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

2016-17 THROUGH 2017-18

Science education has experienced a significant transition over the last decade, catalyzed by a re-envisioning of what students should know and be able to do in science. The new vision—set forth in *A Framework for K–12 Science Education*—lays out the fundamental shift from content knowledge to three-dimensional learning through the integration of Disciplinary Core Ideas, Cross-Cutting Concepts, and Science and Engineering Practices. The release of the Next Generation Science Standards (NGSS) in 2013, which was based on the framework, set off a chain reaction of standards adoption and implementation across states, districts, and schools, including steps taken toward transforming science professional learning, instruction, curriculum, and assessment.

It was in this dynamic context that WestEd's Making Sense of SCIENCE project received an Investing in Innovation (i3) grant. Under the five-year grant, WestEd partnered with Empirical Education Inc. to conduct a two-year impact evaluation and with Heller Research Associates (HRA) to conduct an implementation study and a scale-up study of Making Sense of SCIENCE. What follows is a summary of the report on the impact evaluation. The full report is accessible at https://www.empiricaleducation.com/mss/.

Making Sense of SCIENCE

Making Sense of SCIENCE is a teacher professional learning model aimed at raising students' science achievement 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 Common Core State Standards. The Making Sense of SCIENCE theory of action is based on the premise that professional learning that is situated in an environment of collaborative inquiry and supported by school leadership has a cascade of effects on teachers' content and pedagogical content knowledge, school climate, and opportunities to learn. These effects, in turn, produce improvements in student science achievement and other non-academic outcomes. The key components of the model include leadership professional learning (for site coordinators, a leadership cadre comprising teacher leaders and regional members, and school administrators) and

Making Sense of SCIENCE is an approach to teacher professional learning aimed at raising student science achievement by focusing on the critical connections between science understanding, literacy support, and classroom practices.

professional learning for teachers. The teacher professional learning includes a 5-day course each summer for two summers and six professional learning community (PLC) meetings per year for two years. Making Sense of SCIENCE program developers facilitated the professional learning for the leadership cadre, who then facilitated the summer professional learning courses and school-year PLC meetings for teachers (see the report for a more comprehensive description of the Making Sense of SCIENCE logic model).

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.

The full report is available at empiricaleducation.com/MSS

The Impact Evaluation

Design

The evaluation was a two-year cluster-randomized control trial (RCT) that took place in California and Wisconsin across seven school districts and 66 elementary schools in the 2016–17 and 2017–18 school years. We randomized schools to either receive the Making Sense of SCIENCE professional learning or to the business-as-usual ("control") group, which received the professional learning (delayed-treatment) after the study ended.

The IMPACT EVALUATION was a two-year cluster-randomized control trial that took place in California and Wisconsin across 7 school districts and 66 elementary schools in 2016–17 and 2017–18.

Research Questions

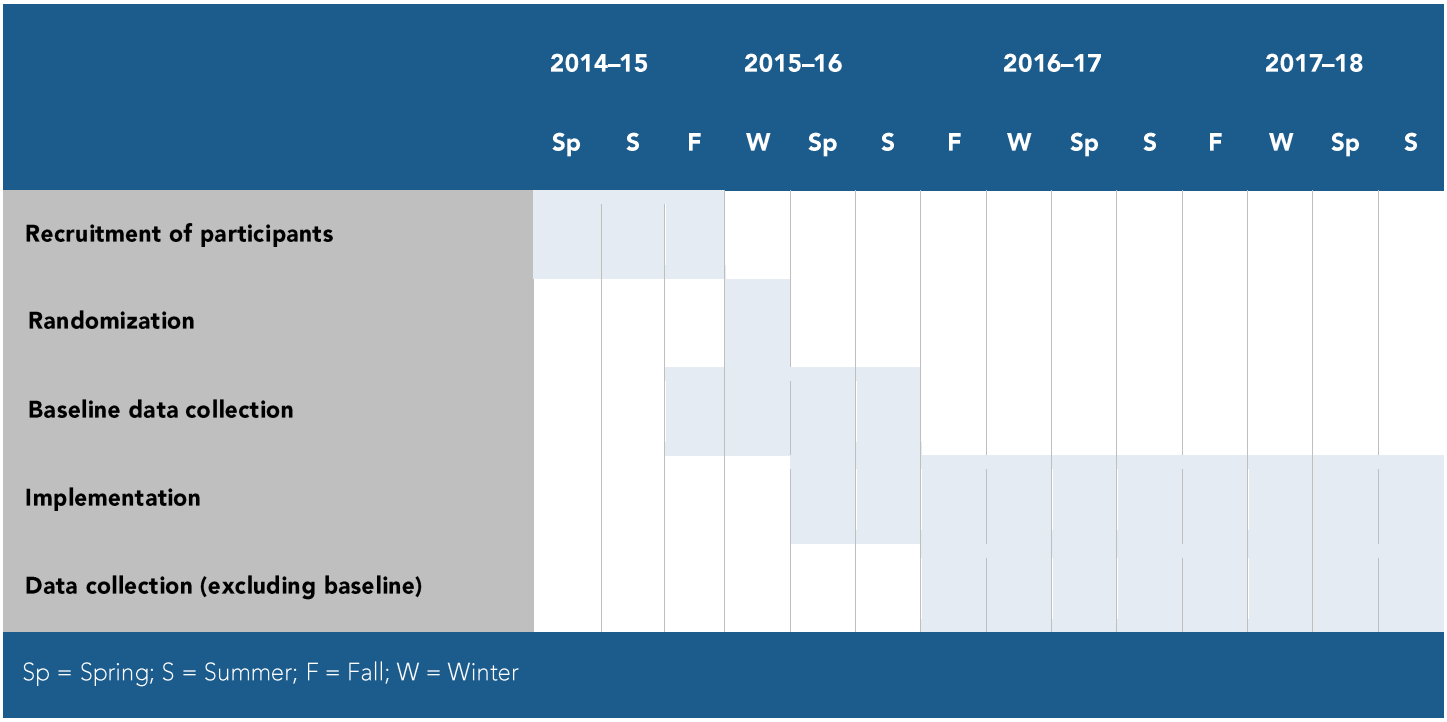
Through this impact study, we aimed to address research questions concerning fidelity of implementation and impacts on teacher, classroom, school, and student outcomes after two years of implementation. Going beyond simply examining average impacts, we explored questions about how impacts vary for different subgroups and under different conditions.

First, we examined the extent to which the program was implemented with fidelity. Then, we examined impacts of Making of Sense of SCIENCE in the order of left to right on the logic model: teacher content and pedagogical content knowledge, teacher attitudes and beliefs, school climate, classroom outcomes and student opportunities to learn, and student achievement and non-academic outcomes. We also examined differential impacts across different teacher characteristics (e.g., pretest score, years of experience, membership in the leadership cadre) and student characteristics (e.g., English learner status, eligibility for Free or Reduced-Price lunch), as well as impacts on subsamples of interest (e.g., by incoming achievement, grade, and state) and on districts with particularly strong implementation.

Study Timeline

The study's timeline is displayed in Table 1. The key milestones include randomization of schools in winter 2015–16, implementation of Making Sense of SCIENCE in 2016–17 and 2017–18, and final outcomes for confirmatory analyses collected in spring 2017–18.

TABLE 1. TIMELINE OF MAJOR ACTIVITIES IN THE STUDY



Measures

The impact evaluation was based on a rich data set collected from students, teachers, and school districts. For the final report, we assessed outcomes in the spring of Year 2 (2017–18) of the trial.

We measured teacher content knowledge and pedagogical content knowledge using an evaluator-developed assessment. Exploratory outcomes such as teacher self-reported attitudes and beliefs about science instruction and learning, opportunity to learn, and school climate were measured using teacher surveys.[1]

For students, we collected data for five academic student outcomes: *science achievement* (confirmatory) based on selected-response items from an evaluator-developed instrument,[2] *communicating about science in writing* based

[1] The study also surveyed administrators, the key findings from which are reported in HRA's implementation report (Wong et al., 2020). Video and audio recordings were also collected for a small subset of classrooms. The sample sizes (due to student consent and limited project resources) were inadequate for inclusion in impact analyses.

[2] The assessment development process, which involved use of general content specifications and established items, was meant to ensure that the instrument is not over-aligned with the intervention. The process is described in Appendix D.

on constructed-response items from the same researcher-developed instrument, and *performance on state assessments in English Language Arts (ELA), math, and science* using data provided by school districts. We also administered a survey to measure student non-academic outcomes, such as enjoyment of science, agency in learning science, and aspirations about future use of science in their adulthood and career. Districts provided student achievement and demographic data.

Full descriptions of measures and the data collection timeline are provided in the report.

Analytic Samples

School Sample. At baseline, in the winter of 2015–2016, we randomly assigned 60 schools to *Making Sense of SCIENCE* or control (see Table 3 in the report for all sample sizes).[3]

Teacher Sample (RCT confirmatory analyses). The confirmatory analysis of impact on teacher content knowledge is based on a sample of 88 teachers who were in the *Retained in Study* sample. That is, they were teachers from the baseline representative sample (BRS)[4] who were still active in the study in spring of Year 2 (2017–18) when the teacher content knowledge assessment was administered. **An additional analysis** was conducted using a sample of teachers comprising these 88 teachers along with 30 additional teachers who were in the study at randomization ("*Mixed* sample"); these 30 teachers were no longer active in the study in spring of Year 2, but agreed to take the teacher content knowledge assessment. They had a range of exposure to Making Sense of SCIENCE professional learning.[5]

Teacher Sample (exploratory analyses). Given that more BRS teachers left the study than we had expected by the second year, we engaged in additional recruitment efforts of mostly teachers who had transitioned into the eligible grade levels at participating schools. This provided a larger sample of 147 teachers for assessing impacts on teacher attitudes and beliefs, opportunity to learn, and school culture, as measured by teacher surveys.

Student Sample (confirmatory analyses). The main analysis of impact on student science achievement comprises 2,140 students of 147 teachers who administered the student science achievement assessment in spring of Year 2 (2017–18). The sample of 147 teachers included not only teachers who were present at randomization, but also teachers who joined the study after randomization ("joiners").[6] To address the second

[3] There were 66 participating schools with 60 units of randomization. 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. At baseline, these schools agreed that if they were randomized to the treatment group, they would work together and implement Making Sense of SCIENCE as if they were one school. Here and henceforth, units of randomization (54 schools and 6 dyads) are referred to as "schools."

[4] The baseline representative sample of teachers consists of those who were in the study at baseline (i.e., at the time of random assignment) and who were randomly selected to participate in data collection. The study collected data from a probability sample, as opposed to all participating teachers, as a cost-saving measure.

[5] Based on the criteria for attrition, the "*Retained in Study* Sample" can at best meet WWC evidence standards with reservations, while the "Mixed Sample" has potential to meet WWC evidence standards without reservations.

[6] Because we cannot rule out that students deliberately selected to be on rosters of study teachers' classes, the confirmatory analysis of impact on student science achievement is likely to at best meet WWC evidence standards with reservations.

confirmatory research question about student science achievement, we also conducted an analysis using the sample of students who were among the lowest third of incoming achievement.

Student samples (exploratory analyses). As part of our exploratory analyses, we identified two additional student samples. The first consists of 1,415 students of 96 teachers who were part of the BRS ("Focused Sample 1"). The second consists of 340 students who were in a *Making Sense of SCIENCE* classroom in both Years 1 and 2 (2016–17 and 2017–18) ("Focused Sample 2").

Samples for other exploratory analyses are described in the report.

Statistical Analysis

Impact analyses used standard methods for cluster randomized trials. We applied multilevel models to estimate the Intent-to-Treat effect of assignment to *Making Sense of SCIENCE* compared to business as usual. We regressed individual scores against baseline covariates, a variable indicating treatment assignment at the school level, and random effects at person and school levels, with block (pair) effects modeled as fixed. We examined the robustness of the effect estimates using alternative model specifications. For specific analyses, we calculated the levels of attrition and differential attrition at the different levels of the study design. We also examined baseline equivalence for the analytic samples for certain outcomes.

Findings

Fidelity of Implementation

The study's measure of fidelity of implementation comprised six components related to the delivery of, and attendance at, leadership and teacher professional learning activities. Within each year, WestEd delivered the summer professional learning as intended, and leadership professional learning activities all met fidelity thresholds for attendance. Notably, we observed strong uptake of Making Sense of SCIENCE within each year among teachers who were in the study early enough to participate in 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 professional learning community meetings. Yet, only 54% of study teachers met the attendance threshold for the summer courses, and 56% of teachers met the attendance threshold for professional learning communities for both years due to the instability of the sample across the two years. Among the 185 participating teachers in *Making Sense of SCIENCE* schools—including those who attrited and 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.

There was strong uptake of Making Sense of SCIENCE professional learning. More than 90% of teachers attended summer courses and school-year meetings in each year. Yet, due to the instability of the sample across the two years, only a little more than half of teachers attended the professional learning activities *across both years.*

Teacher Content Knowledge and Pedagogical Content Knowledge

We observed a positive impact of Making Sense of SCIENCE on teacher content knowledge, with a standardized effect size of 0.56 (p = .006),[7] for the *Retained in Study* sample. For the *Mixed* sample, there was a positive but not statistically significant effect size of 0.22 (p = .165) (Figure 1).

Making Sense of SCIENCE had a positive impact on teacher content knowledge and on the holistic scale of pedagogical content knowledge.

For pedagogical content knowledge,[8] we assessed impact on four scales for each of the two samples. For both samples, we found a significant and positive effect on the holistic scale, with effect sizes of 0.41 (p = .026, *Retained in Study* sample) and 0.36 (p = .049, *Mixed* sample). We also found a marginally significant impact on the PCK-Explanation scale for the *Retained in Study* sample, with an effect size of 0.121 (p = .053) (Figure 1).

FIGURE 2. IMPACTS ON TEACHER CONTENT KNOWLEDGE AND PEDAGOGICAL CONTENT KNOWLEDGE

Note. + p < .1; * p < .05; ** p < .01; ***p < .001

Moderator analyses using the *Retained in Study* sample indicated that there were no differential impacts depending on teachers' incoming level of content knowledge (ES = 0.167, p = .914), years of teaching experience

[7] All reported effect sizes are standardized.

[8] Teachers' responses on the pedagogical content knowledge assessment were rated in terms of (1) Concept Score, (2) Explanation Score, (3) 2-D Score, and (4) Holistic Score. The Concept Score relates to teachers' ability to connect instructional activities to specific conceptual goals. The Explanation Score relates to 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. 2-D score is a measure of teachers' ability to integrate *both* science concepts and explanation practices. Holistic Score is a score based on a holistic assessment of the strength of the response, and took into account whether a teacher's written responses exhibited PCK in conceptual understanding *or* the scientific practice of explanation in any form (Wong et al., 2020).

(ES = -0.012, p = .378), and status as a teacher leader (ES = 0.113, p = .670). No significant differential impacts were found using the *Mixed* sample.

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

We conducted exploratory analysis on the impact of Making Sense of SCIENCE on 30 constructs for teacher attitudes and beliefs (8 outcomes), opportunities to learn (15 outcomes), and school climate (7 outcomes).

Making Sense of SCIENCE teachers reported teaching more science than teachers in the control group. All four outcomes related to students' opportunity to learn as measured by teachers' instructional practices were positive and either significant or marginally significant.

The most promising results are observed for time on science and students' opportunity to learn as measured by teachers' science instructional practices. We observed a positive impact (ES = 0.40, p = .015) for the *Amount of Time Spent on Science*.[9] This is equivalent to an increase of approximately 18 hours in science instructional time over the school year. We also observed positive and either significant or marginally significant impact on all four outcomes related to teachers' science instructional practices (Figure 2).

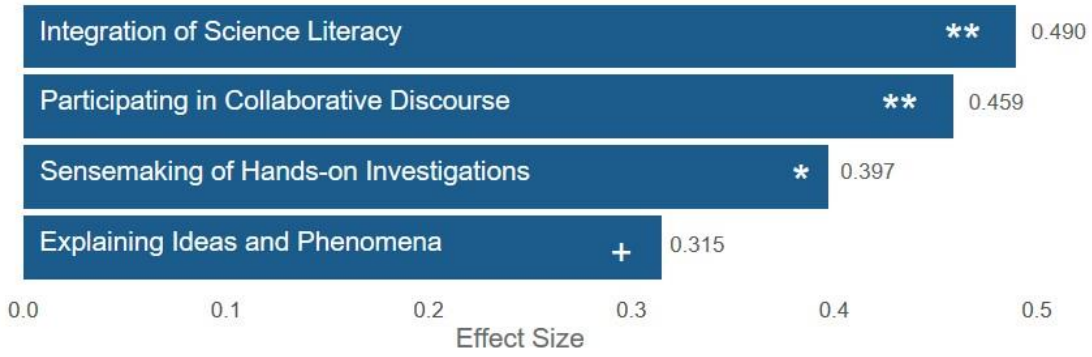


FIGURE 2. IMPACTS ON OPPORTUNITY TO LEARN – INSTRUCTIONAL PRACTICES

Note. + p < .1; * p < .05; ** p < .01; ***p < .001

Other notable findings include *Making Sense of SCIENCE* teachers reporting a greater sense of *Agency in the Classroom* (ES = 0.382, p = .025), and a marginally significant impact on *Confidence in Science Instructional Practices* (ES = 0.261, p = .083). They also reported collaborating with each other more often outside of Making Sense of SCIENCE professional learning events. As for school climate, teacher self-reports yielded a statistically significant impact on *Administrators' Support of Teacher Collaboration* (ES = 0.388, p = .025) and a marginally significant impact on *Administrators Involving Teachers in Science Leadership* (ES = 0.297, p = .058). Teachers also reported spending greater *Amount of Time on Informal Peer Collaboration* (ES = 0.876, p = .003).

[9] Outliers were removed from this outcome.

All ten outcomes related to the opportunity to learn NGSS-aligned content (Disciplinary Core Ideas) were not statistically significant, although eight of the ten outcomes had positive effect sizes. We reflect on this finding further in the discussion section.

For moderator analyses of intermediate outcomes, we focused on differential impact of being a teacher leader and found statistically significant, positive differential impacts for five outcomes: *Belief That Students Are Capable Learners,* three physical science *Disciplinary Core Ideas,* and *Administrators Involving Teachers in Science Leadership*.

With 30 main contrasts of self-reported measures on teacher attitudes and beliefs, opportunities to learn, and school climate, there is a high probability that one or more will reach statistical significance by chance alone. However, the trend of statistically significant or marginally significant results observed in all four instructional practice outcomes measured, and no statistically significant results for all ten opportunity to learn content outcomes give us greater confidence that impacts were observed in one domain but not the other.

Student Science Achievement

For the two confirmatory research questions related to students, we did not observe impact on student science achievement for the full sample (ES = 0.064, p = .494) and for the lowest third of students with incoming ELA achievement (ES = 0.073, p= .567). There was, however, a marginally significant impact for the lowest third of students with incoming math achievement (ES = 0.220, p = .099).

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. We also found no impacts on communicating about science in writing and student science achievement, as measured by the state assessment in science and math, but did observe a marginally significant impact on the state ELA assessment (effect size = 0.090, p = .057).

Effect sizes across different samples and for multiple outcomes are almost all positive, albeit not statistically significant. There were two marginally significant impacts on the lowest third in math incoming achievement and on the state ELA assessment.

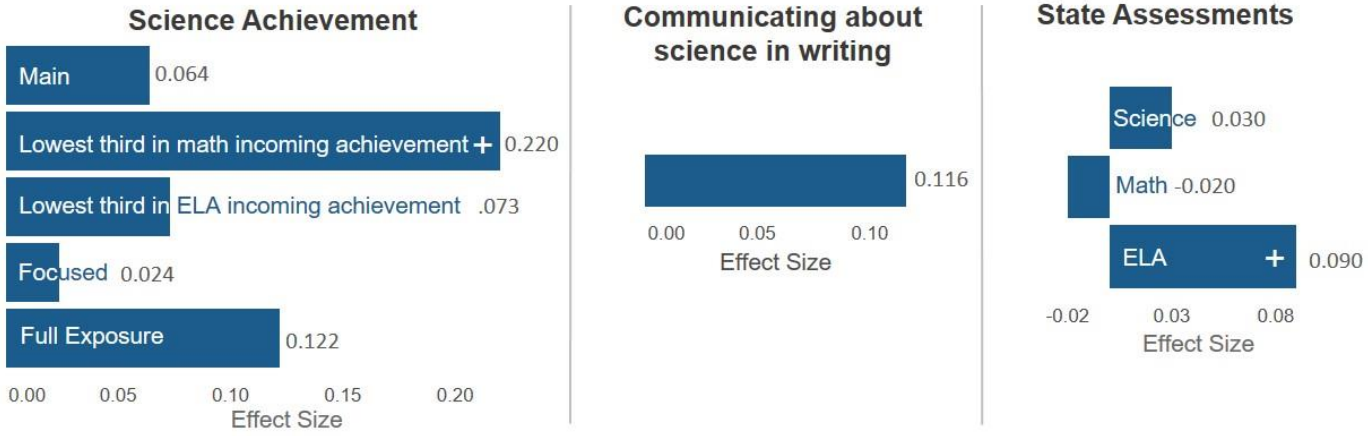


FIGURE 3. IMPACTS ON STUDENT ACHIEVEMENT OUTCOMES

Note. + p < .1; * p < .05; ** p < .01; ***p < .001

For moderator analyses, we evaluated whether impacts on the student science achievement (as measured by the evaluator-developed science assessment) varied by ELA and math pretests, state (WI versus CA), grade (4th versus 5th), student English Language Learner (ELL) status, student eligibility for Free or Reduced-Price lunch (FRPL), and with the moderators combined. We observed that only the differential impact by student ELL status was marginally significant, with a reduction in impact of approximately .150 standard deviations (p = .073) associated with being a ELL, based on the model with just the one interaction, and approximately .230 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.

Turning to districts, we evaluated 1) whether impacts on student science achievement varied by district, and 2) impacts on student science achievement among the districts considered by program developers to be strong implementers.[10] We observed a positive, but not statistically significant, difference in intervention impact between strongly implementing districts and the remaining districts (effect size = 0.080, p = .397). For the high implementing districts, we did not observe an impact of the program (effect size = 0.068, p = .419).

Student Non-Academic Outcomes

Effect sizes for student non-academic outcomes across the eight scales range from -0.225 to 0.135, though there is only one statistically significant and negative impact on *Quality of Science Class – Science Instruction* (p = .030).[11] 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.

Discussion

Findings from this evaluation suggest that Making Sense of SCIENCE has positive impacts on proximal outcomes, such as teacher content knowledge, pedagogical content knowledge, time on science instruction, and teacher instructional practices. Such findings that Making Sense of SCIENCE changes classroom science learning experiences in ways that align with expectations in NGSS, which is a hypothesized precursor to measuring impacts on student achievement, deserve notice.

We did not observe statistically significant impacts on several school climate outcomes, student science achievement and communicating about science in writing as measured by the evaluator-developed assessment, and student non-academic outcomes. However, we did observe a marginally significant effect on the ELA state assessment. This is an encouraging finding given that the Making Sense of SCIENCE is designed

[10] Impacts on student science achievement were measured by the evaluator-developed science assessment.

[11] The remaining seven outcomes were Aspirations, Quality of Science Class – Learning Environment/Classroom Management, Self-Efficacy, Activities in Science Classroom, Agency in Learning, Cognitive Demand, and Enjoyment of Science.

to support connections between science and literacy. Moreover, with a few rare exceptions, all effect sizes were in the positive direction.

Outcomes for which we did not see significant results could be interpreted as less malleable to change. 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.

These findings contribute to a series of studies of Making Sense of SCIENCE from the past few decades. Previous studies were focused on a single grade or specific science content area (e.g., electric circuits or force and motion). This study builds on previous studies by covering a larger grade and content band (i.e., two grades across Earth and space science and physical science) with a focus on NGSS. Results from this study provide suggestive evidence about Making Sense of SCIENCE's ability to scale, particularly in regard to reaching larger grade bands, a broader set of core science ideas, and building a school-level community of practice.

Results from this study are fairly consistent with previous findings on the impact of Making Sense of SCIENCE. 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), with the 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. Heller et al. (2012) found effect sizes ranging between 0.37 to 0.60 (all $p < .001$). Heller et al. (2012) found effect sizes of 0.11 ($p = .04$) for the full sample and 0.31 (p value = .04) for the subset of ELLs, though these results were no longer statistically significant after adjusting for multiple comparisons. The most recent study of Making Sense of SCIENCE (Heller et al., 2017) found mixed results: 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.

The researcher-developed assessment of student science achievement was too difficult and exhibited low test score reliability for students with low incoming achievement.

We identified three potential contributors to the limited impact findings on student science achievement in this study. First, this study was conducted just two years after the release of the NGSS. Valid and reliable student assessments that were aligned with NGSS were not yet available. Researchers decided to administer a researcher-developed assessment, which turned out to be difficult and exhibited limited reliability for students at the low end of the achievement scale. This characteristic of the test has driven us to interpret findings related to science achievement in this study with great caution.

NGSS-aligned curriculum and curriculum resources—important ingredients to student science learning—were not yet available to participating districts during the study years.

A second possible explanation for observing limited impacts on student science achievement is related to the availability (or lack thereof) of curriculum and curriculum materials. NGSS is a set of standards, not a curriculum; and Making Sense of SCIENCE offers teacher professional learning that is not associated with a student curriculum. Therefore, teachers need to have access to a coherent curriculum and corresponding curriculum materials that have within-unit and across-unit coherence, with investigations sequenced in a way that engages students, rather than in traditional sequences that make sense to only science experts (Fortus and Krajick in NRC 2015). As of 2015, while many curricula were being developed, there were not yet any year-long, comprehensive NGSS-aligned curricular resources available at any grade level (NRC, 2015). These resources would inevitably take time to develop and were not yet available to participating districts during the study years. This reality may have played an important factor in the null student impact finding, as research has shown that professional development, when coupled with designated instructional materials, has greater effect than either resource by itself (Bowes & Banilower, 2004). In our case, the lack of a coherent curriculum and up-to-date curricular resources that align with NGSS content and practices was a likely impediment to impact.

A third possible explanation is related to the stability of the sample and fidelity of implementation across the two years of program implementation. While we observe strong uptake of Making Sense of SCIENCE within each year, only a little more than half of teachers were present in the study in both study years and received the amount of professional learning as intended by program developers. This can be attributed to the instability of the study sample, with teachers leaving the school (17% of baseline teachers) or the study-eligible grade or subject (16% of baseline teachers) during the course of the study, or with teachers joining the study (13% of all Making Sense of SCIENCE teachers) after certain professional learning activities were already completed. 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).

Teachers did not meet fidelity of implementation across the two years due to the instability of the sample, with teachers leaving the school or study-eligible grade or subject.

We offer two additional points of reflection that may have broader implications for the field of education evaluation and teacher professional learning that are both related to the fact that the study was a multi-year intervention. 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, Darling-Hammond, Andree, Richardson, Orphanos, 2009; Darling-Hammond, Hyler, Gardner, 2017). An evaluation of such sustained professional learning would require multi-year studies like this one. However, multi-year studies are vulnerable to threats of internal validity related to post-randomization selection of students into schools, and of students into classes within schools, as well as to risks of attrition. But limiting the study to one year would mean a missed opportunity to measure impact of sustained professional learning, thus presenting a research-to-practice dilemma.

A second, related tension—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. The role of teacher leaders, administrators, and district leaders who can be champions of science education and ensure its lateral and vertical coherence also 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 understand 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 greater in both depth and breadth by taking a harder look into what is happening in the classrooms, as well as understanding the ecosystem that encompasses teacher professional learning. It is our hope that such research would ultimately inform and transform science teaching and learning in a way that will afford *all* students the opportunity to earn livable wages and make informed choices as citizens, to compete in the new industries of the 21st century, and to contribute to a society that continues to make new discoveries about ourselves and the universe.

References

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

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

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. 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.

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.

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.

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.

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.