically significant—a thread that runs true throughout the study.

[38] Impact on student science achievement for the California sub-sample (ES = -0.05, p = .736) and impact on the math state standardized assessments using the full sample (ES = -0.02, p = .700)

In the following section, we offer a few perspectives to contextualize these findings within the existing literature about Making Sense of SCIENCE in particular, and about teacher and leadership professional learning more generally. We also offer working hypotheses about why impacts on more proximal outcomes did not translate into statistically significant impacts on student outcomes. We conclude with reflections about tensions and lessons learned from this study.

SITUATING THE FINDINGS IN THE LITERATURE

Three studies of Making Sense of SCIENCE in the last decade—two of which met WWC group design standards without reservations under WWC review standards 3.0 (Heller et al., 2012; Heller, 2012) and a third (Heller et al., 2017) not yet reviewed by WWC—have all shown statistically significant or marginally significant impacts on teacher content knowledge with effect sizes of 1.8 ($p < .001$), 0.38 ($p < .01$), and 0.17 ($p = .09$), respectively.

The same three studies yielded positive, though not all statistically significant, results on student science achievement outcomes. The study of elementary teachers and students yielded effect sizes ranging from 0.37 to 0.60 (all $p < .001$) (Heller et al., 2012). The study of middle school teachers and students produced effect sizes of 0.11 ($p = .04$) for the full sample and 0.31 ($p = .04$) for the subset of English Language learners, though these results were no longer statistically significant after adjusting for multiple comparisons (Heller, 2012). The most recent study of Making Sense of SCIENCE in middle schools found mixed results (Heller et al., 2017). When using pooled data from three project-administered tests over two years of the study, there were no statistically significant results. However, there was suggestive evidence of impact on standardized test performance, with an effect size of 0.17 ($p = .09$) when using the full sample, and an effect size of 0.21 ($p = .04$) after excluding data from one extreme outlier district.

To offer a broader perspective, we try to situate the findings from this i3 study in the larger context of the elementary science education literature. In regard to student science achievement, as pointed out in the Best Evidence Encyclopedia's most recent review of science approaches in elementary schools, there were only 17 studies, among 327 published and unpublished articles that were considered, that met the authors' review standards. The review found that eight inquiry-oriented professional development programs focusing on "effective science teaching...emphasizing conceptual challenge...cooperative learning... science reading integration...teaching scientific vocabulary... and use of an inquiry learning cycle" (but did not provide kits) found significant positive effects, with a weighted mean effect size of 0.30 (Slavin et al., 2012). Due to the small number of studies that qualified for review, the authors cautioned about the tentative nature of conclusions drawn from these findings. For an additional point of reference, we turned to another, and much more recent, Best Evidence Encyclopedia's review of an elementary mathematics study. The meta-analysis included (but is not limited to) nine studies evaluating nine programs focused on teacher professional development aimed at improving "teachers' knowledge of math content and pedagogy" and found a mean effect size on student achievement of 0.03 that was not statistically significant. However, professional development on "classroom management, motivation, and cognition" did yield an average effect size of 0.19 ($p < .01$) (Pellegrini et al., 2020).

We now offer a few working hypotheses on potential reasons for the lack of statistically significant impact findings on student science achievement. We believe this can be attributed to three challenges faced in this study: the lack of an instrument sensitive to measuring the type of three-dimensional learning that aligns with NGSS for low-performing students, the unavailability of NGSS-aligned curriculum and curriculum resources, and the instability of the study sample across the two years of the study.

Assessment

As we discussed at length in other sections of the report, the study took place just as states were adopting and starting their implementation of the NGSS. At the time, the NRC observed that "the assessments that are now in wide use were not designed to meet this vision of science proficiency and cannot readily be retrofitted to do so" (NRC, 2014, page 12). While state science assessments that existed at the time were valid and reliable, they did not measure the type of three-dimensional learning targeted by NGSS. The NRC also noted that developing new assessments would "present[s] complex conceptual, technical, and practical challenges, including cost and efficiency, obtaining reliable results from new assessment types, and developing complex tasks that are equitable for students across a wide range of demographic characteristics" (NRC, 2014, p.16).

Therefore, despite the research team's extensive search for assessments from a variety of sources—including reaching out to state departments of education, university-affiliated assessment centers, and test developers—we could not find an instrument that could measure the type of science achievement targeted in NGSS. Using state assessments was also not an option. In Year 2 of the study (2017–18), when final outcome data would be collected, Wisconsin was still administering the Wisconsin Forward Exam with science items aligned to Wisconsin's Model Academic Standards for Science and enhanced by the NGSS. It was not until spring 2019 that Wisconsin administered the New Science tests aligned with the Wisconsin Standards for Science, which is based on NGSS (Wisconsin DPI, n.d.a). In California, the state was still piloting and field testing the California Science Test in 2016–17 and 2017–18, and the state informed the study team that student scores from these two years would not be available to the research team. This context left us with no choice but to develop our own assessment and to administer it without having had the time to conduct a comprehensive pilot or field-test (see Appendix D for a description of the search for and construction of the assessment). Consequently, the researcher-developed assessment turned out to be difficult and had poor discriminability among students with low achievement. This characteristic of the test has driven us to interpret findings related to science achievement in this study with great caution.

The Availability of Curricula and Curriculum Resources

Another possible explanation for observing null impacts on student science achievement is related to the availability of curricula and curriculum resources. Recall that NGSS is a set of standards. Standards simply define the outcomes, or what students should know and be able to do, from the enacted curriculum. Therefore, as stated in the NRC's guide for implementing NGSS, "teachers need resources that articulate coherent trajectories of questions to investigate or problems to solve that bring together target core ideas, crosscutting concepts, and practices" (NRC, 2015, p.52). Coherence here means that investigations have to be sequenced within-units and across-units such that they engage students in the way that is intended in NGSS, not in traditional sequences that make sense to only science experts (Fortus and Krajick in NRC, 2015). In 2015, the year prior to the start of this study, to the best knowledge of the NRC authors, while many curricula were being developed, there were not yet any year-long, comprehensive NGSS-aligned curriculum resources available at any grade level (NRC, 2015). These resources, they acknowledged, would inevitably take time to develop.

There is no surprise then that during the years of our study, participating districts had not had the opportunity to develop and make available to teachers a coherent curriculum coupled with the corresponding curriculum resources. Some teachers in California were using textbooks published even before the release of the NGSS standards. Teachers in Wisconsin had access to some NGSS-aligned supplementary resources, such as Science A-Z and BrainPOP. However, curriculum resources alone, without a comprehensive curriculum is not sufficient. Certain districts made available

instructional guides during the study years. However, none had gone through a science curriculum adoption cycle since the release of NGSS. The lack of a curriculum and curriculum resources seems to play an important factor in shedding light on the results of this study, as research has shown that professional development, when coupled with designated instructional materials, seems to have greater effect than either resource by itself (Bowes & Banilower, 2004). In our case, the lack of coherent curricula and up-to-date curriculum resources might have even been an impediment to impact.

Stability of the Sample and Fidelity of Implementation across Two Years of the Study

A third possible explanation is related to the stability of the sample and fidelity of implementation across the two years of program implementation. Within each year, WestEd delivered the summer professional learning as intended, and leadership professional learning activities all met fidelity thresholds for attendance. We also observed strong uptake of Making Sense of SCIENCE within each year among teachers who were in the study early enough in the summer to participate in the summer course and were still in the study in the following fall: 94% of teachers in Year 1 (2016–17) and 89% of teachers in Year 2 (2017–18) met the fidelity threshold for attendance at the summer professional learning institutes; 97% of teachers in Year 1 and 90% of teachers in Year 2 met the fidelity threshold for attendance at PLC meetings during the school years.

Yet, only 54% of study teachers met the attendance threshold for the summer courses and 56% of teachers met the attendance threshold for PLC meetings for *both* years. This can be attributed to the instability of the study sample, with teachers leaving the school (17% of baseline teachers) or leaving the study-eligible grade or subject (16% of baseline teachers) during the course of the study. The percentage of teachers leaving the school was congruous with what we observe at the national level: only 84% of teachers stay as a teacher at the same school year-over-year (McFarland et al., 2019). Consequently, only a little more than half of the study teachers received the amount of professional learning as intended by program developers.

While we acknowledge that there could be a number of reasons for the null impact findings on student science achievement (Gerstner et al., 2020), we believe that there is a compelling case to suspect that an insensitive assessment, lack of curriculum and curriculum resources, and instability of the sample were the key contributing factors.

CONSIDERATIONS FOR THE FIELD

We also identified two significant sources of tension that have a broader application in the field. First, it is well documented that teacher professional learning that is sustained over time, offering teachers substantial opportunities to collaborate, is more likely to transform teachers' instructional practices and student learning (Wei et al., 2009; Darling-Hammond et al., 2017). An evaluation of such sustained professional learning would require multi-year studies like this one. However, as discussed in the section about the formation of the sample, multi-year studies leave the study vulnerable to threats of internal validity (related to joiners into schools and within schools in the summers between the study years, when class rosters are formed), as well as to risks of attrition. As we saw in this study, despite our best efforts to inform participants about attrition and to monitor and track participants carefully over time, the reality is that schools are dynamic, open systems that allow for movement of teachers and students. But limiting the study to one-year would mean a missed opportunity to measure impact of sustained professional learning.

A second, related tension that this study raised, particularly for program developers, was whether a two-year professional learning model is possible given the realities of schools—especially those in high-poverty, underserved, transient communities. The frequent transitioning of teachers in and out of grades, subjects, and schools, diminishes the

likelihood of teachers persisting through a two-year professional learning program. Consequently, a point of reflection for program developers may be to consider adapting the program to align with the mobility trajectory of teachers in the targeted populations.

Finally, having identified what we believe to be the most important forces at play in this particular study, we must acknowledge this: student achievement is affected by many factors in a very complex system, of which teacher professional learning is but one critical component. Other factors that could affect what goes on in the classroom (some of which we have touched on above) include instruction, curriculum and curriculum resources, assessment, and leadership at all levels of the school system. Also, the role of teacher leaders, administrators, and district leaders who can be champions of science education and ensure its lateral and vertical coherence cannot be overstated. In this study, we tried to shed light on the impact of Making Sense of SCIENCE, understand the mechanisms driving impact, and determine how and whether impact varies for different groups under different conditions. The field in general, and Making Sense of SCIENCE in particular, would benefit from further research that is both greater in depth and in breadth by taking a harder look into what is happening in the classrooms, as well as understanding the ecosystem that encompasses teacher professional learning.

References

Achieve, Next Gen Science Storylines & STEM Teaching Tools. (2016). *Using Phenomena in NGSS -Designed Lessons and Units*. STEM teaching tools. http://stemteachingtools.org/brief/42

Ainley, M., & Ainley, J. (2011). Student engagement with science in early adolescence: The contribution of enjoyment to students' continuing interest in learning about science. *Contemporary Educational Psychology*, *36*(1), 4-12.

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (2001). Self-efficacy beliefs as shapers of children's aspirations and career trajectories. *Child Development, 72*(1), 187-206.

Banilower, E. R., Smith, P. S., Malzahn, K. A., Plumley, C. L., Gordon, E. M., & Hayes, M. L. (2018). *Report of the 2018 NSSME+*. Horizon Research, Inc.

Bowes, A. S., & Banilower, E. R. (2004). *LSC Classroom Observation Study: An Analysis of Data Collected Between 1997 and 2003.* Horizon Research.

Brahier, D. J., & Schäffner, M. (2004). The Effects of a Study-Group Process on the Implementation of Reform in Mathematics Education. *School Science and Mathematics, 104*(4), 170-178.

Briscoe, C., & Peters, J. (1997). Teacher collaboration across and within schools: Supporting individual change in elementary science teaching. *Science Education, 81*(1), 51-65.

Bryk, A. S. (2010). Organizing schools for improvement. *Phi Delta Kappan*, *91*(7), 23-30.

California Department of Education. (2019). *Smarter Balanced Summative Assessments.* https://www.cde.ca.gov/ta/tg/sa/sbacsummative.asp

California Department of Education Assessment Development & Administration Division. (2019). *California Assessment of Student Performance and Progress California Alternate Assessment for English Language Arts/Literacy and Mathematics Technical Report: 2017–18 Administration.* California Assessment of Student Performance and Progress Technical Reports and Studies. https://www.cde.ca.gov/ta/tg/ca/documents/caa17techrpt.pdf

Carlson, J., Daehler, K. R., Alonzo, A. C., Barendsen, E., Berry, A., Borowski, A., Carperndale, J., Ho Chan, K. K., Cooper, R. Friedrichsen, P., Gess-Newsome, J., Henze-Rietveld, I., Hume, A., Kirschner, S., Liepertz, S., Loughran, J., Mavhunga, E., Neumann, K., Nilsson, P,, … Wilson, C. D. (2019). The refined consensus model of pedagogical content knowledge in science education. In Hume A., Cooper R., Borowski A. (eds) *Repositioning pedagogical content knowledge in teachers' knowledge for teaching science* (pp. 77-92). Springer, Singapore. https://doi.org/10.1007/978-981-13-5898-2_2

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science*, *2*(1), 40.

Casey, P., Dunlap, K., Brown, K., & Davison, M. (2012). Elementary principals' role in science instruction. *Administrative Issues Journal*, *2*(2), 10.

Cavagnetto, A. R., Hand, B., & Premo, J. (2020). Supporting student agency in science. *Theory Into Practice*, *59*(2), 128-138.

Century, J., Rudnick, M., & Freeman, C. (2010). A framework for measuring fidelity of implementation: A foundation for shared language and accumulation of knowledge. *American Journal of Evaluation*, *31*(2), 199-218.

Cervetti, G. N., Barber, J., Dorph, R., Pearson, P. D., & Goldschmidt, P. G. (2012). The impact of an integrated approach to science and literacy in elementary school classrooms. *Journal of research in science teaching*, *49*(5), 631-658.

d'Alessio, M. A. (2018). *Executive Summary: Science Framework for California Public Schools: Kindergarten Through Grade Twelve*. Consortium for the Implementation of the Common Core State Standards.

Daehler, K. R., Wong, N., & Heller, J. I. (2015). Supporting growth of pedagogical content knowledge in science. In Berry, A., Friedrichsen, P., & Loughran, J. (Eds.), *Re-examining pedagogical content knowledge in science education* (pp. 45–59). Routledge Press.

Darling-Hammond, L., Hyler, M. E., & Gardner, M. (2017). *Effective Teacher Professional Development* (research brief). Learning Policy Institute.

Darling-Hammond, L., Wei, R. C., Andree, A., Richardson, N., & Orphanos, S. (2009). *Professional Learning in the Learning Profession: A Status Report on the Teacher Development in the United States and Abroad*. National Staff Development Council and The School Redesign Network at Stanford University.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (Eds.) (2007). *Taking Science to School: Learning and Teaching Science in Grades K-8*. National Academic Press.

Gerstner, C., Allen-Platt, C., Boruch, R., Ruby, A. (2020). *Toward a Science of Failure Analysis: A Narrative Review*. Poster presentation accepted for the annual spring conference of the Society for Research on Educational Effectiveness, Washington, DC. (Conference Canceled).

Graham, P. (2007). Improving teacher effectiveness through structured collaboration: A case study of a professional learning community. *RMLE online*, *31*(1), 1-17.

Hallam, P. R., Smith, H. R., Hite, J. M., Hite, S. J., & Wilcox, B. R. (2015). Trust and collaboration in PLC teams: Teacher relationships, principal support, and collaborative benefits. *NASSP Bulletin*, *99*(3), 193-216.

Heller, J. I., Little, J. W., & Shinohara, M. (2010). *Impact of content-focused and practice-based professional development models on elementary electric circuits teaching and learning.* [Unpublished final report submitted to the National Science Foundation for grant no. 0545445].

Heller, J. I. (2012). *Effects of Making Sense of SCIENCE Professional Development on the Achievement of Middle School Students, Including English Language Learners*. Final Report. NCEE 2012-4002. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. https://eric.ed.gov/?id=ED530414

Heller, J. I., Daehler, K. R., Wong, N., Shinohara, M., and Miratrix, L. W. (2012). Differential effects of three professional development models on teacher knowledge and student achievement in elementary science. *Journal of Research in Science Teaching, 49*(3), 333-362.

Heller, J. I., Wong, N., Limbach, J. O., Yuan L., & Miratrix, L. (2017, September). *Making Sense of SCIENCE: Efficacy Study of a Professional Development Series for Middle School Science Teachers*. U.S. Institute of Education Sciences.

Hermann, M., Clark, M., James-Burdumy, S., Tuttle, C., Kautz, T., Knechtel, V., Dotter, D., Wulsin, C.S., & Deke, J. (2019). *The effects of a principal professional development program focused on instructional leadership (NCEE 2020-0002).* National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*(2), 371–406. http://sitemaker.umich.edu/lmt/files/hillrowanball.pdf

Iveland, A., Tyler, B., Britton, T., Nguyen, K., & Schneider, S. (2017). *Administrators Matter in NGSS Implementation: How School and District Leaders Are Making Science Happen*. WestEd.

Jacob, R., Goddard, K., Miller, R., & Goddard, Y. (2014). Exploring the causal impact of the McREL Balanced Leadership Program on leadership, principal efficacy, instructional climate, educator turnover, and student achievement. *Educational Evaluation and Policy Analysis, 52* 187-220.

Kanter, D. E., & Konstantopoulos, S. (2010). The impact of a project-based science curriculum on minority student achievement, attitudes, and careers: The effects of teacher content and pedagogical content knowledge and inquiry-based practices. *Science Education*, *94*(5), 855-887.

McFarland, J., Hussar, B., Zhang, J., Wang, X., Wang, K., Hein, S., Diliberti, M., Forrest Cataldi, E., Bullock Mann, F., and Barmer, A. (2019). *The Condition of Education 2019* (NCES 2019-144). U.S. Department of Education. National Center for Education Statistics. https://nces.ed.gov/ pubsearch/pubsinfo.asp?pubid=2019144.

McNeill, K. L., Katsh-Singer, R., & Pelletier, P. (2015). Assessing science practices: Moving your class along a continuum. *Science Scope*, *39*(4), 21.

MOSART. (2011). *Test Inventory and Development.* MOSART Self-Service Site. https://www.cfa.harvard.edu/smgphp/mosart/testinventory_2.html

Murphy, C., Neil, P., & Beggs, J. (2007). Primary science teacher confidence revisited: Ten years on. *Educational research*, *49*(4), 415-430.

National Assessment of Educational Progress (NAEP). (2019, June 20). *2015 Science Assessment.* https://www.nationsreportcard.gov/science_2015/

National Center for Education Statistics (NCES). (2020). *American Community Survey*. https://nces.ed.gov/programs/edge/acsdashboard

National Center for Education Statistics (NCES). (2014). *National Assessment of Educational Progress State Profiles*. The Nation's Report Card. http://nces.ed.gov/nationsreportcard/states/

National Research Council (NRC). (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. The National Academies Press. https://doi.org/10.17226/13165.

National Research Council (NRC). (2014). *Developing Assessments for the Next Generation Science Standards*. Committee on Developing Assessments of Science Proficiency in K-12. Board on Testing and Assessment and Board on Science Education, J.W. Pellegrino, M.R. Wilson, J.A. Koenig, and A.S. Beatty, Editors. Division of Behavioral and Social Sciences and Education. The National Academies Press.

National Research Council (NRC). (2015). *Guide to Implementing the Next Generation Science Standards.* Committee on Guidance on Implementing the Next Generation Science Standards. Board on Science Education, Division of Behavioral and Social Sciences and Education. The National Academies Press.

National Research Council (NRC). (1996). *National Science Education Standards.* National Committee on Science Education Standards and Assessment. National Academy Press.

National Research Council (NRC). (2013). *Next Generation Science Standards: For States, by States*.

National Science Board. (2018). *Science and Engineering Indicators 2018*. National Science Foundation (NSB-2018-1).

Pellegrini, M., Lake, C., Neitzel, A., & Slavin, R. E. (2020). *Effective programs in elementary mathematics: A Meta-Analysis*. Manuscript under review. http://www.bestevidence.org/math/elem/elem_math_2020.htm

The President's Council of Advisors on Science and Technology (PCAST). (2010). *Report to the President: Prepare and Inspire: K-12 Education in Science, Technology, Engineering, and Mathematics (STEM) for America's Future*.

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). *What to Do When Data Are Missing in Group Randomized Controlled Trials* (NCEE 2009-0049). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Singer, J. D., & Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford University Press.

Slavin, R. E., Lake, C., Hanley, P., & Thurston, A. (2012). *Effective Programs for Elementary Science: A Best-Evidence Synthesis*. Best Evidence Encyclopedia (BEE). https://eric.ed.gov/?id=ED539695

Sztajn, P., Marrongelle, K., Smith, P., & Melton, B. (2012, March). *Supporting Implementation of the Common Core State Standards for Mathematics: Recommendations for Professional Development.* The William & Ida Friday Institute for Educational Innovation, North Carolina State University College of Education.

Tytler, R., & Osborne, J. (2012). Student attitudes and aspirations towards science. In *Second international handbook of science education* (pp. 597-625). Springer, Dordrecht.

Unlu, F., Bozzi, L., Layzer, C., Smith, A., Price, C., & Hurtig, R. (2016). Linking implementation fidelity to impacts in an RCT. *Treatment Fidelity in Studies of Educational Intervention*, 100.

Urick, A., Wilson, A. S., Ford, T. G., Frick, W. C., & Wronowski, M. L. (2018). Testing a framework of math progress indicators for ESSA: How opportunity to learn and instructional leadership matter. *Educational Administration Quarterly*, *54*(3), 396-438.

Wei, R. C., Darling-Hammond, L., Andree, A., Richardson, N., Orphanos, S. (2009). *Professional learning in the learning profession: A status report on teacher development in the United States and abroad*. National Staff Development Council.

Weis, A. M. (2013). *2012 National Survey of Science and Mathematics Education: Status of Middle School Science.* Horizon Research, Inc.

Wisconsin Department of Public Instruction (DPI). (n.d.a). *Wisconsin Forward Exam*. https://dpi.wi.gov/science/assessment/state-test

Wisconsin Department of Public Instruction (DPI). (n.d.b). *Wisconsin Forward Exam*. https://dpi.wi.gov/assessment/forward

Wisconsin Department of Public Instruction (DPI). (2018). *Wisconsin Forward Exam: Technical Manual 2018*. https://dpi.wi.gov/sites/default/files/imce/assessment/pdf/Forward_Exam_Tech_Report_2018.pdf

Wong, N., Heller, J. I., Kaskowitz, S. R., Burns, S., Limbach, J. O. (2020). *Final report of the Making Sense of Science and Literacy implementation and scale-up studies.* [U.S. Department of Education Project No. U411B140026]. Heller Research Associates.

Reference for this report:

Jaciw, A.P., Nguyen, T., Lin, L., Zacamy, J., Kwong, C., Lau, S. (2020). *Final Report of the i3 Impact Study of Making Sense of SCIENCE, 2016-17 through 2017-18*. (Empirical Education Rep. No. Empirical_MSS-7030-FR1-2020-O.1) San Mateo, CA: Empirical Education, Inc. Retrievable from https://empiricaleducation.com/mss

Appendix

All appendices are accessible at https://www.empiricaleducation.com/mss/.