Assessing the Effect of Project-Based Learning on Science Learning in Elementary Schools

Joseph Krajcik[a], Barbara Schneider[a], Emily Miller, I-Chien Chen[a], Lydia Bradford[a], Kayla Bartz[a], Quinton Baker[c] , Annemarie Palinscar[b], Deborah Peek-Brown[a], and Susan Codere[a]

[a] Michigan State University, [b] University of Michigan, [c]United States Department of Agriculture

Technical Report

January 11, 2021

Michigan State University

A special thanks to Professor Elizabeth Tipton and the three anonymous reviewers.

The research reported here was funded by the George Lucas Educational Foundation. The opinions expressed are those of the authors and not the foundation.

Abstract

The study aimed to determine whether the Multiple Literacies Project-Based Learning (ML-PBL) intervention, which incorporates *A Framework for K–12 Science Education* (National Research Council, 2012) and the Next Generation Science Standards ([NGSS] NGSS Lead States, 2013), improved science academic, social, and emotional learning. Features of project-based learning (PBL) and three-dimensional learning were used to design the ML-PBL units. The intervention included a set of science units and materials, professional development, and unit assessments—all of which were designed to increase students' science knowledge, literacy, and mathematical skills, and support social and emotional learning through self-reflection, collaboration, and ownership of one's work. A randomized control trial in third-grade classrooms was conducted in 46 Michigan schools (23 treatment and 23 control), which included 4 different regions throughout the state, with a total of 2,371 students in the analytic sample. A three-level hierarchical linear model (HLM) was used to assess the difference between the treatment and control students in science achievement to account for the clustering of students within classrooms within schools. Results of the HLM show that the treatment students outperformed the control students by a 0.266 standard deviation on an objective summative test designed to meet the NGSS. This standard deviation corresponds to an eight-percentage-point increase in student science achievement scores. The treatment effect holds when accounting for differing: student reading ability (benchmark) and gender; school level race, ethnicity and SES; and across the regions of the state (which include urban, suburban, and rural areas). A factor analysis conducted on the SEL measure confirms the three constructs: self-reflection, collaboration, and ownership. A three-level HLM was then conducted on each of the constructs. Both reflection and collaboration show positive treatment effect on students' social and emotional learning in science classes. These results suggest that the incorporation of PBL features and three-dimensional learning—which supports students' ability to figure out, make sense of phenomena, and build artifacts that represent responses to the driving question—improve students' science knowledge and social and emotional learning. Additionally, analyses of generalizability indexes show that the average treatment effect is also generalizable to the state of Michigan and the US population. Results of the intervention provide a new approach to science learning that transforms how the students view the world and supports an equitable pathway to learning science.

Introduction

Facing the consequences of a worldwide pandemic and unprecedented climate change, the need for today's students to develop an understanding of scientific ideas and practices is unquestionably more crucial than at any other period in modern history. The growing demand for greater science knowledge is occurring only a few years after the scientific and policy community raised serious concerns around reforming traditional science learning and instruction (see *A Framework for K–12 Science Education* [National Research Council, 2012] and the Next Generation Science Standards [NGSS Lead States, 2013]). Though endorsed by the education and science community, what was missing from these reports was the "how": How should teachers increase their students' science engagement and learning? In response, Multiple Literacies in Project-Based Learning (ML-PBL) was developed, recognizing the need for an empirically tested innovative intervention that would deepen students' use of scientific knowledge and practices to increase their science learning. Beginning early in the students' schooling careers, ML-PBL was designed to support elementary students' science learning, and included high-quality teacher and student curriculum and materials, teacher professional learning (PL) experiences, and classroom-based assessment tasks (Krajcik et al., 2015; Miller & Krajcik, 2019). Anchored in the principles of project-based learning ([PBL], see Krajcik & Shin, 2014)—with its focus on having students investigate questions that they find meaningful and that are aligned with recent policy efforts—ML-PBL worked to transform classrooms into places where students work together to generate knowledge and solve meaningful problems. Reiterating the ML-PBL design process with a pilot and field-test, the most recent initiative was an efficacy cluster randomized trial to determine if the intervention enhanced students' science academic, social, and emotional learning. This report details the 2018–2019 results of the ML-PBL intervention conducted in Michigan with 2,600 third graders and their teachers.

Few curricular innovations have the scope and depth of ML-PBL, which incorporates what is currently known about the teaching and learning of science (National Research Council, 1999; National Research Council, 2007; Sawyer, 2014; National Research Council, 2012; NGSS Lead States, 2013). The NRC 2007, *Framework for K-12 Science Education* advocates changing classroom learning from acquiring disconnected science facts and memorized procedures to environments where students make sense of phenomena and design solutions to complex real-world problems using the three dimensions of scientific knowledge. These three dimensions include disciplinary core ideas (DCIs), science and engineering practices (SEP), and crosscutting concepts (CCCs). DCIs are the fundamental ideas for the scientific disciplines of earth and space sciences, physical science, life science, and engineering design, which focus on the most powerful and generative ideas of science that build across the K–12 spectrum.The SEPs describe how scientists and engineers explore the natural and design world, increasing in complexity across the grades. CCCs are ideas that scientists apply across the disciplines to explore phenomena or solve problems, and serve as a lens for examining phenomena and problems. These dimensions, when used together, are often referred to as "three-dimensional learning," and allow learners to explain a range of natural phenomena and solve engineering problems. Although each of the dimensions is important on its own, together they support students in a figuring-out process central for exploring and explaining phenomena.

The Next Generation Science Standards ([NGSS] NGSS Lead States, 2013) provide a set of performance expectations (PEs) that integrate the *Framework*'s three dimensions of scientific knowledge, i.e., SEPs, DCIs, and CCCs. The NGSS performance expectations require the use of knowledge (Pellegrino & Hilton, 2014), not just the "knowing." What is critical is knowing "how to use knowledge" to make sense of the world. This vision of the *Framework* and NGSS has been widely adopted or adapted in multiple states (National Science Teachers Association, 2019). However, there is a lack of research on evidence-based curricular materials that align with the NGSS. The goal of this study was to fill the limited evidence-based science research at the elementary school level by undertaking a rigorous test of the ML-PBL intervention, designed to answer the following four research questions:

1. What is the main effect of this intervention on third-grade students' science learning? Do ML-PBL treatment students outperform students in the control group on an independent summative science assessment?
2. Does the treatment support more positive responses on an instrument measuring students' social and emotional learning in their science classes compared to the control students?
3. Does the treatment effect differ by student gender, reading proficiency, or school-level characteristics (i.e., proportion of race and ethnic groups and socioeconomic status [SES])?
4. Does fidelity of implementation by the teacher at the classroom level mediate the treatment effect?

The Intervention

The ML-PBL intervention took the learning guidelines of the *Framework* and the performance standards of the NGSS and designed and developed a set of science curricular units and materials, professional learning experiences, and assessments—all of which incorporate multiple literacies to advance student science academic learning and social and emotional development. The design of the curricular resources centers on increasing deep usable science knowledge, building literacy and mathematical skills, and ensuring access and ownership of science learning for all students. Drawing from students' life experiences, ML-PBL-designed learning activities that foster students as active agents in making sense of phenomena. They shift the responsibility of experiential learning primarily to the student. Because the intervention focuses on students and their interests, it is sensitive to the varied needs of their diverse characteristics, including culture, race/ethnicity, and gender. The design of ML-PBL is also structured to support social and emotional learning (SEL), which is defined in terms of self-reflection, capacity for collaboration, and taking ownership and responsibility for one's work (Durlak et al., 2015; Jagers et al., 2018).

Instructional Materials

The instructional materials for teachers and learners are based on theoretical principles and clearly defined teaching methods that are usable and detailed (see Cohen & Ball, 1999). In ML-PBL, teaching and learning are framed by a "driving question" (DQ) that focuses on challenging real-world problems or complex phenomena that create wonder and motivate students to learn. Investigating real-world questions and problems relevant to students' lives has long been embraced as a viable learning method that can be traced back to the progressive ideas of John Dewey (1938). A key aspect of ML-PBL focuses on learners asking their own questions related to the phenomena and the DQs, and finding solutions to those questions as the project progresses. To make sense of phenomena and solve problems, students ask questions and collaboratively plan and conduct investigations. Teachers scaffold lessons to support students in planning and conducting investigations, developing models that show how a phenomenon can occur, analyzing and interpreting data, and building claims with evidence and reasoning. In the end, students create artifacts that represent their emerging understandings and responses to the DQ.

The ML-PBL curricular materials proceeded through four stages: design and development, classroom enactment, testing, and evaluation. The design process began with the selection and unpacking of targeted PEs articulated by the NGSS for this grade level. This process helped the team develop a deep understanding of what students needed to meet for the PEs for third grade (see Harris et al., 2019; Krajcik et al., 2014; Krajcik & Czerniak, 2018). Next, the design team identified compelling and complex phenomena that aligned with the defined PEs, and then developed a DQ for each unit that drove instruction and motivated student learning goals. Following this, the design team turned to a group of practicing teachers who provided feedback on the relevance and grade-level appropriateness of the phenomena, DQs, and PEs.

The next stage of the process was the development of the lesson plans and activities. Each lesson is driven by lesson-level learning goals that include the three dimensions of scientific knowledge (DCIs, SEPs, and CCCs) and culminates with students designing and developing an artifact that responds to the unit DQ and which is responsive to community needs. Although the teacher materials are designed to provide instructional supports for teachers (Drake et al., 2006; Davis & Krajcik, 2005) with suggested questions to prompt students and lesson plans, the lesson sequence is not a prescribed script but rather a flexible roadmap (see Miller et al., 2018).

The completed third-grade curriculum consists of four units, each framed by a DQ and an anchoring phenomenon, culminating in students developing an artifact. The four units are labeled as follows: Squirrels (Adaptation), Toys (Force and Motion), Birds (Biodiversity) and Plants (Weather Climate). The units are then divided into Learning Sets, which are framed by questions that build toward the DQ—these questions occur at the unit, Learning Set, and lesson level. The DQ gradually and purposefully moves students towards using the three dimensions of scientific knowledge to explain and predict a phenomenon by helping students wonder, persist, and make sense of their world with DCIs, SEPs, and CCCs. For example, the Squirrels (or Adaptation) Unit DQ is “Why do I see so many squirrels but I can’t find any stegosauruses?” This unit focuses on learning how species can survive over hundreds of millions of years due to a wide diversity among the species and their adaptability to changes in the environment. The big ideas (DCIs) in this unit are that survival depends on change (adaptation) and that changes in the environment (whether natural or not) can cause changes in the populations of different organisms. The different lessons focus on a different aspect of an organism’s adaptations to environmental changes across time. In the Squirrels Unit, students are asked “What do squirrels need to survive?”, followed by questions about the squirrel’s physiology and environment. Exploring the past through fossils, students learn how scientists use that knowledge to see the change in organisms over time and why some species become extinct while some do not. Throughout the unit, learners build models to explain the various phenomena and questions posed to them. Table 1 below outlines the units, their targeted NGSS performance expectations, and the DQs.

Table 1. Units, Performance Expectations, and Driving Questions for Third Grade

	NGSS performance expectations	
Units	Grade 3	Driving questions
Squirrels/Adaptation	Life Science (3-LS): 4-1, 4-2, 4-3, 4-4, 3-2, 1-1	Why do I see so many squirrels, but I can’t find any stegosauruses?
Toys/Force and Motion	Physical Sciences (3-PS): 2-1, 2-2, 2-3, 2-4	How can we design fun moving toys that any kid can build?
Birds/Biodiversity	Life Science (3-LS): 2-1, 3-1,3-2,4-2 Engineering Design:3-5-ETS1-1	How can we help the birds near our school grow up and thrive?
Plants/Weather Climate	Life Science (3-LS): 1-1,3-1,3-2,4-3, 4-4 Earth and Space Science (3-ESS): 2-1, 2-2, 3-1 Engineering Design: 3-5-ETS1-1 3-5-ETS1-2	How can we design spaces in our community to grow plants for food?

Note. Explanations of NGSS performance expectations and an example of a lesson plan can be found in Appendix A.

Professional Learning

The second key component of the intervention is professional learning that supports teachers in using the ML-PBL materials to promote scientific knowledge, social and emotional development, and sensemaking. The PL uses PBL principles with specific collaborative experiences that underscore the importance of teachers creating classroom environments that stress equity, affirm cultural identity and responsible ownership, and build collaborative productive relationships. Based on best practices suggested by research (National Research Council, 2012), ML-PBL emphasizes the importance of: sustained long-term professional development (Oliveira, 2010); teachers' active participation in learning (Garet et al., 2001); and connections to classroom contexts, collaboration, and reflection (van den Bergh et al., 2014).

The goal of the PL experience was to introduce and help teachers understand the ML-PBL theoretical model, which constitutes three-dimensional scientific knowledge, and gain a familiarity with the NGSS. Several key aspects of the PL were: review the scope of the units; enact some of the experiences the students would engage in during the lessons; explore how to use the materials and their relevant experiential tasks; and learn about the construction of various student artifacts and assessments. The purpose of all the PL activities was to support teachers in conducting the intervention—not as a script that needed to be followed, but rather a roadmap to which adjustments could be made to ensure equitable learning opportunities that were culturally and historically responsive to the students, their families, and their communities.

At the beginning of the school year, all of the participating treatment teachers were invited to a PL session where they learned about the NGSS and PBL and experienced the materials they would be enacting related to the units; these included using the DQ and driving question board, which is where the DQs along with the students' generated questions are housed, making sense of phenomena, engaging in the scientific and engineering practices, and building artifacts. Following these experiences, teachers were asked to share their insights on how they imagined they would be using the ML-PBL materials. In the first of the PL meetings, teachers also learned about the activities the research team would conduct in greater detail, including collecting student work and observing classrooms. These in-person PL sessions occurred three times a year prior to the introduction of new units and were designed as a progression. Not all of the ML-PBL features were introduced during the first session—instead, features were introduced over time, with previous features being reinforced in subsequent sessions. If a teacher was unable to make a session, make-up days were provided.

In addition to the face-to-face meetings, team leaders (who were experienced elementary teachers) met with groups of teachers via video conferencing. These sessions occurred approximately every two weeks to solicit information from the teachers; they included discussions on what worked and what was challenging, as well as questions they and the students had while enacting the lessons. They also previewed the next Learning Set and discussed potential questions regarding its implementation. Some teachers attended these virtual meetings frequently; however, several others did not participate. All the teachers were encouraged to call our hotline with any questions they had. One team member was responsible for the hotline and kept records of these calls, which were then reviewed in weekly team meetings. Overall, among the formal scheduled PL sessions, each treatment teacher received approximately seven days of PL (counting in-person and formal virtual hours) throughout the school year (see Table 2).

Table 2. Scheduled Times for Professional Learning

Date	Hours	Type
Summer		
August 2018	3 days, 7 hours each	Face-to-face
School year		
November–April 2018	3 days, 7 hours each	Face-to-face
November–April 2018	3 days, 1 hour each	Virtual
Total	45	

Control teachers received up to six hours of PL on the NGSS, the state of Michigan's science standards and their relationship with the NGSS, and three-dimensional learning as described in the *Framework for K–12 Science Education* (National Research Council, 2012).

Assessments

Two forms of assessments were implemented in the ML-PBL intervention. At the end of each unit, a posttest assessment was given to all of the treatment students. These posttest assessments were designed upon a method modified for evidence-centered designs by Harris et al. (2015) and the work of Mislevy and Riconscente (2006). Rubrics and scoring protocols were developed, and raters were recruited and trained. The team calculated reliabilities multiple times over the course of each unit. Although the unit assessments were tied to the NGSS performance expectations that students were expected to meet, the post-unit assessments did not include the exact phenomena addressed in the units, but instead one similar in topic and structure. Additionally, an objective summative assessment designed by the Michigan Department of Education was given to both the treatment and control students at the end of the year to assess their science learning. The summative assessment is discussed in the methods section.

The Logic Model

Figure 1 depicts the shared relationships among the three components of the intervention, their provenance, and the outcomes they were expected to impact. Beginning on the left side are the PBL principles (driving question, exploring and explaining phenomena, artifact development, collaboration, equity) and the three dimensions of scientific knowledge described in the *Framework for K–12 Science Education* (National Research Council, 2012). These ideas were incorporated into the three components of ML-PBL materials, PL, and assessments, all of which are expected to directly impact the learning context at the classroom level. Here, the "learning context" should be considered as a moderator, in that the classroom is a stable invariant environmental space. Assuming that the teacher enacts the ML-PBL components with fidelity, the students are more likely to become "engaged." When engaged, students' interest in science increases, they have the requisite competencies to carry out ML-PBL activities, and the questions posed through the DQ spark their curiosity to "figure out" challenging questions to make sense of phenomena. The teachers' enactment of the intervention becomes a variant mediator: suggesting that those teachers who intensively adhere to ML-PBL principles and experiences are more likely to increase student engagement in situ (specifically in their science classes), which leads to a positive impact on students' science academic and SEL. The SEL factors reflect an increasing wonder of how the world works, assuming responsibility for one's and others' contributions to solutions, and the value of working in teams—all recognized as essential competencies for designing complex scientific investigations and models.

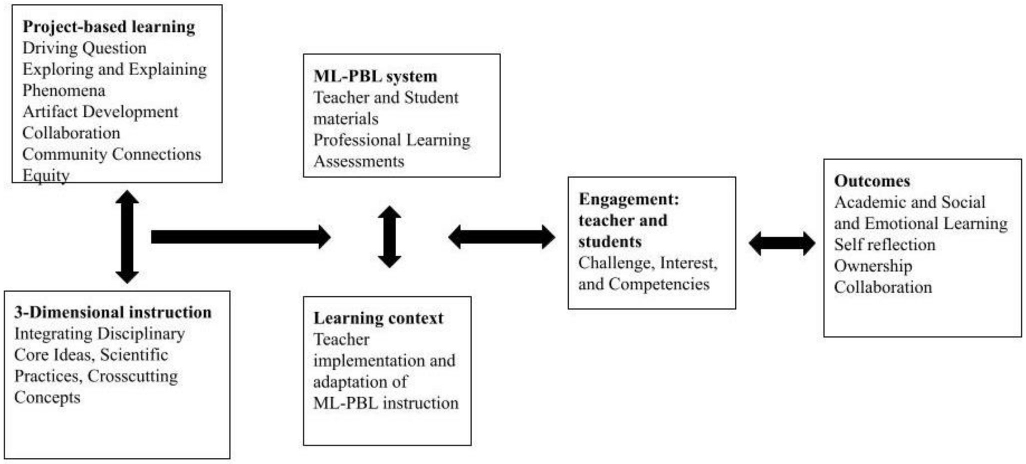

Project-based learning
Driving Question
Exploring and Explaining Phenomena
Artifact Development
Collaboration
Community Connections
Equity
ML-PBL system
Teacher and Student materials
Professional Learning
Assessments
Engagement: teacher and students
Challenge, Interest, and Competencies
Outcomes
Academic and Social and Emotional Learning
Self reflection
Ownership
Collaboration
3-Dimensional instruction
Integrating Disciplinary Core Ideas, Scientific Practices, Crosscutting Concepts
Learning context
Teacher implementation and adaptation of ML-PBL instruction

Figure 1. Logic Model of the ML-PBL System

Methods

The ML-PBL intervention began with teaching experiments, followed first by a pilot and then a field test one year later. Based on the results of the teaching experiments, pilot test, and field test, the intervention was slightly modified and refined for the larger efficacy study, which began in fall 2018 and continued through spring 2019. To test the efficacy of the intervention, the team designed a cluster randomized trial to determine if the intervention improved students' science learning and enhanced their SEL. The methods used to test the intervention included the sample, the instruments and measures, and the plan for analysis.

Pre-hoc Power Analysis

When designing an intervention, it is necessary to estimate how many schools, teachers, and students one needs to detect the true magnitude of the relationship between the outcome and the treatment assignment. To achieve this, a pre-hoc power analysis was conducted to ensure that the sample would be large enough to detect the treatment effect, but not so large that resources would be allocated inefficiently. Typically, one aspires to achieve a power of 0.80.

The following estimates for the study were established using *Optimal Design Plus* (Spybrook et al, 2011) an open-source software package for power calculations in cluster randomized trials. In 2016, Spybrook et al. (2016) used science assessment data from Michigan, Texas, and Wisconsin over multiple years to report design parameters for cluster randomized trials. Based on the estimates from their work, we used an intraclass correlation (ICC) of 0.26 (i.e., 26% of the variance was between schools, and the remaining 74% was between students within schools) and a pre-test R^2 of 0.68 (which explains 68% of the variability of the outcome of interest around the mean) to find a minimum detectable effect size of 0.26 with 0.80 power (see Figure 2). We expected a higher effect size than 0.26. For example, for an effect size of 0.40, the power is 0.99 (an effect size of 0.40 could be achieved with 0.80 power in 22 total schools with 50 students per school—11 treatment and 11 control). However, since a power analysis is only an estimate, it is prudent to be conservative; we therefore estimated the power effect based on 48 schools with an average of 50 students per school. Using all of the same assumptions (ICC = .26; R^2 = 0.68; 50 students per school), if we were only able to include 46 schools (instead of 48) in the analytic sample, the only thing that would change is that the minimum detectable effect size would increase from 0.26 to 0.27.

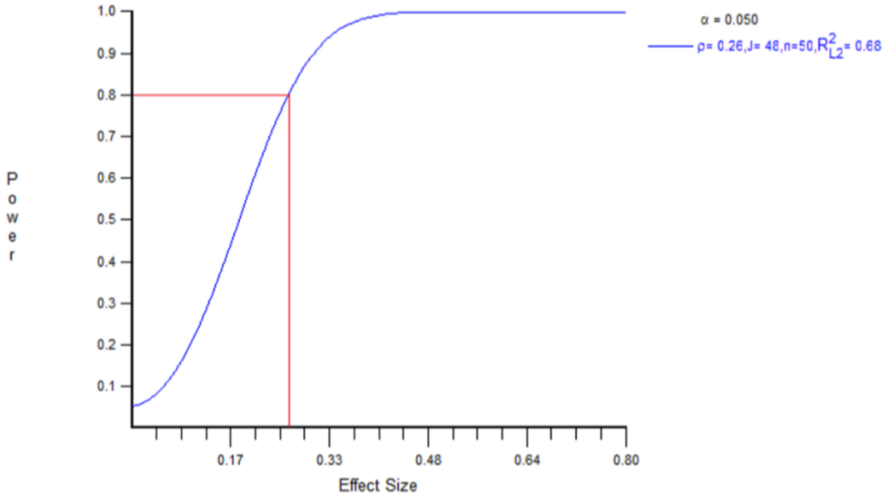


Figure 2. Minimum Detectable Effect Size for 0.80 Power

School Identification and Randomization

One of ML-PBL's major considerations was to provide an intervention that would improve the science academic and SEL of low-income and minority students. To achieve this goal, recruiting a sample of schools from the Detroit Public School Community District (DPSCD), with its large population of Black students, was a high priority. In fall 2017, cooperation with DPSCD was secured, which resulted in access to a concentrated urban population with a large number of low-income and minority students. The inclusion of DPSCD allowed us to examine the impact of the intervention in other regions throughout the state, some of which had student populations that were predominately White and middle-income. The student sample outside DPSCD was organized into three regions: one on the western side, another in the eastern central section, and one in the northern part of the state. All school districts and schools located within each region share the same region code, allowing for the identification of potential sample sites. In the western and eastern central sides of the state, contacts were made with the district science directors about the possibility of conducting the intervention in their districts. Because the northern part of Michigan is not as populated as the rest of the state, and several intermediate school districts (ISD) were contacted, discussions also ensued with the district superintendents about the study and their willingness to participate. Upon their approval, a memorandum of understanding (MOU) was drafted, which they signed to indicate their agreement to cooperate (a sample of this letter is in Appendix B).

Identification of schools eligible to participate in the ML-PBL intervention occurred by taking advantage of the Michigan State Longitudinal Data System, which maintains and continually updates information on all Michigan school students from pre-K through college completion. School eligibility for the ML-PBL program was determined via several factors. These included whether the school was a public non-specialized school with a third-grade enrollment of over 25 students and had a certain proportion of students representing racial and ethnic minorities and/or receiving free and reduced lunch. Using the state school data, a full list of public elementary schools was generated within Genesee, Kent, and DPSCD, as well as schools in several districts in northern Michigan. Once this list of schools in each region was obtained, each school's eligibility to participate in the intervention was determined.

After receiving the school lists in DPSCD and Michigan, the randomization process began. Three of the Michigan regions were combined into one region for randomization—DPSCD was randomized independently because of its homogenous Black and low-income school population characteristics. Using the Stata software package, the sample was re-randomized up to 100 times until the balance p-value exceeded 0.2. The program was also used to check for balance on a specified list of covariates. For DPSCD and all the other schools in the three other regions, adaptive covariate randomization was used to assign the school data set into two groups, checking for balance on the proportion of student racial composition, proportion free and reduced lunch, grade-3 enrollment, and grade-3 Michigan Student Test of Educational Progress (M-STEP) math and reading scores. All randomization was completed by July 2018. Once the class lists were received from the schools, the balance between treatment and control schools was calculated. See Table 3 for the post-randomization numbers for schools, teachers, and science classes in the four regions.

Table 3. Post-Randomization Class List Numbers Before the Start of the Intervention

	Treatment			Control		
	School (N)	Teacher (N)	Science classes (N)	School (N)	Teacher (N)	Science classes (N)
D – DPSCD	9	13	22	11	21	30
G – Genesee	2	2	6	5	11	11
K – Kent	5	15	15	4	10	10
O – Northern MI	7	12	16	4	13	13
Total	23	42	59	24	55	64

One school was dropped because it did not have a third grade due to a recent school consolidation process in DPSCD. Three other schools failed to provide class lists and were thus considered attriting schools. One additional school that provided class lists did not provide the benchmark data nor the summative assessment and was thus also considered an attriting school. (See attrition calculations in Table 8.)

After the treatment and control schools were selected, the principal was contacted, and permission to contact the third-grade teacher was requested. Principals willing to participate signed an MOU; once this was received, the principal was asked for class rosters for each of the third-grade teachers in their school. It is important to note that the MOU did not specify whether the school would be a treatment or control school and that this would be determined after randomization. Schools identified as control schools would receive the treatment the next year (2019–2020). In a few schools, there were multiple third-grade classrooms: in these instances, the principals requested that all of the third-grade classrooms receive the treatment. When this situation occurred, one of the classrooms was identified as the focal classroom. This designation was made in case of future bias problems with respect to attrition and balance issues that could occur with the analytic sample. (In the main effect analyses, we show the treatment effect for just the focal classrooms as well as for the entire sample, see Table 19.)

Table 4. Post-Randomization Numbers After Completion of the Intervention

	Treatment			Control		
	School (N)	Teacher (N)	Science classes (N)	School (N)	Teacher (N)	Science classes (N)
D – DPSCD	9	12	19	10	17	23
G – Genesee	2	2	6	5	10	10

K – Kent	5	15	15	4	10	10
O – MI Other	7	12	16	4	13	13
Total	23	41	56	23	50	56

Note that in Table 4, the number of teachers in the final analytic sample for DPSCD was fewer than in the post-randomization class lists (see Table 3 above). It is important to remember that only in DPSCD were there multiple teachers in the same school teaching multiple sections of third-grade science classes. Checking the balance in the sample including these DPSCD schools, there was not a significant difference between the number of multiple teachers and the number of students within these multiple sections in the treatment and control conditions. In other regions, schools had teachers who had multiple sections but not multiple science teachers with different science classes. The overall demographic characteristics of the schools in the analytic sample are described in Table 5.

Table 5. School-Level Demographic Characteristics

School-level characteristics							
	Free and reduced lunch	Free lunch	American Indian	Asian	Hispanic	Black	White
Treatment	0.622	0.562	0.002	0.032	0.085	0.408	0.446
Control	0.616	0.567	0.004	0.021	0.119	0.421	0.406
Overall	0.619	0.564	0.003	0.027	0.102	0.414	0.426

Region Balance and Generalizability

The balance between the treatment and control samples regarding student characteristics for each of regions can be found in Appendix C. Given the magnitude of the causal treatment effect found for the ML-PBL intervention and the high internal validity within the study, the question became whether this intervention could possibly be generalizable to the rest of the schools in the regions, the rest of the state of Michigan, and even the rest of the country. Although this was not an expectation in the design of this cluster randomized trial, recent methods allowed us to estimate whether our population is generalizable to the larger populations. Relying on the method provided by Tipton et al. (2014), a series of analyses were conducted to estimate the generalizability of the sample to all of the schools in their respective regions, then to the state of Michigan, and finally to third graders in the US population across all 50 states. Table 6 reports the estimates found for the generalizability indexes for the regions, the state of Michigan, and the United States.

As shown in Table 6, the first column shows the school count of the inference population and the second column represents the generalizability index in each respective region, the state, and the nation. The generalizability index represents the degree of similarity between the intervention schools and the inferential population. The third column identifies the magnitude of similarity index from high to low. The index coefficients for the regions and the state of Michigan were obtained using Michigan State Longitudinal Data Files, and the US coefficient is based on Common Core Data (CCD).

For instance, we identified 50 separate inference populations using CCD. For each of the 50 inference populations, we compared the 46 schools in the ML-PBL intervention to the inference population by an estimation of a sampling propensity score. A set of seven school covariates (i.e., student enrollment, proportion of free and reduced lunch, urbanicity, proportion of White, Black, Hispanic, and Asian) were included in the propensity score matching. The generalizability index typically has a value of 0.90 or higher in a random sample. Thus, we can say that when the value is > 0.90, the sample is as similar to the population as a random sample of the same size on the selected covariates. In these situations, if the set of covariates selected includes all those that explain variation in treatment impacts, then the average treatment effect (ATE) estimated in the sample provides an unbiased estimate of the ATE in the population. For example, the degree of similarity is very high in Alabama and Louisiana, 15 other states are high, 27 states are medium, and 4 are low. Importantly, the ML-PBL sample is similar to 66% of states (33 states) if we used 0.70 as the criteria of the generalizability index to be considered approximately similar to a random sample with covariates.

Table 6. Generalizability to Region, State, and Nation

	Inference population school count	Generalizability index	Decision
Region			
Kent	112	0.81	High
Other Michigan Areas	45	0.80	High
DPSCD	56	0.86	High
Genesee	89	0.70	Medium
State of Michigan			
Michigan	1674	0.85	High
Nationwide	46,664	0.85	High
50 States			
Alabama	704	0.94	High
Louisiana	664	0.90	High
Illinois	2,136	0.89	High
Virginia	1,105	0.88	High
Maryland	831	0.88	High
North Carolina	1,327	0.88	High
Pennsylvania	1,487	0.86	High
Missouri	1,092	0.86	High
Indiana	991	0.85	High
Rhode Island	158	0.85	High
South Carolina	616	0.84	High
Florida	1,841	0.84	High
Georgia	1,189	0.84	High
New York	2,276	0.84	High
Wyoming	186	0.82	High
Kentucky	702	0.82	High
Wisconsin	1,009	0.80	High
Kansas	703	0.79	Medium
New Hampshire	260	0.79	Medium
Minnesota	773	0.78	Medium
Vermont	214	0.78	Medium
Iowa	620	0.77	Medium
Arkansas	461	0.77	Medium
Mississippi	422	0.77	Medium
Maine	313	0.76	Medium
Nebraska	513	0.76	Medium
Ohio	1,543	0.75	Medium
Montana	367	0.75	Medium
Nevada	361	0.73	Medium
Oregon	646	0.73	Medium
New Mexico	398	0.73	Medium
Idaho	350	0.72	Medium
West Virginia	405	0.72	Medium
Arizona	896	0.68	Medium
Colorado	923	0.68	Medium
Washington	1,149	0.68	Medium

South Dakota	304	0.67	Medium
North Dakota	257	0.66	Medium
Connecticut	560	0.65	Medium
District of Columbia	76	0.64	Medium
Utah	543	0.63	Medium
Texas	4,156	0.63	Medium
New Jersey	1,296	0.60	Medium
Delaware	96	0.58	Medium
Hawaii	176	0.53	Medium
California	5,203	0.51	Medium
Oklahoma	852	0.44	Low
Massachusetts	930	0.35	Low
Tennessee	953	0.34	Low
Alaska	332	0	Low

Balance of the Analytic Sample

To ensure a valid treatment effect, it is important to examine the balance of the final analytic sample. Due to the clustering of students within schools, a Weighted Least Squares (WLS) was conducted to check the balance including school-level variables. For the student-level variables, an HLM was used to check the balance, given that the final analytics were also conducted with an HLM. Table 7 shows these results as well as the appropriately calculated effect size given the WLS and HLM that was conducted. A WLS with an $\eta^2$was calculated at the school level and a Cohen's f^2 was calculated at the student level.

Table 7. Balance for the Analytic Sample

Panel A. School level—WLS on grade-3 enrollment

	Difference (Treatment—Control)	Effect Size (η2)
Mean reading benchmark	0.35 (4.64)	0.00001324
Proportion free and reduced lunch	0.019 (0.072)	0.00155
Proportion free lunch	0.004 (0.08)	0.00005
Proportion American Indian	-0.003 (0.002)	0.0483
Proportion Asian	0.009 (0.023)	0.003881
Proportion Hispanic	-0.023 (0.073)	0.00242
Proportion Black	-0.064 (0.132)	0.0053
Proportion White	0.083 (0.12)	0.011
Region 1	-0.06 (0.146)	0.004
Region 2	-0.049 (0.18)	0.002
Region 3	0.052 (0.123)	0.004
Region 4	0.057 (0.141)	0.004
DIBELS	0.056 (0.098)	0.0074
F&P	0.081 (0.061)	0.0400
NWEA	-0.237 (0.159)	0.0480
Star	0.178 (0.15)	0.0308
i-Ready	-0.078 (0.15)	0.0063
N	46	

Panel B. Student level		Cohen's $\sqrt{f^2}$
Reading benchmark	1.3 (4.799)	-0.00
Missing reading benchmark	0.063 (0.045)	0.02
Gender	0.163 (0.15)	0.01
Missing gender	-0.111 (0.100)	0.02
Form A	-0.011 (0.007)	0.00
Form B	0.003 (0.006)	0.01
Form C	0.008 (0.007)	0.00
N	2,371	

Note. Benchmark is reported in a percentile ranking. Other covariates are proportions. Standard errors are in parentheses.
*p < 0.05 ** < 0.01 *** < 0.001

Using the *What Works Clearinghouse* ([WWC] 2020) standard of 0.05–0.25 for creating baseline equivalence, all of the estimated effect sizes were compared against these ranges. If the effect sizes are below 0.05 in their absolute value, then equivalence does not need to be established. At the school level, none of the variables exceed an effect size ($\sqrt{\eta^2}$) of 0.05. This is also the case at the student level, where none of the variables exceed an effect size (Cohen's $\sqrt{f^2}$) of 0.05. Given that the NWEA benchmark test and proportion of American Indian students were just below the 0.05 level, they were entered as additional covariates in estimating the main effect of the intervention.

Attrition, Late Joiners, and Baseline Equivalence

To ensure that the treatment effect is not the result of bias from differential sample attrition, school- and student-level attritions were calculated. To examine bias with respect to school attrition, the four attriting schools were removed. These results are reported in Table 8, Panel A. Student attrition was calculated from classroom rosters that were received before the start of the intervention. These results are reported in Table 8, Panel B. Based on the WWC's standard, for an overall attrition of 19% (18.89%), the optimistic boundary for differential attrition is 10.2%. As shown in Table 8, the student differential attrition of 9.14% falls in the WWC's "optimistic" category for primary science interventions, which is considered low attrition and meets the "without reservation" category of potential bias.

Table 8. Attrition Calculations

Panel A. School level				
	Overall	Treatment	Control	Differential
Initial schools	50	25	25	
Final schools	46	23	23	
Attrition	8.00%	8.00%	8.00%	0.00%
Panel B. Student level				
Initial students	2,923	1,518	1,405	
Final students	2,371	1,165	1,206	

Attrition	18.89%	23.30%	14.16%	9.14%

Late joiners also pose an issue of bias if they are aware of the possible benefits of the intervention. However, with respect to late joiners, due to the nature of the ML-PBL intervention being clustered at the school level in non-specialized public schools as well as being a low-profile intervention, the bias due to late joiners is minimal. Table 9 reports the number of students who were on and not on the class rosters. Benchmark and summative assessment information indicates what information was available for those on and off the rosters.

Table 9. Number of Benchmarks and Summative Assessments On and Off Our Roster Lists

	No benchmark	Has benchmark	No summative assessment	Has summative assessment
Roster	388	2,585	552	2,371*
Not on roster	22	72	86	8

*In the analytic sample

There were eight students who were not on the roster that were removed from the analytic sample. Of the eight students, only two had benchmark scores. Of the 22 students with no benchmark scores and not on the rosters, six had a summative assessment. The other 16 students had SEL data that were collected, but no benchmarks or summative assessments.

Based on these numbers and the analytic sample, there were only eight late joiners who were removed. The intent to treat (ITT) model was estimated with and without these eight late joiners. The treatment effect with the 8 late joiners was 0.261 (standard error [SE] = 0.114, significant at the 0.05 level), which is consistent with the estimated treatment effect without these 8 late joiners.

Instruments and Measures

To reliably estimate the impact of the intervention, one of the most important measures is a baseline for student academic achievement. The necessity for this is supported by research in which the power of using academic achievement has been shown to greatly reduce bias. Therefore, requests were made to the treatment and control schools, teachers, and sometimes the districts for students' third-grade fall and winter reading benchmark scores. One school, however, did not administer reading benchmarks, but had math benchmarks; therefore, requests were made for these math benchmarks, which were retrieved. In Michigan, school districts are not required to use the same elementary school benchmark tests. Among the sampled schools, five different benchmark tests were administered: the Northwest Evaluation Association Measure of Academic Progress (NWEA MAP), Star, i-Ready, DIBELS, and Fountas & Pinnell (F&P). Most schools used i-Ready or NWEA MAP and only one school used F&P. Since these tests have different scoring scales, they were normalized using percentile rankings. Using the most recent national norming guides for each test, students' raw scores were converted into percentile rankings and compared across tests (see Table 10).

Table 10. Benchmark Tests by School and Students

	Treatment		Control	
	School (N)	Student (N)	School (N)	Student (N)
DIBELS	3	106	1	46
F&P	1	151	0	0
NWEA	5	324	10	618
Star	4	283	1	72
i-Ready	10	578	11	635

As shown in Table 10, twice as many control students took the NWEA than the treatment group, whereas there are more treatment students than control students who took the Star benchmark. These test differences could produce a bias, especially if one of these benchmarks were either more or less aligned with content in the summative test. To check for the effect that an imbalance of benchmarks might have on the treatment effect, an interaction between the benchmark test type and the treatment was conducted (see Table 11).

Table 11. Interaction of Benchmark Tests with Treatment Effect

Interactions	DIBELS	F&P	NWEA	Star	i-Ready
Treatment	0.282**	NA	0.256	0.205*	0.297*
	(0.105)	NA	(0.136)	(0.102)	(0.118)
Predictor of interest	0.507***	NA	0.232	-0.13	-0.198
	(0.073)	NA	(0.192)	(0.152)	(0.189)
Interaction	-0.377	NA	-0.002	0.545***	-0.082
	(0.2)	NA	(0.184)	(0.109)	(0.199)

Note. Only one school used F&P—it was a treatment school.

Results in Table 11 show a significant relationship between three of the benchmark tests (i.e., DIBELS, Star, and i-Ready) and the treatment effect; one benchmark test (DIBELS) with the predicted treatment effect (the students' score on the type of benchmark taken was related to their summative test score); and an interaction between the treatment and predictor (Star benchmark, though the sample number of students in this category was very small—see Table 10, above). Given that the type of benchmark tests students took could be a proxy for unobservable school-level characteristics affecting the treatment, the type of benchmark taken was entered as a covariate.

Summative Evaluation

To estimate a true treatment effect, the measurement of the academic outcome variable should be independent and objective. To ensure that the intervention was not over-aligned with a summative test which could bias the effect, the team contacted several states (Washington and Nebraska in addition to Michigan) as possibilities, but in all three cases the elementary science tests were still in development. In Michigan, however, the Michigan Department of Education had developed and piloted a summative science test, designed to align with the NGSS and three-dimensional learning. The release of various test items to the ML-PBL team was negotiated with the Michigan Department of Education and Michigan State University legal departments to ensure the secure handling and analysis of the test items. The various test items were released to the ML-PBL team in spring 2019 after formative post-unit assessments had been given to all of the treatment teachers. Designed to be used on a computer, this testing option was not possible, as many of the sampled classrooms did not have computers for all the children. Therefore, the test was converted to paper and pencil.

Specifically created to assess students' learning according to the NGSS for K–5, the Michigan test had many more items than were relevant to third-grade students. Selecting only those items aligned with the third-grade NGSS performance expectations, the test was compiled into three forms, which varied in difficulty levels marked as "easy," "medium," and "hard." To determine if the third-grade summative test met psychometric standards, a series of item response analyses (IRT) were conducted to assure that each form was similar in its ability to differentiate the students' abilities and detect the difficulty of the items. Table 12 shows the mean and standard deviation for the scores of the different forms across treatment and control classrooms. Remember, there are more classrooms than treatment and control schools because of multiple classrooms in some schools. As shown in Table 12, there was no difference in proportion of students taking the different forms between the treatment and the control students.

Table 12. Summative Test Forms by Number of Students by Classrooms

	Treatment classroom		Control classroom		
	(n = 53)		(n = 56)		
	Mean	SD	Mean	SD	T-test
Classroom average student took form B	7.19	1.99	7.07	1.59	-0.349
Classroom average student took form A	7.30	1.56	7.43	1.48	0.447
Classroom average student took form C	7.09	2.02	6.80	1.67	-0.819
Proportion of student took form A	0.34	0.04	0.35	0.03	1.482
Proportion of student took form B	0.33	0.04	0.33	0.03	0.000
Proportion of student took form C	0.33	0.04	0.32	0.03	-1.482

To ensure that all topics had varying level of difficulties, a one-parameter logistic (1PL) IRT was conducted to determine that forms A, B, and C had varying level of difficulty, and that these varying levels of difficulty matched analyses conducted by the Michigan Department of Education. In order to conduct this analysis, the items were dichotomized where an "all or nothing" scenario was created, whereby if any part of the question was wrong, it was marked as a zero, and to receive a score of one, all parts had to be correct. Table 13 presents the results of the IRT analysis, which shows that each form discriminates between each individual student on equal and varying levels of item difficulty.

Table 13. IRT for Each Form

	Coefficient	Standard Error	Z	Significant
Form A				
Discriminant	0.706	0.048	14.72	0.000
Difficulty				
Question 7	0.145	0.125	1.17	0.244
Question 11	0.363	0.127	2.87	0.004
Question 9	0.363	0.127	2.87	0.004
Question 10	0.810	0.136	5.96	0.000
Question 1	0.904	0.139	6.52	0.000
Question 6	0.936	0.140	6.71	0.000
Question 3	1.322	0.154	8.60	0.000
Question 8	1.787	0.176	10.18	0.000
Question 2	2.944	0.247	11.89	0.000
Question 4	4.383	0.369	11.88	0.000
Form B				
Discriminant	1.412	0.097	14.49	0.000
difficulty				
Question 3	-0.878	0.093	-9.47	0.000
Question 4	-0.084	0.079	-1.07	0.000
Question 1	-0.027	0.078	-0.35	0.729
Question 2	0.388	0.081	4.77	0.000
Form C				
Discriminant	1.12	0.073	15.26	0.000
difficulty				
Question 1	0.445	0.094	4.72	0.000
Question 4	0.696	0.099	7.00	0.000
Question 2	0.761	0.101	7.52	0.000
Question 6	1.498	0.129	11.63	0.000
Question 5	1.598	0.134	11.96	0.000
Question 3	1.814	0.145	12.51	0.000
Question 7	3.419	0.271	12.60	0.000

Note. The level of difficulty is in ascending order.

Each form had separate items not repeated in other forms. In only two examples were the items too difficult (shaded in grey): these were Question 4 in Form A and Question 7 in Form C. Given the varying levels of difficulty among the forms administered to the students, there is a possible concern of bias related to the treatment effect and the summative test form. To determine if there was bias in the treatment effect results in the HLM main analysis, the treatment was interacted with each form of the test. As shown in Table 14, there is no interaction between any of the three forms and the treatment, indicating that there was not a difference in the treatment effect based on students taking different forms of the test.

Table 14. Interaction of Forms A, B, and C with Treatment Effect

Interactions	Form A	Form B	Form C
Treatment	0.253*	0.281*	0.252*
	(0.12)	(0.121)	(0.112)
Predictor of interest	-0.011	0.03	-0.017
	(0.052)	(0.058)	(0.056)
Interaction	0.027	-0.057	0.03
	(0.08)	(0.096)	(0.075)

Observation Protocol

To measure the teachers' fidelity of implementation, an observation protocol was used. The protocol is a different type of fidelity of implementation instrument than usually found in interventions. It was not a checklist—rather, the observation form highlighted the principles of PBL, and specific directions were given to the observer to "look for" strategies used not only by the teacher but also the students. Raters had to score how well the teachers were: engaging students in using the DQ, figuring out phenomena, and collaboratively building artifacts; providing opportunities for all students to participate in science; maintaining good classroom management; and implementing all aspects of the PBL principles.

Observers were recruited via recommendations from district science directors, teacher education professors, and the Michigan Science Teachers Association. Most of the observers were retired teachers and familiar with the NGSS and PBL. In-person education of the observers began in September and was conducted by the research team. Watching videos of third-grade teachers conducting ML-PBL lessons, the observers rated the teachers based on a protocol developed by the research team.

Table 15. Observation Training Schedule

Date	Training Hours	Description
Fall		
September 10–28, 2018	8.5	Initial education
September 21, 2018		Observations began
October 3, 2018	3	Check-in (first IRR)
Winter		
January 3, 2019	3	Check-in
January 22, 2019	1	Check-in
January 23, 2019	2	Reliability check (second IRR)
Spring		
March 6, 2019	2	Check-in
March 8, 2019	1	Check-in (third IRR and ICC)
April 19, 2019	1	Check-in
April 30, 2019	1	Review
Total	22.5	

After watching the videos and discussing them, raters independently scored the teachers. Interrater reliabilities (IRRs) were then calculated with two repeated meetings to obtain an IRR of 0.78 (see Table 16 below). When two reviewers observing the same teacher reached a reliability of 0.70, they were ready to rate independently. An ICC analysis was conducted to determine which of our raters were having difficulty distinguishing between the quality of the different teachers (not the students) in the intervention. Typically, some raters tended to score either consistently too low or too high. Special sessions were conducted to help observers achieve a more reliable IRR and ICC. Over the course of the year, we conducted three different rounds of IRRs to maintain consistency across raters.

Table 16. Observer IRRs During the Training Period

Observation Data	IRR (interrater correlation)	Percent of agreement	N of raters in the IRR check
First check 4 raters	0.45	0.48	4
Second check 7 raters (excluding 2 raters who had lower agreement)	0.56 (.70)	0.61 (0.68)	7 (5)
Third check	0.78	0.72	2

Note. In the second and third checks, we also conducted an ICC, paying close attention to differences between the single and average consistent agreement.

Initially, the intent was to obtain five observations of each treatment teacher and visit each control teacher twice. However, there were major problems scheduling observers into the classrooms because of weather conditions, state spring testing, the high costs of recruitment and education, and major distances between school sites. In the end, each treatment teacher was observed twice (though a few were observed more) and only slightly more than half of the control teachers (n = 18) were observed once.

Teacher Surveys

At the beginning of the study, the control teachers received a brief background survey regarding their familiarity with the NGSS and Michigan's adaptation of the standards, PBL, and the types of professional development they had recently received. The intent of the initial background survey was to ensure that both groups of teachers were only somewhat knowledgeable about the theoretical principles of the intervention and had limited PL sessions on PBL.

At the end of year, an exit survey was given to both the treatment and control teachers. The treatment teacher exit survey included items regarding their experiences: specifically with regards to using three-dimensional learning, including scientific practices; the challenges of teaching ML-PBL; their integration of literacy and mathematics in science; their efforts to foster student collaboration, engagement, and SEL; their cultural awareness and equity; and PL for supporting their instruction. It also included questions concerning the coverage of NGSS performance expectations and the quality of resources in the classroom that may have interfered with the execution of some of the intervention lessons. The control teacher exit survey was a modified survey of the one administered to the treatment group, and focused on the teachers' science content coverage, time spent on science instruction, science practices, exposure to the NGSS and PBL professional learning during the year, and the quality of science resources.

SEL Instrument

To create an instrument that would measure third graders' SEL for their science classes, the team consulted with multiple sources, including psychological research studies on SEL, developmentally appropriate questions for third graders, and items from other national assessments (e.g., the Early Childhood Longitudinal Study [ECLS]) used to measure differences in SEL with both oral and written formats. The intent was not to measure personality traits but rather SEL constructs that could be observed when students are involved in science lessons. Accounting for differences in literacy skills among students, a drawn thumbs-up, thumbs-down, and closed fist were used to measure agreement. Students circled their feelings on a paper/pencil form administered to both treatment and control groups in fall

2018 and spring 2019. The SEL instrument was designed and field-tested in the year prior to the efficacy study. Based on results from a confirmatory factor analysis (CFA), the instrument was slightly revised to add 7 new items to the 11 original items in the field test. Results from the 18-item instrument show a more robust measurement of the original three constructs: self-reflection, ownership, and collaboration. Table 17 shows the SEL items and their CFA.

Table 17. Item Factor Loading for the 18 SEL Items

	Factor 1			Factor 2			Factor 3		
	Estimate	SE		Estimate	SE		Estimate	SE	
Reflection									
In science, I ask and explore questions that I don't know the answer to.	0.289	(.050)		-0.017	(.026)	a	0.103	(.043)	a
In science, I figure out how things work.	**0.456**	**(.041)**		-0.011	(.013)	a	-0.080	(.043)	a
Even when I don't know the answer, I like to keep working in science.	**0.344**	**(.040)**		0.107	(.031)		0.001	(.023)	a
In science, talking about my ideas helps me learn.	**0.248**	**(.071)**		0.038	(.030)	a	0.183	(.063)	a
Doing investigations helps me figure out how things work.	**0.432**	**(.042)**		-0.005	(.025)	a	-0.003	(.038)	a
In science, I enjoy asking questions and wondering about things.	0.207	(.062)		0.202	(.034)		0.130	(.047)	a
Ownership									
The ideas I am learning in science are important to me.	0.305	(.060)		**0.266**	**(.037)**		0.021	(.043)	a
In class, I enjoy doing science.	0.115	(.063)	a	**0.529**	**(.051)**		-0.013	(.011)	a
I wish we spent more time doing science.	-0.003	(.010)	a	**0.605**	**(.022)**		0.022	(.039)	a
We can use science ideas to help our community.	0.338	(.062)		0.105	(.039)	a	-0.076	(.052)	a
When doing science in school, I feel smart.	0.318	(.043)		**0.202**	**(.034)**		0.013	(.028)	a
Collaboration									
In science, I work with others to figure things out.	0.290	(.047)		-0.048	(.023)	a	**0.170**	**(.045)**	
In science, reading helps me learn.	0.195	(.058)	a	0.058	(.031)	a	0.165	(.049)	a
In science, I help my class figure out how things work.	0.128	(.058)	a	0.053	(.031)	a	**0.324**	**(.046)**	
In science, listening to others' ideas helps me learn.	0.103	(.079)	a	-0.015	(.014)	a	**0.385**	**(.071)**	
In science, I enjoy doing investigations with a partner.	0.292	(.050)		0.027	(.027)	a	0.231	(.050)	a
In science, I use ideas from my partner to solve problems.	0.004	(.034)	a	0.007	(.026)	a	**0.431**	**(.037)**	
In science, I feel good when others use my ideas.	-0.002	(.035)	a	0.073	(.032)	a	**0.289**	**(.036)**	

Note. a indicates that the factor loadings are not significant at the p = <. 001 level.

Table 18 reports the results of the exploratory structural equation model. Several analyses were conducted to confirm the factor analysis for the SEL constructs; descriptions of these methods can be found in Appendix D. One of the most critical analyses was the comparative fit index (CFI), which describes the model fit index.

Table 18. Latent Mean Differences Between Treatment and Control Groups Using Nested Chi-Square Difference Tests

Model	χ^2	DF	p-value	$\Delta\chi^2$	Δ df	p	Equivalent between treatment and control
Final two group CFA	-43736	118	0.000				
Reflection (F1)	-45488	117	0.000	1752	1	0.000	No
Ownership (F2)	-45605	117	0.000	1869	1	0.000	No
Collaboration (F3)	-45631	117	0.000	1895	1	0.000	No

After the invariance tests and identification of the CFA model, latent SEL construct means were analyzed to access the differences of three latent constructs between the treatment and control groups. The control group is the reference group when comparing the latent means with the treatment group. In so doing, the latent means of the control group are fixed at zero, the latent means of the treatment group represent the mean differences between the two groups. The results of using the nested Chi-Square difference test are reported in Table 18. Results show significant difference for three latent constructs between the treatment and control groups at the 0.001 level. The three latent means of the constructs of reflection, ownership, and collaboration of the treatment students are significantly higher than those of the control students.

Analysis

To assess the difference between the treatment and control conditions in science achievement and to account for the clustering that occurs as a result of the assignment of schools to treatment, a three-level hierarchical linear model (HLM) was used (Raudenbush & Bryk, 2002; Bloom, 2005; Raudenbush, 1997). Because students are not only nested within schools, but also nested within classrooms, it was hypothesized that there may be classroom-level effects.[3] Therefore, six different three-level models with students nested within classrooms within schools were estimated. Below are Models 1 and 2, which provides the treatment effect shown in Table 19.

Models for the Main Effect

Models 1 and 2:

$$Y_{ijk} = \gamma_{000} + \gamma_{001} Treatment_k + \epsilon_{ijk} + r_{0jk} + \mu_{00k}$$

In the first two models, the treatment effect is given in a single predictor multi-level model. Additionally, as noted above, the intervention included teachers who taught multiple sections; however, they were only given the support to teach one class of ML-PBL. Thus, each model—both with the full sample of students and then only with those students who were in the focal classrooms—was estimated. For the following two models, we included controls for school-level race and ethnicity proportions, as well as a regional fixed effect.

Models 3 and 4:

$$\begin{aligned} Y_{ijk} = \gamma_{000} + \gamma_{001} Treatment_k &+ \gamma_{002} Region2_k + \gamma_{003} Region3_k + \gamma_{004} Region4_k \\ &+ \gamma_{005} PropAm_k + \gamma_{006} PropAs_k + \gamma_{007} PropHs_k + \gamma_{008} PropBl_k + \gamma_{009} PropWh_k \\ &+ \epsilon_{ijk} + r_{0jk} + \mu_{00k} \end{aligned}$$

Models 5 and 6 add individual-level variables, such as the student's benchmark scores as well as the type of benchmark taken.

Models 5 and 6:

$$\begin{aligned} Y_{ijk} = \gamma_{000} + \gamma_{001} Treatment_k &+ \gamma_{002} Region2_k + \gamma_{003} Region3_k + \gamma_{004} Region4_k \\ &+ \gamma_{005} PropAm_k + \gamma_{006} PropAs_k + \gamma_{007} PropHs_k + \gamma_{008} PropBl_k + \gamma_{009} PropWh_k \\ &+ \gamma_{0010} FP_k + \gamma_{0011} NWEA_k + \gamma_{0012} Star_k + \gamma_{0013} Iready_k + \gamma_{100} benchmark_{ijk} \\ &+ \epsilon_{ijk} + r_{0jk} + \mu_{00k} \end{aligned}$$

Where $Y_{ijk}$is the standardized summative science assessment.
γ_{000} is the mean outcome.
γ_{001} is the difference between the treatment and control groups.
γ_{002}, γ_{003}, γ_{004}, γ_{005}, γ_{006}, γ_{007}, γ_{008}, γ_{009}, γ_{0010}, γ_{0011}, γ_{0012}, γ_{0013}, γ_{100} are the differences of other covariates.
Finally, $\mu_{00k}$is the school-level error term, r_{0jk} is the classroom-level error term, and $\epsilon_{ijk}$is the student-level error term.

Heterogeneity Models

To determine if student improvement differed across school-level characteristics, interactions between school-level race/ethnicity and SES with treatment at the school level were also conducted. A similar analysis with the students' gender and reading proficiency was conducted, with a cross-level interaction with the treatment at the school level and the gender and reading proficiency at the student level. For these interactions, the following models were used.

Proportion free and reduced lunch:

$$Y_{ijk} = \gamma_{000} + \gamma_{001}Treatment_k + \gamma_{002}Proportion\ free\ and\ reduced\ lunch_k + \gamma_{003}Treatment_k \times Proportion\ free\ and\ reduced\ lunch_k + \epsilon_{ijk} + r_{0jk} + \mu_{00k}$$

Proportion free lunch:

$$Y_{ijk} = \gamma_{000} + \gamma_{001}Treatment_k + \gamma_{002}Proportion\ free\ lunch_k + \gamma_{003}Treatment_k \times Proportion\ free\ lunch_k + \epsilon_{ijk} + r_{0jk} + \mu_{00k}$$

Proportion American Indian:

$$Y_{ijk} = \gamma_{000} + \gamma_{001}Treatment_k + \gamma_{002}PropAm_k + \gamma_{003}Treatment_k \times PropAm_k + \epsilon_{ijk} + r_{0jk} + \mu_{00k}$$

Proportion Asian:

$$Y_{ijk} = \gamma_{000} + \gamma_{001}Treatment_k + \gamma_{002}PropAs_k + \gamma_{003}Treatment_k \times PropAs_k + \epsilon_{ijk} + r_{0jk} + \mu_{00k}$$

Proportion Hispanic:

$$Y_{ijk} = \gamma_{000} + \gamma_{001}Treatment_k + \gamma_{002}PropHs_k + \gamma_{003}Treatment_k \times PropHs_k + \epsilon_{ijk} + r_{0jk} + \mu_{00k}$$

Proportion Black:

$$Y_{ijk} = \gamma_{000} + \gamma_{001}Treatment_k + \gamma_{002}PropBl_k + \gamma_{003}Treatment_k \times PropBl_k + \epsilon_{ijk} + r_{0jk} + \mu_{00k}$$

Proportion White:

$$Y_{ijk} = \gamma_{000} + \gamma_{001}Treatment_k + \gamma_{002}PropWh_k + \gamma_{003}Treatment_k \times PropWh_k + \epsilon_{ijk} + r_{0jk} + \mu_{00k}$$

Where $Y_{ijk}$is the standardized summative science assessment.
γ_{000} is the mean outcome.
γ_{001} is the difference between the treatment and control groups for the mean variable of interest.
γ_{002} is the slope of the variable of interest of the control group.
γ_{003} is the slope of the variable of interest of the treatment group.
Finally, $\mu_{00k}$is the school-level error term, r_{0jk} is the classroom-level error term, and $\epsilon_{ijk}$is the student-level error term.

Region 1:

$$Y_{ijk} = \gamma_{000} + \gamma_{001}Treatment_k + \gamma_{002}Region1_k + \gamma_{003}Treatment_k \times Region1_k + \epsilon_{ijk} + r_{0jk} + \mu_{00k}$$

Region 2:

$$Y_{ijk} = \gamma_{000} + \gamma_{001}Treatment_k + \gamma_{002}Region2_k + \gamma_{003}Treatment_k \times Region2_k + \epsilon_{ijk} + r_{0jk} + \mu_{00k}$$

Region 3:

$$Y_{ijk} = \gamma_{000} + \gamma_{001}Treatment_k + \gamma_{002}Region3_k + \gamma_{003}Treatment_k \times Region3_k + \epsilon_{ijk} + r_{0jk} + \mu_{00k}$$

Region 4:

$$Y_{ijk} = \gamma_{000} + \gamma_{001} Treatment_k + \gamma_{002} Region4_k + \gamma_{003} Treatment_k \times Region4_k + \epsilon_{ijk} + r_{0jk} + \mu_{00k}$$

Where $Y_{ijk}$is the standardized summative science assessment.

γ_{000} is the mean outcome.

γ_{001} is the difference between the treatment and control groups for the regions other than the one of interest.

γ_{002} is the difference of the region of interest of the control group.

γ_{003} is the difference of the region of interest of the treatment group.

Finally, $\mu_{00k}$is the school-level error term, r_{0jk} is the classroom-level error term, and $\epsilon_{ijk}$is the student-level error term.

Gender:

$$Y_{ijk} = \gamma_{000} + \gamma_{001} Treatment_k + \gamma_{010} Gender_{ijk} + \gamma_{011} Treatment_k \times Gender_{ijk} + \epsilon_{ijk} + r_{0jk} + \mu_{00k}$$

Where $Y_{ijk}$is the standardized summative science assessment.

γ_{000} is the mean outcome.

γ_{001} is the difference between the treatment and control groups for the males.

γ_{010} is the difference between males and females in the control group.

γ_{011} is the difference between males and females in the treatment group.

Finally, $\mu_{00k}$is the school-level error term, r_{0jk} is the classroom-level error term, and $\epsilon_{ijk}$is the student-level error term.

Reading benchmark:

$$Y_{ijk} = \gamma_{000} + \gamma_{001} Treatment_k + \gamma_{010} Reading\ Benchmark_{ijk} + \gamma_{011} Treatment_k \times Reading\ Benchmark_{ijk} + \epsilon_{ijk} + r_{0jk} + \mu_{00k}$$

Where $Y_{ijk}$is the standardized summative science assessment.

γ_{000} is the mean outcome.

γ_{001} is the difference between the treatment and control groups for the mean of the reading benchmark scores.

γ_{010} is the slope of the reading benchmark of the control group.

γ_{011} is the slope of the reading benchmark of the treatment group.

Finally, $\mu_{00k}$is the school-level error term, r_{0jk} is the classroom-level error term, and $\epsilon_{ijk}$is the student-level error term.

SEL models

To examine if the treatment supported students' SEL in comparison to the control students, difference tests were used to measure individual item responses between treatment and comparison groups. The process of identifying and testing the constructs is presented above under the instruments. After having employed the structural equation modeling and CFA, three-level HLMs were used to test a treatment effect on each of the three constructs. However, to conduct these HLMs, it was necessary to ensure the unbiasedness of the constructs. Therefore, for this analysis, the Bartlett factor score was used as the most robust and unbiased true factor score (DiStefano et al., 2009). The Bartlett descriptive factor scores can be found above. For these constructs, the following models were estimated.

Additionally, there was suspicion of differences in SEL outcomes by gender as well as by school context. In particular, given that students from low-income and minority schools may have never experienced the kinds of learning experiences, materials, and classroom culture provided by ML-PBL, greater differences in SEL may be observed. To test this assumption, a cross-level interaction of treatment and gender was analyzed, as well as the interaction of treatment with school proportion of free and reduced lunch and race/ethnicity.

Base model:

$$Y_{ijk} = \gamma_{000} + \gamma_{001}Treatment_k + \epsilon_{ijk} + r_{0jk} + \mu_{00k}$$

Where $Y_{ijk}$is the factor score on reflection, ownership, or collaboration.

γ_{000} is the mean outcome.

γ_{001} is the difference between the treatment and control groups.

Finally, $\mu_{00k}$is the school-level error term, r_{0jk} is the classroom-level error term, and $\epsilon_{ijk}$is the student-level error term.

Gender interaction:

$$Y_{ijk} = \gamma_{000} + \gamma_{001}Treatment_k + \gamma_{010}Gender_{ijk} + \gamma_{011}Treatment_k \times Gender_{ijk} + \epsilon_{ijk} + r_{0jk} + \mu_{00k}$$

Where $Y_{ijk}$is the Factor score on reflection, ownership, or collaboration.

γ_{000} is the mean outcome.

γ_{001} is the difference between the treatment and control groups for the males.

γ_{010} is the difference between males and females in the control group.

γ_{011} is the difference between males and females in the treatment group.

Finally, $\mu_{00k}$is the school-level error term, r_{0jk} is the classroom-level error term, and $\epsilon_{ijk}$is the student-level error term.

Region interaction:

Region 1:

$$Y_{ijk} = \gamma_{000} + \gamma_{001}Treatment_k + \gamma_{002}Region1_k + \gamma_{003}Treatment_k \times Region1_k + \epsilon_{ijk} + r_{0jk} + \mu_{00k}$$

Region 2:

$$Y_{ijk} = \gamma_{000} + \gamma_{001}Treatment_k + \gamma_{002}Region2_k + \gamma_{003}Treatment_k \times Region2_k + \epsilon_{ijk} + r_{0jk} + \mu_{00k}$$

Region 3:

$$Y_{ijk} = \gamma_{000} + \gamma_{001}Treatment_k + \gamma_{002}Region3_k + \gamma_{003}Treatment_k \times Region3_k + \epsilon_{ijk} + r_{0jk} + \mu_{00k}$$

Region 4:

$$Y_{ijk} = \gamma_{000} + \gamma_{001}Treatment_k + \gamma_{002}Region4_k + \gamma_{003}Treatment_k \times Region4_k + \epsilon_{ijk} + r_{0jk} + \mu_{00k}$$

Where $Y_{ijk}$is the factor score on reflection, ownership, or collaboration.

γ_{000} is the mean outcome.

γ_{001} is the difference between the treatment and control groups for the regions other than the one of interest.

γ_{002} is the difference of the region of interest of the control group.
γ_{003} is the difference of the region of interest of the treatment group.
Finally, $\mu_{00k}$is the school-level error term, r_{0jk} is the classroom-level error term, and $\epsilon_{ijk}$is the student-level error term.

Although interacting region with treatment may not give the full insight into what is going on in the students' SEL scores, it does give insight into how different contexts may affect the treatment effect on students' SEL scores.

Fidelity of Implementation:

The analysis for estimating the fidelity of implementation is based on a 3-2-1 mediation effect model (Pituch et al., 2009).

Hausman Test Result:

To ensure that using a random effect model as opposed to a teacher fixed effect model was appropriate, a Hausman test was conducted to show that the probability limits of the two coefficients were equal. The resulting test (Chi-square with df of 4 =1.88, p = 0.597) did not reject the null hypothesis that a random effect estimation should not be used. This is consistent with the idea that, in a randomized control trial, the unobserved heterogeneity is uncorrelated with treatment.

Results

Results of the HLM show that the treatment students outperformed the control students by a .266 standard deviation on the summative science assessments. The results from the analysis are reported in Table 19. Columns 1 and 2 show this ITT estimation of the treatment effect on student science achievement. Column 2 is the estimation of the treatment effect without covariates but with only the focal classrooms. Columns 3 and 4 include additional school-level covariates. Columns 5 and 6 include the covariates, the baseline reading benchmark, and the benchmark test type. Column 5 includes all the students and Column 6 includes only the focal classrooms. Across all the estimations, the treatment effect remains and is statistically significant. The largest effect is shown in Columns 5 and 6, which include the additional covariate of benchmark scores and benchmark types; however, there were some students who did not have corresponding benchmark scores and thus the sample size decreases. To ensure that a missing benchmark was not biasing these results, an analysis of the relationship between the summative assessment and the dummy for having a benchmark was conducted. This relationship was not significant (t = 1.42, p = .156). Regardless, given the decrease in sample size and missing students when using the benchmark scores as a covariate, Column 3 is the reported treatment effect of .266, which corresponds to an 8-percentage point increase on a standardized science summative test as compared to the control group.

Table 19. Estimated Treatment Effect of ML-PBL

	(1)	(2)	(3)	(4)	(5)	(6)
Treatment effect	.262*	.304**	.266**	.33**	.356**	.415***
	(0.114)	(0.109)	(0.096)	(0.098)	(0.105)	(0.108)
All classrooms	X		X		X	
Focal classrooms only		X		X		X
Additional school-level covariates			X	X	X	X
Reading benchmark					X	X
N	2,371	1,975	2,371	1,975	2,186	1,833

Note. Treatment effect is the difference between the treatment and control group, measured in standard deviations. Standard errors are in parentheses. The additional covariates include proportion of races in a school and the region of the school. Reading benchmark includes the student's percentile ranking as well as the type of test taken.

*p < 0.05 **p < 0.01 ***p < 0.001

A sensitivity analysis was conducted to check the robustness of the findings. Using Estimate 3—which includes additional school-level covariates for Frank et al. (2013)'s framework for evaluating the robustness of an inference—to invalidate the inference, 29.228% of the estimate would have to be due to bias. This means that 693 cases would have to be replaced with cases for which there was a zero effect.

SEL Results

For the SEL outcomes, difference tests on all the social and emotional questions were conducted. Next, a factor analysis was conducted that indicated the validity of the three constructs: reflection, ownership, and collaboration. A three-level HLM was then conducted on each of the three constructs. An interaction of the treatment effect taking into consideration gender and region was also conducted and is reported in Table 20.

Table 20. Estimated Treatment Effect on Reflection, Ownership, and Collaboration

Outcome of interest					Gender	
Outcome	CFA Factor 1: Reflection	CFA Factor 2: Ownership	CFA Factor 3: Collaboration	CFA Factor 1: Reflection	CFA Factor 2: Ownership	CFA Factor 3: Collaboration
Treatment	.427**	0.225	.403**	.385*	0.183	.369*
	(0.137)	(0.136)	(0.131)	(0.155)	(0.158)	(0.144)
Outcome of interest				0.058	0.062	0.06
				(0.033)	(0.034)	(0.036)
Interaction				.105*	0.076	0.093
				(0.053)	(0.06)	(0.054)
Constant	-.648**	-0.456	-.601**	-.298*	-.307*	-.272*
	(0.243)	(0.245)	(0.229)	(0.132)	(0.134)	(0.121)
Random effects–Variances						
School	0.169	0.164	0.153	0.183	0.178	0.169
	(0.061)	(0.062)	(0.054)	(0.065)	(0.065)	(0.061)
Classroom	0.039	0.045	0.035	0.042	0.05	0.039
	(0.019)	(0.026)	(0.014)	(0.021)	(0.029)	(0.019)
Student	0.387	0.382	0.387	0.418	0.408	0.387
	(0.067)	(0.064)	(0.062)	(0.073)	(0.069)	(0.067)
N	1,843	1,843	1,843	1,569	1,569	1,843

Note. Coefficients are in standard deviations. Standard errors are in parentheses. *p < .05 **p < .01 ***p < .001

Predictor of interest	Free and reduced lunch			Free lunch			American Indian		
Outcome	CFA Factor 1: Reflection	CFA Factor 2: Ownership	CFA Factor 3: Collaboration	CFA Factor 1: Reflection	CFA Factor 2: Ownership	CFA Factor 3: Collaboration	CFA Factor 1: Reflection	CFA Factor 2: Ownership	CFA Factor 3: Collaboration
Treatment	.417**	0.207	.398**	.427**	0.21	.396**	.412**	0.221	.382**
	(0.133)	(0.131)	(0.127)	(0.132)	(0.13)	(0.127)	(0.132)	(0.134)	(0.125)
Predictor of interest	-0.496	-0.499	-0.45	-0.494	-0.5	-0.448	-2.43	-1.97	-3.06
	(0.332)	(0.327)	(0.32)	(0.35)	(0.344)	(0.336)	(12.61)	(12.54)	(12.32)
Interaction	0.715	1.01*	0.483	0.789	1.053*	0.573	-8.4	1.51	-12.11
	(0.441)	(0.429)	(0.436)	(0.455)	(0.433)	(0.449)	(14.68)	(16.76)	(14.01)
Constant	-0.216	-.227*	-0.194	-0.215	-.226*	-0.193	-0.217	-.228	-0.194
	(0.112)	(0.114)	(0.105)	(0.112)	(0.114)	(0.105)	(0.113)	(0.114)	(0.105)
Random effects–Variances									
School	0.162	0.151	0.148	0.16	0.148	0.147	0.168	0.164	0.151
	(0.06)	(0.061)	(0.053)	(0.06)	(0.061)	(0.0540	(0.061)	(0.061)	(0.054)
Classroom	0.039	0.046	0.035	0.039	0.045	0.035	0.039	0.045	0.035
	(0.019)	(0.026)	(0.014)	(0.019)	(0.026)	(0.014)	(0.019)	(0.026)	(0.014)
Student	0.387	0.382	0.387	0.387	0.382	0.387	0.387	0.382	0.387
	(0.067)	(0.064)	(0.062)	(0.067)	(0.064)	(0.062)	(0.067)	(0.064)	(0.062)
N	1,843	1,843	1,843	1,843	1,843	1,843	1,843	1,843	1,843

Note. Coefficients are in standard deviations. Standard errors are in parentheses. *p < .05 **p < .01 ***p < .001

Predictor of interest	Proportion Asian			Proportion Hispanic			Proportion Black			Proportion White		
Outcome	CFA Factor 1: Reflection	CFA Factor 2: Ownership	CFA Factor 3: Collaboration	CFA Factor 1: Reflection	CFA Factor 2: Ownership	CFA Factor 3: Collaboration	CFA Factor 1: Reflection	CFA Factor 2: Ownership	CFA Factor 3: Collaboration	CFA Factor 1: Reflection	CFA Factor 2: Ownership	CFA Factor 3: Collaboration
Treatment	.403**	0.198	.380**	.431**	0.228	.408**	.402***	0.198	.381***	.407***	0.205	.385**
	(0.127)	(0.125)	(0.122)	(0.132)	(0.132)	(0.126)	(0.11)	(0.104)	(0.104)	(0.116)	(0.113)	(0.113)
Outcome of interest	2.35*	2.39*	2.20*	.726**	.754**	.675*	-0.87**	-.899**	-.799**	.715*	.735*	.652*
	(0.975)	(0.975)	(0.927)	(0.279)	(0.282)	(0.262)	(0.279)	(0.28)	(0.265)	(0.284)	(0.287)	(0.269)
Interaction	-1.44	-1.27	-1.44	-.744*	-.798*	-.623*	1.05**	1.17***	.917**	-.937**	-1.05**	-.821*
	(1.020)	(1.000)	(0.977)	(0.334)	(0.351)	(0.317)	(0.338)	(0.318)	(0.331)	(0.339)	(0.33)	(0.327)
Constant	-0.2	-.210*	-0.18	-.226*	-.237*	-0.203	-0.200*	-.211**	-.18*	-.199	-.209*	-.178
	(0.106)	(0.107)	(0.1)	(0.113)	(0.114)	(0.106)	(0.08)	(0.081)	(0.08)	(0.093)	(0.094)	(0.088)
Random effects–Variances												
School	0.157	0.149	0.143	0.156	0.15	0.142	0.098	0.087	0.093	0.129	0.118	0.12
	(0.058)	(0.059)	(0.052)	(0.054)	(0.053)	(0.048)	(0.036)	(0.034)	(0.033)	(0.042)	(0.04)	(0.037)
Classroom	0.039	0.046	0.035	0.039	0.045	0.035	0.04	0.046	0.036	0.039	0.045	0.035
	(0.019)	(0.027)	(0.014)	(0.019)	(0.026)	(0.014)	(0.02)	(0.027)	(0.014)	(0.019)	(0.026)	(0.014)
Student	0.387	0.382	0.387	0.387	0.382	0.387	0.387	0.382	0.387	0.387	0.382	0.387
	(0.067)	(0.064)	(0.062)	(0.067)	(0.064)	(0.062)	(0.067)	(0.064)	(0.062)	(0.067)	(0.064)	(0.062)
N	1,843	1,843	1,843	1,843	1,843	1,843	1,843	1,843	1,843	1,843	1,843	1,843

Note. Coefficients are in standard deviations. Standard errors are in parentheses. *p < .05 **p < .01 ***p < .001

Predictor of interest	City			Suburban			Rural		
Outcome	CFA Factor 1: Reflection	CFA Factor 2: Ownership	CFA Factor 3: Collaboration	CFA Factor 1: Reflection	CFA Factor 2: Ownership	CFA Factor 3: Collaboration	CFA Factor 1: Reflection	CFA Factor 2: Ownership	CFA Factor 3: Collaboration
Treatment	0.111	-0.114	0.117	.605**	0.402*	.567**	.484**	0.296	.45***
	(0.089)	(0.092)	(0.088)	(0.194)	(0.192)	(0.185)	(0.157)	(0.155)	(0.15)
Outcome of interest	-0.489*	-.493*	-.455*	.317	0.318	.293	.357*	.363*	.335*
	(0.212)	(0.215)	(0.199)	(0.167)	(0.169)	(0.158)	(0.139)	(0.141)	(0.133)
Interaction	.663*	.713**	.598*	-.524*	-.524*	-.482*	-0.324	-0.395	-0.28
	(0.256)	(0.25)	(0.247)	(0.209)	(0.211)	(0.202)	(0.248)	(0.248)	(0.238)
Constant	0.021	0.013	0.027	-.324	-.335*	-0.294	-.285*	-.296*	-.258*
	(0.019)	(0.019)	(0.023)	(0.166)	(0.168)	(0.156)	(0.137)	(0.139)	(0.128)
Random effects–Variances									
School	0.137	0.129	0.126	0.155	0.150	0.141	0.16	0.155	0.145
	(0.044)	(0.043)	(0.039)	(0.052)	(0.052)	(0.046)	(0.055)	(0.056)	(0.049)
Classroom	0.038	0.045	0.035	0.038	0.045	0.035	0.039	0.045	0.035
	(0.019)	(0.026)	(0.014)	(0.019)	(0.026)	(0.014)	(0.019)	(0.026)	(0.014)
Student	0.387	0.324	0.387	0.387	0.382	0.387	0.387	0.382	0.387
	(0.067)	(0.064)	(0.062)	(0.067)	(0.064)	(0.062)	(0.067)	(0.064)	(0.062)
N	1,843	1,843	1,843	1,843	1,843	1,843	1,843	1,843	1,843

Across two of the three constructs (reflection and collaboration), there was a positive treatment effect on students' SEL in their science classes. While there is not an overall significant effect on ownership, when inspecting the interaction between treatment and free and reduced lunch, proportion Black, and urbanicity, there is a significant interaction between the treatment and these variables on ownership. When looking at the interaction between gender and treatment effect on reflection and ownership, there is an even stronger treatment effect for the girls than for the boys. When looking across proportion race, there are interactions for proportion of Hispanic, Black, and White students. The Hispanics and Whites across the board do better on the constructs, but the interaction with treatment is negative, indicating that they have a higher base response on the construct, so there is not as much to increase compared to the other races. We see the opposite in the proportion Black, where schools that have high proportion of Black students typically score low on the constructs unless they were in the treatment group, where the difference between the treatment and control groups is much higher. We find similar findings when comparing urbanicity of city versus suburban and rural.

Heterogeneity

For the tests of heterogeneity, interactions with the treatment effect were first conducted for school proportion free and reduced lunch and race and ethnicity and then the regional effects, followed by cross-level interactions on student gender and reading benchmark with treatment. The summary of the school- and student-level heterogeneity is reported in Table 21.

Table 21. Heterogeneity School-Level Effects

	model 1	Proportion free and reduced lunch	Proportion free lunch	Proportion American Indian	Proportion Asian	Proportion Hispanic	Proportion Black	Proportion White	City	Suburban	Rural
Treatment	.262*	.300***	.287***	.273*	.235*	.263*	.260**	.241**	.329**	.133	.301*
	(.114)	(.084)	(.081)	(.111)	(.108)	(.110)	(.091	(.090)	(.099)	(.159)	(.124)
Variable of interest		-.883***	-.901***	17.7***	3.06***	-.345*	-.388**	.521***	-.272*	.009	.454***
		(.215)	(.201)	(2.02)	(.603)	(.163)	(.121)	(.123)	(.125)	(.102)	(.097)
Treatment x variable of interest		-.376	-.314	-29.4	-3.29***	.416	-.253	.119	-.208	.353*	-.248
		(.356)	(.338)	(20.3)	(.671)	(.295)	(.249)	(.225)	(.206)	(.179)	(.221)
Constant	-.433**	-.162**	-.160**	-.195**	-.142*	-.170*	-.153*	-.148**	-.030	-.123	-.251
	(.165)	(.053)	(.050)	(.061)	(.056)	(.066)	(.061)	(.055)	(.062)	(.099)	(.069)
Random effects–Variances											
School	.109	.049	.042	.095	.095	.105	.060	.053	.072	.095	.089
	(.028)	(.019)	(.020)	(.027)	(.026)	(.028)	(.030)	(.027)	(.029)	(.026)	(.029)
Classroom	.033	.033	.034	.033	.033	.033	.034	.035	.034	.033	.034
	(.014)	(.014)	(.014)	(.014)	(.014)	(.014)	(.015)	(.015)	(.015)	(.014)	(.015)
Student	.850	.850	.850	.851	.850	.850	.850	.850	.850	.850	.850
	(.032)	(.032)	(.032)	(.032)	(.032)	(.032)	(.032)	(.032)	(.032)	(.032)	(.032)
N	2,371	2,371	2,371	2,371	2,371	2,371	2,371	2,371	2371	2371	2371

Note. Coefficients are in standard deviations. Standard errors are in parentheses. *p < .05 **p < .01 ***p < .001

When looking at heterogeneity effects, only proportion Asian has an interaction, which would indicate that, on average, the treatment effect had less effect for Asians; however, when taking a closer look at the proportion of Asian students in the sample, there is a large interaction between the benchmark for reading and for those who are Asian. These results seem to indicate that there is a baseline difference in reading scores on the proportion Asian and treatment status. Therefore, the interaction reported in the above table (Table 21) is likely the result of this underlying difference between schools and could be accounted for by the high proportion of Asians in northern Michigan who received high scores on the NWEA benchmark. In fact, when controlling for reading scores, the interaction on proportion of Asian students and treatment decreases and is no longer significant at the 0.05 (coef = -1.03, p-value = .103) level. When looking at heterogeneity by urbanicity, there is a strong treatment effect by the suburban group; otherwise, for the city and rural schools, there is no interaction and the treatment effect remains.

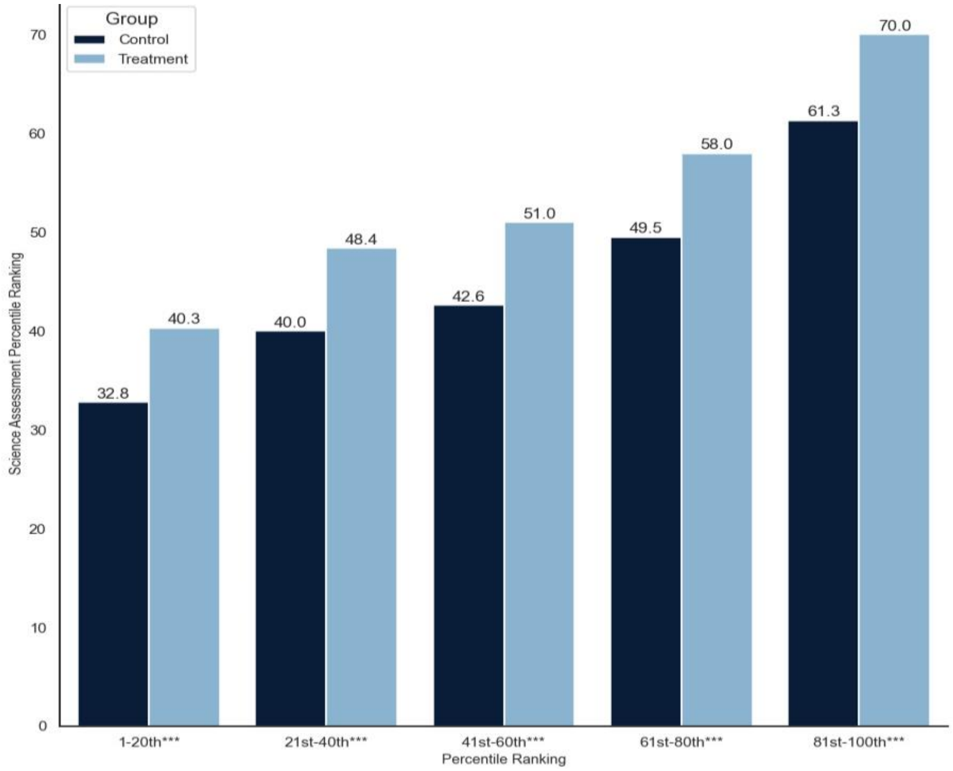
Table 22. Heterogeneity Student-Level Effects

	Gender	**Reading benchmark**
Treatment	.328**	.277**
	(.121)	(.084)
Variable of interest	.117	.014***
	(.063)	(.001)
Treatment x variable of interest	-.028	-.000
	(.086)	(.002)
Constant	-.288	-.147**
	(.075)	(.044)
Random effects–Variances		
School	.093	.044
	(.029)	(.019)
Classroom	.037	.025
	(.017)	(.011)
Student	.834	.742
	(.034)	(.027)
N	1,922	2,186

Note. Coefficients are in standard deviations. Standard errors are in parentheses. *p < .05 **p < .01 ***p < .001

Results show no interaction between gender or reading benchmark, indicating that the treatment was effective across gender and level of reading.

To check if the treatment students outperformed control students on the science assessment across all the pre-test reading percentile rankings, an interaction of the percentiles with treatment was conducted. These results are depicted in Figure 3. Across all quintiles, the treatment effect held, with the treatment students outperforming the control students when controlling for pre-test third-grade reading state-administered standardized test scores. These results are consistent with the heterogeneity analysis of reading scores described above.



Notes: a. ***p-value < 0.001
b. Standard Errors: For the treatment group, the standard error for the 1-20th percentile is 0.03, 21st-40th is 0.023, 41st-60th is 0.023, 61st-80th is 0.019, 81st-100th is 0.018. For the control group, the standard error for the 1st-20th percentile is 0.016, 21st-40th is 0.014, 41st-60th is 0.017, 61st-80th is 0.021, and 81st-100th is 0.017.

Figure 3. Third-Grade Treatment Effect Controlling for Reading Benchmark Scores

Fidelity of Implementation

A series of mediation analyses were conducted, and treatment effects were not detected. However, it must be noted that the mediation analyses conducted were based on a reduced sample set for only those teachers for whom observations or exit surveys were available. When running the treatment effects on this reduced data set, we found no treatment effects. Due to concerns that this sub-sample may have been biased, we included a dummy variable for inclusion in this reduced data set and ran the treatment effect on the full data set including the dummy variable. The dummy coefficient was not significant, indicating that there was no evidence that this sub-sample was biased. Therefore, the reduced data set is not powered enough to detect a mediation effect, as opposed to bias affecting the results of the mediation analyses. This means that if observations and exit surveys were available for all teachers, then there may in fact be a mediation effect that is currently unable to be observed.

We conclude that the variance of the treatment effect cannot be explained largely by the treatment teachers' level of implementation of the intervention. This is not unexpected, given that the intervention was a complex system, containing three major components, high-quality teacher and student materials and experiences, PL, and formative assessments that enhance student thinking and performance.

Discussion

ML-PBL is an unusual intervention in that it includes not only curriculum and lesson materials, but also professional learning and experiential embedded assessments designed to be used formatively throughout the units. Taking into account the holistic nature of ML-PBL suggests why other elementary school science reform interventions have shown somewhat limited effects (Klager, 2017). Typically, science reforms do not have the scope and depth of ML-PBL, nor are they based on PBL principles, the National Research Council's (2012) *Framework* for three-dimensional learning, or the NGSS; moreover, they are usually not conducted with an efficacy trial. ML-PBL results show a significant intervention treatment effect and one that is higher than what has been reported in other science studies (see Harris, 2015; Lynch et al., 2012; Wilson et al., 2010). The main effect that shows a significant difference on an objective science assessment between the treatment and control student is considerable, even taking into account developmental differences in reading (as shown in Figure 3). (See also Table 19, Column 3.) What is particularly noteworthy is that these main effects hold for: students of differing reading abilities and gender; school-level race, ethnicity, and SES; and across the major geographic regions of the state.

One way to interpret the ML-PBL main effect is to imagine a school district standardizing science achievement measures on a 100% scale, where the proficiency cutoff is 70%. By participating in ML-PBL, third-grade students who score 65%, which is below the proficiency cutoff, would be expected to have an 8% gain, moving them into the proficiency level. Additionally, students in the treatment effect might increase their letter grade by more than half a letter grade, from a C+ to a B.

Another positive treatment effect was found for social and emotional learning during science classes. The constructs used for this analysis reflect the components of the intervention practices, such as supporting students when they take on the role of being the driver for asking questions and figuring out phenomena. Finally, collaboration is a key feature of ML-PBL and one that ensures that all students work with others as they pursue questions. It underscores the importance of allowing all students to participate in experiences that encourage equitable practices that support science learning. These positive effects are largely being driven by females and urban schools where there are larger proportions of Black students.

Limitations

There are several limitations to this work. First, we would have preferred to have been able to have a larger number of teacher observations throughout the school year, but this was not possible because of cost constraints and observer and teacher availability. Second, we were not able to obtain student-level demographic information and had to rely on the school measures. Every study wishes that they had better measures: in this case, individual student information on key demographic variables—such as race and ethnicity, socioeconomic indicators, and family composition—would have been key to more closely analyzing the heterogeneity effects, but that was not available. This meant that we were forced to at least have a school indicator of the composition of the demographic characteristics, i.e. free and reduced lunch and proportion of racial and ethnic diversity. The good news is that we do not find any moderating effects on the treatment condition (except on proportion Asian, which is explained above); this shows that the treatment worked for all students, regardless of their fall academic benchmarks. We continue to work on possible avenues for obtaining this information; however, the school-level indicators, for the most part, reflect the characteristics of the overall racial and ethnic diversity within each school's student population. Third, other than through the observations and teacher reports, we did not have an indicator for the quality of the PL that the teachers received. In the future, we plan to obtain more information directly from the teachers engaged in PL.

Implications for Practice

Our interest in creating the ML-PBL intervention was in addressing the need for high-quality instructional and learning materials, teacher professional learning, and assessments that provide guidelines for school professionals engaging in transforming their science education programs to be aligned with the National Research Council's (2012) *A Framework for K–12 Science Education* and the NGSS (NGSS Lead States, 2013). Our intervention was designed not as a curricular guide with a script, but rather as a new approach to science learning and instruction that captured how to bring experiential learning into elementary classrooms, which supports equitable science academic, social, and emotional learning. Overall, results suggest that the integration of PBL features and three-dimensional learning, along with professional learning and assessments, together promote science learning and SEL for third-grade elementary school students.

What accounts for these effects? We argue that it is the entirety of the treatment—including key components of learning coupled with professional learning and assessments—that drives our increase in academic science learning and social and emotional development. Some important key components of our intervention are: the focus on students making sense of phenomena they find meaningful by using various science and engineering practices, disciplinary core ideas, and crosscutting concepts; collaborating on this figuring out process; building artifacts that represent responses to questions the students ask; and the DQ that leads to these outcomes.

Overall, the careful design-based research of the curriculum materials coupled with professional learning has allowed teachers from multiple settings to support students in developing academic and SEL outcomes. The data and results suggest that, for this population as well as the entire state of Michigan and the United States as a whole (see Table 6), our intentions were and could be fulfilled. We look forward to providing additional materials throughout the grade spans and with larger populations of students and teachers, including additional supports that address students with various learning disabilities. Another of our interests is to develop a strong coalition of teacher leaders who can help facilitate more than content knowledge but also how to use knowledge to figure out and understand phenomena, solve problems, and inspire curiosity about how the world works.

References

Bloom, H. (2005). *Randomizing groups to evaluate place-based programs*. MDRC.

Cohen, D. K., & Ball, D. L. (1999). *Instruction, capacity, and improvement* (CPRE Research Report Series RR-43). Consortium for Policy Research in Education.

Davis, E. A., & Krajcik, J. S. (2005). Designing educative curriculum materials to promote teacher learning. *Educational Researcher*, *34*(3), 3–14. https://doi.org/10.3102/0013189X034003003

Dewey, J. (1938). *Experience and education*. Macmillan Publishers.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, *14*(20), 1–11. https://doi.org/10.7275/da8t-4g52

Drake, C., & Sherin, M. G. (2006). Practicing change: Curriculum adaptation and teacher narrative in the context of mathematics education reform. *Curriculum Inquiry*, *36*(2), 153–187. https://doi.org/10.1111/j.1467-873X.2006.00351.x

Durlak, J. A., Domitrovich, C. E., Weissberg, R. P. & Gullotta, T. P. (Eds.). (2015). *Handbook of social and emotional learning: Research and practice*. Guilford Publications.

Frank, K. A., Maroulis, S. J., Duong, M. Q., & Kelcey, B. M. (2013). What would it take to change an inference? Using Rubin's causal model to interpret the robustness of causal inferences. *Educational Evaluation and Policy Analysis*, *35*(4), 437–460. https://doi.org/10.3102/0162373713493129

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

Harris, C. J., Krajcik, J., Pellegrino, J. & DeBarger, A. H. (2019). Designing knowledge-in-use assessments to promote deeper learning. *Educational Measurement: Issues and Practice,* *38*(2), 53–67. https://doi.org/10.1111/emip.12253

Harris, C., Penuel, W., D'Angelo, C., DeBarger, A., Gallagher, L., Kennedy, C., Cheng, B., & Krajcik, J. (2015). Impact of project-based curriculum materials on student learning in science: Results of a randomized controlled trial. *Journal of Research in Science Teaching, 52*(10), 1362–1385. https://doi.org/10.1002/tea.21263

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Jagers, R. J., Rivas-Drake, D., & Borowski, T., (2018). Equity & social and emotional learning: A cultural analysis. *Frameworks Briefs* (November). https://casel.org/wp-content/uploads/2020/04/equity-and-SEL-.pdf

Klager, C. (2017). Project-based learning in science: A meta-analysis of science achievement effects (Working Paper). Michigan State University.

Krajcik, J., Codere S., Dahsah, C., Bayer, R., & Mun, K. (2014). Planning instruction to meet the intent of the Next Generation Science Standards. *Journal of Science Teacher Education*, 25(2), 157–175. https://doi.org/10.1007/s10972-014-9383-2

Krajcik, J., & Czerniak, C. (2018). *Teaching science in elementary and middle school: A project-based approach* (5th ed.). Taylor & Francis Group.

Krajcik, J. S., Palincsar, A., & Miller, E., (2015). *Multiple literacies in project-based learning*. George Lucas Educational Foundation.

Krajcik, J. S., & Shin, N. (2014). Project-based learning. In R. K. Sawyer (Ed.), *The Cambridge handbook of learning sciences* (2nd ed., pp. 275–297). Cambridge University Press.

Lynch, S. J., Pyke, C., & Grafton, B. H. (2012). A retrospective view of a study of middle school science curriculum materials: Implementation, scale-up, and sustainability in changing policy environment. *Journal of Research in Science Teaching, 49*(3), 305–332, https://doi.org/10.1002/tea.21000

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the big five factor structure through exploratory structural equation modeling. *Psychological Assessment, 22*(3), 471–491. https://doi.org/10.1037/a0019227

Miller, E., Codere, S., & Krajcik, J. (2018). Developing assessment tasks to promote student sensemaking of phenomena and flexible thinking. In J. Kay & R. Luckin (Eds.). *Rethinking learning in the digital age: Making the learning sciences count*. 13th International Conference of the Learning Sciences (ICLS) 2018, Volume 3. International Society of the Learning Sciences.

Miller, E. C., & Krajcik, J. S. (2019). Promoting deep learning through project-based learning: A design problem. *Disciplinary and Interdisciplinary Science Education Research*, *1*(7). https://doi.org/10.1186/s43031-019-0009-6

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, structures, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 61–90). Lawrence Erlbaum Associates, Inc.

National Academy of Engineering and National Research Council. (2014). *STEM integration in K–12 education: Status, prospects, and an agenda for research*. National Academies Press. https://doi.org/10.17226/18612

National Research Council. (1999). *How people learn: Brain, mind, experience, and school.* National Academies Press.

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K–8*. National Academies Press. https://doi.org/10.17226/11625

National Research Council. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.

National Science Teachers Association. (2019). *About the Next Generation Science Standards*. https://ngss.nsta.org/About.aspx

NGSS Lead States. (2013). *Next Generation Science Standards: For states, by states*. National Academies Press.

Oliveira, A. W. (2010). Improving teacher questioning in science inquiry discussions through professional development. *Journal of Research in Science Teaching*, *47*(4), 422–453. https://doi.org/10.1002/tea.20345

Pellegrino, J. W., & Hilton, M. L. (Eds.). (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. National Academies Press.

Pituch, K. A., Murphy, D. L., & Tate, R. L. (2009). Three-level models for indirect effects in school- and class-randomized experiments in education. *The Journal of Experimental Education*, *78*(1), 60–95. https://doi.org/10.1080/00220970903224685

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*(2), 173–185. https://doi.org/10.1037/1082-989X.2.2.173

Raudenbush, S. W., & Bryk, A. (2002). *Hierarchical linear models* (2nd ed.). SAGE Publications.

Raykov, T., & Marcoulides, G. A. (2000). A method for comparing completely standardized solutions in multiple groups. *Structural Equation Modeling: A Multidisciplinary Journal*, *7*(2), 292–308. https://doi.org/10.1207/S15328007SEM0702_9

Sawyer, R. K. (2014). The future of learning: Grounding educational innovation in the learning sciences. In R.K. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences* (2nd ed., pp. 1–20). Cambridge University Press.

Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., & Raudenbush, S. W. (2011). *Optimal design plus empirical evidence: Documentation for the "Optimal Design" Software Version 3.0.* http://www.hlmsoft.net/od/od-manual-20111016-v300.pdf

Spybrook, J., Westine, C. D., & Taylor, J. A. (2016). Design parameters for impact research in science education: A multistate analysis. *AERA Open*, *2*(1), 1–15. https://doi.org/10.1177/2332858415625975

Tipton, E. (2014). How Generalizable Is Your Experiment? An Index for Comparing Experimental Samples and Populations. *Journal of Educational and Behavioral Statistics*, *39*(6), 478–501. https://doi.org/10.3102/1076998614558486

van den Bergh, L., Ros, A., & Beijaard, D. (2014). Improving teacher feedback during active learning. *American Educational Research Journal*, *51*(4), 772–809. https://doi.org/10.3102/0002831214531322

What Works Clearinghouse. (2020). *What Works Clearinghouse: Standards handbook version 4.1*. Institute of Education Sciences, US Department of Education.

Wilson, C. D., Taylor, J. A., Kowalski, S. M., & Carlson, J. (2010). The relative effects and equity of inquiry-based and commonplace science teaching on students' knowledge, reasoning, and argumentation. *Journal of Research in Science Teaching*, *47*(3), 276–301. https://doi.org/10.1002/tea.20329

Appendix A.
NGSS Standards

Squirrels/Adaptation

3-LS4-1:
Analyze and interpret data from fossils to provide evidence of the organisms and the environments in which they lived long ago.

3-LS4-2:
Use evidence to construct an explanation for how the variations in characteristics among individuals of the same species may provide advantages in surviving, finding mates, and reproducing.

3-LS4-3:
Construct an argument with evidence that in a particular habitat some organisms can survive well, some survive less well, and some cannot survive at all.

3-LS4-4:
Make a claim about the merit of a solution to a problem caused when the environment changes and the types of plants and animals that live there may change.

3-LS3-2:
Use evidence to support the explanation that traits can be influenced by the environment.

3-LS1-1:
Develop models to describe that organisms have unique and diverse life cycles but all have in common birth, growth, reproduction, and death.

Toys/Forces and Motion

3-PS2-1:
Plan and conduct an investigation to provide evidence of the effects of balanced and unbalanced forces on the motion of an object.

3-PS2-2:
Make observations and/or measurements of an object’s motion to provide evidence that a pattern can be used to predict future motion.

3-PS2-3:
Ask questions to determine cause and effect relationships of electric or magnetic interactions between two objects not in contact with each other.

3-PS2-4:
Define a simple design problem that can be solved by applying scientific ideas about magnets.

Birds/Biodiversity

3-LS2-1:
Construct an argument that some animals form groups that help members survive.

3-LS3-1:
Analyze and interpret data to provide evidence that plants and animals have traits inherited from parents and that variation of these traits exists in a group of similar organisms.

3-LS3-2:
Use evidence to support the explanation that traits can be influenced by the environment.

3-LS4-2:
Use evidence to construct an explanation for how the variations in characteristics among individuals of the same species may provide advantages in surviving, finding mates, and reproducing.

3-5-ETS1-1:
Define a simple design problem reflecting a need or a want that includes specified criteria for success and constraints on materials, time, or cost.

Plants

3-LS1-1:
Develop models to describe that organisms have unique and diverse life cycles but all have in common birth, growth, reproduction, and death.

3-LS3-1:

Analyze and interpret data to provide evidence that plants and animals have traits inherited from parents and that variation of these traits exists in a group of similar organisms.
3-LS3-2:
Use evidence to support the explanation that traits can be influenced by the environment.
3-LS4-3:
Construct an argument with evidence that in a particular habitat some organisms can survive well, some survive less well, and some cannot survive at all.
3-LS4-4:
Make a claim about the merit of a solution to a problem caused when the environment changes and the types of plants and animals that live there may change.
3-ESS2-1:
Represent data in tables and graphical displays to describe typical weather conditions expected during a particular season.
3-ESS2-2:
Obtain and combine information to describe climates in different regions of the world.
3-ESS3-1:
Make a claim about the merit of a design solution that reduces the impacts of a weather-related hazard.
3-5-ETS1-1:
Define a simple design problem reflecting a need or a want that includes specified criteria for success and constraints on materials, time, or cost.
3-5-ETS1-2:
Generate and compare multiple possible solutions to a problem based on how well each is likely to meet the criteria and constraints of the problem.

Why Do I See So Many Squirrels But I Can't Find Any Stegosauruses?

Learning Set 2: How is the squirrel's structure unique and important?

Lesson 2.2: How does a squirrel balance? SEL Focus: How can we challenge ourselves?

Lesson Overview (Est. time: 60 min.)	**L2.2:** How does a squirrel balance? **Lesson Snapshot** 1. <u>Introduction:</u> Watch the balancing video and introduce the DQ and the challenge. 2. <u>Planning and Investigation:</u> Introduce students to another use for models, to test their ideas. Students develop a plan for how they will balance on and walk across the rope (pool noodle). As students test out their plans, plan 'checkers' make sure they follow the plan. 3. <u>Analyzing Data and Discussion:</u> Students observe and analyze the squirrel skeletal structure: tail, low to the ground and light-weight body and compare with the marmot's skeletal structure. They watch a video of the marmot "balancing." Students make claims with evidence about the need for balance, the squirrel's structure, and how the structure helps the eastern grey squirrel meet its needs for survival. 4. <u>Wrap Up:</u> Students analyze the marmot video, photo, body, and skeleton of a marmot and discuss if the marmot needs to balance and how they know. **<u>Learning Performance</u>** Students will develop claims with evidence that the squirrel's structures are related to its survival in its environment and that a person can tell the behaviors of an organism by looking at its skeleton.(through the lenses of *structure and function*, *patterns*, and *cause and effect*). **SEL Learning Goal** **Identity development during challenges:** We can learn to be comfortable with uncertainty and take on the challenge of collaborating with others and new ways of engaging with content. **Building toward PEs** **3-LS3-1** Analyze and interpret data to provide evidence that plants and animals have **traits i**nherited from parents and that variation of these traits exists in a group of similar organisms. (in birds) **3-LS4-1** Analyze and interpret data from fossils to provide evidence of the organisms and the environment in which they lived long ago. (in L3.6) **3-LS4-3** Construct an argument with evidence that in a particular habitat, some organisms survive well, some less well, and some cannot survive at all. (in L3.6) **Math Standards: Measurement and Data** - Represent and Interpret Data. CCSS.MATH.CONTENT.**3.MD.B.4** Generate measurement data by measuring lengths using rulers marked with halves and fourths of an inch. Show the data by making a line plot, where the horizontal scale is marked off in appropriate units— whole numbers, halves, or quarters. **Math Competency Statements** I can generate data by measuring lengths to the half and fourth of an inch/cm.

Materials and Prep	***Materials*** - Driving Question Board (DQB) Learning Set 2, Driving Question Slides - 2 strands of thick rope (or pool noodles) - Science notebooks - Pictures and diagrams of the 1. squirrel's skeletons and body structure and 2. hoary marmot body Pictures and diagrams of Eastern Grey Squirrel and Marmot including skulls and skeletons - Video Squirrel relaxing on telephone wire; Squirrel on obstacle course; Squirrel on a Phone Line; Squirrel jumping; marmot mother and baby walking on rocks - Object or weights of 10 lbs, and an object that weighs just over a pound for demonstration - Measuring tape/yardstick or meter stick - Pillow for landing ***Preparation*** - Have the videos ready - Decide in advance if there will be rules for balancing (shoes/no shoes, etc.) - Reading (for teachers) about balancing and squirrels http://www.nutsaboutsquirrels.com/2830/how-do-squirrels-walk-along-wires/ - Hoary marmot (in the rocky mountains)https://en.wikipedia.org/wiki/Hoary_marmot ***Embedded Language Supports*** - Negotiation of meaning through authentic peer dialogue - Support language of meaning through realia, video and action - Explicit support in developing claims - Discourse tools from WIDA
Lesson Component	**How to Implement**
What are kids figuring out?	**Students are figuring out** that squirrels have special structures that allow them to balance and that they need to balance to survive. Students also make a claim that a person could tell that an animal needed to balance by looking at its skeleton. **SEL:** Students are figuring out that they are comfortable with taking risks, making mistakes, and trying out new ideas. **Look Fors** 1.**Look for** 1. students basing claims (LS2, Driving Question slides—claims about structure) on structure and function and cause and effect, and using evidence from the skeleton and body structure and, 2. throughout the lesson, **(SEL) look for** students trying new challenges and risking making mistakes.
1 **Introduction** (10 min)	**Introduce Phenomenon and DQ, "How does a squirrel balance?"** 1. Have a student read the unit LS 2 Driving Question. Remind the students that they will still explore squirrel survival, but now they will look at the structures of the animals' bones and bodies to examine how that helps them to survive where they live outside. Have a student read the lesson Driving Question. Discuss in *turn-and-talk*, "how could we find out how the squirrel balances?" 2. Review the student definition of structures from L2.1 slides. Have one student read it to the class. ***Turn-and-talk***, "Will structures be important in this lesson, too?" Have students share what their partner said. 3. Ask if they've ever seen a squirrel jump or balance. (Solicit some responses.) Watch video of a squirrel balancing and jumping. Videos: Squirrel relaxing on telephone wire; Squirrel on obstacle course; Squirrel on a phone line; Squirrel jumping 4. Ask students to review the purpose of the survival model that they drew in Learning Set 1. *What were they trying to communicate? Why did they make it?* Tell students that they will be using a different type of model today. Present the challenge: Students will try to balance on the piece of rope (or pool noodle) while walking from one end to the other. Then they will jump from the end of the rope to the pillow and "land." Ask students to brainstorm what they think the rope and the yardstick "stand" for. They are going to use this physical model to test their ideas about squirrels' structures and how they

	help squirrels. - Let the students know that they will have more than one chance to try their plan (attempt). Tell them there is no one ‘right way’ to plan and encourage them to try something new. 5. Write down the lesson 2.1 Driving Question: “***How does a squirrel balance***?” Have students do a *turn-and-talk* about what they think they will figure out today. Have two students share what their partner said. Add questions to the DQB.
2 **Planning and Investigation** (25 min)	**Describe the challenge for balancing and conducting the investigation.** 1. Students design an investigation of themselves, humans, balancing on the thick ropes (pool noodles) and jumping and landing. They need to make a careful plan, because the class will check their plan for accuracy as they attempt to balance. 2. Students may use a yardstick for balance to stretch behind them like a squirrel’s tail, crouch down low, go sideways, crawl, etc. 3. Students draw their plan of themselves balancing on the rope and jumping and landing. 4. After 5-10 minutes, students will come up one by one and explain their plan to the class. Then, they will try to balance on the rope as outlined in their plan. 5. Students compare how they balanced and jumped to how squirrels balance (i.e., squirrels flip their tails back and forth, stay low to the ground, are light). This might be a good time to rewatch one of the videos from the Introduction segment. Ask students how the physical model of balancing on the rope as if on a tree branch helps them understand a squirrel’s structure. 6. If there is time, give students the opportunity to “correct their plans and try again (make adjustments so the plan matches what the students do in the trial).
3 **Analyzing Data and Discussion** (15 min)	**Compare a squirrel and a marmot and how they may, or may not, need to balance to survive.** 1. Show a picture of a marmot and its environment. Compare the marmot to the eastern grey squirrel. Ask, “How are the structures the same or different?” - Skeletal and Body Diagrams of Eastern Grey Squirrel and Marmot including skulls and skeletons 2. Write down measurements of a hoary marmot and an eastern grey squirrel and demonstrate with objects and tape measure how much these two animals weigh and how long they are. If there is time, students can work in small groups to measure the length of the two animals. Pass around objects to the students. 3. In North America - Adult hoary marmots weigh **10 pounds** (**4.5 kg**) or more and may exceed **30 inches (76 cm)** in total length. - Adult eastern grey squirrels can weigh up to about **1 ⅓ pounds (600 grams or 20 ounces)** and they are **18 to 20 inches (46 to 51 cm)** long (including the tail.) 4. Ask students to use evidence to predict if the marmot needs to balance for survival. They should do this in groups of three and then share out in large group. Help students make sense of each other's ideas. - **Suggested Questions:** “*How can you use the marmot’s structure and body to make a prediction about the marmot’s needs for survival? Do they need to climb trees to get their food and hide from predators?” “Could the marmot have* ***different ways of meeting its needs*** *for survival than the eastern grey squirrel? How could this be?”* 5. Watch the video of the marmot to check and discuss claims: marmot mother and baby walking on rocks

	 4. Help students listen carefully and think about one another's ideas **Suggested Prompts** *What is a difference that your friend noticed about the marmot's structure compared to the squirrel's? How do they think this structure helps or doesn't help the marmot balance and jump in the trees? Does their idea make sense to you?*
4 **Wrap Up** (10 min)	**Wrap up and use questions for reflection in science notebooks** 1. As a large group, construct an explanation about the need to balance, the structure of the squirrel, and how the structure helps it meet its needs for survival. Once a claim is agreed upon, write the Claim on the Driving Question slides 2. If there is time, look over the Learning Set 2, Driving Question slides; decide if any questions have been answered.
Formative Assessment	**Look Fors** When developing the final explanation, **look for** 1. students basing claims on structure and function and cause and effect, and using evidence from the skeleton and body structure and, 2. throughout the lesson, **(SEL) look for** students trying new challenges and risking making mistakes. **Evidence Statement** The claim will include evidence from the investigation of balancing and a connection to the skeleton. The second claim is the reverse, that the skeleton gives clues to how an organism might have needed to balance. **SEL:** Students will try more than one time.

Image attribution: https://westbridgfordwire.com/rspca-says-dont-forget-pets-snowy-weather-arrives/

Appendix B
MOU

Treatment MOU

Memorandum of Understanding between the __________ School and the Multiple-Literacies in Project-Based Learning Project, CREATE for STEM Institute at Michigan State University, Concerning a Third-Grade Efficacy Study

The Multiple Literacies in Project-Based Learning (ML-PBL) project is funded through the George Lucas Educational Research Foundation. The goal of ML-PBL is to design, develop and test elementary school science materials to meet the Next Generation Science Standards (NGSS), and start children on a path of lifelong learning. The ML-PBL materials align with the new Michigan Science Standards, NGSS and the Common Core State Standards for English Language Arts/Literacy and Mathematics. The ML-PBL project team is committed to supporting the learning of all students regardless of their first language or background experiences. Our ultimate goal is to produce project-based curricular resources that teachers and students will find engaging, and that will support teachers in enabling their learners to reach ambitious and standards-aligned science, language arts, and mathematics goals. We are entering the 4th year of the project, which involves conducting an efficacy study to provide evidence that students learn in project-based environments.

Below we share what the district, school and teachers will receive from ML-PBL for participating in the efficacy study, and what ML-PBL will need in return, so that the research is informed by *our joint efforts*.

School and Teacher Selection:

1. From 16-24 schools in the Genesee/Kent County area, we matched pairs of schools based on demographic characteristics and achievement data. From the pairs we randomly assigned schools to treatment and control conditions. Your school was assigned to the Treatment Condition. All grade 3 teachers in the treatment schools will have the opportunity to participate.

In SY 2018-19, the ML-PBL project will provide treatment sites with the following:

1. All third-grade ML-PBL materials: four units, including teacher and student materials, all provided online. The project will provide post-unit and end-of-year assessments.
2. Teachers will receive professional development, including an orientation to all of the teaching supports, and how to access them.
3. The student and teacher materials are designated Creative Commons-Open Source, so that after the project, the district is able to use the curriculum resources free of charge in perpetuity.
4. Optional student access to online resources. **Note: ML-PBL curriculum resources are designed to be used with or without student access to technology**.
5. We have contracted with ECA Science Kit Services to provide kits containing a majority of the materials students will need to engage in first-hand science experiences. These kit materials will be delivered to the school offices at identified times.
6. Professional Learning will be provided throughout the school year: 3 days during the summer, 1 day prior to each new unit , and video conference support every two weeks during enactment. If substitutes are necessary for PL, the project will provide funding for them. Teachers will be reimbursed for travel to and from professional learning if it is outside of their home district/ISD.

District, School and Teacher Responsibilities at ML-PBL treatment sites:

1. Enact the full curriculum. Teachers will need to dedicate at least 45 minutes of instructional time to science instruction at least four days each week.
2. Print out and copy student notebook sheets and other daily printed materials, if desired.

3. Attend all professional learning activities. Schools/District will need to arrange for substitute teachers or provide facilities for professional learning outside of school hours.
4. Administer post-unit assessments and student interest surveys for each unit, and submit copies (i.e., either hard copies or electronic copies) of student responses to project site leaders for scoring and analysis. Teachers may also use these assessments for their own evaluation purposes.
5. Allow project team to administer end-of-year, new Michigan Science Standards-aligned assessment.
6. Allow ML-PBL researchers to collect student work and interview a small number of students from treatment classrooms.
7. Pack up materials and consumables that were not used at the end of each unit, so that ECA can pick the kits up at designated times.
8. Provide access to State reading and mathematics scores (M-STEP) and to district reading and mathematics benchmark scores (e.g., NWEA, STAR, etc.) on an individual student basis.
9. Provide access to all ML-PBL classrooms to make observations and videos of the enactment.
10. Participate in teacher surveys and interviews.
11. Provide reliable Internet connections and effective IT support for students using optional computing devices. Additional support is available via our "hotline" at MLPinfo@umich.edu.
12. Submit for each participating class a list of students, including full name of student, District Permanent ID #, and DOB. For classes using computing devices, include students' gmail addresses and passwords.

PRIVACY AND CONFIDENTIALITY

- The data for this project and information about all participating teachers and students will be kept confidential.
- De-identified data (with no names or other information that could identify student, school, or district) will be shared with research partners at other institutions collaborating on this research for coding and analysis.
- Audio and videotapes will be kept in a secure locked cabinet and only available to ML-PBL researchers.
- The results of this study may be published or presented at professional meetings, but the identities of all research participants will remain anonymous.
- The ML-PBL project has been deemed "exempt" by the Institutional Review Board of Michigan State University, which means that it involves minimal or no risk for participating students and teachers.
- We will prepare letters informing parents/guardians about this research, including contact information if they should have any questions.

Questions should be directed to: Professor Joe Krajcik (krajcik@msu.edu), the Principal Investigator, or Sue Codere (Coderesu@msu.edu), the ML-PBL Project Manager, or Nathan Burroughs (Burrou25@msu.edu), the Efficacy Study Project Manager.

Signatures: Date:

_____ Public Schools ML-PBL Project
CREATE for STEM, MSU

Control MOU

Memorandum of Understanding between the __________ School District and the Multiple-Literacies in Project-Based Learning Project, CREATE for STEM Institute at Michigan State University, Concerning a Third-Grade Efficacy Study

The Multiple Literacies in Project-Based Learning (ML-PBL) project is funded through the George Lucas Educational Research Foundation. The goal of ML-PBL is to design, develop and test elementary school science materials to meet the Next Generation Science Standards (NGSS), and start children on a path of lifelong learning. The ML-PBL materials align with the new Michigan Science Standards, NGSS and the Common Core State Standards for English Language Arts/Literacy and Mathematics. The ML-PBL Team is committed to supporting the learning of all students regardless of their first language or background experiences. Our ultimate goal is to produce project-based curricular resources that teachers and students will find engaging, and that will support teachers in enabling their learners to reach ambitious and standards-aligned science, language arts, and mathematics goals. We are entering the 4th year of the project, which involves conducting an efficacy study to provide evidence that students learn in project-based environments.

Below we share what the district, school and teachers will receive from ML-PBL for participating in the efficacy study, and what ML-PBL will need in return, so that the research is informed by *our joint efforts*.

School and Teacher Selection:

1. From 16-24 schools in the Genesee/Kent County area, we matched pairs of schools based on demographic characteristics and achievement data. From the pairs we randomly assigned schools to treatment and control conditions. Your school was assigned to the control condition. All grade 3 teachers in the control schools will have the opportunity to participate as controls in SY2-18-19 and as treatment schools in SY2019-20.

The ML-PBL project will provide to Control (non-treatment) sites:

1. During SY 2018-19, teachers in the non-treatment schools will receive one day of professional development on the NGSS.
2. During SY 2019-20, teachers in the non-treatment sites will receive all of the third-grade units and kits that teachers in the treatment schools received in the SY 2018–19, and professional learning similar to what was offered to treatment schools in SY 2018–19.
   a. All third-grade ML-PBL materials: four units, including teacher and student materials, all provided online. The project will provide post-unit and end-of-year assessments.
   b. Teachers will receive professional development, including an orientation to all of the teaching supports, and how to access them.
   c. The student and teacher materials are designated Creative Commons-Open Source, so that after the project, the district is able to use the curriculum resources free of charge in perpetuity.
   d. Optional student access to online resources. **Note: ML-PBL curriculum resources are designed to be used with or without student access to technology**.
   e. We have contracted with ECA Science Kit Services to provide kits containing a majority of the materials students will need to engage in first-hand science experiences. These kit materials will be delivered to the school offices at identified times.
   f. Professional Learning will be provided throughout the school year: 3 days during the summer, 3 additional days throughout the school year (1 day prior to each new unit), and video conference support every two weeks during enactment. If substitutes are necessary for PL, the project will provide funding for them. Teachers will be reimbursed for travel to and from professional learning if it is outside of their home district/ISD.

Responsibilities of District and Teachers at Control (non treatment) sites:

1. Allow project team to administer end-of-year, new Michigan Science Standards-aligned assessment.
2. Identify, working with teachers and researchers, a limited number of focus students from whom we will collect work samples and conduct student interviews about their science work.
3. Provide access to State reading and mathematics scores and to district reading and mathematics benchmark scores (e.g., NWEA, STAR, etc.) on an individual student basis.
4. Provide access to classrooms to make observations and to conduct surveys and interviews. We would like permission to videotape the teaching of science three times during the school year.

PRIVACY AND CONFIDENTIALITY

- The data for this project and information about all participating teachers and students will be kept confidential.
- De-identified data (with no names or other information that could identify student, school, or district) will be shared with research partners at other institutions collaborating on this research for coding and analysis.
- Audio and videotapes will be kept in a secure locked cabinet and only available to ML-PBL researchers.
- The results of this study may be published or presented at professional meetings, but the identities of all research participants will remain anonymous.
- The ML-PBL project has been deemed “exempt” by the Institutional Review Board of Michigan State University, which means that it involves minimal or no risk for participating students and teachers.
- We will prepare letters informing parents/guardians about this research, including contact information if they should have any questions.

Questions should be directed to: Professor Joe Krajcik (krajcik@msu.edu), the Principal Investigator, or Sue Codere (Coderesu@msu.edu), the ML-PBL Project Manager, or Nathan Burroughs (Burrou25@msu.edu), the Efficacy Study Project Manager.

Signatures: Date:

_____ Public Schools

ML-PBL Project
CREATE for STEM, MSU

Appendix C.
Region Balance

The following tables show our sample compared with the schools that share the same region code according to the Michigan Department of Education. This comparison was done post-hoc because our intervention was never intended to be generalizable; however, we found that it would be useful to understand our schools in their district and community context. In the future, we hope to scale our intervention with a generalizable national sample with a nested randomized treatment and control sample.

Table C.1 Detroit Region Balance

	Detroit school population	Detroit school sample
Total eligible schools	236	19
Average N of third-grade teachers (per school)	2.82	1.53
Average N of third-grade students (per school)	72	60
Number of student enrollment	456	498
% of free-reduced lunch students	89%	77%
% of Native American students	0.98%	0.55%
% of Asian students	1.42%	2.48%
% of Hispanic students	15.10%	13.32%
% of Black students	80.90%	82.06%
% of White students	1.60%	1.59%
% of minority students	98.40%	98.41%

Source: Data are from the Michigan Consortium for Educational Research (MCER)

Table C.2. Region: Genesee

	Genesee school population	Genesee school sample
Total eligible schools	66	7
Average N of third-grade teachers (per school)	2.82	2.43
Average N of third-grade students(per school)	73	94
Number of student enrollment	423	510
% of free-reduced lunch students	62.67%	52.35%
% of Native American students	0.34%	1.96%
% of Asian students	1.12%	1.79%
% of Hispanic students	4.50%	3.97%
% of Black students	19.90%	12.19%
% of White students	74.14%	80.09%
% of minority students	25.86%	19.91%

Table C.3. Region: Kent

	Kent school population	Kent school sample
Total eligible schools	112	9
Average N of third-grade teachers (per school)	2.92	2.89
Average N of third-grade studetns (per school)	73	86
Number of student enrollment	464	442
% of free-reduced lunch students	54.21	59.29%
% of Native American students	0.80%	2.15%
% of Asian students	4.00%	3.84%
% of Hispanic students	20.89%	19.35%
% of Black students	16.43%	13.67%
% of White students	57.88%	61.00%
% of minority students	42.12%	39.00%

Table C.4. Region: Northern Michigan

	Northern MI school population	Northern MI school sample
Total eligible schools	98	11
Average N of third-grade teachers (per school)	2.87	2.63
Average N of third-grade students (per school)	79	87
Number of student enrollment	432	455
% of free-reduced lunch students	49.00%	41.05%
% of Native American students	2.82%	2.83%
% of Asian students	2.20%	3.52%
% of Hispanic students	6.00%	4.22%
% of Black students	8.00%	10.50%
% of White students	80.98%	78.93%
% of minority students	19.02%	21.07%

As explained above, the Northern MI region includes three districts from northern Michigan because each district is very small. Overall, the samples for the treatment and control are similar to other schools sharing the same region codes. The exception is Genesee, which is predominately White. Given what we know about our regions, this is why we chose to use region and school race/ethnicity as additional covariates.

Teacher Fixed Effect Model

For a sensitivity check, we also employed a teacher-level fixed effect model as shown in Table 16. Using a fixed effect dummy for teachers did not affect our treatment effect. This finding suggested to us that potential unobserved differences between teachers did not bias our treatment effect.

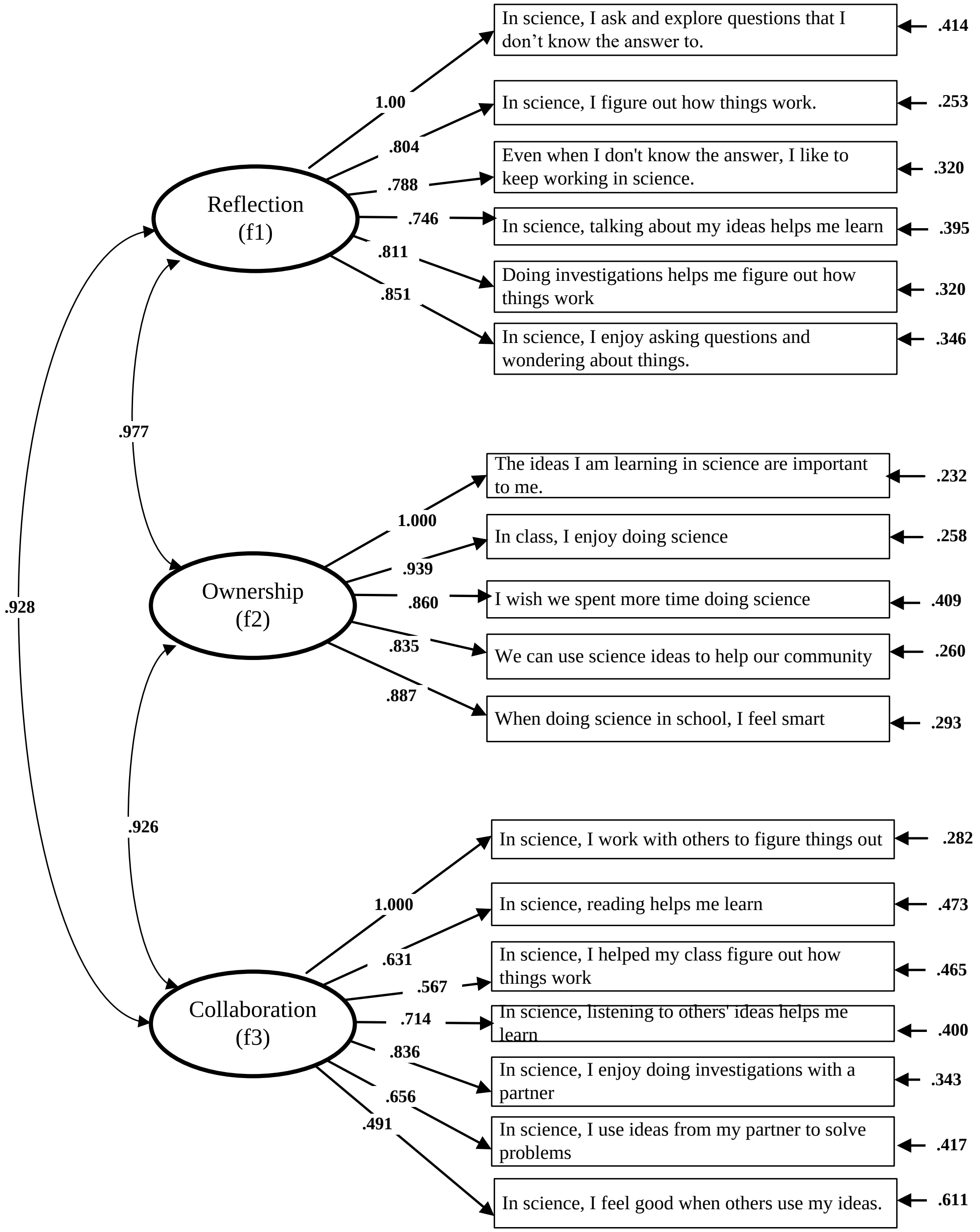
Table C.5. Treatment Effect Using a Fixed Effect Model

	Fixed effect (Model 1)	Fixed effect (including student-level covariates)
	b/se	b/se
Treatment	0.362***	0.308***
	(0.074)	(0.072)
Benchmark		0.016***
		(0.002)
Gender		-0.240
		(0.411)
Gender flag		-0.015
		(0.231)
Form B		-1.750**
		(0.634)

Form C		-1.326+
		(0.672)
Constant	-.141	0.284
	(.059)	(0.373)
N	2,371	2,371

Note. Treatment effect is the difference between the treatment and control group, measured in standard deviations. Standard errors are in parentheses.
+p < 0.1 *p < 0.05 **p < 0.01 ***p < 0.001

Appendix D.
SEL measure analysis


Reflection (f1)
Ownership (f2)
Collaboration (f3)
.977
.928
.926
1.00
In science, I ask and explore questions that I don't know the answer to.
.414
.804
In science, I figure out how things work.
.253
.788
Even when I don't know the answer, I like to keep working in science.
.320
.746
In science, talking about my ideas helps me learn
.395
.811
Doing investigations helps me figure out how things work
.320
.851
In science, I enjoy asking questions and wondering about things.
.346
1.000
The ideas I am learning in science are important to me.
.232
.939
In class, I enjoy doing science
.258
.860
I wish we spent more time doing science
.409
.835
We can use science ideas to help our community
.260
.887
When doing science in school, I feel smart
.293
1.000
In science, I work with others to figure things out
.282
.631
In science, reading helps me learn
.473
.567
In science, I helped my class figure out how things work
.465
.714
In science, listening to others' ideas helps me learn
.400
.836
In science, I enjoy doing investigations with a partner
.343
.656
In science, I use ideas from my partner to solve problems
.417
.491
In science, I feel good when others use my ideas.
.611

ESEM Model goodness-of-fit: χ^2 (df) = 297.97 (102), p < .000, RMSEA = .028, CFI = .97, SRMR = 0.018

Figure D.1. Three Factor (ESEM Plot)

Table D.1.

	Unstandardized estimate				Standardized estimate			
	Treatment		Control		Treatment		Control	
	Estimate (λ)	SE.	Estimate (λ)	SE.	Estimate (λ)	SE.	Estimate (λ)	SE.
Reflection								
v1_In science, I ask and explore questions that I don't know the answer to.	0.934	(.083)	1.715	(.096)	0.222	(.020)	0.425	(.021)
v2_In science, I figure out how things work.	0.745	(.063)	1.97	(.091)	0.177	(.015)	0.489	(.018)
v6_Even when I don't know the answer, I like to keep working in science.	1.398	(.078)	1.927	(.095)	0.333	(.019)	0.478	(.020)
v7_In science, talking about my ideas helps me learn.	1.477	(.089)	1.823	(.100)	0.351	(.021)	0.452	(.021)
v9_Doing investigations helps me figure out how things work.	1.228	(.078)	1.978	(.096)	0.292	(.019)	0.491	(.020)
v16_In science, I enjoy asking questions and wondering about things.	1.621	(.086)	2.075	(.101)	0.386	(.020)	0.515	(.021)
Ownership								
v3_The ideas I am learning in science are important to me.	0.949	(.057)	2.213	(.123)	0.300	(.018)	0.541	(.018)
v5_In class, I enjoy doing science.	1.578	(.064)	2.398	(.131)	0.499	(.020)	0.586	(.020)
v11_I wish we spent more time doing science.	1.742	(.078)	2.198	(.131)	0.551	(.025)	0.537	(.022)
v15_We can use science ideas to help our community.	0.748	(.056)	2.128	(.121)	0.237	(.018)	0.520	(.019)
v17_When doing science in school, I feel smart.	1.179	(.082)	2.261	(.129)	0.373	(.026)	0.552	(.020)
Collaboration								
v4_In science, I work with others to figure things out.	1.084	(.071)	2.074	(.113)	0.274	(.018)	0.498	(.020)
v8_In science, reading helps me learn.	1.165	(.082)	1.643	(.115)	0.295	(.021)	0.395	(.023)
v10_In science, I helped my class figure out how things work.	1.395	(.083)	1.474	(.109)	0.353	(.021)	0.354	(.022)
v12_In science, listening to others' ideas helps me learn.	1.455	(.079)	1.862	(.114)	0.368	(.020)	0.447	(.022)
v13_In science, I enjoy doing investigations with a partner.	1.204	(.082)	2.161	(.121)	0.305	(.021)	0.519	(.021)
v14_In science, I use ideas from my partner to solve problems.	1.434	(.088)	1.709	(.110)	0.363	(.022)	0.411	(.022)
v18_In science, I feel good when others use my ideas.	1.14	(.099)	1.286	(.115)	0.289	(.025)	0.309	(.025)
Latent Mean of Three Constructs								
Reflection	2.742	(.244)	1.371	(.078)				
Ownership	2.804	(.171)	1.131	(.064)				
Collaboration	2.463	(.161)	1.209	(.066)				

Note. Estimate λ is the factor loading for each item in the construct. All factor loadings and latent means are significant at the 0.001 level. The model goodness-of-fit index: RMSEA = 0.049, CFI = 0.92

Based on our SEL theoretical constructs in ML-PBL and the exploration of the ESEM, Table F.1 showed that the three-factor structure model generally fitted the data better than two-factor model or four-factor model in the overall sample and the sample separated by treatment conditions. The CFI was larger than .90 for the two-factor and three-factor solutions except for the four-factor solution in the treatment condition. The reduction of the CFI index indicated less fitting between factor solutions and the data structure. The root means square error of approximation (RMSEA) index also increased in the four-factor solution across samples, which suggested less fitting between the observed data and the four-factor model.

Table D.2. Exploratory Structure Equation Model

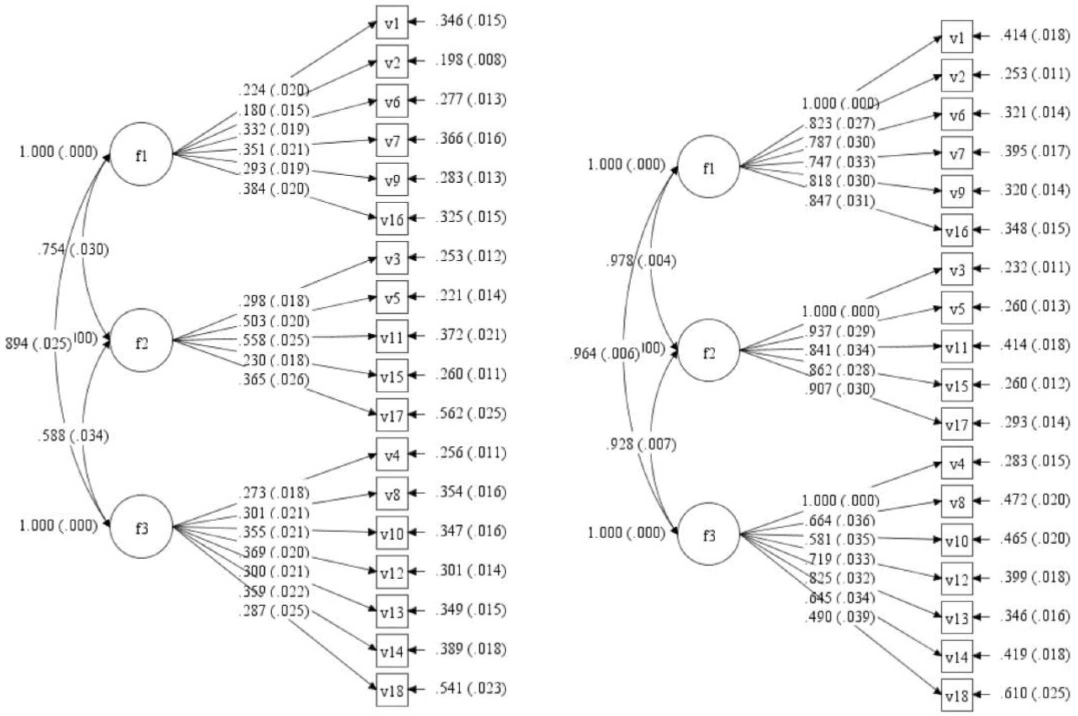
Date	Model	χ (df), p =	RMSEA	CFI	SRMR
Analytic data	ESEM 2 factor	451.430 (118), p< 0.001	0.034	0.94	0.024
	ESEM 3 factor	**297.970 (102), p<0.001**	**0.028**	**0.97**	**0.018**
	ESEM 4 factor	264.729 (87), P < 0.001	0.031	0.96	0.017
Treatment group	ESEM 2 factor	208.771 (118), p<0.001	0.028	0.95	0.025
	ESEM 3 factor	**181.123(102), p <0.001**	**0.025**	**0.97**	**0.022**
	ESEM 4 factor	868.436(87), p < 0.001	0.086	0.67	0.132
Control group	ESEM 2 factor	363.444(118), p < 0.001	0.041	0.94	0.028
	ESEM 3 factor	**248.605(102), p <0.001**	**0.023**	**0.98**	**0.021**
	ESEM 4 factor	165.609(87), p < 0.001	0.027	0.98	0.020

Table D.3. Item Factor Loading for the 18 SEL Items

Factor 1			Factor 2			Factor 3		
Estimate	SE		Estimate	SE		Estimate	SE	
0.289	(0.050)		-0.017	(0.026)	a	0.103	(0.043)	a
0.456	**(0.041)**		-0.011	(0.013)	a	-0.080	(0.043)	a
0.344	**(0.040)**		0.107	(0.031)		0.001	(0.023)	a
0.248	**(0.071)**		0.038	(0.030)	a	0.183	(0.063)	a
0.432	**(0.042)**		-0.005	(0.025)	a	-0.003	(0.038)	a
0.207	(0.062)		0.202	(0.034)		0.130	(0.047)	a
0.305	(0.060)		**0.266**	**(0.037)**		0.021	(0.043)	a
0.115	(0.063)	a	**0.529**	**(0.051)**		-0.013	(0.011)	a
-0.003	(0.010)	a	**0.605**	**(0.022)**		0.022	(0.039)	a
0.338	(0.062)		0.105	(0.039)	a	-0.076	(0.052)	a
0.318	(0.043)		**0.202**	**(0.034)**		0.013	(0.028)	a
0.290	(0.047)		-0.048	(0.023)	a	**0.170**	**(0.045)**	
0.195	(0.058)	a	0.058	(0.031)	a	0.165	(0.049)	a
0.128	(0.058)	a	0.053	(0.031)	a	**0.324**	**(0.046)**	
0.103	(0.079)	a	-0.015	(0.014)	a	**0.385**	**(0.071)**	
0.292	(0.050)		0.027	(0.027)	a	0.231	(0.050)	a
0.004	(0.034)	a	0.007	(0.026)	a	**0.431**	**(0.037)**	
-0.002	(0.035)	a	0.073	(0.032)	a	**0.289**	**(0.036)**	

Note. a indicate the factor loadings are not significant at p = <0.001 level.

To determine the model fit, we used multiple indices (Raykov & Marcoulides, 2000). The two-group CFA model fit was assessed using the chi-square approximation of the discrepancy function (χ^2), the CFI, the root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR) as guides in assessing fit. For the CFI, values above 0.90 generally indicate models with acceptable fit. For the RMSEA, a value below 0.08 usually indicates a reasonable fit, with a threshold of 0.05 reflecting a close fit to the data (Marsh et al., 2010). Hu and Bentler (1999) suggest the use of combining several goodness-of-fit measures to obtain a robust assessment of the model fit.



Treatment Student: CFA Structure

Control Student: CFA Structure

CFA Model goodness-of-fit: χ^2 (df) = 1163.50 (265), p < .000, RMSEA = .049, CFI = .92

Note. Plot presents the unstandardized path coefficients.

Figure D.2. Two-Group Unconstraint Three-Factor CFA

Table D.4. Bartlett Factor Scores and Eigenvalues on SEL Constructs

	Treatment			**Control**		
	Bartlett factor scores: Reflection	Bartlett factor scores: Ownership	Bartlett factor scores: Collaboration	Bartlett factor scores: Reflection	Bartlett factor scores: Ownership	Bartlett factor scores: Collaboration
Mean	0.191	0.101	0.171	-0.173	-0.135	-0.130
SD	0.835	0.873	0.896	1.124	1.119	1.052
Eigenvalues	2.252	2.362	2.473	3.121	3.169	2.876

Note. Bartlett scores were estimated by using maximum likelihood estimates, which produce estimates that are the most likely to represent the “true” factor scores.

Appendix E.

Table E.1. Estimated Treatment Effect Full Table

Model	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Treatment	0.262*	0.304**	0.271**	0.332***	0.356**	0.415***
	(0.114)	(0.109)	(0.091)	(0.093)	(0.105)	(0.108)
Proportion American Indian			11.11	10.0	6.18	6.41
			(6.57)	(6.52)	(3.37)	(3.73)
Proportion Asian			2.70	2.90	1.17	1.35
			(1.69)	(1.59)	(1.32)	(1.42)
Proportion Hispanic			2.3	2.32	1.12	1.05
			(1.58)	(1.46)	(1.32)	(1.41)
Proportion Black			2.21	2.30	0.958	0.992
			(1.57)	(1.46)	(1.34)	(1.42)
Proportion White			2.44	2.50	0.990	1.04
			(1.40)	(1.29)	(1.26)	(1.34)
Region 2			0.469*	0.456*	-0.015	-0.127
			(0.226)	(0.226)	(0.203)	(0.235)
Region 3			0.435*	0.401	0.058	-0.041
			(0.221)	(0.218)	(0.182)	(0.208)
Region 4			0.329	0.291	-0.273	-0.421
			(0.214)	(0.208)	(0.209)	(0.238)
Benchmark					0.014***	0.013***
					(0.001)	(0.001)
F&P					0.064	0.016
					(0.151)	(0.156)
NWEA					0.374**	0.385*
					(0.144)	(0.154)
Star					0.269	0.230
					(0.150)	(0.159)
i-Ready					-0.012	-0.038
					(0.164)	(0.186)
Constant	-0.433**	-0.463**	-2.98	-3.10*	-2.20	-2.18
	(0.165)	(0.164)	(1.53)	(1.42)	(1.32)	(1.42)
Random effects						
School-level	0.109	0.081	0.031	0.020	0.024	0.019
	(0.028)	(0.027)	(0.030)	(0.034)	(0.022)	(0.019)
Classroom-level	0.033	0.040	0.039	0.047	0.024	0.027
	(0.014)	(0.018)	(0.017)	(0.023)	(0.010)	(0.011)
Individual-level	0.850	0.848	0.850	0.848	0.741	0.744
	(0.032)	(0.033)	(0.032)	(0.033)	(0.027)	(0.029)
N	2,371	1,975	2,371	1,975	2,186	1,833