REPORT APPENDIX

AN EMPIRICAL EDUCATION REPORT APPENDIX

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

2016-17 THROUGH 2017-18

APPENDIX

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.

Table of Contents

Appendix A. Making Sense of SCIENCE Logic Model Terminology and Definitions

This appendix provides the key terminologies that were used in the Making Sense of SCIENCE logic model and their definitions, as supported by extant literature. This appendix was written and provided by the WestEd program team.

Logic Model Construct Descriptions

Leadership Outcomes

We posit that the Making Sense of SCIENCE Leadership Development component has a direct impact on state, regional, district, school, and teacher leaders and has more distal outcomes on school culture and teacher attitudes and beliefs.

State and Regional Leader Outcomes

- *Deeper knowledge of standards implementation (e.g., NGSS)* — State and regional leaders with greater knowledge of reform-based standards and best practices associated with standards implementation are better equipped to build an infrastructure for developing and sustaining improvements for science education in the long term (Penuel, et al., 2014).
- *Greater ability to support implementation of school/district professional learning* — With technical assistance, state and regional leaders are able to set priorities and adequately align resources to support professional learning, and science teaching and learning.

Administrator Outcomes

We posit that Making Sense of SCIENCE has an impact on school principals, coaches, and district administrators.

- *Deeper knowledge of instructional shifts in science and standards implementation supports* — The literature suggests that when administrators have deeper understanding of reform-based standards and the instructional shifts required, they have a better understanding of how demanding this work is and the kinds of supports their administrators and teachers need. Subsequently, they are able to provide the appropriate supports for standards implementation (Iveland, et al., 2017).
- *Shifted beliefs that learning science is as important as other subjects* — Teachers often cite that the biggest barrier to teaching science is time, due to the demands to meet accountability requirements for math, reading, and writing. When administrators shift their belief that science is also an important subject, they are able to signal to teachers that science is a priority and allocate more time and resources to support science teaching and learning.
- *Increased philosophical alignment with standards* — When administrators understand the instructional shifts required by reform-based standards, they have a better understanding of what that will look like in the classroom and will give teachers permission to grow and fail as they try to incorporate these instructional shifts in their classroom.

Teacher Leaders

Making Sense of SCIENCE grows the leadership capacities of teachers through professional learning and coaching that strengthens the knowledge and skills needed to be effective in their own classrooms. We intentionally build the skills and confidence of teacher leaders to facilitate the professional growth of their peers. We posit that MSS has an impact on the skills and knowledge of teacher leaders.

- *Deeper knowledge of standards implementation (e.g., NGSS)* — When teacher leaders have greater knowledge of standards, they are able to take on the role of a curriculum specialist and can serve as a catalyst of change to bring about the implementation of science standards in a school (Harrison & Killion, 2007).
- *Greater skill in facilitating teacher learning and collaboration* — When teacher leaders develop their content and pedagogical content knowledge through Making Sense of SCIENCE courses, and also develop their facilitation skills through the Making Sense of SCIENCE Facilitation Academies, they are able to facilitate communities of learning through school-wide approved processes, particularly professional learning communities (PLCs). In PLCs, when teachers learn with and from one another, they can focus on what most directly improves student learning (Harrison & Killion, 2007).

Teacher Outcomes

We posit that the MSS Teacher Professional Learning component has a direct impact on teachers' content knowledge and pedagogical content knowledge, and may have a distal impact on teacher attitudes and beliefs.

Content Knowledge

Teacher content knowledge is used to describe the body of knowledge that teachers teach and that students are expected to learn in a content area. The focus on teacher content knowledge is aligned with the literature that provides clear evidence on the critical role that teacher content knowledge plays in raising student achievement (Hill et al., 2005; Kanter & Konstantopoulos, 2010).

Pedagogical Content Knowledge

Pedagogical Content Knowledge (PCK) is used to describe the knowledge that teachers use to transform particular subject matter for student learning. We are guided by the definition of PCK as identified in the Revised Consensus Model (RCM) of PCK (Carlson & Daehler, 2019). This model identifies three distinct realms of knowledge that teachers have that ultimately mediate student outcomes: 1) *collective PCK,* which is described as the specialized professional knowledge held by educators in the field; 2) *personal PCK,* which is the cumulative and procedural pedagogical content knowledge and skills of an individual teacher; and 3) *enacted PCK,* which refers to a teacher's practice of engaging with teaching during planning, instruction, and reflection on instruction and student outcomes.

Attitudes and Beliefs

Attitudes and beliefs are amplifiers to how teachers develop their personal PCK. The literature on attitudes and beliefs documents the connection between 1) teacher attitudes and beliefs, and 2) teachers' thought process, classroom practices, change, and pedagogical practices used to teach (Porter and Freeman, 1985, as cited in Pajares, 1992). Additionally, attitudes and beliefs shape the way teachers react, and choose to respond to reforms (Jones & Carter, 2013). We hypothesize that with the Making Sense of SCIENCE professional learning courses and PLCs, we can expect to see some shifts in teachers' implicit knowledge and beliefs about students, teaching as identified by the constructs below. The first three constructs below are explicitly supported by Making Sense of SCIENCE professional learning, and the remaining constructs posit more distant expected teacher outcomes.

- *Belief that students are capable learners* — Literature suggests that teacher expectations of student abilities and the changeability of student abilities and their potential interacts with their behavior in the classroom. The National Science Education Standards (National Research Council, 1996) are based on five key assumptions, one of which is "Actions of teachers are deeply influenced by their understanding of and relationships with students."
- *Philosophically aligned with standards* — For teachers to be able to make the shifts required by three-dimensional science standards, they need to develop themselves, and understand that students need deeper understanding of science and engineering content through making sense of phenomena and designing solutions. Students also need opportunities to integrate science content and practices. In order for teachers to guide students to making sense of phenomena, teachers need to see a) their role shift to being a facilitator in learning, and b) students taking on the process of learning like scientists and engineers through active exploration and sense-making processes.
- *Values being a reflective practitioner* — Science teachers need to engage in reflective practices to assess their teaching of science as promoted by the National Science Education Standards: Teachers of science engage in ongoing assessment of their teaching of student learning. In doing this, teachers use student data, observations of teaching, and interactions with colleagues to reflect on and improve teaching practice (National Research Council, 1996). Making Sense of SCIENCE supports teacher reflective cycles of their practice by examining student data and interacting with other teachers through PLCs.
- *Confidence* — As elementary science teachers often express severe lack of confidence in science teaching (Murphy, et al., 2007), science professional learning is hypothesized to impact teacher confidence to:
    - teach science;
    - teach with science instructional practices; and
    - support literacy.

- *Self-efficacy* — When teachers become knowledgeable about a particular subject, it increases their belief in his or her capability to organize and execute courses of action required to successfully accomplish a specific teaching task in a particular context (Tschannen-Moran et al., 1998). Teachers who have higher self-efficacy, content knowledge, and attitudes have students with higher achievement than do teachers who have lower levels of self-efficacy (Evans, 2011).
- *Agency in the classroom* — Teachers who are given autonomy to teach science by supportive districts and administrators have the capacity to act intentionally in setting instructional goals in their classrooms (Calvert, 2016).
- *Agency in science leadership* — The capacity of teachers to act purposefully to direct their professional growth and contribute to the growth of their colleagues depends on a teacher's internal traits and supportive structural conditions that support professional learning (Calvert, 2016).
- *Professional aspirations* — The extent to which teachers stay in a school or school system as teachers and their aspirations to pursue leadership positions are influenced by school cultures that value and respect teachers and develop teaching and leadership expertise (Cameron & Lovett, 2015).

School Climate

We posit that when Making Sense of SCIENCE works with state and regional coordinators, districts, and school principals through our partnership and leadership development offerings, we can see positive changes that trickle down to create a positive district and school climate that is conducive for science teaching and learning.

District Support

An essential element of reform in science education is district support for science teaching and learning. We posit that Making Sense of SCIENCE contributes to the following improvements at the district level.

- *Providing guidelines on science instruction* — District guidelines that outline the expected shifts to happen in elementary science including developing curriculum frameworks, evaluation criteria for instructional materials in science, and outlining the scope and sequence for science teaching science.
- *Allocating resources for professional learning in science* — District guidelines that allocate coaching resources, professional learning time, teacher pay, or substitute time for science professional learning.
- *Allocating resources for science materials* — District leaders make investments and allocate resources to purchase science curriculum, instructional materials, laboratory equipment, and technology supports for science teaching and learning.

- *Prioritizing support for science learning* — Superintendents and other district leaders signal the importance of science education by outlining guidelines for time on science and putting science on the agenda for administrators to take to their school sites.
- *Participating in science-related conversations/activities* — Superintendents and other district leaders actively attend meetings to get informed on standards-based science implementation and take part in discussions that shape science education.
- *Involving teachers in district science decisions* — Superintendents and district leaders actively invite teachers to create or provide input on science standards-based implementation in Local Control and Accountability Plans and involve them in the selection of instructional materials.
- *Building capacity for science professional learning* — Superintendents and other district leaders invest in building leadership capacity and material support for teacher professional learning in science.

Administrative Support

We posit that Making Sense of SCIENCE contributes to following improvements at the administrator level.

- *Providing science resources and supplies* — Administrators approve teacher requests and increase the availability of science resources and supplies in a school (Ivelandet al., 2017).
- *Supporting teacher collaboration* — Administrators forge the conditions that make PLCs a priority by changing the structure of the school day, and providing the financial support needed to make PLCs happen (Iveland et al., 2017).
- *Acting as an instructional leader* — School principals can play an important role as instructional leaders when they spend time to support and collaborate with other teachers on science content and instruction. When administrators participate in professional learning alongside teachers, they are more likely to support and compel teachers to improve their practice and to learn new skills (Jenkins, 2009; Casey et al., 2012).
- *Prioritizing support for science learning and teacher professional learning* — When school principals and administrators participate in professional learning alongside teachers, they are more likely to allocate more time for science instruction, extracurricular science activities, and teacher collaboration (Iveland et al., 2017) and allocate time and resources for teacher professional learning in science.
- *Involving teachers in science leadership* — Principal actions and the relationship amongst adults in a school are determining factors in developing sustaining science leadership, particularly among teachers. When principals empower teachers to take on additional science leadership responsibilities, teachers are able to lead within and beyond the classroom, identify with and contribute to a community of teacher learners, and influence others towards improved instructional practice (Katzenmeyer & Moller, 2009).

Collaboration

Sustained, job-embedded, and collaborative teacher learning can occur in PLCs. In PLCs, teachers collaborate and work together in continual dialogues to examine their practice and student performance and to develop and implement more effective instructional practices (Darling-Hammond & Richardson, 2009). We posit that when teachers and administrators participate in Making Sense of SCIENCE, it contributes to improvement in the following areas.

- *Teacher-to-teacher collaboration* — When teachers collaborate in functional PLCs, they allow for teachers to take risks in teaching and changes in instruction that are reform-oriented and student-centered (Briscoe & Peters, 1997; Brahier & Schäffner, 2004). The expected changes associated with PLCs result from increases in the amount of time allocated for collaboration in PLCs and the type of substantive activities that teachers engage in the PLCs around content learning and instruction (Graham, 2007).
- *Administrator-to-teacher collaboration* — Similar to collaboration between peers, when administrators gradually take on the role of instructional leaders and increase the amount and type of substantive activities in which they collaborate with teachers in PLCs, teachers become encouraged to take on risks and change their instruction towards reform-oriented practices (Urick et al., 2018).

School Culture

A positive school culture promotes cooperative learning, group cohesion, respect, and mutual trust which can directly improve a school's learning environment (Thapa et al., 2013). With the two-pronged approach of Making Sense of SCIENCE in providing teacher and leadership professional learning, we posit to see improvements at the distal level at the school level.

- L*earning climate* — Schools experience changes towards a conducive learning climate that is student-centered and endorses ambitious academic work coupled with adequate support for all students (Bryk, 2010).
- *Trust and respect among peers and among peers and administrators* — In schools where teachers and teachers and administrators increasingly trust and respect each other, learning becomes conducive for both teachers and students. Principals supportive of science as a priority play a critical role in influencing the levels of trust and respect between teachers (Hallam et al., 2015).

Opportunity to Learn Science in the Classroom

Opportunity to learn is a multi-dimensional construct central to quality teaching and a prerequisite to student achievement (Elliott & Bartlett, 2016). It is composed of the amount of instructional time on science and the content taught; instructional quality of science that reflects the shifts in three-dimensional science standards. Conducive classroom cultures also facilitate student-centered learning of science.

We posit that students with Making Sense of SCIENCE teachers are likely to see the following changes in their opportunities to learn science in the classroom.

Time on Science

Measures of instructional time in literature have been grouped into four ranges—years, days, hours, and minutes (Frederick & Walberg 1980). In our definition of instructional time, we define time on science by the time allocated to science in minutes and the frequency in which it is taught during the week.

- Amount of time on science in minutes of science learning per week and integrated science-literacy time
- Frequency of the amount of time science is taught per week and per year

The types of tasks and activities teachers use to engage students in science look different in an NGSS-aligned classroom (Tekkumru-Kisa et al., 2020). Consequently, we hypothesize to see teachers' shifts in the types of instructional tasks assigned and the content students engage with that are aligned with three-dimensional learning as listed under the ***Instructional Changes*** and ***Content of Science Taught*** sections below.

Instructional Changes

- *Sense-making of hands-on investigations* — Sensemaking is the process that students and teachers undertake to think together and to make sense of what things mean. When students conduct investigations and produce and/or come up with data, they work together to analyze this data by looking for patterns and relationships to develop explanations and models (McNeill et al., 2015)
- *Engaging in scientific argumentation* — Our definition is aligned with the National Research Council (2013), which outlines that when students engage in scientific argumentation, they are expected to listen to, compare, and evaluate competing ideas and methods based on their merits.
- *Explaining ideas and phenomena* — Phenomena are events that are observable and repeatable and can be explained or predicted using science knowledge. The instructional shifts required by three-dimensional standards use phenomena as the starting point for learning. Students are taught to develop ideas, based on evidence, to explain phenomena (Achieve, 2017).
- *Integration of science and literacy* — Literacy is the ability to read, write, and engage with scientific texts because when students engage in these activities, they are able to deepen their conceptual understanding of science (Cervetti et al., 2012).
- *Integration of science and mathematics* — Our definition is aligned with the National Research Council (2013), which outlines that the integration of mathematics is fundamental in providing students with opportunities to engage in a range of tasks such as constructing simulations; statistically analyzing data; and recognizing, expressing, and applying quantitative relationships.
- *Participating in collaborative discourse* — When teachers create classroom discourse structures, they enable both the students and the teacher to engage in collaborative knowledge building.

These collaborative processes also use discourse structures that move away from the IRE usual mode of classroom discourse, in which the teachers follow the pattern of *initiating, responding to, and evaluating (IRE)* responses (Hmelo-Silver & Barrows, 2008).

- *Reflecting on learning* — Metacognitive inquiries and formative assessment practices are powerful learning tools. Metacognition is defined in terms of student understanding of their processes of learning, in terms of how and what they learn. When students engage in metacognitive discourse, they engage in the process of making explicit their tacit reasoning and problem-solving strategies (Greenleaf et al., 2011). Formative assessments also help students understand their learning, but it is important for both teachers and students to reflect on what they learn about student understanding.
- *Participating in cognitively challenging tasks* — Cognitively challenging tasks refer to the depth of student engagement in conceptual thinking. We are guided by the definition of Elliott & Bartlett (2016) which prescribes that teachers must dedicate instructional time to addressing a range of cognitive processes, instructional practices, and grouping formats when covering content.

Content of Science Taught

- Standards-aligned science concepts in science, life, and physical science disciplinary core ideas are taught with breadth and depth.
- Science practices of developing and using models, arguing from evidence, constructing explanations, and analyzing data and representations of data are taught with breadth and depth.
- Cross-cutting concepts of cause and effect, energy and matter, and systems and systems models are taught with breadth and depth — The identified cross-cutting concepts provide the connections and tools to understand the science concepts taught in Making Sense of SCIENCE in science, life, and physical sciences.
- Literacy skills related to science-specific ways of reading, writing, and engaging in scientific discourse are taught in breadth and depth — According to the National Science Education Standards, scientific literacy means that a person can ask questions, and is able to read about, describe, explain, and write about natural phenomena. For students to acquire science-specific literacy skills, they need to learn and observe how to read, write, and engage in discourse using science-specific conventions that model how scientists work every day (NRC, 2013; Wright et. al., 2016).

Conducive Classroom Cultures

Classroom culture is influenced by teacher attitudes and approaches, and teacher participation in professional learning is linked to investigative classroom cultures (Supovitz & Turner 2000). We posit that students with Making Sense of SCIENCE teachers will show improvements in the following characteristics of conducive classroom cultures.

- *Student-centered learning* — Student-centered classrooms are characterized by teachers who know and communicate that they do not need to know everything and who value student ideas in making sense of phenomena. Student-centered classrooms are also characterized by sustained engagement with student questions and ideas. These classrooms are characterized by a safe classroom culture, in which students and teachers celebrate risk taking in learning.
- *Student agency* — We define agency as students' choice and capacity to take responsibility for their own learning. Classroom cultures also promote student agency when students feel that they can share ideas without being held up for ridicule and recognize the dialogical opportunities available when they consider and value each other's ideas in the process of learning (Cavagnetto et al., 2020).
- *High expectations of students* — When teachers raise their expectations and increase the rigor of their instructions, they facilitate student learning. Teacher expectations also contribute to the whole-class teaching environment through grouping choices, a continuum of cognitively challenging tasks, and student agency in what they learn (Rubie, 2009).
- *Environment conducive to learning with appropriate classroom management* — Classroom management plays an important role in creating a safe and conducive learning environment for learning science. Making Sense of SCIENCE staff model how inquiry-based learning can be facilitated in the classroom during teacher professional learning.
- *Active student engagement* — When students are actively engaged in a classroom, they are seen participating in discussion and showing understanding of the purpose of the lesson goals.

Student Achievement and Attitudes

We posit that Making Sense of SCIENCE has a distal impact on student achievement, student attitudes, and dispositions towards science. Specifically, we hypothesize seeing improvements in the following student outcomes.

Science and English Language Arts Achievement

- *Science knowledge* — improved student science content knowledge in earth and physical sciences
- *Science practices* —practices used by scientists as they investigate models and build theories about the world (National Research Council, 2012). We are particularly interested in looking at how Making Sense of SCIENCE improves student skill with developing and using models; arguing from evidence; constructing explanations; and analyzing data and representations of data.
- *Communicating science ideas* —reading and writing are essential skills in science (National Research Council, 2012). Making Sense of SCIENCE improves student skills in communicating science ideas through writing and sustained productive scientific discourse.
- *English Language Arts Achievement* — improved student achievement in reading, writing, speaking and listening

Student Attitudes

- *Aspirations* — improved student dispositions and attitudes towards science and the development of an interest in pursuing a career in science or science related work (Tytler & Osborne, 2012)
- *Self-efficacy* — students who judge themselves to be efficacious in science and their academic capabilities in science also foster a sense of efficacy to pursue careers in science (Bandura et al., 2001).
- *Agency in learning* —students' choice and capacity to take responsibility for their own learning. Classroom cultures also promote student agency when students feel that they can share ideas without being held up for ridicule and recognize the dialogical opportunities available when they consider and value each other's ideas in the process of learning (Cavagnetto et al., 2020).
- *Enjoyment of science* — greater student enjoyment of science is found to be predictive of students' interest in engaging further with science topics (Ainley & Ainley, 2011).

Appendix B. Survey Scales of Teacher Attitudes and Beliefs, Opportunities to Learn, and School Climate

This appendix provides information on the 30 intermediate outcomes analyzed and reported in Chapter 5. For each outcome, we list the survey items that were used to construct the outcome, the scale of the items, the data source, the number of items, the resulting Cronbach alphas, and the method used to create the outcome.

TABLE B1. COMPOSITE CREATION FOR INTERMEDIATE OUTCOMES

Outcome on which to assess impact and items included	Scale of items (post recoding)	Data source	No. of items	Cronbach's alpha	Method of composite creation
Belief that students are capable learners					
• The majority of my students are capable of learning rigorous science even if they come from a challenging environment. • The majority of my students are capable of going to science-related careers. • Given the right supports, my low performing students typically are able to learn challenging science concepts.	5-pt; agree	Tch: W 18	3	0.853	Average over items
Philosophically aligned with NGSS					
• To what extent do you think that NGSS is aligned with your pedagogical practices?	5-pt; align	Tch: Spr 18	1	NA	Average over items
Values life-long learning					
• I am confident I can learn science given the right support. • I frequently seek out information or learning opportunities to strengthen my teaching. • I actively seek input from colleagues to improve my teaching. • It's okay if I don't feel confident in science. I can build off of my current understanding.	5-pt; agree	Tch: W 18	4	0.612	Average over items
Confidence in addressing student performance expectations					
• Analyzing and interpreting data from maps to describe patterns of Earth's features • Using evidence to construct an explanation relating the speed of an object to the energy of that object • Developing a model of waves to describe patterns in terms of amplitude and wavelength and that waves can cause objects to move	5-pt; conf	Tch: Spr 18	6	0.887	Average over items

TABLE B1. COMPOSITE CREATION FOR INTERMEDIATE OUTCOMES

Outcome on which to assess impact and items included	Scale of items (post recoding)	Data source	No. of items	Cronbach's alpha	Method of composite creation
• Developing a model using an example to describe ways the geosphere, biosphere, hydrosphere, and/or atmosphere interact • Developing a model to describe that matter is made of particles too small to be seen • Supporting an argument that the gravitational force exerted by Earth on objects is directed down					
Confidence in science instructional practices					
• Teach science in engaging ways • Teach students to do hands-on science activities or investigations • Get students to use scientific terms accurately • Teach students to collect data • Teach students to represent data (e.g., graphs, images, simulations, physical models) • Teach students to identify evidence or data that support a claim • Use a variety of models (e.g., graphs, images, simulations, physical models) to support students' science learning • Have students use existing models (e.g., graphs, images, simulations, physical models) to explain something that has been observed • Help students develop their own models to explain a phenomenon • Help both your high and low achieving students learn challenging science • Get students to reflect on their learning and then to revise their thinking • Help students understand the world in terms of interacting systems • Foster discussions among students that help them learn science • Explicitly teach students how to have productive science conversations that are grounded in evidence • Teach science in a way that meets the NGSS expectations	5-pt; conf	Tch: Spr 18	15	0.969	Average over items

TABLE B1. COMPOSITE CREATION FOR INTERMEDIATE OUTCOMES

Outcome on which to assess impact and items included	Scale of items (post recoding)	Data source	No. of items	Cronbach's alpha	Method of composite creation
Confidence in supporting literacy in science					
• Assign writing tasks that help students learn science • Help students understand how to use reading strategies to make sense of science texts • Help students communicate science ideas in writing • Explicitly teach students how to read complex informational texts that include graphs, diagrams, symbols, and data tables • Explicitly teach students how to write scientific explanations • Teach students to articulate clear, convincing reasons for their answers	5-pt; conf	Tch: Spr 18	6	0.947	Average over items
Self-efficacy					
• I understand science concepts well enough to be effective in teaching elementary students. • I am typically able to answer students' questions related to the science they are studying. • I am typically able to respond effectively to students' ideas about most science topics. • I am effective at explaining to students scientific reasons for outcomes of science experiments. • I am skilled at identifying what science concepts my students find confusing.	5-pt; agree	Tch: Spr 18	5	0.925	Average over items
Agency in the classroom					
• Setting performance standards for students • Selecting science curriculum • Determining the pedagogical techniques that you use to teach science • Choosing criteria for grading student performance • Amount of time science is taught • Pacing of science instruction	5-pt; infl	Tch: W 18	6	0.840	Average over items

TABLE B1. COMPOSITE CREATION FOR INTERMEDIATE OUTCOMES

Outcome on which to assess impact and items included	Scale of items (post recoding)	Data source	No. of items	Cronbach's alpha	Method of composite creation
Amount of time spent on science instruction					
• Think about the times you spent teaching science during these past four school weeks. Approximately, how many total hours of science did you teach per week in those weeks?	Number of hours	Tch: F/W/Spr 18	1	NA	For each survey, summed across 4 weeks; Average over the three survey
Sensemaking of hands-on investigations					
• Working collaboratively in small groups • Engaging in hands-on science activity • Making a claim based on a hands-on activity or data	5-pt; emph	Tch: F/W/Spr 18	3	0.852 - Fl 0.767 - W 0.852 - Spr	Average over items
Integration of science and literacy					
• Constructing a written scientific explanation for a "how" or "why" question • Constructing a verbal scientific explanation for a "how" or "why" question • Listening to and building on other peoples' ideas • Writing to support learning from a hands-on activity • Reading to support learning from a hands-on activity • Discussing to support learning from a hands-on activity	5-pt; emph	Tch: F/W/Spr 18	6	0.890 - Fl 0.882 - W 0.891 - Spr	Average over items
Participating in collaborative discourse					
• Listening to and building on other peoples' ideas • Discussing to support learning from a hands-on activity	5-pt; emph	Tch: F/W/Spr 18	2	0.738 - Fl 0.710 - W 0.724 - Spr	Average over items

TABLE B1. COMPOSITE CREATION FOR INTERMEDIATE OUTCOMES

Outcome on which to assess impact and items included	Scale of items (post recoding)	Data source	No. of items	Cronbach's alpha	Method of composite creation
Explaining ideas and phenomena					
• Exploring real-world phenomena • Making sense of science ideas	5-pt; emph	Tch: F/W/Spr 18	2	0.694 - Fl 0.700 - W 0.861 - Spr	Average over items
NGSS-aligned ES topics (teacher only) (DCIs): Earth's place in the universe					
• Evidence of change in landscape over time • Relationship between fossils and rock layers • Explaining the brightness of the Sun relative to other stars • Explaining day and night • Explaining changing positions of the Sun, moon, and stars	3-pt; Did teach	Tch: Spr 18	5	0.788	Average over items
NGSS-aligned ES topics (teacher only) (DCIs): Earth's systems					
• Effects of weathering • Factors affecting rates of erosion • Defining Earth's systems (e.g., atmosphere, biosphere, geosphere, hydrosphere) • How living things affect their physical environments • Interactions affecting Earth's systems (e.g., landforms, climate, weather) • Interpreting maps of Earth's features (e.g., mountains, ocean trenches, volcanoes) • Distribution of water on Earth (e.g., oceans, glaciers, atmosphere)	3-pt; Did teach	Tch: Spr 18	7	0.813	Average over items
NGSS-aligned ES topics (teacher only) (DCIs): Earth and human activity					
• Renewable and nonrenewable resources • Types of natural hazards • Human impact on Earth systems	3-pt; Did teach	Tch: Spr 18	3	0.866	Average over items

TABLE B1. COMPOSITE CREATION FOR INTERMEDIATE OUTCOMES

Outcome on which to assess impact and items included	Scale of items (post recoding)	Data source	No. of items	Cronbach's alpha	Method of composite creation
NGSS-aligned PS topics (teacher only) (DCIs): Motion and stability - forces and interactions					
• How unbalanced forces affect motion • Using patterns in motion to predict future motion • Contact forces between objects • Factors that affect the size of electric and magnetic forces • Direction of gravitational force	3-pt; Did teach	Tch: Spr 18	5	0.855	Average over items
NGSS-aligned PS topics (teacher only) (DCIs): Definitions of energy					
• Energy associated with objects moving at different speeds • Energy associated with sound, light, and electrical currents	3-pt; Did teach	Tch: Spr 18	2	0.744	Average over items
NGSS-aligned PS topics (teacher only) (DCIs): Conservation of energy and energy transfer					
• Transfer of light energy • Transfer of electrical energy • Transfer of energy when objects collide • Change in motion when objects collide • Transfer of energy from the Sun to plants to animals • Conversion of stored energy to other types • Conservation of energy	3-pt; Did teach	Tch: Spr 18	7	0.874	Average over items
NGSS-aligned PS topics (teacher only) (DCIs): Waves					
• Defining waves • Amplitude and wavelength • Light and observing objects • Transmitting digitized information	3-pt; Did teach	Tch: Spr 18	4	0.917	Average over items

TABLE B1. COMPOSITE CREATION FOR INTERMEDIATE OUTCOMES

Outcome on which to assess impact and items included	Scale of items (post recoding)	Data source	No. of items	Cronbach's alpha	Method of composite creation
NGSS-aligned PS topics (teacher only) (DCIs): Matter and its interactions					
• Particulate nature of matter • Identifying substances based on their properties • Identifying chemical reactions • Conservation of matter	3-pt; Did teach	Tch: Spr 18	4	0.911	Average over items
Science and Engineering Practices (SEPs) (teacher only)					
• Asking and defining problems • Developing and using models • Planning and carrying out investigations • Analyzing and interpreting data • Using mathematics and computational thinking • Constructing explanations and designing solutions • Engaging in argument from evidence • Obtaining, evaluating, and communicating information	3-pt; Did teach	Tch: Spr 18	8	0.896	Average over items
CCCs (teacher only)					
• Patterns • Cause and effect: mechanism and explanations • Scale, proportion, and quantity • Systems and system models • Energy and matter: flows, cycles, and conservation • Structure and function • Stability and change	3-pt; Did teach	Tch: Spr 18	7	0.890	Average over items

TABLE B1. COMPOSITE CREATION FOR INTERMEDIATE OUTCOMES

Outcome on which to assess impact and items included	Scale of items (post recoding)	Data source	No. of items	Cronbach's alpha	Method of composite creation
Teacher collaboration - amount					
• Approximately how much total time, if any, in the past four school weeks did you participate in informal peer collaboration for science instruction (for example, sharing lesson plans or resources, discussing student work, informally observing a colleague's science lesson etc.)?	Select 1 of 4 or 5 options	Tch: Fl, W, Spr 18	1	NA	Average across 3 surveys
Culture of peer collaboration					
• Collaboration happens organically among teachers. • Teachers find peer collaboration helpful.	5-pt; agree	Tch: Spr 18	2	0.707	Average across items
Trust and respect among teachers					
• Teachers trust each other in this school. • Teachers regularly observe each other teaching classes. • It's okay to discuss feelings, worries, and frustrations with other teachers. • Teachers are supported by their colleagues to try out new ideas in teaching. • Teachers respect other teachers who take the lead in school improvement efforts. • Coaches and/or mentors are respected by teachers at this school.	5-pt; agree	Tch: W 18	6	0.815	Average across items
Trust and respect between teachers and administrators					
• There is an atmosphere of trust and mutual respect between teachers and school administrators. • Teachers feel comfortable raising issues and concerns with school administrators. • The school administration consistently supports teachers.	5-pt; agree	Tch: W 18	3	0.906	Average across items
Supporting teacher collaboration					
• Peer collaboration is supported by administrators at my school.	5-pt; agree	Tch: Spr 18	1	NA	None needed

TABLE B1. COMPOSITE CREATION FOR INTERMEDIATE OUTCOMES

Outcome on which to assess impact and items included	Scale of items (post recoding)	Data source	No. of items	Cronbach's alpha	Method of composite creation
Prioritizing support for teacher professional learning in science					
• Our principal/assistant principal provides support for professional learning. • Our principal/assistant principal provides the support teachers need to improve our science instruction.	5-pt; agree	Tch, W 18	2	0.752	Average across items
Administrator support involving teachers in science leadership					
• Teachers are relied upon to make decisions about educational issues. • Teachers are encouraged to participate in school leadership roles (e.g., leader of a professional learning community (PLC), mentor, member of the School Improvement Team).	5-pt; agree	Tch: W 18	2	0.675	Average across items

Note. Under the *Data Source* column: Tch = Teacher, F = fall, W = winter, Spr = spring, 18 indicates that the survey was administered in Year 2 (2017–18).

Appendix C. Teacher Content Knowledge Assessment: Item-Level Information

This appendix presents additional information about the teacher content knowledge assessment, including the source, brief description, proportion correct, biserial correlation, and item difficulty and discrimination for a 2-Parameter Logistic (2PL) model.

TABLE C1. ITEM-LEVEL INFORMATION FOR THE TEACHER CONTENT KNOWLEDGE ASSESSMENT

Item ID	Proportion correct	Biserial correlation	2PL difficulty	2PL discrimination	Source	Description
1	.797	.317	-1.800	0.871	NY Regents ES June 2016	Water cycle/Energy (diagram)
2	.949	.393	-1.946	2.977	NY Regents ES June 2016	Water cycle/Process (diagram)
3	.653	.293	-0.880	0.816	MOSART Physics	Gravity on objects
4	.703	.163	-2.437	0.365	MCAS	compare two waves
5	.881	.151	-3.311	0.653	NY Regents ES Jan 2017	Reflecting insolation

TABLE C1. ITEM-LEVEL INFORMATION FOR THE TEACHER CONTENT KNOWLEDGE ASSESSMENT

Item ID	Proportion correct	Biserial correlation	2PL difficulty	2PL discrimination	Source	Description
6	.788	.405	-1.399	1.178	MOSART Astronomy	Earth's rotation
7	.873	.262	-2.107	1.109	MCAS	ID which part is wavelength
8	.720	.246	-1.289	0.840	MOSART Earth Science	Evaporation
9	.551	.425	-0.190	1.415	MOSART Physical Science	Density
10	.881	.175	-3.821	0.555	MOSART Earth Science	Mountains/tectonics
11	.254	.327	1.165	1.164	MOSART Physics	Transfer of energy/open system
12	.534	.166	-0.362	0.387	MOSART Astronomy	Sun & ice/temp
13	.525	.270	-0.150	0.742	NY Regents ES June 2016	Heat transfer through conduction (diagram)
14	.305	.207	1.770	0.491	NY Regents PS June 2015	Transfer of energy in a system (block on table)
15	.949	.168	-4.787	0.651	NY Regents ES June 2016	Intensity of insolation (diagram)
16	.881	.323	-2.056	1.218	NAEP	Ice melts
17	.831	.411	-1.557	1.334	MCAS	Phase change/physical change
18	.534	.524	-0.097	2.150	NY Regents ES June 2016	Molecules
19	.576	.250	-0.552	0.600	NY Regents ES June 2016	Sun movement
20	.881	.408	-1.725	1.652	NAEP	Sun warms water
21	.636	.348	-0.792	0.796	NY Regents PS June 2016	Light waves
22	.466	.162	0.309	0.466	MOSART Physical Science	Chemical change
23	.415	.224	0.704	0.518	NY Regents PS June 2016	Diagram path of ball thrown through the air.

TABLE C1. ITEM-LEVEL INFORMATION FOR THE TEACHER CONTENT KNOWLEDGE ASSESSMENT

Item ID	Proportion correct	Biserial correlation	2PL difficulty	2PL discrimination	Source	Description
24	.619	.308	-0.607	0.938	NY Regents ES Jan 2017	Heat transfer (diagram)
25	.839	.328	-1.936	1.013	NECAP	temp of spoons in water--heat transfer
26	.432	.347	0.376	0.850	MOSART Chemistry	Molecular structure, phys change
27	.729	.407	-1.021	1.246	MOSART Earth Science	Earth closed system
28	.924	.236	-3.007	0.950	NECAP	hammer nail, energy transfer
29	.619	.300	-0.819	0.644	NY Regents PS June 2016	acceleration due to gravity (graph)

Note. 2PL = 2-Parameter Logistic Item Response Theory score calibration; ES = Earth science; PS = physical science; NECAP = The New England Common Assessment Program; NAEP = National Assessment of Educational Progress; NY = New York; MOSART = Misconception-Oriented Standards-Based Assessment Resources for Teachers; MCAS = Massachusetts Comprehensive Assessment System

Appendix D. Assessment of Student Science Achievement: Construction, Forms, Administration, Item Statistics, and Approaches to Scaling

This appendix provides the details on the student science achievement assessment.[1] We describe the assessment's construction, test forms, administration, and approaches to scaling. We also provide select item-level statistics based on Classical Test Theory (CTT) and Item Response Theory (IRT). In describing the test forms, we provide brief descriptions of other types of items that were included in the science assessment, such as the constructed-response items and student survey scales, in order to fully present what was asked of students in spring Year 2 (2017–18).[2]

Construction of the Student Science Assessment

This evaluation was conducted in the 2016–17 and 2017–18 school years, just three years after the Next Generation Science Standards (NGSS) were rolled out. Therefore, we faced the immense challenge of finding an established NGSS-aligned assessment to evaluate impacts of Making Sense of SCIENCE on student science achievement in Grades 4 and 5.

In the summer of 2014–2015 and throughout the 2015–16 school year, we conducted a search for NGSS-aligned instruments to measure student science achievement. We short-listed potential instruments and reached out to several assessment developers, including 1) Education Testing Service (ETS) about the Cognitively Based Assessment of, for, and as Learning (CBAL®), 2) Northwest Evaluation Association (NWEA) for their Measures of Academic Progress (MAP) assessment, and 3) the California Department of Education about their new NGSS-aligned science assessment. We found that these instruments were not far along enough in the development process. In fall 2015–2016, at the suggestion the Making Sense of SCIENCE Technical Working Group (TWG), we opened discussions with a university-based center that at the time was partnering with the state department of education to administer an NGSS-aligned science assessment. When we approached the center, the science assessment had been field tested the prior school year (2015–16) and was operational in 2016–17. The study team made a few adjustments to the assessment to be suitable for administration in this study, such as supplementing grade-appropriate items for students in the study and shortening the test in order to administer it within one hour. Prior to the administration of the assessment, but when it was too late to change course, evaluators and program developers were provided the full assessment—as opposed to just the sample of items that we were shown previously—for review. The team then recognized that the assessment was inadequate due to the inaccuracies in the science content and the verboseness of the questions, which would have been especially problematic for English learner students in the study. Despite this recognition, the lack of

[1] Note that the "student science achievement assessment" refers to the selected-response items of the general science assessment administered to students, which included both selected-response and constructed-response items.

[2] This appendix focuses on the selected-response items of the student science assessment. For item-level statistics of the constructed-response items, contact the developers (Heller Research Associates).

options at the time compelled the team to administer the test to students in the spring of Year 1 (2016–17). In the fall of Year 2 (2017–18), we faced a difficult decision: either to go with an assessment in its second year of operation that was problematic in ways described above, or to proceed with developing an assessment with no opportunity to pilot. Without yet knowing the result from the exploratory year, evaluators and program developers jointly decided to not continue using the assessment and to instead develop an assessment with the guidance of TWG members.

The study team constructed the student science assessment using selected-response items from publicly available sources and constructed-response items developed by HRA. The selected-response items originated from the following publicly-available tests and item banks: Massachusetts Comprehensive Assessment System (MCAS), The New England Common Assessment Program (NECAP), the Misconceptions-Oriented Standards-Based Assessment Resources for Teachers (MOSART), American Association for the Advancement of Science (AAAS), and National Assessment of Educational Progress (NAEP). Two external reviewers with content and test-development expertise reviewed items for accuracy, clarity, and alignment with NGSS. To preserve the original items as much as possible, we asked the external reviewers to select, reject, or suggest only minor revisions to each item. We only made minor revisions to selected items.

Test Forms of the Student Science Assessment

We ultimately assembled a pool of 49 selected-response items aligned with fourth and fifth grade standards in Earth and space science (10 for fourth grade, 9 for fifth grade), physical science (10 for fourth grade, 10 for 5th grade), and scientific inquiry (10 items common to fourth and fifth grade).

The basic organization of test forms is displayed in Figure D1.

Forms A-type (A1 – A4) were administered to fourth graders and contain the same 30 selected-response items across the four forms. Forms A1 – A4 are differentiated in terms of the survey questions they ask. Forms A-type include 10 selected-response inquiry items that are common across all 16 forms of the assessment.

Forms B-type (B1 – B4) were administered to fifth graders and contain the same 29 selected-response items across the four forms. Forms B1 – B4 are differentiated in terms of the survey questions they ask. Forms B-type include 10 common selected-response inquiry items (the same as those in the fourth-grade A-type forms).

A random sample of about 20% of fifth-grade students received the remaining eight forms (Forms C to J). These forms contained specific combinations of constructed-response items designed to test "communication of science ideas in writing", as discussed in Chapter 8 and Appendix R. A random sample of fourth grade students received form E, the only form with constructed-response items deemed appropriate for fourth graders. All students responding to forms C to J also responded to the 10 common selected-response inquiry items.

The 10 inquiry items were included with the goal of potentially linking science scores across all form types. Over the course of the project, we determined that selected- and constructed-response items were measuring different skills and should not be combined. Two of the inquiry items were problematic (see next subsection), and three of the inquiry items tapped the life science content strand. This made the set inappropriate for linking Earth and space science and physical science scores across grades.

TABLE D1. FORMS FOR THE STUDENT SCIENCE ASSESSMENT AND SURVEY

			Selected-response items			Survey items				Constructed-response items							
Form		Grades administered	4th grade (20 items)	5th grade (19 items)	4th and 5th inquiry (10 items)	Set A	Set B	Set C	Set D	CR 1	CR 2	CR 3	CR 4	CR 5	CR 6	CR 7	CR 8
A	A1	4	■		■	■											
	A2	4	■		■		■										
	A3	4	■		■			■									
	A4	4	■		■				■								
B	B1	5		■	■	■											
	B2	5		■	■		■										
	B3	5		■	■			■									
	B4	5		■	■				■								
C		5			■					■		■			■	■	
D		5			■					■			■		■		■
E		4,5			■					■				■	■	■	
F		5			■						■	■			■		■
G		5			■						■		■		■	■	
H		5			■						■			■	■		■
I		5			■					■		■		■		■	
J		5			■						■		■		■		■

Administration of the Student Science Assessment

In spring Year 2 (2017–2018), teachers administered the student science assessment and survey to their students. Students received computer access and a 60-minute period to complete the assessment. Per district requests, we emphasized to teachers that all required district and state testing must be prioritized over the Making Sense of SCIENCE assessment. We closely monitored teachers' completion progress and sent periodic reminders.

The assessment was administered on the Quest platform, an online testing system developed by the 3-C Institute for Social Development. We selected the Quest platform because its design incorporates Universal Design principles, including accommodations such as text-to-speech. We wanted to include the text-to-speech feature in order to accommodate English learner students, students with reading disabilities, and or students with limited literacy. Therefore, we provided class sets of headphones to teachers who needed them for their students to use during testing. Additionally, we asked teachers to not administer the science assessment and survey to students who take the alternative or modified state assessments. For students who need testing accommodation (other than voice-over) per their Individualized Education Programs, we asked teachers to use their discretion in deciding whether it was feasible to test such students.

Identifying Items to be Removed Prior to Impact Analysis

After we collected the student science achievement assessment data, an initial analysis of the items led us to remove several items prior to further analysis. We removed 3 life science items from the 10 inquiry items because Making Sense of SCIENCE did not address this content strand. We also removed one item because of an abnormally high level of non-response and one item because one of the incorrect response options was a strong distractor selected by many students. Both of these items led to instability of item calibration using IRT, so we removed them from the assessment for both grades. Analysis of Differential Item Functioning (DIF), conducted by an independent contractor, revealed none of the items to be problematic. Therefore, we removed no additional items. The final forms included 25 selected-response items in Grade 4 and 24 selected-response items in Grade 5. We used these items to estimate achievement.

We calculated item parameter values and IRT-based scale scores using the full analytic sample used to run confirmatory impact analyses (N = 2,140). The goal was to have the closest possible correspondence between the sample used for both score calibration and impact estimation. Tables D2 and D3 display select item statistics for the fourth- and fifth-grade science achievement tests.

TABLE D2. SUMMARY OF CLASSICAL TEST THEORY AND ITEM-RESPONSE THEORY PARAMETERS FOR 4TH GRADE

Item ID	Factor 1 loadings (Oblique rotation)	Factor 2 loadings (Oblique rotation)	Percent correct	Biserial correlation	Difficulty (1PL)	Difficulty (2PL)	Discrimination (2PL)	Difficulty (3PL)	Discrimination (3PL)	Strand	Source
1	0.476	0.103	.652	.398	-1.076	-0.611	1.453	-0.298	1.749	ESS	NECAP
2	0.450	-0.040	.439	.366	0.419	0.266	1.135	0.552	1.462	PS	NAEP
3	0.435	-0.039	.443	.364	0.388	0.256	1.066	0.653	1.738	PS	NAEP
4	0.426	0.082	.541	.358	-0.284	-0.196	1.091	0.111	1.295	PS	NECAP
5	0.415	-0.010	.390	.343	0.760	0.532	1.001	0.806	1.296	ESS	NECAP
6	0.415	-0.014	.388	.330	0.773	0.532	1.025	0.861	1.631	ESS	MCAS
7	0.401	0.166	.600	.352	-0.696	-0.492	1.020	-0.106	1.219	ESS	MCAS
8[b]	0.395	0.118	.727	.322	-1.669	-1.063	1.169	-0.753	1.288	ESS	NAEP
9	0.381	-0.109	.321	.313	1.272	0.974	0.891	1.230	1.540	ESS	NECAP
10	0.361	-0.091	.410	.283	0.620	0.512	0.805	0.866	1.051	PS	invented
11[b]	0.336	-0.140	.366	.273	0.932	0.844	0.720	1.221	1.003	ESS	NAEP
12	0.300	0.150	.572	.241	-0.499	-0.464	0.702	-0.044	0.760	PS	MCAS
13	0.295	-0.069	.334	.243	1.171	1.155	0.650	1.431	1.322	PS	MOSART-r
14	0.281	0.260	.464	.246	0.244	0.243	0.639	0.774	0.793	PS	AAAS
15	0.277	-0.011	.325	.221	1.246	1.256	0.633	1.556	0.899	ESS	MCAS
16[b]	0.252	0.170	.418	.222	0.562	0.661	0.531	1.224	0.701	ESS	NAEP
17	0.235	-0.001	.263	.203	1.747	1.948	0.564	1.767	1.915	ESS	MOSART
18	0.148	-0.113	.305	.109	1.398	2.564	0.329	1.880	2.179	ESS	NAEP
19	0.130	0.066	.394	.106	0.730	1.604	0.272	2.616	0.412	PS	NECAP
20[b]	0.101	0.005	.449	.084	0.346	1.024	0.201	2.622	0.344	PS	NAEP
21	0.097	-0.014	.227	.080	2.076	5.339	0.232	4.539	0.545	ESS	NAEP

TABLE D2. SUMMARY OF CLASSICAL TEST THEORY AND ITEM-RESPONSE THEORY PARAMETERS FOR 4TH GRADE

Item ID	Factor 1 loadings (Oblique rotation)	Factor 2 loadings (Oblique rotation)	Percent correct	Biserial correlation	Difficulty (1PL)	Difficulty (2PL)	Discrimination (2PL)	Difficulty (3PL)	Discrimination (3PL)	Strand	Source
22	0.081	-0.019	.272	.080	1.677	5.704	0.174	2.309	2.147	ESS	MOSART-r
23	0.086	0.328	.615	.077	-0.804	-2.121	0.224	-0.353	0.267	PS	AAAS
24[b]	0.136	-0.178	.389	.098	0.768	1.692	0.272	2.570	0.725	PS	NAEP
25	0.134	-0.168	.326	.105	1.234	2.784	0.265	2.505	1.122	PS	MCAS
Proportion variance explained [a]	2.411	0.411									
Inter-factor correlations	-.007										
Reference axis correlations	.007										

Note. 1/2/3 PL = 1/2/3-Parameter Logistic; NECAP = The New England Common Assessment Program; NAEP = National Assessment of Educational Progress; MOSART = Misconception-Oriented Standards-Based Assessment Resources for Teachers; MCAS = Massachusetts Comprehensive Assessment System; AAAS = American Association for the Advancement of Science; ESS = Earth and space science; PS = physical science

[a] eliminating other factors

[b] item was included in both fourth-grade and fifth-grade forms

TABLE D3. SUMMARY OF CLASSICAL TEST THEORY AND ITEM-RESPONSE THEORY PARAMETERS FOR 5TH GRADE

Item ID	Factor 1 loadings (Oblique rotation)	Factor 2 loadings (Oblique rotation)	Percent correct	Biserial correlation	Difficulty (1PL)	Difficulty (2PL)	Discrimination (2PL)	Difficulty (3PL)	Discrimination (3PL)	Strand	Source
1	0.434	-0.021	.377	.296	1.079	0.557	1.113	0.81	1.445	PS	MOSART
2	0.412	-0.007	.353	.316	1.309	0.724	1.009	1.031	1.718	PS	NECAP
3	0.396	0.056	.483	.307	0.149	0.081	0.95	0.564	1.438	ESS	MOSART
4	0.386	-0.009	.491	.277	0.074	0.041	0.912	0.462	1.201	ESS	MCAS
5	0.325	0.121	.315	.246	1.671	1.134	0.769	1.445	0.999	PS	MOSART
6	0.284	0.138	.289	.217	1.934	1.473	0.669	1.607	1.402	ESS	MOSART
7	0.283	-0.187	.66	.189	-1.428	-1.116	0.651	-0.592	0.719	PS	NECAP
8[b]	0.392	-0.190	.785	.269	-2.779	-1.306	1.292	-0.974	1.556	ESS	NAEP
9	0.227	0.144	.219	.155	2.734	2.466	0.55	2.268	1.176	PS	AAAS
10	0.207	0.175	.326	.161	1.568	1.632	0.469	1.818	1.403	PS	MOSART
11[b]	0.349	-0.054	.403	.247	0.842	0.573	0.766	0.973	1.008	ESS	NAEP
12	0.180	0.019	.302	.135	1.8	2.141	0.405	2.551	0.689	PS	MOSART
13	0.145	-0.100	.336	.11	1.464	2.426	0.286	2.934	0.543	ESS	MCAS
14	0.135	0.006	.405	.102	0.825	1.369	0.285	2.635	0.399	ESS	MOSART
15	0.123	-0.011	.389	.087	0.967	1.673	0.274	2.661	0.426	PS	MOSART
16[b]	0.298	-0.107	.416	.209	0.727	0.565	0.653	1.07	0.93	ESS	NAEP
17	0.059	0.009	.309	.044	1.732	8.259	0.098	6.784	0.299	PS	MOSART
18	0.152	0.264	.285	.147	1.972	2.58	0.366	2.078	1.719	ESS	NAEP
19	-0.048	0.235	.236	-.005	2.521	-11.005	-0.107	3.073	1.955	PS	NECAP
20[b]	0.051	-0.083	.425	.018	0.65	3.358	0.09	5.015	0.206	PS	NAEP
21	0.114	0.227	.418	.112	0.712	1.326	0.253	2.016	1.527	ESS	NAEP

TABLE D3. SUMMARY OF CLASSICAL TEST THEORY AND ITEM-RESPONSE THEORY PARAMETERS FOR 5TH GRADE

Item ID	Factor 1 loadings (Oblique rotation)	Factor 2 loadings (Oblique rotation)	Percent correct	Biserial correlation	Difficulty (1PL)	Difficulty (2PL)	Discrimination (2PL)	Difficulty (3PL)	Discrimination (3PL)	Strand	Source
22	0.086	0.213	.302	.115	1.807	4.516	0.187	2.701	1.514	ESS	MOSART
23	0.103	-0.130	.335	.057	1.478	3.553	0.195	3.817	0.468	ESS	MCAS
24[b]	0.181	0.051	.396	.147	0.911	1.168	0.374	1.952	0.956	PS	NAEP
Proportion variance explained [a]	1.570	0.433									
Inter-factor correlations	.030										
Reference axis correlations	-.030										

Note. 1/2/3 PL = 1/2/3-Parameter Logistic; NECAP = The New England Common Assessment Program; NAEP = National Assessment of Educational Progress; MOSART = Misconception-Oriented Standards-Based Assessment Resources for Teachers; MCAS = Massachusetts Comprehensive Assessment System; ESS = Earth and space science; PS = physical science

[a] eliminating other factors

[b] item was included in both fourth-grade and fifth-grade forms

Choosing an Approach to Scaling

Under the guidance of a psychometrician, we examined the characteristics of the assessment. It was notably difficult based on examination of item percent-correct scores. In Tables D4 and D5, we display averages of percent-correct scores by decile of ELA and math third-grade pretest scores. Figures D1 and D2 show the average percent-correct scores on the science achievement assessment (by treatment and control) across deciles of the ELA pretest and math pretest distributions. We observe that proportions correct are low, with percent-correct scores below 50% across most of the pretest distributions.

TABLE D4. MEAN PERCENT CORRECT SCORES ON THE STUDENT SCIENCE ACHIEVEMENT ASSESSMENT BY DECILE OF ELA PRETEST ACHIEVEMENT (AVERAGED ACROSS GRADES 4 AND 5)

Decile	*n*	Mean	Std dev	Minimum	Maximum
1	214	0.30	0.12	0.00	0.80
2	214	0.32	0.11	0.04	0.64
3	214	0.32	0.12	0.04	0.72
4	214	0.35	0.12	0.04	0.76
5	214	0.40	0.13	0.08	0.80
6	214	0.41	0.13	0.16	0.84
7	214	0.44	0.14	0.13	0.84
8	214	0.46	0.13	0.16	0.84
9	214	0.51	0.15	0.17	0.88
10	214	0.57	0.14	0.24	0.88

TABLE D5. MEAN PERCENT CORRECT SCORES ON THE STUDENT SCIENCE ACHIEVEMENT ASSESSMENT BY DECILE OF MATH PRETEST ACHIEVEMENT (AVERAGED ACROSS GRADES 4 AND 5)

Decile	*n*	Mean	Std dev	Minimum	Maximum
1	214	0.30	0.12	0.00	0.80
2	214	0.32	0.11	0.04	0.64
3	214	0.32	0.12	0.04	0.72
4	214	0.29	0.11	0.04	0.80
5	214	0.31	0.12	0.04	0.80
6	214	0.34	0.13	0.00	0.72
7	214	0.37	0.12	0.12	0.79
8	214	0.39	0.14	0.08	0.76
9	214	0.41	0.14	0.08	0.84
10	214	0.43	0.12	0.17	0.76

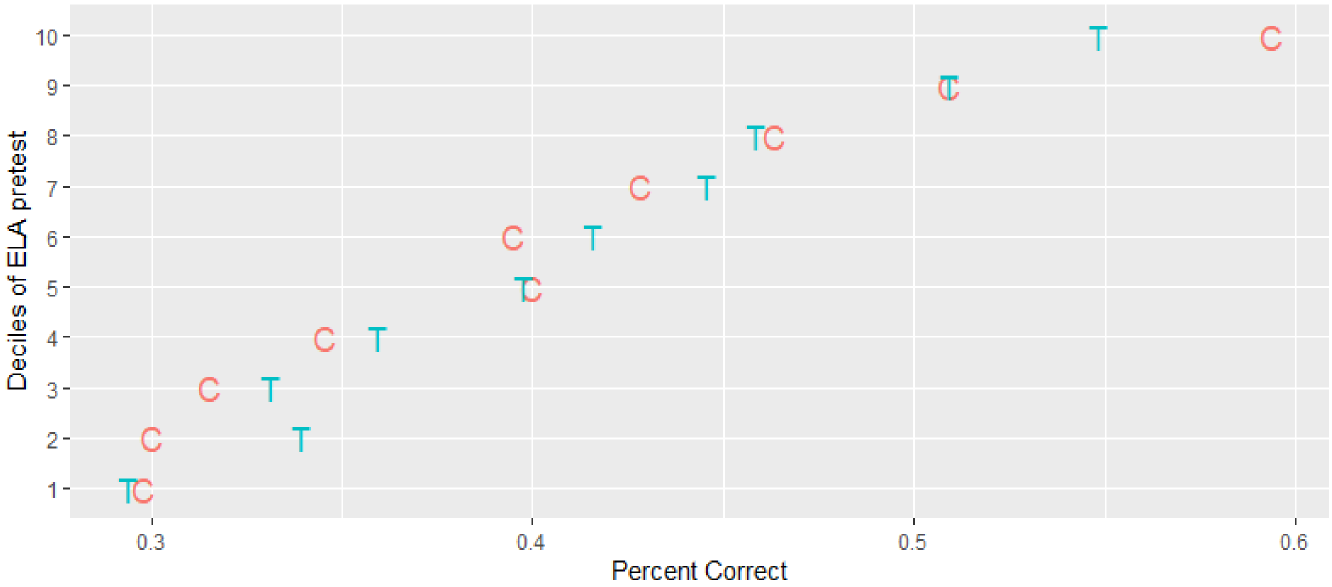

FIGURE D1. DIFFERENCES IN PERCENT CORRECT ON THE STUDENT SCIENCE ACHIEVEMENT ASSESSMENT BETWEEN CONDITIONS BY DECILE OF ELA PRETEST

Note. C is control; T is treatment

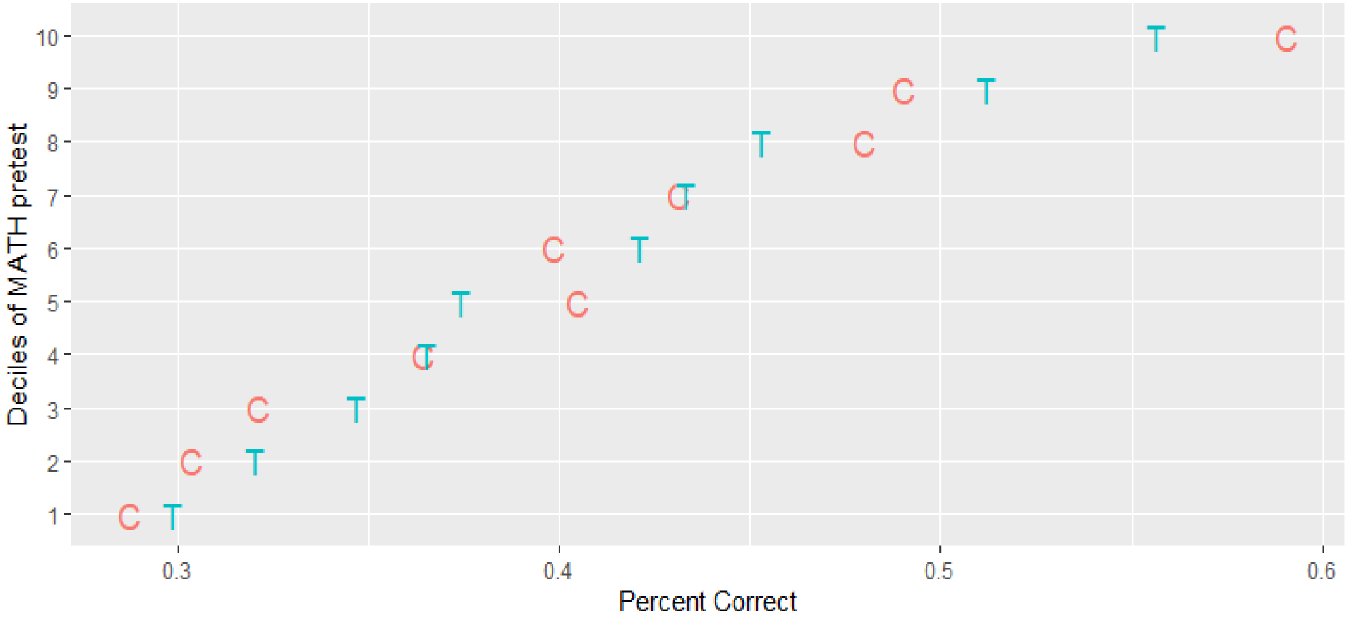

FIGURE D2. DIFFERENCES IN PERCENT CORRECT ON THE STUDENT SCIENCE ACHIEVEMENT ASSESSMENT BETWEEN CONDITIONS BY DECILE OF MATH PRETEST

Note. C is control; T is treatment

There was debate among advisors as to the best approach to scaling. One opinion was that a 3-Parameter Logistic (3PL) model was justified given apparent levels of student guessing. Others objected that students would not outright guess, even with difficult items, and that at least some of the less proficient students would likely use strategies to, for instance, narrow the number of response options.

Facing a complex choice and recognizing alternative rationales for using different IRT models, we opted to analyze impacts with four approaches to scaling: as percent-correct and using three standard IRT-based models, 1PL, 2PL, and 3PL.[3] Our reasoning was that if impacts are robust to the choice of scaling, it would add more confidence to our result. We also viewed this as an opportunity to conduct research on an interesting question: whether different approaches to scaling would lead to similar results that support the same conclusion about program impact. The Test Characteristic and Test Information Curves for the three IRT models (displayed in Figures D3, D4, D5) show different patterns, with the 3PL model reflecting minimal information on the low end of the achievement scale. This is not surprising given that students responded correctly only slightly above the guessing rate at the low end of the scale. This suggested the scale would be non-discriminative of ability in that range, potentially limiting reliability of individual scores, as well as the precision of average achievement scores and impact estimates in that interval. On the other hand, the correlations among the scores with the four approaches to scaling were high (see Figures D6 and D7), possibly making little difference to average scores and estimates of impact. Thus, it was not clear what the effect of the different approaches to scaling would be on student science achievement and on the impact on that outcome. (We show the results for fourth grade; very similar results were obtained for fifth grade.)

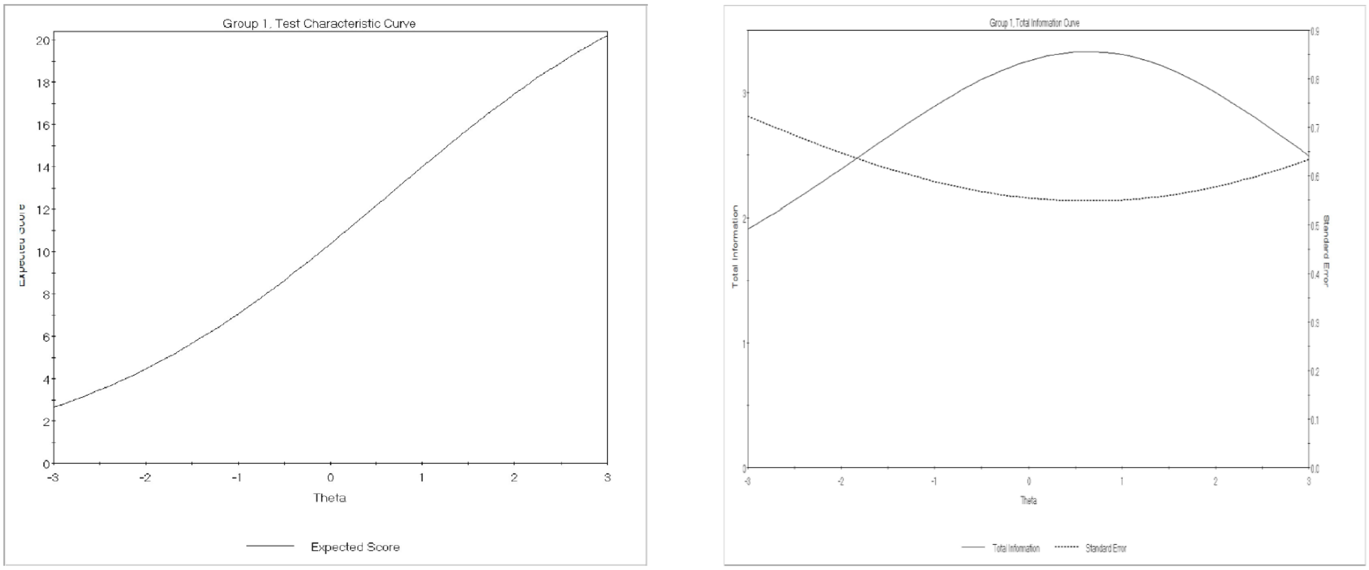

FIGURE D3. TEST CHARACTERISTIC CURVE AND TEST INFORMATION CURVE FOR 1PL MODEL IN FOURTH GRADE (N = 1,220)

[3] We used the software IRTPRO (Cai, Thissen, & du Toit, 2011). We conducted separate score calibrations in Grade 4 and Grade 5. We used the Bock-Aitkin EM algorithm in IRTPRO to obtain item parameter and student score estimates.

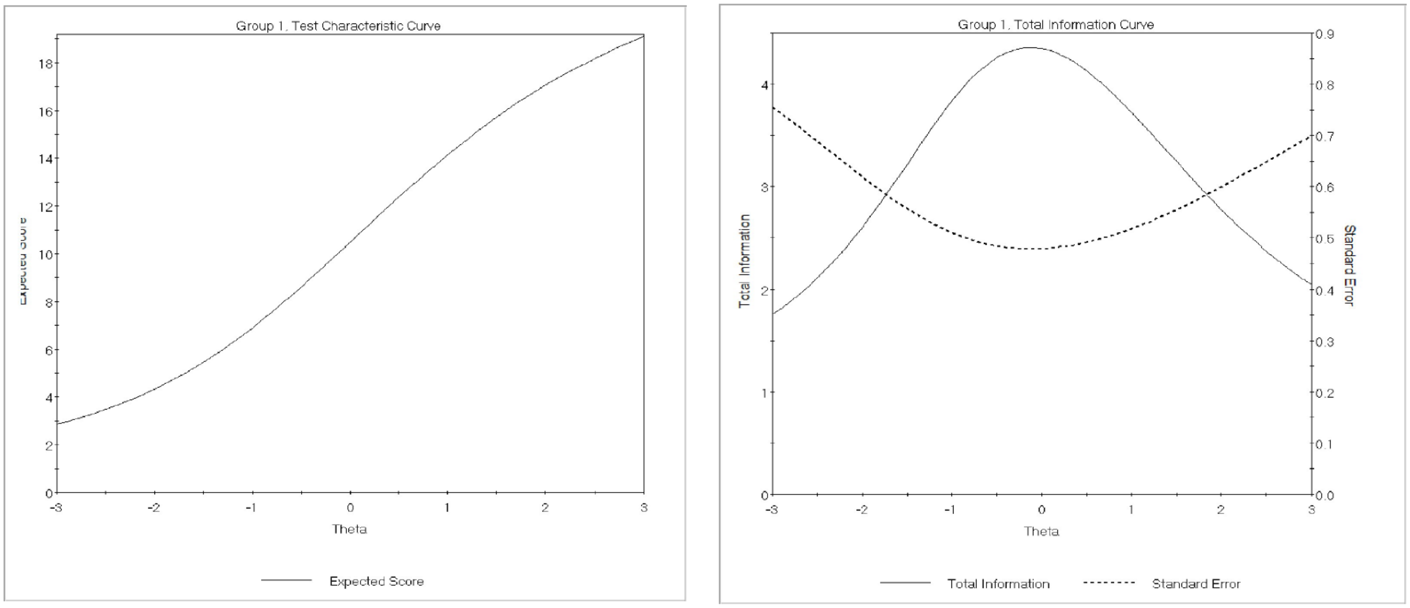

FIGURE D4. TEST CHARACTERISTIC CURVE AND TEST INFORMATION CURVE FOR 2PL MODEL IN FOURTH GRADE (N = 1,220)

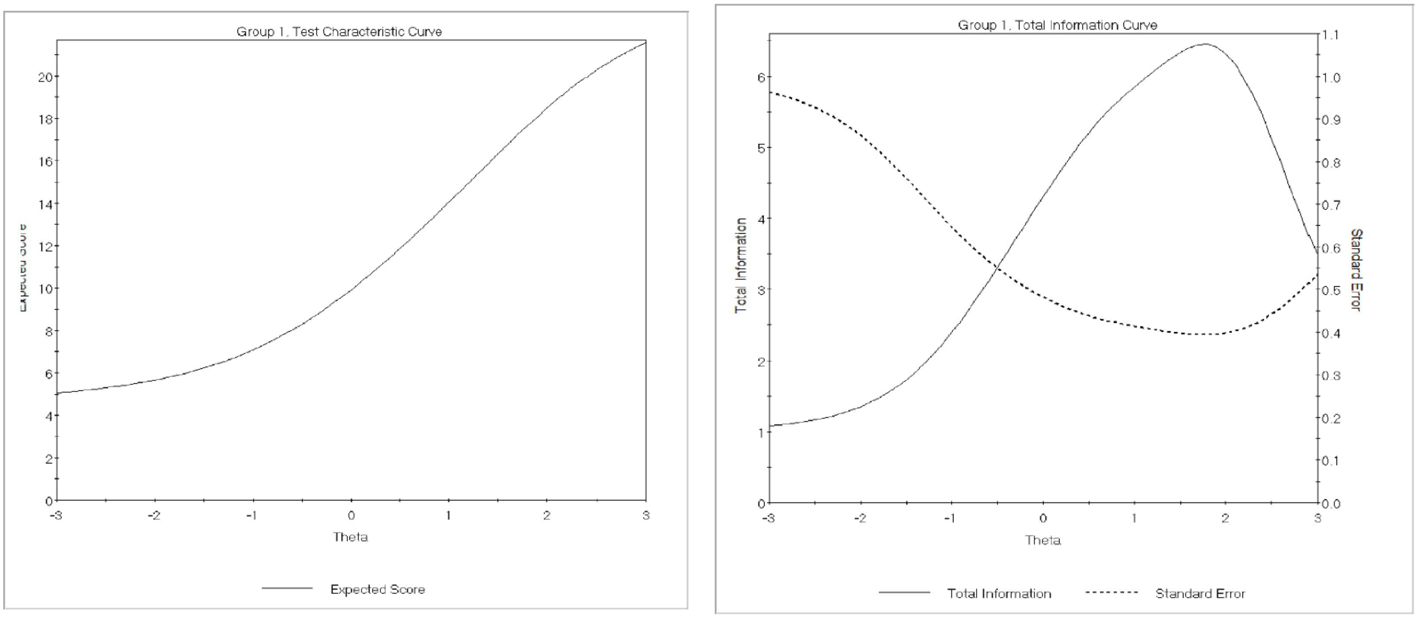

FIGURE D5. TEST CHARACTERISTIC CURVE AND TEST INFORMATION CURVE FOR 3PL MODEL IN FOURTH GRADE (N = 1,220)

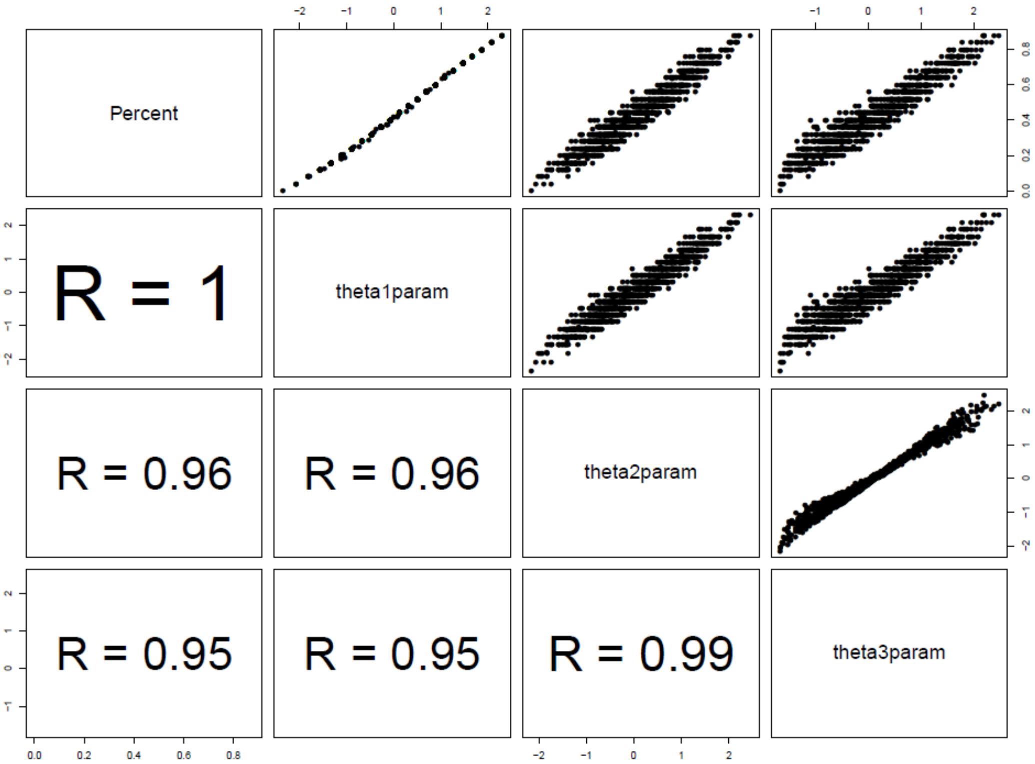

FIGURE D6. CORRELATIONS AMONG SCORES ON THE STUDENT SCIENCE ACHIEVEMENT ASSESSMENT CALIBRATED USING DIFFERENT APPROACHES
(N = 1,220)

As shown in Appendix M, among our sensitivity analyses for the confirmatory test of impact on student science achievement, we examined results using 24 impact models: 3 covariate sets (no covariates, pretest as the only covariate, and with a full set of covariates) × 2 ways of modeling randomized blocks (fixed or random) × 4 calibration methods (percent-correct and 1PL, 2PL, and 3PL scaling). All impact models included random effects for schools (the unit of random assignment). Then using Type-3 tests of fixed effects, we examined whether, for the 24 approaches, impacts varied depending on the three main criteria informing the impact model: the approach to scaling, the covariates used in analysis, and whether randomized blocks were modeled as fixed or random.

None of the impact estimates reached statistical significance (all p values were greater than .30). The Type-3 tests of fixed effects revealed that among the 24 combinations of approaches to modeling impact, estimates did not vary beyond chance depending on scaling (p = .996), but they did vary depending on which covariates were used (p < .001), and depending on whether the randomized blocks were modeled as fixed or random (p < .001). A notable result is that impact estimates ranged in values between -.028 and .081 standard deviations, which is a substantial difference considering that impacts as small as .05 standard deviations are considered substantively important (Bloom et al., 2008).

Appendix E. All Student Survey Scales Measuring Opportunity to Learn and Non-Academic Outcomes

This appendix provides the student survey scales that were administered to students to measure students' opportunity to learn and non-academic outcomes in spring of Year 2 (2017–18). The set of survey scales consisted of six scales modified 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 answer scales. We also created two survey scales to measure cognitive demand and agency in learning.

Items with an "*" were reverse coded before analysis.

ITEM SET 1 (FORMS A1 AND B1)

Aspirations

To what extent do you agree or disagree with the following statements about science (5-point scale: Strongly disagree, Disagree, Neither disagree or agree, Agree, Strongly agree)

a) I expect to use science when I am an adult.
b) Knowing science will help me get a job.
c) I would consider having a job in science.
d) Knowing science will help me in my work when I am an adult.

Source: Friday Institute for Educational Innovation (2012). *Elementary School STEM - Student Survey.* Raleigh, NC: Author.

Quality of Science Class – Learning Environment/Classroom Management

How often do the following things happen **in your science class?** (5-point scale: Almost never, Once in a while, Sometimes, Frequently, Almost always)

a) Our class stays busy and does not waste time when doing science.
b) Students in my class are respectful to our teacher during science class.
c) Students in my class behave the way my teacher wants them to during science class.
d) Students in my class know what they are supposed to be doing and learning in science.
e) Students in my class listen to each other when someone is sharing their ideas about science.
f) Students like raising their hands and asking questions in science.
g) The behavior of other students in my science class helps my learning of science.
h) Students share their science ideas in class.
i) The teacher respects students' science ideas in my class.
j) Rules are used in our science class to make sure everyone is treated fairly.
k) The teacher trusts students to take care of science materials.
l) Our teacher treats us fairly.

Source: Colorado Education Initiative (2013). *Colorado's Student Perception Survey*. Denver, CO: Author.

ITEM SET 2 (FORMS A2 AND B2)

Self-Efficacy

To what extent do you agree or disagree with the following statements about science (5-point scale: Strongly disagree, Disagree, Neither disagree or agree, Agree, Strongly agree)

a) I usually do well in science.
b) Science is harder for me than for many of my classmates.*
c) I am just not good at science.*
d) I learn things quickly in science.
e) My teacher tells me I am good at science.
f) Science is harder for me than any other subject.*
g) Science makes me confused.*

Source: TIMSS 2015 Student Questionnaire. Copyright © 2014 International Association for the Evaluation of Educational Achievement (IEA).

Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Activities in Science Classroom

How often do you do these things **in your class when you are learning science?** (5-point scale: Never, Almost Never, Sometimes, Very Often, Always)

a) I watch the teacher do a science experiment.
b) I plan or do a science experiment or project on my own.
c) I work with other students in a small group on a science experiment or project.
d) I read about science.
e) I write an explanation for something I am studying in science.
f) I discuss with other students the things I am studying in science.
g) I discuss with my science teacher the things I am studying in science.

Source: TIMSS 2007 Student Questionnaire. Copyright © 2007 International Association for the Evaluation of Educational Achievement (IEA).

Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College

ITEM SET 3 (FORMS A3 AND B3)

Quality of Science Class – Science Instruction

To what extent do you agree or disagree with the following statements about **your class when you are learning science?** (5-point scale: Strongly disagree, Disagree, Neither disagree or agree, Agree, Strongly agree)

a) My teacher plans interesting things for us to do.
b) My teacher makes us think.
c) My teacher wants us to talk about what we think.
d) My teacher asks us to write down what we do, think, and observe.
e) My teacher thinks we can learn challenging science.
f) My teacher tells us it is okay to be wrong sometimes in science.
g) My teacher asks interesting questions.

Source: TIMSS 2015 Student Questionnaire. Copyright © 2014 International Association for the Evaluation of Educational Achievement (IEA).

Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Agency in Learning

How often do the following things happen **in your class when you are learning science?** (5-point scale: Almost never, Once in a while, Sometimes, Frequently, Almost always)

a) The teacher asks me to share my ideas in science.
b) I have choices about what I learn in science.
c) The teacher tells us what to do in science class.*
d) Students get to figure things out in my science class.
e) The teacher does most of the explaining in my science class.*
f) The teacher asks students to lead science activities.

ITEM SET 4 (FORMS A4 AND B4)

Cognitive Demand

How often do the following things happen **in your class when you are learning science?** (5-point scale: Almost never, Once in a while, Sometimes, Frequently, Almost always)

a) I learn challenging things in science class.
b) I have to think hard to figure things out in science class.
c) The teacher asks me to explain my ideas in science class.
d) The teacher encourages me to work hard in science class.
e) The teacher has high expectations for me in science class.

Enjoyment of Science

To what extent do you agree or disagree with the following statements about learning science (5-point scale: Strongly disagree, Disagree, Neither disagree or agree, Agree, Strongly agree)

a) I enjoy learning science.
b) I wish I did not have to study science.*
c) Science is boring.*
d) I learn many interesting things in science.
e) I like science.
f) I look forward to learning science in school.
g) Science teaches me how things in the world work.
h) I like to do science experiments.
i) Science is one of my favorite subjects.

Source: TIMSS 2015 Questionnaire. Copyright © 2015 International Association for the Evaluation of Educational Achievement (IEA).

Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Appendix F. Description of the Principal Surveys

This appendix provides a description of the principal surveys, which were administered in spring 2016–17 and spring 2017–18. Heller Research Associates (HRA) analyzed and reported the survey responses by (Wong et al., 2020).

The purpose of the administrator survey (intended for principals or vice principals) was to capture information about each school and its leadership, particularly in relation to science instruction at baseline and throughout the course of the study. The surveys covered a range of topic areas, including:

- how **science instruction** is prioritized compared to other subjects at the school, barriers and supports for science instruction, and resources available for science instruction;
- **philosophy** about and confidence in teaching and learning science and attitude toward change;
- **perceived influence in and capacity** to support teachers in a number of areas such as presenting opportunities for professional learning, supporting collaboration, and giving instructional feedback;
- familiarity with and attitudes toward **NGSS**;
- teacher and administrator turnover rates, **school climate** and the dynamics among administrators and teachers at their school, and the culture of collaboration;
- **professional learning** implemented at the school;
- **education and teaching background** including years of experience teaching and in school leadership positions; and
- **demographic information** such as race/ethnicity and gender (on baseline survey only).

For each survey, either the principal or the vice principal (but not both) would complete the survey. Administrators who joined the study after randomization did not receive the baseline survey, but did answer a subset of questions, including demographic and teaching background information.

Appendix G. Description of the Pilot of the Video and Audio Recordings

This appendix documents the pilot of the classroom video recordings conducted in spring of Year 1 (2016–17) and the audio recordings collected in spring of Year 2 (2017–18).

Pilot of Classroom Video Recordings

During spring 2017–18, researchers piloted a process to collect data on classroom instructional practices and students' discourse patterns through video recorded classroom observations. The pilot process included obtaining active and passive parental consent, training local camera operators to set up classroom sets of video/audio equipment, scheduling the observations, and collecting the data from a subset of study teachers. The purpose of the pilot was to estimate parental consent response rates and determine the feasibility of scheduling for the full sample of schools. The pilot also allowed researchers to test if the type and quality of the video and audio captured would be sufficient for use with the classroom observation scoring protocol, which was also in development.

Parental Consent Process

The parental consent process was piloted in 21 schools (9 schools in districts that required active parental consent and 12 schools in districts that required passive parental consent). In the districts that required active parental consent, approximately 35% (8/23) of teachers had a somewhat acceptable number of students (more than 10) who agreed to be video recorded. In the districts that required passive parental consent, 88% (23/26) of teachers had an acceptable number of students who could be video recorded. However, several teachers expressed that it would be too burdensome to remove the students who were not allowed to be recorded from class on the day of the observations. Teachers reported that they did not want these students to miss the lesson and did not have a central place for the students to go during this time. Logistically, this was a challenge for teachers and researchers.

Scheduling and Set Up

The study team piloted the scheduling and data collection process with 15 teachers from two districts in California that required passive parental consent. Researchers sent teachers the following instructions regarding recording.

- We intend to record your classroom when science instruction is taking place. This means that dates and times during which students are taking tests, watching movies, etc. should not be included as potential video observation sessions.
- We intend to record your classroom when the teacher who is participating in the Making Sense of SCIENCE study is teaching, not a teaching assistant or instructional specialist.
- We would prefer to record an earth science or physical science lesson, but recognize that this may not be possible given your existing plans for science instruction.
- We would like to see how teachers support students' science dialogue.

- There will be two sets of cameras at each school during the week they are being recorded. This means that two teachers can be recorded on the same day at the same time. There will be a camera operator who will come to your classroom at the time of the scheduled observation to set up and take down equipment. However, this person will not stay in the classroom during the recording.

Ideally, we'd like to record your class for an entire lesson arc for a scenario where the class is introduced to something, do or observe an investigation/lab/or demo, and then talk about it. At a minimum, we'd like to record you for two consecutive lessons for a total of 60 minutes of science instruction. For example:

- If you teach a 90-minute lesson, we'd record one lesson.
- If you teach 45-minute or 60-minute lessons, we'd record two consecutive lessons.
- If you teach 20-minute lessons, we'd record three consecutive lessons.

We understand that your science instruction may not fit into these three examples, so please let us know if you have a different structure, and we will work with you to figure out what is best. If possible, we would prefer to finish recordings at a school within one school week.

We hired local camera operators and trained them to set up with the Swivl units, iPads, and microphones. In each classroom, one Swivl rig tracked the movement of the teacher's microphone (attached to a lanyard around their neck). We set up the second Swivl rig in the back of the classroom to capture the board/projector, and placed the other microphones on the right or left side of the classroom, out of reach of students. Camera operators set up and removed the equipment but were not present in the classroom during the recording. Camera operators also collected, if available, a pre-observation form, lesson plans, and photos of student artifacts at the end of each recording. We uploaded all data to a central repository for viewing and coding.

Results of Video Pilot

The pilot produced approximately 19 hours of video from 12 teachers (3 teachers were unable to be recorded for various reasons after scheduling). Given the issues with consent response rates and the resource intensive process, the study team decided not to collect video from the full set of teachers in year 2 of the study.

Summary of the Audio Study

Researchers continued to investigate classroom instructional practices through a modified data collection plan that did not include video, or the need for scheduling data collection during specific times. In spring 2018, the Making Sense of SCIENCE research team collected classroom data through audio recordings, which were supplemented by a survey, teacher self-recorded interview, artifacts, and photos of instructional materials and classroom activities. The purpose of the audio recordings and interviews was to capture NGSS-aligned instructional practices and decisions and to determine if Making Sense of SCIENCE impacted the enacted instruction.

Study teachers in six of the seven districts were invited to participate in this data collection. Of the 105 teachers who were invited to participate, 26 agreed (15 in control schools and 11 from *Making Sense of SCIENCE* schools). Of those, 19 teachers completed the data collection activity (9 control and 10 *Making Sense of SCIENCE*). The remaining seven teachers reported scheduling issues and were unable to complete the activity. The research team mailed "audio recording kits" to the teachers, which included parental consent forms, an audio recorder, a disposable camera, an information survey about the recorded lesson, and instructions.

Teachers were given the following guidelines for deciding which lesson(s) to record.

- Plan to record 90 minutes of science instruction in one or more consecutive lessons.
- Pick a lesson that shows how you include next generation science learning (NGSS) in your classroom.
- Focus on Earth science or physical science topics throughout the recorded lessons, if possible. If it is not possible, other science topics are acceptable.
- Do not select times when students are primarily taking tests, watching movies, doing non-science work, etc.
- It is not necessary to create a lesson solely for the purpose of this recording.

Teachers wore the USB audio recorder on a lanyard around their neck during the recorded lessons. They turned the recorder off when speaking to students who did not have parental consent to be recorded. Additionally, we asked teachers to provide photocopies of lesson plans, notes, handouts, or materials used by students during the lesson(s), and slides or overheads projected for students. They also used the disposable camera to take photos of student or teacher writings or drawings done on the board during the lesson(s), as well as any posted instructions, diagrams, and guidelines referred to during the lesson.

After the lesson, we asked teachers to complete a Classroom Information Survey about their class and the recorded lesson(s). We also asked them to record a post-lesson reflection interview in response to the questions on a teacher interview protocol asking them to reflect on the lesson (what they did, what was effective, how they would modify the lesson in the future) and ways in which the lesson included aspects of NGSS. Once they completed their audio recording study, they mailed back their completed audio kit to researchers for analysis. HRA analyzed the data from the Classroom Information Survey, which focused on content, and teachers' attitude and beliefs before, during, and after the lesson (Wong et al., 2020). They state that a secondary analysis of audio-recorded classes and teacher interviews would offer more insight into the conceptual orientation of Making Sense of SCIENCE versus control classrooms, as well as the nature of student group discussions.

Appendix H. Hierarchical Linear Model for the Analysis of Impact on Teacher Content Knowledge

This appendix presents the hierarchical linear model used to evaluate impact on teacher content knowledge (Equation H1).

$$y_{jk} = \beta_0 + \beta_1 treatment_k + \sum_{q=1}^{Q} \lambda_q X_{jq} + \sum_{r=1}^{R} \gamma_r Z_{jkr} + \sum_{s=1}^{S} BLOCK_s D_{ks} + \xi_k + \varepsilon_{jk}$$

(H1)

y_{jk} is the outcome for teacher j in school k. *Treatment*$_k$ is a binary variable at the school level, with 0 indicating assignment to control and 1 indicating assignment to *Making Sense of SCIENCE*. The effect of the intervention is assessed in terms of the statistical significance of the estimate of β_1. The model includes effects of covariates at the school level (λ_q), and at the teacher level (γ_r), as well as fixed effects for randomized blocks (we assume S blocks with $BLOCK_{js}$ taking on the value 1 if school k is in block s and 0 otherwise.) ξ_k and ε_{jk} represent school- and teacher-level random effects.

Appendix I. Detailed Results of the Benchmark Analysis of Impacts on Teacher Content Knowledge

This appendix presents the detailed results of the benchmark analysis on teacher content knowledge for the *Mixed* and *Retained in Study* samples.

Results of the Benchmark Analysis of Impacts on Teacher Content Knowledge for the Mixed Sample (N = 118)

TABLE I1. ESTIMATES OF FIXED EFFECTS – MIXED SAMPLE

Fixed effects	Coefficient	Standard error	*df*	t-ratio	*p* value
Intercept	-2.958	1.058	27	-2.79	.010
Treatment status	0.191	0.134	27	1.43	.165
Content knowledge pretest	3.844	0.522	39	7.36	<.0001
Ethnicity is White	-0.084	0.334	39	-0.25	.802
Ethnicity is Hispanic	-0.159	0.321	39	-0.49	.624
Ethnicity is Black	-0.498	0.351	39	-1.42	.164
Ethnicity is Unknown	0.266	0.325	39	0.82	.418
Ethnicity is Mixed	0.046	0.445	39	0.10	.918
Ethnicity is Native American (reference category)	0.000	.	.	.	.
Missing ethnicity	-0.628	0.647	39	-0.97	.338
Teacher gender is female	-0.029	0.196	39	-0.15	.884
Teacher gender is male (reference category)	0.000	.	.	.	.
Missing gender	1.497	1.050	39	1.43	.162
Certificate in Early Childhood Ed.	-0.207	0.255	39	-0.81	.422
Certified in Eng. Language Dev.	-0.070	0.174	39	-0.41	.687
Has Higher Ed. Degree	-0.143	0.113	39	-1.27	.212
Years of teaching experience	0.005	0.009	39	0.63	.535
Missing years teaching	0.014	0.199	39	0.07	.944
Taught science previous year	0.233	0.206	39	1.13	.265
Indicates little time for science instruction	0.088	0.217	39	0.40	.688
Pretest scale: School context teacher / admin culture	-0.220	0.208	39	-1.06	.298
Pretest scale: School context teacher culture	0.090	0.173	39	0.52	.605
Pretest scale: Use of NGSS activities	-0.047	0.041	39	-1.14	.262
Pretest scale: confidence with science content	-0.018	0.101	39	-0.18	.857
Pretest scale: confidence in science instruction	0.200	0.137	39	1.46	.152
Pretest scale: confidence with literacy and discourse	0.002	0.147	39	0.01	.990
Pretest scale: Perceived level of influence	0.112	0.097	39	1.15	.256

TABLE I1. ESTIMATES OF FIXED EFFECTS – MIXED SAMPLE

Fixed effects	Coefficient	Standard error	*df*	t-ratio	*p* value
Pretest Scale: Beliefs about teaching	0.136	0.148	39	0.92	.361
Pretest Scale: Has mentor available at school	-0.162	0.244	39	-0.66	.512

Note. We do not include estimates for pair fixed effects in the table.

TABLE I2. ESTIMATES OF LEVEL-1 AND LEVEL-2 VARIANCE COMPONENTS (RANDOM EFFECTS) – MIXED SAMPLE

Random effect	Variance component	Standard error	Z value	*p* value
School	.047	0.113	0.42	.337
Residual (teacher)	.502	0.107	4.70	< .001

Results of the Benchmark analysis for of Impacts in Teacher Content Knowledge for the Retained in Study Sample (N = 88)

TABLE I3. ESTIMATES OF FIXED EFFECTS – RETAINED IN STUDY SAMPLE

Fixed effects	Coefficient	Standard error	*df*	t-ratio	*p* value
Intercept	-0.061	0.790	21	-0.08	.939
Treatment status	0.483	0.158	21	3.07	.006
Content knowledge pretest	3.954	0.544	16	7.27	<.0001
Ethnicity is White	0.548	0.366	16	1.50	.153
Ethnicity is Hispanic	0.138	0.356	16	0.39	.704
Ethnicity is Black	-0.458	0.431	16	-1.06	.304
Ethnicity is Unknown	0.270	0.434	16	0.62	.542
Ethnicity is Mixed	0.856	0.462	16	1.85	.083
Ethnicity is Native American (reference category)	0.000	.	.	.	.
Missing ethnicity	0.492	0.675	16	0.73	.476
Teacher gender is female	-0.154	0.189	16	-0.82	.426
Teacher gender is male (reference category)	0.000	.	.	.	.
Missing gender	-2.242	1.064	16	-2.11	.051
Certificate in Early Childhood Ed.	0.266	0.297	16	0.90	.384
Certified in Eng. Language Dev.	-0.117	0.161	16	-0.73	.478

TABLE I3. ESTIMATES OF FIXED EFFECTS – RETAINED IN STUDY SAMPLE

Fixed effects	Coefficient	Standard error	*df*	t-ratio	*p* value
Has Higher Ed. Degree	-0.293	0.110	16	-2.66	.017
Years of Teaching Experience	0.004	0.008	16	0.44	.665
Missing Years Teaching	0.390	0.253	16	1.54	.143
Taught science previous year	-0.177	0.305	16	-0.58	.570
Indicates no time science instruction	-0.283	0.192	16	-1.48	.159
Pretest scale: School context teacher / admin culture	0.113	0.174	16	0.65	.526
Pretest scale: School context teacher culture	-0.309	0.166	16	-1.86	.082
Pretest scale: Use of NGSS activities	-0.076	0.046	16	-1.66	.117
Pretest scale: confidence with science content	-0.029	0.101	16	-0.28	.780
Pretest scale: confidence in science instruction	0.162	0.178	16	0.91	.377
Pretest scale: confidence with literacy and discourse	0.011	0.148	16	0.08	.941
Pretest scale: Perceived level of influence	0.111	0.100	16	1.11	.284
Pretest scale: Beliefs about teaching	-0.239	0.118	16	-2.02	.060
Has mentor available at school	0.189	0.282	16	0.67	.513

Note. We do not include estimates for pair fixed effects in the table.

TABLE I4. ESTIMATES OF LEVEL-1 AND LEVEL-2 VARIANCE COMPONENTS (RANDOM EFFECTS) – RETAINED IN STUDY SAMPLE

Random effect	Variance component	Standard error	Z value	*p* value
School	.211	0.188	1.12	.131
Residual (teacher)	.355	0.113	3.15	<.001

Appendix J. Sensitivity Analysis for the Analysis of Impact on Intermediate Outcomes

This appendix presents the sensitivity analyses conducted to assess the robustness of results of analysis of impacts on intermediate outcomes.

Because our priori selected benchmark model (model 1 in Table J1) yields an estimate of zero for the school-level random effect, as part of the sensitivity analysis, we remove pair effects altogether to free up variance to allow the school variance component to be estimated (model 2). A result of zero variance with no *p* value means the estimation procedure has reached a boundary condition for estimating the corresponding effect (Singer & Willett, 2003), often implying that the variance component is trivially different from zero. However, we prioritized including the school level in analysis because schools are the unit of random assignment. As expected, with this change, in most cases, the school variance component becomes estimable. However, by excluding block effects, our impact estimates are less precise, with several of the results no longer reaching statistical significance (comparing model 1 to model 2). However, we also see that magnitudes of the impact estimate do not fluctuate much between model 1 and model 2, indicating that reaching the boundary condition in estimating the school effect in model 1 is not inducing any major bias in the impact estimates. In further assessing the sensitivity of the benchmark result, we evaluate impact using the benchmark model specification but with restricted maximum likelihood estimation instead of full maximum likelihood (model 3). Many school-level variance components become estimable, but there is an accompanying loss of precision. Several results that were statistically significant under model 1 ceased to be so with other models (models 4 and 5). In total, we show results from five approaches to modeling impact.

TABLE J1. MODELS FOR ASSESSING IMPACT ON INTERMEDIATE OUTCOMES

	School effects	Pair effects	Restricted maximum likelihood
Model 1	Random	Fixed	Maximum likelihood
Model 2	Random	--	Maximum likelihood
Model 3	Random	Fixed	Restricted
Model 4	Random	Random	Restricted
Model 5	Random	--	Restricted

TABLE J2. RESULTS FOR MODELING IMPACT ON INTERMEDIATE OUTCOMES

	Model 1			Model 2			Model 3			Model 4				Model 5		
	Point est.	*p* value	School var.	Point est.	*p* value	School var.	Point est.	*p* value	School var.	Point est.	*p* value	School var.	Pair var.	Point est.	*p* value	School var.
Belief that students are capable learners	-0.085	.519	0.000	-0.059	.673	0.000	-0.085	.613	0.000	-0.063	.679	0.000	0.042	-0.059	.706	0.000
Philosophically aligned with NGSS	0.214	.215	0.000	0.239	.236	0.093	0.214	.349	0.000	0.244	.289	0.134	0.016	0.246	.287	0.151
Values life-long learning	0.022	.773	0.000	0.026	.752	0.002	0.010	.927	0.029	0.025	.789	0.011	0.007	0.026	.785	0.016
Confidence in addressing student performance expectations	0.208	.133	0.000	0.198	.182	0.029	0.205	.282	0.040	0.201	.238	0.057	0.004	0.201	.237	0.061
Confidence in science instructional practices	0.204	.074	0.000	0.249	.062	0.035	0.204	.164	0.000	0.239	.074	0.000	0.089	0.251	.100	0.060
Confidence in supporting literacy in science	0.187	.206	0.000	0.312	.035	0.000	0.187	.325	0.000	0.312	.059	0.000	0.000	0.312	.059	0.000
Self-efficacy	0.110	.285	0.000	0.091	.465	0.046	0.110	.405	0.000	0.088	.484	0.009	0.064	0.091	.521	0.069
Agency in the classroom	0.352	.020	0.000	0.410	.014	0.038	0.355	.103	0.093	0.407	.033	0.067	0.012	0.410	.031	0.077
Amount of time spent on science instruction	1.746	.016	0.000	1.075	.171	0.770	1.713	.087	1.138	1.045	.246	1.729	0.000	1.045	.246	1.729
Sensemaking of hands-on investigations	0.518	.019	0.000	0.237	.362	0.187	0.518	.067	0.000	0.300	.249	0.003	0.243	0.233	.428	0.276
Integration of science and literacy	0.593	.003	0.000	0.342	.134	0.066	0.593	.022	0.000	0.392	.095	0.000	0.199	0.337	.196	0.135

TABLE J2. RESULTS FOR MODELING IMPACT ON INTERMEDIATE OUTCOMES

	Model 1			Model 2			Model 3			Model 4				Model 5		
	Point est.	*p* value	School var.	Point est.	*p* value	School var.	Point est.	*p* value	School var.	Point est.	*p* value	School var.	Pair var.	Point est.	*p* value	School var.
Participating in collaborative discourse	0.583	.005	0.000	0.347	.146	0.089	0.583	.029	0.000	0.399	.100	0.000	0.226	0.343	.206	0.161
Explaining ideas and phenomena	0.392	.068	0.000	0.188	.440	0.130	0.374	.202	0.091	0.217	.399	0.064	0.134	0.179	.518	0.213
NGSS-aligned PS topics (DCIs): Motion and stability - forces and interactions	0.554	.216	0.000	0.142	.762	0.092	0.556	.345	0.092	0.156	.774	0.453	0.000	0.156	.774	0.453
NGSS-aligned PS topics (DCIs): Definitions of energy	0.117	.610	0.000	0.045	.845	0.000	0.117	.691	0.000	0.045	.861	0.000	0.000	0.045	.861	0.000
NGSS-aligned PS topics (DCIs): Conservation of energy and energy transfer	0.958	.138	0.000	0.766	.269	0.000	0.958	.247	0.000	0.734	.332	0.000	1.402	0.761	.333	0.235
NGSS-aligned PS topics (DCIs): Waves	0.395	.414	0.000	0.299	.530	0.000	0.395	.524	0.000	0.299	.574	0.000	0.000	0.299	.574	0.000

TABLE J2. RESULTS FOR MODELING IMPACT ON INTERMEDIATE OUTCOMES

	Model 1			Model 2			Model 3			Model 4				Model 5		
	Point est.	*p* value	School var.	Point est.	*p* value	School var.	Point est.	*p* value	School var.	Point est.	*p* value	School var.	Pair var.	Point est.	*p* value	School var.
NGSS-aligned PS topics (DCIs): Matter and its interactions	0.503	.313	0.000	0.118	.811	0.000	0.497	.447	0.144	0.136	.809	0.154	0.000	0.136	.809	0.155
NGSS-aligned ES topics (DCIs): Earth's place in the universe	-0.351	.471	0.000	-0.944	.090	0.981	-0.452	.524	1.344	-0.950	.131	1.372	0.000	-0.950	.131	1.372
NGSS-aligned ES topics (DCIs): Earth's systems	0.464	.452	0.000	-0.205	.780	2.612	0.413	.674	4.135	-0.204	.804	3.301	0.000	-0.204	.804	3.301
NGSS-aligned ES topics (DCIs): Earth and human activity	-0.133	.674	0.000	-0.455	.232	0.682	-0.182	.716	1.072	-0.454	.286	0.866	0.000	-0.454	.286	0.866
Science and Engineering Practices (SEPs)	0.602	.349	0.000	0.670	.350	1.097	0.602	.465	0.000	0.634	.392	0.080	1.629	0.666	.413	1.725
Crosscutting Concepts (CCCs)	0.907	.193	0.000	0.855	.209	0.000	0.907	.310	0.000	0.855	.261	0.000	0.000	0.855	.261	0.000
Teacher collaboration – amount	0.593	.000	0.000	0.594	.000	0.019	0.592	.001	0.049	0.590	.000	0.037	0.000	0.590	.000	0.037
Culture of peer collaboration	0.175	.155	0.000	0.049	.748	0.086	0.175	.279	0.003	0.067	.667	0.041	0.079	0.047	.784	0.117
Trust and respect among teachers	0.048	.584	0.000	-0.103	.338	0.042	0.037	.772	0.039	-0.090	.443	0.042	0.020	-0.104	.394	0.057
Trust and respect between teachers and administrators	0.070	.644	0.000	0.044	.785	0.002	0.046	.828	0.084	0.037	.835	0.009	0.047	0.032	.865	0.051
Supporting teacher collaboration	0.328	.037	0.000	0.281	.099	0.000	0.328	.105	0.000	0.265	.155	0.000	0.093	0.278	.150	0.008

TABLE J2. RESULTS FOR MODELING IMPACT ON INTERMEDIATE OUTCOMES

	Model 1			Model 2			Model 3			Model 4				Model 5		
	Point est.	*p* value	School var.	Point est.	*p* value	School var.	Point est.	*p* value	School var.	Point est.	*p* value	School var.	Pair var.	Point est.	*p* value	School var.
Prioritizing support for teacher professional learning in science	0.128	.383	0.000	0.054	.742	0.017	0.128	.494	0.000	0.060	.727	0.000	0.106	0.048	.800	0.056
Administrator support involving teachers in science leadership	0.190	.111	0.000	0.139	.319	0.050	0.190	.211	0.000	0.142	.327	0.019	0.063	0.138	.384	0.075

Note. *p* values < .05 are highlighted in red. Est. = Estimate; Var. = Variance; NGSS = Next Generation Science Standards; PS = Physical Science; DCI = Disciplinary Core Ideas; ES = Earth and Space Science; SEP = Science and Engineering Practices; CCC = Crosscutting Concepts

Appendix K. Hierarchical Linear Model Associated with the Confirmatory Impacts on Student Science Achievement (Selected-Response Items)

This appendix presents the hierarchical linear model used to evaluate impact on student science achievement (Equation K1).

$$y_{ijk} = \beta_0 + \beta_1 treatment_k + \sum_{q=1}^{Q} \lambda_q \, X_{ijkq} + \sum_{r=1}^{R} \gamma_r \, Z_{jkr} + \sum_{s=1}^{S} \alpha_s \, Z_{ks} + \sum_{t=1}^{T} BLOCK_t \, D_{ts} + \xi_k + \varepsilon_{ijk}$$

(K1)

y_{ijk} is the outcome for student *i* belonging to the class of teacher *j* (in the 2017/18 school year) in school *k*. *Treatment*$_k$ is a binary variable at the school level, with 0 indicating assignment to control and 1 indicating assignment to *Making Sense of SCIENCE*. The effect of the intervention is assessed in terms of the statistical significance of the estimate of β_1. The model includes effects of covariates at the students level (λ_q), at the teacher level (γ_r), and at the school level (α_s) as well as fixed effects for randomized blocks (we assume *T* blocks with $BLOCK_T$ taking on the value 1 if school *k* is in block *t* and 0 otherwise.) ξ_k and ε_{ijk} represent school- and student-level random effects, respectively.

Appendix L. Full Estimates of the Benchmark Impact Model for the Confirmatory Analysis of Impacts on Student Science Achievement (Selected-Response Items)

This appendix provides the full estimates of the benchmark impact model for the confirmatory analysis of impact on student science achievement (full sample N = 2,140) as measured by selected-response items on the student science assessment.

TABLE L1. ESTIMATES OF FIXED EFFECTS

Fixed effects	Coefficient	Standard error	t-ratio	*df*	*p* value
Intercept	-0.675	0.489	-1.379	23	.181
Treatment Status	0.062	0.089	0.696	23	.494
School Size	0.000	0.000	0.244	23	.809
School in City	0.115	0.165	0.696	23	.493
Title 1 Status	-0.035	0.230	-0.152	23	.880
State CA	0.305	0.394	0.774	23	.447
ELA pretest	0.318	0.028	11.207	2049	<.001
Math pretest	0.266	0.028	9.460	2049	<.001
Grade 4	-0.031	0.043	-0.731	2049	.465
Male	0.064	0.033	1.915	2049	.056
ELL	-0.130	0.046	-2.810	2049	.005
Asian	-0.012	0.067	-0.183	2049	.855
Black	-0.160	0.066	-2.412	2049	.016
Hispanic	-0.097	0.053	-1.837	2049	.066
Native Indian	-0.135	0.183	-0.739	2049	.460
Gender, ELL & Ethnicity Missing	-0.229	0.166	-1.379	2049	.168
Ethnicity Unspecified	0.015	0.068	0.222	2049	.824
FRPL Eligible	0.076	0.050	1.519	2049	.129
FRPL missing	0.185	0.101	1.836	2049	.066
With White teacher	-0.018	0.055	-0.334	2049	.738
Teacher ethnicity missing	0.353	0.158	2.239	2049	.025
Certificate in elementary education	0.248	0.095	2.603	2049	.009
Certificate in English Language development	0.097	0.058	1.673	2049	.094
Highest level of education	0.020	0.036	0.562	2049	.574
Missing Highest level of education	-0.218	0.235	-0.927	2049	.354
Years of classroom teaching	0.001	0.003	0.407	2049	.684
Missing in years of classroom teaching	0.168	0.112	1.502	2049	.133

TABLE L1. ESTIMATES OF FIXED EFFECTS

Fixed effects	Coefficient	Standard error	t-ratio	*df*	*p* value
Missing Teacher pretest	0.442	0.149	2.965	2049	.003
Teacher content pretest	0.749	0.217	3.450	2049	<.001
Missing Survey	-0.522	0.250	-2.090	2049	.037
NGSS Missing	-0.010	0.103	-0.096	2049	.924
Teacher's gender	-0.285	0.067	-4.256	2049	<.001
Taught Science in previous year	0.058	0.079	0.727	2049	.467
Not enough time for Science instruction	-0.074	0.058	-1.282	2049	.200
Composite of school context culture between admins and teachers	0.091	0.061	1.496	2049	.135
Composite of school context culture among teachers	-0.043	0.058	-0.744	2049	.457
NGSS-related activities participated in	0.034	0.016	2.185	2049	.029
Composite of confidence on specific science content	-0.063	0.037	-1.700	2049	.089
Composite of confidence in literacy and discourse	-0.024	0.039	-0.612	2049	.541
Composite of perceived level of influence	-0.024	0.037	-0.639	2049	.523
Composite of teaching philosophies	-0.080	0.049	-1.637	2049	.102
Having coachers or mentors for science instruction	0.056	0.066	0.844	2049	.399

Note. We do not include estimates for pair fixed effects in the table.

CA = California; ELA = English Language Arts; ELL = English Language Learner; FRPL = Free or Reduced Price Lunch; NGSS = Next Generation Science Standards

TABLE L2. ESTIMATES OF LEVEL-1 AND LEVEL-2 VARIANCE COMPONENTS (RANDOM EFFECTS)

Random effect	Standard deviation	Variance component	*df*	Chi-squared	*p* value
School	0.236	0.056	23	106.194	<.001
Student	0.739	0.546			

Note. The analysis was conducted using HLM software, which does not provide *df*, test statistic, and *p* value at level-1.

Appendix M. Sensitivity Analyses for the Confirmatory Impacts on Student Science Achievement (Selected-Response Items)

This appendix presents the results of the sensitivity analyses for the confirmatory impacts on student science achievement as measured by the selected-response items on the science assessment. The sensitivity analyses include scores derived from different score calibration approaches (percent-correct, 1PL, 2PL, and 3PL) and different model specifications. Results in Tables M1 and M2 are based on the score calibrations that included all items. Results in Table M3 are based on "reduced item" forms where we excluded items with factor loading less than .20 on the principal dimension. Scores are calculated as percent-correct.

TABLE M1. IMPACT FINDINGS BASED ON 3PL SCALING WITH ALTERNATIVE MODELS N (SCHOOLS) = 55, N (STUDENTS) = 2,140

	Impact (SE)	t	*df*	*p* value	Effect size	Change percentile rank
Benchmark analysis full covariate set [a]	0.062 (.089)	.696	23	.494	.064	2.5%
Sensitivity Analysis (Alternative Models)						
Like benchmark, no covariates, no blocks	-0.044 (.102)	-.432	53	.667	-.045	-1.8%
Like benchmark, no covariates	0.004 (.089)	0.041	26	.967	.004	0.1%
Like benchmark, pretests are only covariates	0.050 (.067)	.751	26	.459	.052	2.1%
Include teacher random effect	0.065 (0.091)	0.72	23	.478	0.067	2.7%
Use pair random (instead of fixed) effects	0.022 (0.075)	0.29	23	.773	0.023	0.9%
Ignore pair level	0.022 (0.075)	0.29	49	.772	0.023	0.9%
Ignore school, model random intercept and treatment at pair level	0.024 (0.075)	0.32	22	.756	0.025	1.0%
Use ML instead of REML	.053 (.054)	.980	23	.339	0.055	2.2%
OLS	.051 (.048)	1.06	2072	.288	0.053	2.1%
Multiple Imputation to address missing [b]	.069 (.080)	.86	519.35	.390	.071	2.9%

Note. Most covariates are modeled at the student level.

OLS is Ordinary Least Squares. ML is Full Maximum Likelihood. REML is Restricted Maximum Likelihood.

Mean performance in the treatment group was 0.00 units (1.00 *sd*) in the treatment group, and -0.07 units (0.94 *sd*) in the control group.

[a] Full results of benchmark model are provided in Appendix L.

[b] The sample consisted of 2,544 students. This included the 2,140 students used in the benchmark analysis plus students with spring 2018 posttests who had been excluded because they were missing one of the two pretests. Therefore, all students have posttests and some may be missing one or both pretests. Imputation is of all missing covariates including the pretests. Students without posttests are listwise deleted. The imputation regression model included an indicator variable for intervention status, included all covariates that were used for statistical adjustment in the impact estimation model, and included the outcome when imputing missing baseline data. Results were based on 10 round of imputation. Each analysis adjusted for the nesting of individual outcomes in schools. Analysis was conducted using PROC MI and PROC MIANALYZE in SAS. PROC MIXED was used for each imputation cycle.

TABLE M2. IMPACT FINDINGS BASED ON PERCENT-CORRECT, 1PL, 2PL AND 3PL SCALING ALTERNATIVE MODELS N (SCHOOLS) = 55, N (STUDENTS) = 2,140

	Impact (SE)	t	*df*	*p* value	Effect size	Change percentile rank
Random Pair, No covariates, %correct	-.020 (.084)	-0.237	26	.814	-0.021	-0.8%
Random Pair, No covariates, 1 PL	-.019 (.084)	-.222	26	.826	-0.020	-0.8%
Random Pair, No covariates, 2 PL	-.019 (.087)	-.217	26	.830	-0.020	-0.8%
Random Pair, No covariates, 3 PL	-.027 (.086)	-.310	26	.759	-0.028	-1.1%
Random Pair, pretest only covariate, %correct	.038 (.057)	.668	26	.510	0.039	1.6%
Random Pair, pretest only covariate, 1 PL	.039 (.057)	.688	26	.497	0.040	1.6%
Random Pair, pretest only covariate, 2 PL	.043 (.059)	.739	26	.467	0.044	1.8%
Random Pair, pretest only covariate, 3 PL	.036 (.060)	.600	26	.554	0.037	1.5%
Random Pair, all covariates, %correct	.015 (.061)	.250	23	.804	0.015	0.6%
Random Pair, all covariates, 1 PL	.020 (.060)	.335	23	.741	0.021	0.8%
Random Pair, all covariates, 2 PL	.034 (.066)	.510	23	.615	0.035	1.4%
Random Pair, all covariates, 3 PL	.026 (.068)	.386	23	.703	0.027	1.1%
Fixed Pair, No covariates, %correct	.010 (.084)	.118	26	.907	0.010	0.4%
Fixed Pair, No covariates, 1 PL	.012 (.084)	.138	26	.891	0.012	0.5%
Fixed Pair, No covariates, 2 PL	.011 (.088)	.127	26	.900	0.011	0.4%
Fixed Pair, No covariates, 3 PL	.004 (.089)	.041	26	.967	0.004	0.2%
Fixed Pair, pretest only covariate, %correct	.052 (.061)	.849	26	.404	0.054	2.2%
Fixed Pair, pretest only covariate, 1 PL	.053 (.061)	.878	26	.388	0.055	2.2%
Fixed Pair, pretest only covariate, 2 PL	.057 (.065)	.886	26	.384	0.059	2.4%
Fixed Pair, pretest only covariate, 3 PL	.050 (.067)	.751	26	.459	0.052	2.1%
Fixed Pair, all covariates, %correct	.065 (.082)	.790	23	.438	0.067	2.7%
Fixed Pair, all covariates, 1 PL	.067 (.082)	.824	23	.419	0.069	2.8%
Fixed Pair, all covariates, 2 PL	.077 (.087)	.883	23	.386	0.079	3.1%
Fixed Pair, all covariates, 3 PL	.062 (.089)	.696	23	.494	0.064	2.5%

Note. Most covariates are modeled at the student-level.

1/2/3 PL = 1/2/3-Parameter Logistic

TABLE M3. IMPACT FINDINGS BASED ON THE PERCENT-CORRECT METRIC WITH REDUCED-ITEMS FORMS N (SCHOOLS) = 55, N (STUDENTS) = 2,140

	Impact (SE)	t	*df*	*p* value	Effect size	Change percentile rank
Benchmark analysis full covariate set [a]	0.078 (0.091)	0.86	23	0.398	0.080	3.2%
Sensitivity Analysis (Alternative Models)						
Like benchmark, no covariates, no blocks	-0.033 (0.100)	-0.33	53	0.743	-0.034	1.4%
Like benchmark, no covariates	0.011 (0.086)	0.12	26	0.903	0.011	0.5%
Like benchmark, pretests are only covariates	0.056 (0.064)	0.88	26	0.388	0.058	2.3%
Include teacher random effect	0.078 (0.091)	0.85	23	0.402	0.080	3.2%
Use pair random (instead of fixed) effects	0.035 (0.072)	0.49	23	0.632	0.036	1.4%
Ignore pair level	0.035 (0.072)	0.49	49	0.629	0.036	1.4%
Ignore school, model random intercept and treatment at pair level	0.039 (0.072)	0.53	23	0.600	0.040	1.6%
Use ML instead of REML	0.065 (0.053)	1.22	23	0.234	0.067	2.7%
OLS	0.062 (0.049)	1.29	2072	0.199	0.064	2.6%

Note. Most covariates are modeled at the student level.

Mean performance in the treatment group was 0 units (1.00 *sd*) in the treatment group, and -0.06 units (0.94 *sd*) in the control group.

OLS is Ordinary Least Squares. ML is Maximum Likelihood. REML is Restricted Maximum Likelihood.

[a] Full results of benchmark model are provided in Appendix L.

Appendix N. Full Estimates of the Benchmark Impact Model for the Confirmatory Analysis of Impacts on Science Achievement of Students in the Lowest Third of Incoming Achievement

TABLE N1. ESTIMATES OF THE BENCHMARK IMPACT MODEL FOR STUDENTS IN THE LOWEST THIRD OF INCOMING ELA ACHIEVEMENT (N = 715)

Fixed effects	Coefficient	Standard error	t-ratio	*df*	*p* value
Intercept	-1.773	0.708	-2.504	23	.020
Treatment Status	0.054	0.093	0.581	23	.567
School Size	0.000	0.000	0.103	23	.919
School in City	0.138	0.174	0.793	23	.436
Title 1 Status	-0.038	0.249	-0.153	23	.879
State CA	0.728	0.447	1.627	23	.117
ELA pretest	0.022	0.067	0.323	624	.747
Math pretest	0.248	0.040	6.143	624	<.001
Grade 4	-0.026	0.077	-0.342	624	.733
Male	-0.030	0.055	-0.554	624	.580
ELL	-0.046	0.072	-0.641	624	.522
Asian	-0.036	0.131	-0.277	624	.782
Black	-0.070	0.111	-0.635	624	.525
Hispanic	-0.070	0.102	-0.687	624	.492
Native Indian	0.082	0.331	0.249	624	.803
Gender, ELL & Ethnicity Missing	-0.335	0.269	-1.246	624	.213
Ethnicity Unspecified	0.171	0.132	1.298	624	.195
FRPL Eligible	-0.035	0.092	-0.377	624	.706
FRPL missing	0.214	0.202	1.058	624	.290
With White teacher	-0.116	0.084	-1.391	624	.165
Teacher ethnicity missing	0.151	0.245	0.617	624	.537
Certificate in elementary education	0.131	0.181	0.723	624	.470
Certificate in English Language development	0.087	0.091	0.954	624	.341
Highest level of education	-0.029	0.063	-0.460	624	.646
Missing Highest level of education	-0.241	0.385	-0.625	624	.532
Years of classroom teaching	0.008	0.005	1.664	624	.097

TABLE N1. ESTIMATES OF THE BENCHMARK IMPACT MODEL FOR STUDENTS IN THE LOWEST THIRD OF INCOMING ELA ACHIEVEMENT (N = 715)

Fixed effects	Coefficient	Standard error	t-ratio	*df*	*p* value
Missing in years of classroom teaching	-0.080	0.162	-0.494	624	.621
Missing Teacher pretest	0.637	0.232	2.747	624	.006
Teacher content pretest	1.153	0.335	3.438	624	<.001
Missing Survey	0.203	0.380	0.535	624	.593
NGSS Missing	0.120	0.143	0.836	624	.404
Teacher's gender	-0.244	0.100	-2.440	624	.015
Taught Science in previous year	0.000	0.122	-0.001	624	.999
Not enough time for Science instruction	-0.013	0.093	-0.135	624	.893
Composite of school context culture between admins and teachers	0.027	0.095	0.287	624	.774
Composite of school context culture among teachers	0.016	0.088	0.176	624	.861
NGSS-related activities participated in	0.044	0.026	1.709	624	.088
Composite of confidence on specific science content	-0.077	0.060	-1.275	624	.203
Composite of confidence in literacy and discourse	0.006	0.067	0.096	624	.924
Composite of perceived level of influence	0.031	0.058	0.541	624	.589
Composite of teaching philosophies	0.067	0.069	0.959	624	.338
Having coaches or mentors for science instruction	-0.086	0.112	-0.772	624	.440

Note. We do not include estimates for pair fixed effects in the table.

CA = California; ELA = English Language Arts; ELL = English Language Learner; FRPL = Free or Reduced Price Lunch; NGSS = Next Generation Science Standards

TABLE N2. ESTIMATES OF LEVEL-1 AND LEVEL-2 VARIANCE COMPONENTS (RANDOM EFFECTS)

Random effect	Standard Deviation	Variance Component	*df*	Chi-squared	*p* value
School	0.150	0.022	23	26.922	.259
Student	0.689	0.475			

Note. The analysis was conducted using HLM software, which does not provide *df*, test statistic, and *p* value at level-1.

TABLE N3. ESTIMATES OF THE BENCHMARK IMPACT MODELS FOR STUDENTS IN THE LOWEST THIRD OF INCOMING MATH ACHIEVEMENT (N=713)

Fixed effects	Coefficient	Standard error	t-ratio	*df*	*p* value
Intercept	-0.923	0.705	-1.309	23	0.203
Treatment Status	0.162	0.094	1.718	23	0.099
School Size	0.000	0.000	-1.07	23	0.296
School in City	0.101	0.173	0.585	23	0.564
Title 1 Status	-0.655	0.264	-2.479	23	0.021
State CA	0.379	0.454	0.834	23	0.413
ELA pretest	0.211	0.049	4.331	622	<0.001
Math pretest	0.131	0.056	2.348	622	0.019
Grade 4	0.034	0.077	0.446	622	0.655
Male	0.019	0.055	0.345	622	0.730
ELL	-0.039	0.071	-0.541	622	0.589
Asian	0.125	0.135	0.927	622	0.354
Black	-0.082	0.111	-0.738	622	0.461
Hispanic	-0.063	0.101	-0.625	622	0.532
Native Indian	0.158	0.368	0.43	622	0.668
Gender, ELL & Ethnicity Missing	-0.253	0.279	-0.907	622	0.365
Ethnicity Unspecified	0.010	0.136	0.074	622	0.941
FRPL Eligible	0.102	0.094	1.09	622	0.276
FRPL missing	0.278	0.197	1.414	622	0.158
With White teacher	-0.030	0.082	-0.365	622	0.715
Teacher ethnicity missing	0.148	0.227	0.65	622	0.516
Certificate in elementary education	0.280	0.164	1.707	622	0.088
Certificate in English Language development	0.081	0.092	0.873	622	0.383
Highest level of education	-0.042	0.065	-0.65	622	0.516
Missing Highest level of education	-0.049	0.372	-0.132	622	0.895
Years of classroom teaching	0.004	0.004	0.929	622	0.353
Missing in years of classroom teaching	0.021	0.183	0.113	622	0.910
Missing Teacher pretest	0.455	0.240	1.895	622	0.059
Teacher content pretest	0.773	0.346	2.232	622	0.026
Missing Survey	0.095	0.401	0.236	622	0.813

TABLE N3. ESTIMATES OF THE BENCHMARK IMPACT MODELS FOR STUDENTS IN THE LOWEST THIRD OF INCOMING MATH ACHIEVEMENT (N=713)

Fixed effects	Coefficient	Standard error	t-ratio	*df*	*p* value
NGSS Missing	0.132	0.149	0.882	622	0.378
Teacher's gender	-0.306	0.103	-2.969	622	0.003
Taught Science in previous year	0.055	0.125	0.442	622	0.659
Not enough time for Science instruction	-0.046	0.096	-0.481	622	0.631
Composite of school context culture between admins and teachers	0.165	0.100	1.655	622	0.098
Composite of school context culture among teachers	-0.100	0.095	-1.052	622	0.293
NGSS-related activities participated in	0.028	0.025	1.144	622	0.253
Composite of confidence on specific science content	-0.052	0.062	-0.848	622	0.397
Composite of confidence in literacy and discourse	0.028	0.068	0.413	622	0.680
Composite of perceived level of influence	0.013	0.058	0.229	622	0.819
Composite of teaching philosophies	-0.003	0.073	-0.037	622	0.971
Having coaches or mentors for science instruction	-0.025	0.111	-0.224	622	0.823

Note. We do not include estimates for pair fixed effects in the table.

CA = California; ELA = English Language Arts; ELL = English Language Learner; FRPL = Free or Reduced Price Lunch; NGSS = Next Generation Science Standards

TABLE N4. ESTIMATES OF LEVEL-1 AND LEVEL-2 VARIANCE COMPONENTS (RANDOM EFFECTS)

Random effect	Standard Deviation	Variance Component	*df*	Chi-squared	*p* value
School	0.165	0.027	23	29.59186	0.161
Student	0.689	0.475			

Note. The analysis was conducted using HLM software, which does not provide *df*, test statistic, and *p* value at level-1.

Appendix O. Sensitivity Analyses for the Confirmatory Impacts on Science Achievement of Students in the Lowest Third of Incoming Achievement

This appendix presents the sensitivity analyses for confirmatory impacts on student science achievement for students in the lowest third of incoming ELA and math achievement. Student science achievement is measured using selected-response items on the science assessment. The sensitivity analyses include scores derived from different score calibration approaches (percent-correct, 1PL, 2PL, and 3PL) and different model specifications.

ELA Pretest

TABLE O1. IMPACT FINDINGS BASED ON 3PL SCALING ALTERNATIVE MODELS N (SCHOOLS = 55, N (STUDENTS) = 715

	Impact (SE)	t	*df*	*p* value	Effect size	Change percentile rank
Benchmark analysis full covariate set [a]	.054 (.093)	0.581	23	.567	.073	2.9%
Sensitivity Analysis (Alternative Models)						
Like benchmark, no covariates, no blocks	.072 (.068)	1.061	53	.293	.098	3.9%
Like benchmark, no covariates	.094 (.069)	1.378	26	.180	.128	5.1%
Like benchmark, pretests are only covariates	.083 (.066)	1.254	26	.221	.113	4.5%
Include teacher random effect	0.054(0.093)	0.58	24	.565	0.073	2.9%
Use pair random (instead of fixed) effects	0.022 (0.072)	0.30	24	.766	0.030	1.2%
Ignore pair level	0.022 (0.072)	0.30	50	.764	0.030	1.2%
Ignore school, model random intercept and treatment at pair level	0.022 (0.072)	0.31	22	.760	0.030	1.2%
Use ML instead of REML	0.047 (0.074)	0.64	24	.530	0.064	2.5%
OLS	0.047 (0.078)	0.61	647	.545	0.064	2.5%

Note. Most covariates are modeled at the student level. Mean performance in the treatment group was -0.607 units (.747 *sd*) in the treatment group, and -0.688 units (0.723 *sd*) in the control group.

OLS is Ordinary Least Squares. ML is Full Maximum Likelihood. REML is Restricted Maximum Likelihood.

[a] Full results of benchmark model are provided in Appendix N.

TABLE O2. IMPACT FINDINGS BASED ON PERCENT-CORRECT METRIC WITH REDUCED-ITEMS FORMS N (SCHOOLS) = 55, N (STUDENTS) = 715

	Impact (SE)	t	*df*	*p* value	Effect size	Change percentile rank
Benchmark analysis full covariate set [a]	0.041 (0.093)	0.44	24	.662	0.056	2.2%
Sensitivity Analysis (Alternative Models)						
Like benchmark, no covariates, no blocks	0.061 (0.067)	0.91	53	.369	0.083	3.3%
Like benchmark, no covariates	0.084 (0.063)	1.33	26	.196	0.115	4.6%
Like benchmark, pretests are only covariates	0.070 (0.061)	1.15	26	.260	0.096	3.8%
Include teacher random effect	0.042 (0.094)	0.45	24	.659	0.057	2.3%
Use pair random (instead of fixed) effects	0.030 (0.069)	0.45	23	.658	0.041	1.6%
Ignore pair level	0.031 (0.069)	0.45	50	.655	0.042	1.7%
Ignore school, model random intercept and treatment at pair level	0.031 (0.069)	0.45	23	.655	0.042	1.7%
Use ML instead of REML	0.038 (0.073)	0.52	24	.608	0.052	2.1%
OLS	0.038 (0.077)	0.49	647	.621	0.052	2.1%

Note. Most covariates are modeled at the student level. Mean performance in the treatment group was -0.671 units (0.745 *sd*) in the treatment group, and -0.599 units (0.720 *sd*) in the control group.

OLS is Ordinary Least Squares. ML is Full Maximum Likelihood. REML is Restricted Maximum Likelihood.

[a] Full results of benchmark model are provided in Appendix N.

Math Pretest

TABLE O3. IMPACT FINDINGS BASED ON 3PL SCALING ALTERNATIVE MODELS N (SCHOOLS) = 55, N (STUDENTS) = 713

	Impact (SE)	t	*df*	*p* value	Effect size	Change percentile rank
Benchmark analysis full covariate set [a]	.162 (0.094)	1.718	23	.099	.220	8.7%
Sensitivity Analysis (Alternative Models)						
Like benchmark, no covariates, no blocks	.060 (0.078)	0.769	53	.445	.081	3.2%
Like benchmark, no covariates	.092 (0.074)	1.246	26	.224	.125	5.0%
Like benchmark, pretests are only covariates	.091 (0.073)	1.258	26	.220	.123	4.8%
Include teacher random effect	0.151 (0.096)	1.57	23	.129	0.204	8.1%
Use pair random (instead of fixed) effects	0.052 (0.071)	0.73	23	.473	0.070	2.8%
Ignore pair level	0.052 (0.071)	0.73	49	.469	0.070	2.8%
Ignore school, model random intercept and treatment at pair level	0.053 (0.072)	0.74	22	.465	0.072	2.9%

TABLE O3. IMPACT FINDINGS BASED ON 3PL SCALING ALTERNATIVE MODELS N (SCHOOLS) = 55, N (STUDENTS) = 713

	Impact (SE)	t	*df*	*p* value	Effect size	Change percentile rank
Use ML instead of REML	0.147 (0.073)	2.01	23	.056	0.200	7.9%
OLS	0.147 (0.077)	1.92	645	.056	0.200	7.9%

Note. Most covariates are modeled at the student level. Mean performance in the treatment group was -0.600 units (.758 *sd*) in the treatment group, and -0.678 units (0.720 *sd*) in the control group.

OLS is Ordinary Least Squares. ML is Full Maximum Likelihood. REML is Restricted Maximum Likelihood.

[a] Full results of benchmark model are provided in Appendix N.

TABLE O4. IMPACT FINDINGS PERCENT CORRECT METRIC WITH REDUCED-ITEMS FORMS N (SCHOOLS) = 55, N (STUDENTS) = 713

	Impact (SE)	t	*df*	*p* value	Effect size	Change percentile rank
Benchmark analysis full covariate set [a]	0.162 (0.102)	1.58	24	.127	0.220	8.7%
Sensitivity Analysis (Alternative Models)						
Like benchmark, no covariates, no blocks	0.070 (0.080)	0.88	53	.383	0.095	3.7%
Like benchmark, no covariates	0.088 (0.080)	1.11	26	.276	0.120	4.8%
Like benchmark, pretests are only covariates	0.089 (0.078)	1.14	26	.265	0.121	4.8%
Include teacher random effect	0.162 (0.102)	1.58	24	.127	0.220	8.7%
Use pair random (instead of fixed) effects	0.074 (0.073)	1.01	23	.323	0.101	4.0%
Ignore pair level	0.074 (0.073)	1.01	49	.318	0.101	4.0%
Ignore school, model random intercept and treatment at pair level	0.076 (0.073)	1.04	23	.309	0.104	4.1%
Use ML instead of REML	0.146 (0.073)	2.00	24	.569	0.200	7.9%
OLS	0.146 (0.077)	1.90	645	.058	0.200	7.9%

Note. Most covariates are modeled at the student level. Mean performance in the treatment group was -0.671 units (0.737 *sd*) in the treatment group, and -0.580 units (0.730 *sd*) in the control group.

OLS is Ordinary Least Squares. ML is Full Maximum Likelihood. REML is Restricted Maximum Likelihood.

[a] Full results of benchmark model are provided in Appendix N.

Appendix P. Sample Sizes and Baseline Equivalence for the Impact on Student Science Achievement for Specific Subsamples

This appendix presents the sample sizes and baseline equivalence for the impact on student science achievement (selected-response items) for specific subsamples (Focused Samples 1 and 2, by state, and by grade).

FOCUSED SAMPLE 1

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. Counts are shown in Table P1.

TABLE P1. FOCUSED SAMPLE 1

	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	15	432	13	287
Control	12	382	11	314
Total N	27	814	24	601

Note. *MSS* stands for the group of students of teachers receiving the Making Sense of SCIENCE program. CA is California. WI is Wisconsin. Posttest is the student science achievement assessment.

We tested baseline equivalence for (a) ELA pretest, (b) math pretest. Results are in Table P2. We observed that baseline equivalence is established for both ELA pretest and math pretest.

TABLE P2. TESTS OF BASELINE EQUIVALENCE BETWEEN CONDITIONS ON ELA AND MATH FOR FOCUSED SAMPLE 1

	ELA pretest	Math pretest
N (schools)	51	
N (students)	1,415	
Point estimate for difference between conditions	-0.099	-0.077
Standard error	0.106	0.115
***p* value**	.360	.510
Standardized effect size	-0.101	-0.081

FOCUSED SAMPLE 2

The sample included 340 students (167 treatment, 173 control) who had both grade 3 state ELA and math pretests, with 178 students from California and 162 students from Wisconsin. Counts are shown in Table P3.

TABLE P3. FOCUSED SAMPLE 2

	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	12	113	6	60
Control	5	65	8	102
Total N	17	178	14	162

Note. *MSS* stands for the group of students of teachers receiving the Making Sense of SCIENCE program. CA is California. WI is Wisconsin. Posttest is the student science achievement assessment.

We tested baseline equivalence for (a) ELA pretest and (b) math pretest. Results are in Table P4. We observed that baseline equivalence is established for both ELA pretest and math pretest.

TABLE P4. TESTS OF BASELINE EQUIVALENCE BETWEEN CONDITIONS ON ELA AND MATH FOR FOCUSED SAMPLE 2

	ELA pretest	Math pretest
N (schools)	31	
N (students)	340	
Point estimate for difference between conditions	0.001	-0.045
Standard error	0.166	0.166
***p* value**	.993	.788
Standardized effect size	0.001	-0.047

IMPACT BY STATE

TABLE P5. SAMPLE FOR ANALYSIS OF IMPACT BY STATE

	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	1,303	25	837

Note. *MSS* stands for the group of students of teachers receiving the Making Sense of SCIENCE program. CA is California. WI is Wisconsin. Posttest is the student science achievement assessment.

We tested baseline equivalence for both ELA pretest and math pretest in both California and Wisconsin samples. Results are in Table P6. We observed that baseline equivalence is established for both the ELA pretest and the math pretest in both California and Wisconsin samples.

TABLE P6. TESTS OF BASELINE EQUIVALENCE BETWEEN CONDITIONS ON ELA AND MATH FOR SAMPLE USED IN ANALYSIS OF IMPACT BY STATE

	California sample		Wisconsin sample	
	ELA pretest	Math pretest	ELA pretest	Math pretest
N (Schools)	30		25	
N (Students)	1,303		837	
Point estimate for difference between conditions	-0.101	-0.049	-0.101	-0.043
Standard error	0.122	0.132	0.135	0.147
***p* value**	.420	.719	.471	.776
Standardized effect size	-0.106	-0.051	-0.102	-0.045

IMPACT BY GRADE

TABLE P7. SAMPLE FOR ANALYSIS OF IMPACT BY GRADE

	Grade 4		Grade 5	
	Count of schools	Count of students with posttest	Count of schools	Count of students with posttest
MSS	27	611	28	527
Control	24	609	21	393
Total N	51	1,220	49	920

Note. *MSS* stands for the group of students of teachers receiving the Making Sense of SCIENCE program.

We tested baseline equivalence for both ELA pretest and math pretest in both Grade 4 and Grade 5 samples. Results are in Table P8. Baseline equivalence is established for both the ELA pretest and math pretest in both Grade 4 and Grade 5 samples.

TABLE P8. TESTS OF BASELINE EQUIVALENCE BETWEEN CONDITIONS ON ELA AND MATH FOR SAMPLE USED IN ANALYSIS OF IMPACT BY STATE

	Grade 4 sample		Grade 5 sample	
	ELA pretest	Math pretest	ELA pretest	Math pretest
N (Schools)	51		49	
N (Students)	1220		920	
Point estimate for difference between conditions	-0.083	-0.022	-0.078	0.035
Standard error	0.104	0.096	0.103	0.124
***p* value**	.434	.818	.455	.780
Standardized effect size	-0.086	-0.023	-0.080	0.035

Appendix Q. Supplemental Analysis on the Impact on Student Science Achievement under High Fidelity of Implementation

There are several alternatives for evaluating the impact of a program under the condition of high fidelity of implementation. We adapted an approach by Unlu et al. (2010). Assessing impact on student achievement under high fidelity of implementation (FOI) required following these steps.

1. Specify a rule for identifying teachers who are above a specific threshold of actual implementation (in the treatment group), whom we refer to as "high implementers."
2. Apply a model to predict high implementation in the treatment group using a set of teacher baseline covariates.
3. Apply the model developed under 2 to identify a matched sample of control teachers who plausibly would have implemented at the same above-threshold levels had they been randomly assigned to treatment.
4. Assess the impact for students of teachers who are either strongly implementing (in treatment) or selected as potentially high implementers using model-based results in (in control).

While we explored several variants of the method, analysis was fundamentally limited in two ways. First, it was difficult to obtain an adequately powered estimate of the relationship between baseline (endogenous) characteristics and FOI. FOI was assessed based on attendance in professional learning events. There was variability in attendance over the two-year implementation. Attendance was determined in large part by assignment of teachers to study-eligible classes. Teachers who joined the study late would receive less than full professional learning. While some of the variation on FOI could be attributable to teacher-level endogenous factors (including, for example, lower motivation leading to late joining), many of the differences in FOI (professional learning attendance) were based on mobility resulting from a combination of teacher- and school organizational factors that could not be easily captured through surveys of teacher baseline characteristics. Factors affecting joining and professional learning (and FOI) levels were likely not sufficiently exogenous to serve as an instrument, while also noisy enough that we could not, with precision, relate teacher baseline characteristics to FOI outcomes. Essentially, administration-controlled mobility would add a lot of noise to the variability in FOI, in a way that would not be predictive of achievement, producing a highly underpowered analysis of the effects of dosage.

To address the first limitation, we limited the sample to teachers in both conditions who remained in the study for both years. This eliminates the influence of teacher movement between eligible and non-eligible grades/subjects as a source of variance in FOI. This, however, reduced the sample size of teachers, and most remaining treatment teachers were fully or close-to-fully compliant with full FOI, making it impossible to model the relationship between teacher-endogenous variables and FOI. This precludes using the methods above.

We therefore relied on analysis of impacts on student science achievement using focused samples 1 and 2 (see Chapter 6). This approach at least held constant the length of time the teacher spent in the study, or the exposure of students to teachers in the study in both conditions.

Appendix R. Detailed Impact Analysis Findings for the Constructed Response Items

This appendix presents the details of the analysis of the impact on student communication of science in writing (Chapter 8), as measured by constructed-response items on the student science assessment, at the item-level.

TABLE R1. SANDSTONE (N = 449 STUDENTS) GRADES 4 AND 5

	Model 1	Model 2	Model 3 [a]	Model 4	Model 5	Model 6 [a]
Fixed effects						
Treatment	-0.014 (0.029) *p* = .704	0.002 (0.030) *p* = .952	0.036 (0.029) *p* = .210	-0.004 (0.031) *p* = .898	0.0003 (0.027) *p* = .991	0.043 (0.030) *p* = .153
Pretests (ELA and math)		X	X		X	X
Other covariates			X			X
Matched pairs				X	X	X
Standardized effect size	-0.042	0.006	0.109	-0.012	0.001	0.130
Random effects						
Pair	.002 (.004) *p* = .309	.002 (.003) *p* = .203	0 [c]			
School [b]	.007 (.005) *p* = .076	.002 (.003) *p* = .292	0 [c]	0 [c]	0 [c]	0 [c]
Student	.076 (.005) *p* < .001	.076 (.005) *p* < .001	.066 (.004) *p* < .001	.092 (.006) *p* < .001	.073 (.005) *p* < .001	.062 (.004) *p* < .001

Note. Performance on this item was rated in terms of five ordinal categories. Quantities in parentheses are standard errors; All effect sizes are regression-adjusted impact estimates divided by the control *sd* for the outcome distribution for the control group.

[a] We also evaluated impact varies across grades for Model 3 and 6. We observed no differential effect across for Model 3 (*p* = .843) or for Model 6 (*p* = .960)

[b] The school is the unit randomized.

[c] Zero effect estimate with no *p* value indicates that estimation met boundary condition for quantity. This often indicates that the quantity is trivially different from zero.

The difference between treatment and control in the cumulative log odds of a higher-rated response for models like 3 and 6, but where we modeled student responses as ordinal, are as follows:

Model 3: .257 (*p* = .219), LOR(Cox) = .160

Model 6: .326 (*p* = .148), LOR(Cox) = .198

TABLE R2. BASKETBALL (N = 266 STUDENTS) 5TH GRADE ONLY

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Fixed effects						
Treatment	0.079 (0.034) *p* = .020	0.088 (0.031) *p* = .004	0.073 (0.035) *p* = .041	0.091 (0.034) *p* = .009	0.088 (0.032) *p* = .006	0.109 (0.042) *p* = .010
Pretests		X	X		X	X
Other Covariates			X			X
Matched Pairs				X	X	X
Standardized ES	0.304	0.338	0.281	0.350	0.338	0.419
Random effects						
Pair	.002 (.004) *p* = .249	.002 (.002) *p* = .227	0**			
School [a]	.0004 (.004) *p* = .453	0 [b]	0 [b]	0 [b]	0 [b]	0 [b]
Student	.067 (.006) *p* <.001	.056 (.005) *p* <.001	.050 (.004) *p* <.001	.060 (.005) *p* <.001	.050 (.004) *p* <.001	.044 (.004) *p* <.001

Note. Performance on this item was rated in terms of eight ordinal categories (zero and blanks were combined). Quantities in parentheses are standard errors; All effect sizes are regression adjusted impact estimates divided by the control *sd* for the outcome distribution for the control group.

[a] The school is the unit randomized.

[b] Zero effect estimate with no p value indicates that estimation met boundary condition for quantity. This often indicates that the quantity is trivially different from zero.

The difference between treatment and control in the cumulative log odds of a higher-rated response for models like 3 and 6 (the main models), but where we modeled student responses as ordinal, are as follows:

Model 3: .492 (p = .083), LOR(Cox) = .298

Model 6: .827 (p = .021), LOR(Cox) = .501

TABLE R3. BIRDFOOD (N = 427 FOURTH- AND FIFTH-GRADES STUDENTS)

	Model 1	Model 2	Model 3 [a]	Model 4	Model 5	Model 6 [a]
Fixed effects						
Treatment	0.002 (0.023) *p* = .920	0.009 (0.020) *p* = .673	0.008 (0.022) *p* = .728	0.007 (0.023) *p* = .754	0.006 (0.021) *p* = .779	0.012 (0.023) *p* = .604
Pretests		X	X		X	X
Other Covariates			X			X
Matched Pairs				X	X	X
Standardized ES	0.009	0.039	0.035	0.030	0.026	0.052
Random effects						
Pair	.001 (.001) *p* = .171	0 [c]	0 [c]			
School [b]	.0001 (.002) *p* = .465	0 [c]	0 [c]	0 [c]	0 [c]	0 [c]
Student	.052 (.004) *p* <.001	.043 (.003) *p* <.001	.037 (.002) *p* <.001	.049 (.003) *p* <.001	.041 (.003) *p* <.001	.035 (.002) *p* <.001

Note. Performance on this item was rated in terms of five ordinal categories (zero and blanks were combined). Quantities in parentheses are standard errors; All effect sizes are regression adjusted impact estimates divided by the control *sd* for the outcome distribution for the control group.

[a] We also evaluated whether impact varies across grades for Model 3 and 6. We observed no differential effect across for Model 3 (*p* = .580) or for Model 6 (*p* = .397).

[b] The school is the unit randomized

[c] Zero effect estimate with no p value indicates that estimation met boundary condition for quantity. This often indicates that the quantity is trivially different from zero.

The difference between treatment and control in the cumulative log odds of a higher-rated response for models like 3 and 6, but where we modeled student responses as ordinal, are as follows:

Model 3: .069 (p = .778), LOR(Cox) = .041

Model 6: .181 (p = .507), LOR(Cox) = .109

TABLE R4. CLOUDY DAYS (N = 638 FOURTH- AND FIFTH-GRADES STUDENTS)

	Model 1	Model 2	Model 3 [a]	Model 4	Model 5	Model 6 [a]
Fixed effects						
Treatment	0.015 (0.025) *p* =.532	0.036 (0.023) *p* =.131	0.032 (0.027) *p* =.235	0.024 (0.026) *p* =.350	0.029 (0.025) *p* =.244	0.032 (0.030) *p* =.280
Pretests		X	X		X	X
Other Covariates			X			X
Matched Pairs				X	X	X
Standardized ES	0.048	0.116	0.103	0.077	0.094	0.103
Random effects						
Pair	0 [c]	0 [c]	0 [c]			
School [b]	0 [c]	0 [c]	0 [c]	0 [c]	0 [c]	0 [c]
Student	.101 (.005) *p* <.001	.087 (.005) *p* <.001	.081 (.005) *p* <.001	.097 (.005) *p* <.001	.084 (.005) *p* <.001	.078 (.004) *p* <.001

Note. Performance on this item was rated in terms of eight ordinal categories (zero and blanks were combined). Quantities in parentheses are standard errors; All effect sizes are regression adjusted impact estimates divided by the control *sd* for the outcome distribution for the control group.

[a] We also evaluated whether impact varies across grades for Model 3 and 6. We observed no differential effect across for Model 3 (*p* = .653) or for Model 6 (*p* = .711).

[b] The school is the unit randomized

[c] Zero effect estimate with no p value indicates that estimation met boundary condition for quantity. This often indicates that the quantity is trivially different from zero.

The difference between treatment and control in the cumulative log odds of a higher-rated response for models like 3 and 6, but where we modeled student responses as ordinal, are as follows:

Model 3: .306 (*p* = .156), LOR(Cox) = .185

Model 6: .321 (*p* = .202), LOR(Cox) = .194

TABLE R5. PEA SEEDS (N = 260 FIFTH GRADE STUDENTS)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Fixed effects						
Treatment	-0.027 (0.029) *p* = .347	-0.002 (0.026) *p* = .932	-0.004 (0.031) *p* = .891	0.004 (0.029) *p* = .886	0.012 (0.027) *p* = .653	0.004 (0.037) *p* = .919
Pretests		X	X		X	X
Other Covariates			X			X
Matched Pairs				X	X	X
Standardized ES	-0.113	-0.008	-0.017	0.017	0.050	0.017
Random effects						
Pair	0 [b]	0 [b]	0 [b]			
School [a]	.0007 (.002) *p* = .356	0 [b]	0 [b]	0 [b]	0 [b]	0 [b]
Student	.048 (.005) *p* < .001	.041 (.004) *p* < .001	.036 (.003) *p* < .001	.042 (.004) *p* < .001	.038 (.003) *p* < .001	.032 (.003) *p* < .001

Note. Performance on this item was rated in terms of four ordinal categories (zero and blanks were combined). Quantities in parentheses are standard errors; All effect sizes are regression adjusted impact estimates divided by the control *sd* for the outcome distribution for the control group.

[a] The school is the unit randomized

[b] Zero effect estimate with no p value indicates that estimation met boundary condition for quantity. This often indicates that the quantity is trivially different from zero.

The difference between treatment and control in the cumulative log odds of a higher-rated response for models like 3 and 6, but where we modeled student responses as ordinal, are as follows:

Model 3: .091 (*p* = .828), LOR(Cox) = .055

Model 6: .403 (*p* = .584), LOR(Cox) = .244

TABLE R6. BOILING WATER (N = 187 FIFTH GRADE STUDENTS)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Fixed effects						
Treatment	-0.030 (0.062) *p* = .626	0.011 (0.057) *p* = .850	0.062 (0.066) *p* = .354	0.012 (0.060) *p* = .838	0.033 (0.056) *p* = .558	0.090 (0.076) *p* = .234
Pretests		X	X		X	X
Other Covariates			X			X
Matched Pairs				X	X	X
Standardized ES	-0.075	0.028	0.155	0.030	0.083	0.225
Random effects						
Pair	.011 (.013) *p* = .206	009 (.008) *p* = .145	0 [b]			
School [a]	.001 (.013) *p* = .465	0 [b]	0 [b]	0 [b]	0 [b]	0 [b]
Student	.161 (.019) *p* < .001	.139 (.015) *p* < .001	.122 (.012) *p* < .001	.136 (.014) *p* < .001	.118 (.012) *p* < .001	.095 (.010) *p* < .001

Note. Performance on this item was rated in terms of three ordinal categories (zero and blanks were combined). Quantities in parentheses are standard errors; All effect sizes are regression adjusted impact estimates divided by the control *sd* for the outcome distribution for the control group.

[a] The school is the unit randomized

[b] Zero effect estimate with no p value indicates that estimation met boundary condition for quantity. This often indicates that the quantity is trivially different from zero.

The difference between treatment and control in the cumulative log odds of a higher-rated response for models like 3 and 6, but where we modeled student responses as ordinal, are as follows:

Model 3: .337 (*p* = .421), LOR(Cox) = .204

Model 6: 1.651 (*p* = .036), LOR(Cox) = 1.00

TABLE R7. ICE CUBE (N = 195 FIFTH GRADE STUDENTS)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Fixed effects						
Treatment	0.031 (0.078) *p* =.692	-0.005 (0.075) *p* =.994	-0.003 (0.083) *p* =.971	0.073 (0.065) *p* =.262	0.006 (0.062) *p* =.919	-0.023 (0.083) *p* =.785
Pretests [2]		X	X		X	X
Other Covariates [38]			X			X
Matched Pairs [27]				X	X	X
Standardized ES	0.069	-0.011	-0.007	0.162	0.013	-0.051
Random effects						
Pair	0 [b]	0 [b]	0 [b]			
School [a]	.025 (.013) *p* =.030	.025 (.012) *p* =.019	.015 (.011) *p* =.102	0 [b]	0 [b]	0 [b]
Student	.173 (.020) *p* <.001	.144 (.016) *p* <.001	.129 (.015) *p* <.001	.161 (.016) *p* <.001	.139 (.014) *p* <.001	.107 (.011) *p* <.001

Note. Performance on this item was rated in terms of three ordinal categories (zero and blanks were combined). Quantities in parentheses are standard errors; All effect sizes are regression adjusted impact estimates divided by the control *sd* for the outcome distribution for the control group.

[a] The school is the unit randomized

[b] Zero effect estimate with no p value indicates that estimation met boundary condition for quantity. This often indicates that the quantity is trivially different from zero.

The difference between treatment and control in the cumulative log odds of a higher-rated response for models like 3 and 6, but where we modeled student responses as ordinal, are as follows:

Model 3: -.163 (p = .764), LOR(Cox) = -.100

Model 6: -.267 (p = .740), LOR(Cox) = - .162

TABLE R8. MINERAL SCRATCH (N = 378 FOURTH- AND FIFTH-GRADE STUDENTS)

	Model 1	Model 2	Model 3 [a]	Model 4	Model 5	Model 6 [a]
Fixed effects						
Treatment	-0.008 (0.040) *p* = .832	0.004 (0.037) *p* = .910	-0.024 (0.042) *p* = .564	0.010 (0.040) *p* = .793	0.013 (0.038) *p* = .730	-0.017 (0.045) *p* = .711
Pretests		X	X		X	X
Other Covariates			X			X
Matched Pairs				X	X	X
Standardized ES	-0.021	0.011	-0.063	0.026	0.034	-0.045
Random effects						
Pair	.007 (.004) *p* = .058	.002 (.003) *p* = .257	0 [c]			
School [b]	0 [c]	0 [c]	0 [c]	0 [c]	0 [c]	0 [c]
Student	.140 (.011) *p* < .001	.123 (.009) *p* < .001	.110 (.008) *p* < .001	.114 (.008) *p* < .001	.084 (.005) *p* < .001	.103 (.007) *p* < .001

Note. Performance on this item was rated in terms of three ordinal categories (zero and blanks were combined). Quantities in parentheses are standard errors; All effect sizes are regression adjusted impact estimates divided by the control *sd* for the outcome distribution for the control group.

[a] We also evaluated whether impact varies across grades for Model 3 and 6. We observed no differential effect across for Model 3 (*p* = .908) or for Model 6 (*p* = .265).

[b] The school is the unit randomized

[c] Zero effect estimate with no p value indicates that estimation met boundary condition for quantity. This often indicates that the quantity is trivially different from zero.

The difference between treatment and control in the cumulative log odds of a higher-rated response for models like 3 and 6, but where we modeled student responses as ordinal, are as follows:

Model 3: -.143 (*p* = .568), LOR(Cox) = -.086

Model 6: -.096 (*p* = .730), LOR(Cox) = -.049

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.

Bloom, H. S., Hill, C. J., Black, A. B., and Lipsey, M. W. (2008). Performance Trajectories and Performance Gaps as Achievement Effect-Size Benchmarks for Educational Interventions. *Journal of Research on Educational Effectiveness, 1*(4), 289-328.

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.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]*. Lincolnwood, IL: Scientific Software International.

Calvert, L. (2016). The power of teacher agency. *The Learning Professional*, *37*(2), 51.

Cameron, M., & Lovett, S. (2015). Sustaining the commitment and realising the potential of highly promising teachers. *Teachers and Teaching, 21*(2), 150-163.

Carlson, J., & Daehler, K. R. (2019). The refined consensus model of pedagogical content knowledge in science education. In *Repositioning pedagogical content knowledge in teachers' knowledge for teaching science* (pp. 77-92). Springer, Singapore.

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.

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.

Darling-Hammond, L., & Richardson, N. (2009). Research review/teacher learning: What matters. *Educational Leadership*, *66*(5), 46-53

Elliott, S. N., & Bartlett, B. J. (2016). *Opportunity to Learn*. Oxford Handbooks Online. Oxford University Press.

Evans, B. R. (2011). Content Knowledge, Attitudes, and Self-Efficacy in the Mathematics New York City Teaching Fellows (NYCTF) Program. *School Science and Mathematics*, *111*(5), 225-235.

Frederick, W. C., & Walberg, H. J. (1980). Learning as a function of time. *The Journal of Educational Research, 73*(4), 183–194. https://doi.org/10.1080/00220671.1980.10885233

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

Greenleaf, C. L., Litman, C., Hanson, T. L., Rosen, R., Boscardin, C. K., Herman, J., ... & Jones, B. (2011). Integrating literacy and science in biology: Teaching and learning impacts of reading apprenticeship professional development. *American Educational Research Journal*, *48*(3), 647-717.

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.

Harrison, C., & Killion, J. (2007). Ten roles for teacher leaders. *Educational Leadership*, *65*(1), 74.

Hmelo-Silver, C. E., & Barrows, H. S. (2008). Facilitating collaborative knowledge building. *Cognition and Instruction*, *26*(1), 48-94.

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.

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.

Jenkins, B. (2009). What it takes to be an instructional leader. *Principal*, *88*(3), 34-37.

Jones, M. G., & Carter, G. (2013). Science teacher attitudes and beliefs. In *Handbook of research on science education* (pp. 1081-1118). Routledge.

Katzenmeyer, M., & Moller, G. (2009). *Awakening the Sleeping Giant: Helping Teachers Develop as Leaders*. Corwin Press.

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.

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

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

National Research Council. (1996). *National Science Education Standards.* National Academic Press. https://doi.org/10.17226/4962.

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. National Academies Press.

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

Pajares, M. F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research*, *62*(3), 307-332.

Penuel, W. R., Harris, C. J., & DeBarger, A. H. (2014). *Implementing the Next Generation Science Standards: Strategies for Educational Leaders*.

Rubie-Davies, C. (2009). Teacher expectations and labeling. In *International handbook of research on teachers and teaching* (pp. 695-707). Springer.

Singer, J. D., & Willett, (2003). *Applied longitudinal data analysis*. New York: Oxford University Press.

Supovitz, J. A., & Turner, H. M. (2000). The effects of professional development on science teaching practices and classroom culture. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, *37*(9), 963-980.

Tekkumru-Kisa, M., Kisa, Z., & Hiester, H. (2020). Intellectual work required of students in science classrooms: Students' opportunities to learn science. *Research in Science Education*, 1-15.

Tschannen-Moran, M., Hoy, A. W., & Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Review of Educational Research, 68*(2), 202-248

Thapa, A., Cohen, J., Guffey, S., & Higgins-D'Alessandro, A. (2013). A review of school climate research. *Review of Educational Research, 83*(3), 357-385.

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

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.

Wright, K. L., Franks, A. D., Kuo, L. J., McTigue, E. M., & Serrano, J. (2016). Both theory and practice: Science literacy instruction and theories of reading. *International Journal of Science and Mathematics Education*, *14*(7), 1275-1292.

Unlu, F., Bozzi, L., Layzer, C., Smith, A., Price, C., & Hurtig, R. (2010, October). Linking implementation fidelity to impacts in an RCT: A matching approach. In *symposium: Using matching methods to analyze RCT impacts on program-related subgroups, Association for Public Policy Analysis and Management* (Vol. 13).

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.