CHALLENGER CENTER FOR SPACE SCIENCE EDUCATION

ENGILEARN

DEPARTMENT OF EDUCATION INVESTING IN INNOVATION FUND

ENGILEARN IMPACT EVALUATION REPORT

JUNE 2018

Submitted by:
The Policy & Research Group
www.policyandresearch.com

8434 Oak St.
New Orleans, LA 70118

1631 15th Ave. W., Suite 113
Seattle, WA 98119

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The Challenger Center for Space Science Education (Challenger Center) implemented its EngiLearn program, after two years of development, during the 2016/2017 school year through the four-year *Investing in Innovation* (*i3*) grant from the *U.S. Department of Education* (ED). The goal of the i3 grant program is to test innovative educational practices to assess their impact on improving student achievement or student growth, closing achievement gaps, decreasing dropout rates, increasing high school graduation rates, and increasing college enrollment and completion rates. Challenger Center was awarded an i3 grant in 2013 to develop and implement a classroom-specific simulation experience appropriate to fifth-grade science objectives and produce professional development modules that ensure teachers understand the science concepts taught in their classrooms.

The i3 grant requires rigorous evaluation and in 2013, Challenger Center contracted with The Policy & Research Group (PRG) to conduct the evaluation of the EngiLearn program. In consultation with leaders from Challenger Center, PRG developed the *i3 Evaluation Design Summary*. The *Design Summary* specifies the research questions and methods to be used to answer those questions. Based on this plan, PRG has assessed the impact and implementation of the EngiLearn program. The purpose of the *EngiLearn Impact Evaluation Report* is to report and explain the results of the impact study and provide formative feedback that can be used to further develop the intervention.

The audience for this report is Challenger Center staff alone. The intent is to provide them with formative feedback that can be used to develop the intervention. To this end we not only summarize the results from the primary research questions, but also report on any exploratory results that are informative or revealing.

THE PROGRAM – ENGILEARN

EngiLearn consists of a hands-on, experiential science simulation for fifth-grade students and professional development training for fifth-grade science teachers. The program is an adaptation of Challenger Center's site-based simulation technology for use in classroom settings.

Challenger Center developed the *Aquatic Investigators Teacher's Guide*, which outlines the five-day curriculum content and structure of delivery for the EngiLearn program. Teachers lead students through two days of pre-mission educational activities to prepare for the ocean simulation activity on day three. The core of the intervention is a 2 ½ hour hands-on, computer-based ocean simulation, or "mission," that is led by teachers in the classroom. The simulation experience on the third day includes "embedded assessments," which collect data on students' progression through the mission, providing real-time feedback to teachers, who can then adjust the pacing to students' needs and offer support and assistance as necessary. Teachers then guide students through two days of mission reflection and educational activities on days four and five.

Challenger Center provided all teachers who instructed treatment classrooms with professional development training, which took place in the semester that EngiLearn was to be implemented in each district. Challenger Center trained intervention teachers on the STEM (science, technology, engineering, and math) content included in the intervention and the intervention software.

IMPACT STUDY

The impact evaluation aims to answer five primary research questions that are concerned with EngiLearn's effect on outcomes identified by the program's theory of change. In particular, it seeks to estimate EngiLearn's impact on students' academic (science achievement) and noncognitive (academic engagement, self-efficacy in science, academic intentions in STEM, and digital literacy) outcomes. In addition to these primary outcomes, we also investigate the impact of EngiLearn on math achievement and environmental awareness as a secondary study.

To investigate the effect of EngiLearn on these outcomes, we compare outcomes for students assigned to the EngiLearn and teaching-as-usual (TAU) condition. The study is a cluster randomized control trial (CRCT) in which the student is the unit of analysis and the classroom is the unit of assignment. This means that students were randomly assigned into either EngiLearn or TAU conditions by classroom, but impacts are estimated using individual-level data. We estimate the impact of the EngiLearn intervention within the intent-to-treat (ITT) framework, which means that the analysis aims to include all of the students initially enrolled into the study, regardless of their actual exposure to the intervention.

At all participating schools, classrooms were randomly assigned to participate in the EngiLearn intervention or receive the traditional ocean science curriculum (TAU). Therefore, the impact that we are assessing is the interactive, hands-on simulation experience (EngiLearn) as compared with the standard science curriculum delivered through some combination of traditional instruction, readings, and small group activities.

Ocean science knowledge and noncognitive outcomes are measured using the *Ocean Science Assessment*, which is an instrument developed by PRG and Challenger Center comprised of 50 closed-ended questions that assess student knowledge of ocean science concepts, academic self-efficacy and engagement, future STEM intentions, digital literacy, and environmental awareness. The ocean science test score represents the number of knowledge items (out of 26) that a student answered correctly on ocean science concepts. Noncognitive outcomes are constructed as individual scale scores by taking the average of an individual's ratings for all items in a scale.

Outcome data were collected directly from 2,546 fifth-grade students in 123 science classrooms using the *Ocean Science Assessment*. Administrative student characteristic and math achievement data were collected from each of the four participating school districts.[1] Data collection procedures were the same for students enrolled in both the treatment and the control groups. We use multilevel linear regression to estimate the impact of the EngiLearn program on all outcomes and control for the baseline measure of the outcome variables to increase the precision of our estimates.[2] A multilevel modeling approach is used to account for the nested structure of the data and to account for clustering effects.

KEY FINDINGS

PRIMARY STUDY

- Statistical estimates indicate that the EngiLearn program had no effect on students' ocean science achievement. Students who were offered the EngiLearn intervention had virtually

[1] PRG did not receive student characteristic data from Hanover County School District.

[2] We initially planned to include a series of individual-level demographic variables in our regression model to increase the precision of our estimates; however, Hanover County School District refused to provide PRG with demographic data for their students enrolled in the study and prohibited PRG from adding demographic questions to the *Ocean Science Assessment*. Therefore, we are unable to include these variables in our analytic model.

identical post-intervention scores on the ocean science posttest assessment as students who were offered the teaching-as-usual curriculum. Although both groups improved from pre- to posttest, students in the control group improved at the same rate as EngiLearn students, suggesting that the EngiLearn experience, as implemented, provided no additional academic gains compared with the TAU experience.

- The EngiLearn program also had no discernible effect on noncognitive outcomes. Results from the statistical models indicate that students who were offered EngiLearn had virtually identical post-intervention scores on the noncognitive assessments as students who were offered the teaching-as-usual curriculum. For each of the noncognitive scales the difference between the EngiLearn and TAU posttest scale scores was not meaningfully different.

- The average post-intervention *Self-Efficacy in Science scale* score is significantly lower for the EngiLearn group postintervention; however, we don't infer much from this. The estimated effect is so small in magnitude for the full sample of students that we have concluded that this is a statistical distinction without a meaningful difference, but rather an artifact of the large sample size and small variation in responses to the self-efficacy scale.

SECONDARY STUDY

- Model estimates indicate that the EngiLearn program had no effect on students' math achievement. Students who were offered the EngiLearn intervention had virtually identical scores on their fifth-grade math assessment as students who were offered the teaching-as-usual curriculum.

- The EngiLearn program also had no detectable effect on students' environmental awareness. Students in both groups had statistically indistinguishable post-intervention scores on the *Environmental Awareness scale*.

The EngiLearn intervention did not have the hypothesized impact on primary or secondary outcomes and there is little variation to these findings. Although null results are not ever desirable, they are inevitable in applied research and they can also be useful for program development. Null results can be especially constructive when the findings are used by the development team to adjust program components or revise the theory of change. To assist the team at Challenger Center in their future development of the EngiLearn program, PRG conducted a range of exploratory analyses to discover whether the program was more or less effective for some subgroups (e.g., male or female; high- or low-achieving students).[3]

Most of the exploratory analyses we conducted were statistically insignificant or substantively inconsequential. However, some findings that were significant pose interesting questions for future development:

- High-achieving students (i.e., students who scored higher on the ocean science knowledge test at baseline) who participated in the EngiLearn intervention scored significantly lower on their ocean science posttest assessment compared to their teaching-as-usual counterparts.

[3] It's worth pointing out that some of these analyses are not causal in interpretation because they don't involve a comparison of the randomly assigned sample (but rather an endogenous subgrouping). There is also a risk in running a multitude of statistical tests and drawing inferences from those tests. Probability sampling dictates that significant findings will appear by chance alone and the risk of a spurious finding increases with multiple tests on an analytic sample.

- Female students who participated in EngiLearn reported significantly lower self-efficacy in science on their posttest than the TAU females. We do not see a similar negative effect for male students. Looking further, this negative effect is magnified when we limit the sample to female, high-achieving students.

These findings are unforeseen, but useful and they suggest possible avenues for adjustments to the program.

- The feedback that PRG and Challenger Center received from the teachers who implemented EngiLearn indicated that, although the curriculum was rigorous, and students enjoyed the intervention experience, there was not enough time built into the curriculum to get through all the components. In addition, the intervention is designed to be provided to students within one week, which is a relatively short length of time. Similar hands-on, collaborative, and immersive STEM curriculums that have proven to increase student science achievement when compared with the standard science curriculum offered in schools are typically longer in length.

- Some research finds that high school female students do not report consistent gains in self-efficacy in science and math after applied STEM coursework compared to their male counterparts.[4] Although EngiLearn is not a technical education intervention, it is a hands-on, skills-building learning experience where students can apply the information they learn in class (abstract and theoretical) to real-world problem solving (applied learning).

- Group work may also be a factor. Although the literature on STEM and group work in elementary school settings is limited, research on college-age students suggests that females experience collaborative group work differently than their male counterparts, may be less likely to take on a leadership role in group projects, and may be less comfortable sharing their ideas in small groups.[5, 6]

The research literature is instructive but not definitive. EngiLearn is a group-centered and computer-mediated hands-on experience for fifth-grade students, not an applied course in high school technical education or a small group exercise in college. Challenger Center, nevertheless, has an opportunity as it develops EngiLearn make alterations. One advantage at this stage is that EngiLearn is modest in size and scope. Future developmental research should include qualitative feedback from students and teachers. This may help identify specific components that need attention.

[4] Sublett, C., & Plasman, J. S. (2017). How does applied STEM coursework relate to mathematics and science self-efficacy among high school students? Evidence from a national Sample. *Journal of Career and Technical Education*, *32*(1), 29–50.
[5] Eddy, S. L., Brownell, S. E., Thummaphan, P., Lan, M. C., & Wenderoth, M. P. (2015). Caution, student experience may vary: Social identities impact a student's experience in peer discussions. *CBE Life Sciences Education*, *14*(4), 1–17.
[6] Micari, M., & Drane, D. (2011). Intimidation in small learning groups: The roles of social-comparison concern, comfort, and individual characteristics in student academic outcomes. *Active Learning in Higher Education*, *12*(3), 175–187.

INTRODUCTION

The Challenger Center for Space Science Education (Challenger Center) implemented its EngiLearn program, after two years of development, during the 2016/2017 school year through the four-year *Investing in Innovation* (*i3*) grant from the *U.S. Department of Education* (ED). The goal of the i3 grant program is to test innovative educational practices to assess their impact on improving student achievement or student growth, closing achievement gaps, decreasing dropout rates, increasing high school graduation rates, and increasing college enrollment and completion rates. Challenger Center was awarded an i3 grant in 2013 to develop and implement a classroom-specific simulation experience appropriate to fifth-grade science objectives and produce professional development modules that ensure teachers understand the science concepts taught in their classrooms.

The i3 grant requires rigorous evaluation and in 2013, Challenger Center contracted with The Policy & Research Group (PRG) to conduct the evaluation of the EngiLearn program. In consultation with leaders from Challenger Center, PRG developed the *i3 Evaluation Design Summary*. The *Design Summary* specifies the research questions and methods to be used to answer those questions. The *Design Summary* was completed before the evaluation team examined any outcome data and it provides a detailed explication of how the evaluation team will assess program impact and implementation, including design specification, data collection protocols, instruments, variable definitions, outcome measures, and analytic methods.

Based on this plan, PRG has assessed the impact and implementation of the EngiLearn program. We have provided the ED with a summative formal report of these findings. This is a formal requirement of the i3 grant program. The ED collects evaluation results from all i3 grantees as part of the evidence review process to identify promising practices.

The purpose of the *EngiLearn Impact Evaluation Report* is to report and explain the results of the impact study in a more comprehensive and deliberate way.[7] The audience for this report is Challenger Center staff alone. The intent is to provide them with formative feedback that can be used to develop the intervention. To this end we not only summarize the results from the primary research questions, but also report on any exploratory results that are informative or revealing.

OVERVIEW

In conducting the impact evaluation of the EngiLearn program, PRG used questionnaire and test data collected from 2,546 fifth-grade study participants in 123 science classrooms, and administrative data collected from 30 elementary schools in 4 participating school districts. This report describes the intervention that was implemented, the research questions that guided the study design, definitions of the outcome measures examined, the procedures used to collect data, the methods used to measure and evaluate the effects of the EngiLearn program, the results of the evaluation, and a discussion of the findings and limitations of this study.

[7] In addition to this report and the findings presented to ED, PRG completed two additional reports as part of our evaluation. In the *EngiLearn Teacher Feedback Memo*, PRG provided Challenger Center with the qualitative and quantitative findings from a questionnaire administered to teachers who implemented the EngiLearn intervention to science classrooms, as well as two focus group discussions that expounded questionnaire responses. Additionally, PRG provided each of the four implementing school districts (Frederick County, Hanover County, Albemarle County, and Powhatan County) with a descriptive tabulation of results from the *Ocean Science Assessment* pre- and posttest administrations, which are the primary source of data collected for this evaluation. These reports contained no discussion about the EngiLearn program impact on student outcomes. These four reports were provided to each school district with the permission of Challenger Center.

Overall, we find that the EngiLearn program had no effect on primary student outcomes. Students assigned to the EngiLearn program did not demonstrate improved ocean science knowledge and noncognitive skills relative to students assigned to the standard Virginia curriculum. The results are consistent and have strong internal validity due to randomized assignment and large sample size.[8] Because results do not conform to the program's theory of change, we conduct additional, exploratory analyses to better understand what may be motivating these results. Exploratory analyses are not necessarily causal in interpretation and are improvised on the basis of available data, implementation results, and confirmatory empirical findings.[9] Exploratory analyses suggest that high-achieving students (i.e., students who scored higher on the ocean science knowledge test at baseline) who participated in EngiLearn scored lower on their ocean science posttest compared with their teaching-as-usual (TAU) counterparts, and that female students who participated in EngiLearn scored lower on self-efficacy in science than female TAU students. These findings, although surprising, are discussed to provide Challenger Center with information that might be helpful with the development of the program and its theory of change.

STRUCTURE OF THE REPORT

In this *Impact Evaluation Report*, we first present an overview of EngiLearn as it was intended to be implemented in four districts in Virginia during the 2016/2017 school year. We then itemize the research questions, outline the study design, eligibility criteria, assignment procedures, definitions of outcomes, and data collection methods, all of which were specified in the *i3 Evaluation Design Summary*. A description of the study sample and the baseline equivalence of the treatment and control samples, as well as a brief overview of our analytic methods follows. We report on the confirmatory results for all five primary research questions, as well as the two secondary research questions. The report concludes with an interpretive discussion of the primary findings along with additional exploratory analyses that examine potential mediating and moderating factors. The report concludes with a list of potential limitations to the methods and procedures employed in this study.

This report also contains appendices that provide additional details on information presented in the body of the report. Appendix A explicates the random assignment procedures and the analytic methods used to answer primary and secondary research questions. Appendix B provides the results of the baseline equivalence tests and benchmark statistical models for the primary and secondary research questions. Appendix C explains outcome variable operationalization. Appendix D summarizes data collection and data management procedures. Appendix E outlines the EngiLearn implementation study data collection and findings. Appendix F provides a graphical representation of the EngiLearn program's logic model.

PROGRAM DESCRIPTION – ENGILEARN

The EngiLearn program consists of a hands-on, experiential science simulation for fifth-grade students and professional development training for fifth-grade science teachers. The program is an adaptation of Challenger Center's site-based simulation technology for use in classroom settings.

[8] For a detailed discussion of the EngiLearn impact study design, see the Study Design section of this report.

[9] It's important to note that these exploratory findings, because they are improvised and are conducted on subgroups, may not necessarily provide the same causal inference as the primary results, nor do they provide a conclusive picture of the underlying dynamics of the EngiLearn program; much of that lies beyond the scope of the data. They do, however, provide some relevant empirical evidence that is meaningful for the continued development of the program.

The core of the intervention is a 2 ½ hour hands-on, computer-based ocean simulation, or "mission," that is led by teachers in the classroom over the course of one week (or five class periods). Students work in small teams to complete the mission, learning ocean science content, conducting experiments, and answering questions. The teacher provides instruction throughout the experience acting as "mission control." The simulation includes "embedded assessments," which collect data on students' progression through the mission, providing real-time feedback to teachers, who can then adjust the pacing to students' needs and offer support and assistance as necessary. The classroom simulation experience is guided by the *Aquatic Investigators Teacher's Guide*, which outlines the curriculum content and structure of delivery, and is supported by professional development provided by Challenger Center to all intervention teachers, along with teacher-led pre- and post-mission educational activities for students.

The professional development training took place during the semester that EngiLearn was to be implemented in each district; it was designed to train intervention teachers on the STEM content included in the intervention and the intervention software. The science simulation intervention activities took place during a single week in the 2016/2017 school year.

PROFESSIONAL DEVELOPMENT

Challenger Center delivered professional development to the teachers of the intervention classes at the beginning of the 2016/2017 school year. Challenger Center provided three online sessions that focused on general STEM content, training on use of the simulation software and embedded assessments, and instructions on implementing pre- and post-simulation activities. Each of the three online sessions were comprised of three modules. Participating teachers were encouraged to complete the online sessions within three weeks, and prior to their participation in a two-day in-person professional development training.

Participating teachers were expected to complete three online sessions and attend both in-person training days prior to implementing EngiLearn in their classrooms. Attendance data collected by Challenger Center and sent to PRG indicate that of the 55 teachers who taught EngiLearn to treatment classrooms, 52 (95%) completed both in-person trainings and all three online modules. Three teachers did not complete the online modules but did attend both in-person training days. In some cases, science teachers taught multiple classrooms within a school, some of which were assigned to the intervention experience and some assigned to the control experience. In these cases, teachers who received the professional development training provided by Challenger Center also taught control classrooms.

ENGILEARN CURRICULUM

Throughout intervention week, teachers guided fifth-grade students in the intervention classes through a series of activities outlined in the *Aquatic Investigators Teacher's Guide* curriculum. The curriculum outlined the core activities and learning objectives to be covered over the course of five days (two pre-simulation days, one mission simulation day, and two post-simulation days). Days 1 and 2 consisted of approximately 90 minutes' worth of activities, including an introduction to the intervention and the simulation, practice using the simulation software, and science content related to the simulation.

The objectives of Lesson One: The Carbon Cycle, are for students to be able to identify the parts of the carbon cycle and the human behaviors that have an effect on that cycle, as well as for students to identify carbon as an important building block of life on Earth. The content of the lesson involves an introduction to the EngiLearn software, a brief PowerPoint slideshow introducing the carbon cycle, a 30-

minute group activity and discussion further exploring the carbon cycle, and a jigsaw puzzle activity instructing students on the spheres of the Earth.

Lesson Two: The Water Cycle objectives are for students to identify the part of the water cycle, and identify the biosphere, lithosphere, hydrosphere, geosphere, atmosphere, and ocean as having unique characteristics that effect life on Earth. Students move through a series of activities including video and classroom discussion introducing the water cycle, a question and answer team-building activity, a 40-minute station rotation activity where students learn about the different spheres of Earth, and an individual EngiLearn software and postcard activity.

Lesson Three: The Mission is the day on which intervention students complete the 2 ½ hour interactive simulation experience in ocean content. The learning objectives of this lesson are for students to successfully navigate the EngiLearn software to complete their mission and identify and problem-solve issues in the ocean and environment to benefit the Hawaiian monk seal. Students are assigned to work in teams of three to six during the simulation. Teachers "fly" the mission, guiding students through the experience. The lesson activities include a pre-simulation reflection on the previous days' content, a mission introduction video, a 25-minute teacher-guided mission exploration activity where students can apply math, science, and engineering knowledge to mend a scientific sea base, a 30-minute mission continuation activity where students work in small groups to answer ocean science questions, and a 20-minute ocean debris tracking activity, followed by a brief end-of-day reflection activity.

Challenger Center provides teachers with embedded assessments during the simulation that allow them to access data during and after the mission. The assessments collect data on how much time students are spending on each question posed in the mission and what questions they are answering correctly and incorrectly. The assessments are designed to allow teachers to assess students' progression through the mission, including the concepts they are learning and those they are struggling to learn. Teachers may use the assessments in real time to guide the mission appropriately to the needs of their students.

The aim of Lesson Four: Engineering Solutions is for students to solve an engineering challenge by evaluation design, power needs, and strengths and weaknesses of a sea base, and then to create an updated sea base that can better withstand an earthquake. Students begin with a short reflection video on the previous day's mission experience, and then break into small groups to complete a 20-minute earthquake activity. Teachers bridge the earthquake activity with the previous day's mission experience before students begin a 20-minute power, water, and oxygen group activity. Finally, students complete a 30-minute base redesign activity within the EngiLearn software.

On the final day of the EngiLearn intervention, students complete Lesson Five: Communicating Knowledge, where the aim is for students to understand the importance of communicating their knowledge of the oceans through writing, art, and oral communication and are provided with an opportunity to practice these different forms of information sharing. Students complete a series of activities where they learn the value of writing research findings, practice writing a research article, share their findings through music, compiling a newsletter showcasing each student's articles, and creating and presenting a poster discussing what they have learned about the ocean and the Hawaiian monk seal population.

The aim of the teacher professional development and the simulation activities is to increase teachers' understanding of science concepts, engagement and self-efficacy in teaching science concepts, and understanding of simulation theory and competency in simulation tools (mediators), which in turn will

lead to increased science achievement, academic self-efficacy, intentions to pursue STEM learning, engagement, and digital literacy for participating students (short-term outcomes). In the long-term, the intervention is posited to increase students' math achievement and enrollment in STEM classes and careers. See Appendix F for a graphical representation of the program's logic model.

CONTROL EXPERIENCE – TEACHING AS USUAL

In contrast with those students offered EngiLearn, students assigned to the control group had the opportunity to experience ocean science instruction that they would normally expect to receive in fifth-grade classrooms in the four study districts in Virginia. This is a TAU control. The TAU experience would have naturally varied by teacher, school, and district – there were 4 school districts, 30 schools, and 83 teachers who were involved – but students assigned to the control condition would have been taught the same concepts within the same standards of learning as the EngiLearn group but by different methods. For example, students in the EngiLearn group experienced the hands-on, problem-solving, computer-simulated learning experience that is described above. Comparison students, meanwhile, would have been taught the same concepts through some combination of traditional instruction, readings, and small group activities. Additionally, TAU teachers would not have had the ability to use embedded assessments to assess student comprehension and participation levels in real time, as described above.[10]

Using a TAU contrast rather than a no-intervention control is useful because it contrasts the experimental intervention (in this case EngiLearn) with current practice. This means the outcomes that result from the TAU experience are not just any counterfactual – the expected outcomes in the absence of EngiLearn. They represent the expected outcomes for the current classroom experience. The impact estimate produced by this study, therefore, is best conceived as the average "improvement" in outcomes that are attributable to EngiLearn above and beyond what is being achieved by current practice on average. In other words, a null outcome does not represent no improvement in student outcomes, but rather no relative improvement over what is currently being achieved by TAU. A positive impact represents a desirable improvement over current practice and a negative impact represents an undesirable decline over current practice.

RESEARCH QUESTIONS

By design, the impact evaluation was developed to answer five primary research questions that are concerned with EngiLearn's effect on outcomes identified by the program's theory of change. As delineated by the questions below, our analysis seeks to estimate EngiLearn's impact on students' ocean science achievement, as well as their self-reported academic engagement, academic self-efficacy in science, STEM academic intentions, and digital literacy. In addition to these primary outcomes, we also investigate the impact of EngiLearn on math achievement and environmental awareness. The math and environmental awareness questions are classified as secondary because Challenger's theoretical model of change does not specifically hypothesize change for these two domains. Although they serve slightly different purposes, the estimates produced in response to the primary and secondary research questions are based on the same rigorous methods and therefore are both causal in interpretation.

[10] Teachers whose classrooms were randomly assigned to the EngiLearn condition were also provided with a multiday in-person and online professional development to enhance confidence in teaching the core ocean science learning concepts within the curriculum. Within many schools, however, there were teachers who taught multiple or all fifth-grade science classes, which meant that there was a substantial amount of crossover in the receipt of professional development. In all, 37 of the 65 control classes were taught by a teacher who received professional development. This means that professional development cannot properly be considered part of the treatment-control contrast.

PRIMARY

ACADEMIC OUTCOMES

1. What is the effect of EngiLearn on the science achievement of 5th grade students compared to the teaching-as-usual condition?

NONCOGNITIVE OUTCOMES

2. What is the effect of EngiLearn on academic engagement of 5th grade students compared to the teaching-as-usual condition?
3. What is the effect of EngiLearn on 5th grade students' academic self-efficacy in science compared to the teaching-as-usual condition?
4. What is the effect of EngiLearn on 5th grade students' academic intentions in STEM compared to the teaching-as-usual condition?
5. What is the effect of EngiLearn on 5th grade students' digital literacy compared to the teaching-as-usual condition?

SECONDARY[11]

1. What is the effect of EngiLearn on the math achievement of 5th grade students compared to the teaching-as-usual condition?
2. What is the effect of EngiLearn on the environmental awareness of 5th grade students compared to the teaching-as-usual condition?

STUDY DESIGN

This impact evaluation investigates the effect of EngiLearn on participating students' academic achievement and noncognitive skills. We do this by comparing outcomes for students assigned to the EngiLearn and TAU condition. The study is a cluster randomized control trial (CRCT) in which the student is the unit of analysis and the classroom is the unit of assignment. This means that students were randomly assigned into either EngiLearn or TAU conditions by classroom, but impacts are estimated at the individual level, using individual-level data.

Outcome data were collected directly from students using the *Ocean Science Assessment* whereas student characteristic and math achievement data were collected from each participating school district. Data collection procedures were the same for students enrolled in both the treatment and the control groups. We use multilevel linear regression to estimate the impact of the EngiLearn program on outcomes and control for the baseline measure of the outcome variables to increase the precision of our estimates.[12] A multilevel modeling approach is used to account for the nested structure of the data and to account for clustering effects.

INITIAL ELIGIBILITY CRITERIA

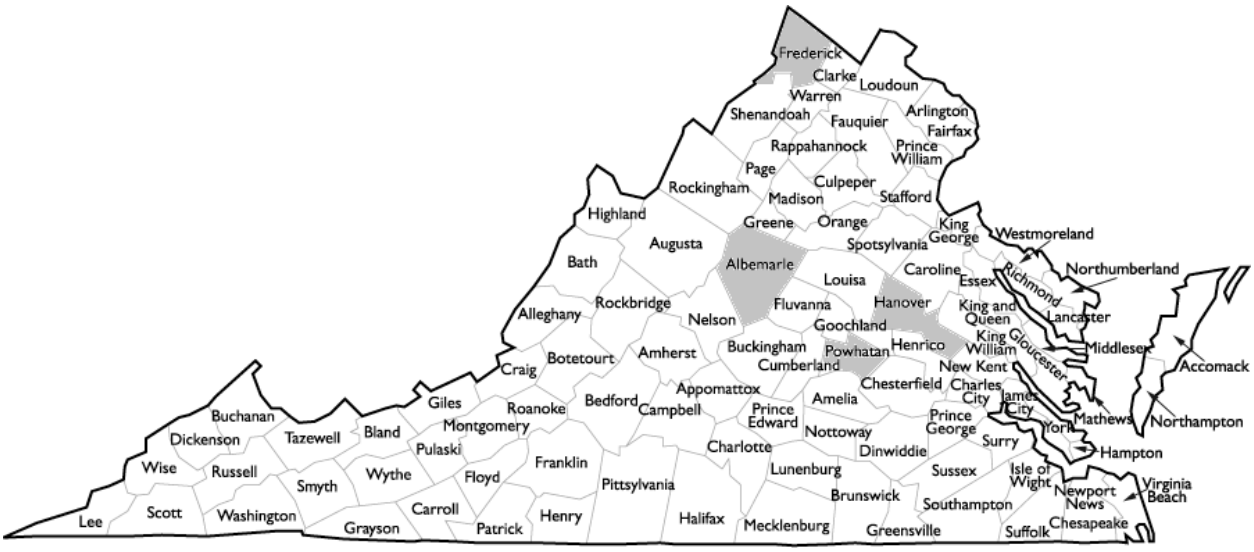
The study examined the effects of EngiLearn among fifth-grade students in selected schools in the state of Virginia. PRG and Challenger Center signed Memorandums of Understanding (MOUs) with four school districts: Albemarle County, Frederick County, Hanover County, and Powhatan County. Figure 1 presents

[11] PRG did not report the results of the secondary study outcomes to the i3 analysis and reporting team (i.e., Abt Associates).

[12] We initially planned to include a series of individual-level demographic variables in the regression model to increase the precision of the estimates, however, Hanover County School District refused to provide PRG with demographic data for their students enrolled in the study and prohibited PRG from adding demographic questions to the *Ocean Science Assessment*. Therefore, we are unable to include these variables in the analytic model.

a map of Virginia counties with the four implementation districts highlighted. To be eligible for inclusion in the study, districts were required to be located in the state of Virginia, include elementary schools, and formally agree to participate in the study. Challenger Center selected four school districts that met basic criteria and were proximate to Richmond, Virginia, where a Challenger Learning Center is located. In the fall of 2015, participating school districts contacted all schools with at least four fifth-grade classes in the 2015/2016 school year, inviting them to participate in the study.[13] To be included in the study, schools were required to teach the standard Virginia elementary school curriculum, school principals were required to sign an MOU agreeing to participate in the study and not provide similar enrichment programs during the same semester that EngiLearn is implemented, and obtain consent to participate in the study from science teachers of at least two eligible classes of fifth-grade students.

Figure 1. *Implementation School Districts in Virginia*



All students in the impact study were assigned to classes before randomization occurred, and therefore we do not include joiners in the analytic sample. To be eligible for inclusion in the study, classes were required to follow the Virginia fifth-grade science curriculum, take the Virginia Standards of Learning (SOL) test, be taught by teachers who are eligible and have consented to participate in the study, and have their classroom roster submitted to PRG at the end of the third week of class during the implementation semester. Finally, to be eligible to participate in the study, students were required to attend a study school at the time of study enrollment, be enrolled in a study class by the end of the third week of the implementation semester, and be eligible to take the unmodified Virginia SOL test.

ASSIGNMENT PROCEDURES

This impact study is a CRCT where the unit of assignment is the classroom, blocked by school. Class rosters were fixed before randomization occurred and any students who joined the class after randomization occurred were able to participate in classroom (treatment or control) activities but were excluded from the study sample. During the third week of the implementation semester and prior to

[13] If, after the school district contacted the school to invite them to participate in the study, the school decreased the number of fifth-grade classes to fewer than four, the school remained in the sample, provided it had at least two fifth-grade classes at the beginning of the 2016/2017 school year.

randomization, schools provided PRG with classroom lists that identified the number of students enrolled in each class and the science teacher and homeroom teacher assigned to each class.[14]

Table 1 presents the total number of classrooms and students that were randomized into the study in each district, overall and by treatment condition. A total of 123 classrooms within 30 schools were included in the study; 58 classes were assigned to the treatment condition and 65 were assigned to the control condition. A total of 2,546 students were randomized; 1,194 were assigned to the treatment condition and 1,352 were assigned to the control condition. On average, classrooms contained between 20 and 21 students, ranging from a minimum of 9 to a maximum of 33 students.

Table 1. Number of Classrooms and Students Randomized, by District and Treatment Assignment

	Classrooms			Students		
County	Number Randomized	Treatment	Control	Number Randomized	Treatment	Control
Frederick County	39	19	20	938	462	476
Hanover County	52	24	28	1,009	448	561
Albemarle County	17	7	10	272	112	160
Powhatan County	15	8	7	327	172	155
Total	**123**	**58**	**65**	**2,546**	**1,194**	**1,352**

OUTCOME MEASURES

The *Ocean Science Assessment* is an instrument developed by PRG and Challenger Center. It is comprised of 50 closed-ended questions that assess student knowledge of ocean science concepts, academic self-efficacy and engagement, future STEM intentions, digital literacy, and environmental awareness.

PRG worked with the Challenger Center and fifth-grade teachers to develop the ocean science test. Challenger Center identified eight "learning concepts" under three Virginia Fifth-Grade Science Standards of Learning that the EngiLearn intervention aimed to address. Item developers were recruited from the Richmond Challenger Center to create test items that would assess student ocean science knowledge. Noncognitive outcome scales that had some evidence of validity and reliability were identified in the educational literature. PRG then adapted the scales for the population of interest.[15, 16]

OCEAN SCIENCE TEST SCORE

The ocean science test score represents the number of knowledge items (out of 26) that a student answered correctly on ocean science concepts. Each multiple-choice item had four response options from which to choose. Students were instructed to choose only one option; blank or skipped responses are coded as incorrect answers. Scores for the ocean science cognitive score are only calculated if a student answered at least one question on the assessment; observations with no responses to the ocean science cognitive items were excluded from calculations.

[14] Additional details on classroom random assignment procedure can be found in Appendix A.
[15] Face validity of the ocean science test is established on the basis that the content met Virginia Fifth-Grade Science Standards of Learning requirements. PRG conducted cognitive testing with students from the target population to get feedback on clarity and content of test questions, as well as general layout of the assessment. PRG then revised assessment questions based on cognitive testing feedback and conducted field testing of the instrument with four classrooms of fifth graders in Frederick County School District to collect data on item performance. PRG finalized the *Ocean Science Assessment* prior to the first questionnaire administration in the fall semester of 2016.
[16] See Appendix C for additional details about the operationalization of the outcome measures.

NONCOGNITIVE OUTCOME MEASUREMENT SCALES

To measure the four noncognitive skills identified by Challenger Center as primary outcomes of the EngiLearn program, PRG selected scales that had been found to be reliable and valid in previous research. PRG adapted individual scale items based on feedback received by fifth-grade students during cognitive interviews and pilot testing. For the four noncognitive scales (self-efficacy in science, academic engagement, STEM academic intentions, and digital literacy), PRG constructs individual scale scores by taking the average of an individual's ratings for all items in a particular scale. Scale scores are only calculated if a student answered all of the scale items; observations missing one or more items are excluded from calculations.

DATA COLLECTION

OCEAN SCIENCE KNOWLEDGE AND NONCOGNITIVE OUTCOMES

The *Ocean Science Assessment* was administered to all participating fifth-grade students at schools in four counties in Virginia during the 2016/2017 school year.[17] Science teachers at each school administered the questionnaire to all program participants before and after implementation of the program. Teachers administered the questionnaires to each science class during their regularly scheduled class time. Teachers administered make-up questionnaires during the week following the initial administrations for any absent students. Students completed paper questionnaires, which were returned to PRG's office for data cleaning, entry, and analysis. See Appendix D for additional details on outcome data collection procedures.

STUDENT CHARACTERISTICS AND MATH ACHIEVEMENT DATA

PRG entered into a data sharing agreement with the district office of each of the four counties in which EngiLearn was implemented to receive math SOL scores and demographic data for all students participating in the study at the end of the fourth- and fifth-grade years. A comprehensive list of the variables requested from each district is shown in Appendix D. We requested and received demographic data and math SOL scores from Frederick, Albemarle, and Powhatan School Districts. Hanover County School District did not agree to send PRG any demographic data for participating students but did provide fourth- and fifth-grade math SOL scores.[18, 19]

SAMPLE DESCRIPTION AND BASELINE EQUIVALENCE

In this section, we first present a description of the demographic characteristics of study participants from three districts - Frederick, Albemarle, and Powhatan Counties. Hanover County School District did not provide any demographic data and prohibited PRG from including any demographic questions on the *Ocean Science Assessment*. Therefore, we are unable to provide a comprehensive description of the

[17] Participating Frederick County schools administered the questionnaires during the fall 2016 semester, while Hanover County, Albemarle County, and Powhatan County schools administered the questionnaire during the spring 2017 semester. The decision to implement EngiLearn in the fall or spring semester was determined by the semester in which districts taught the science concepts that are covered by the EngiLearn curriculum.

[18] Frederick, Albemarle, and Powhatan School Districts provided PRG with the test name and test score for the test the student took to satisfy the math SOL requirement in Grades 4 and 5. For example, some students may have taken a higher- or lower-level test than the grade they were in. Hanover School District only provided PRG with test scores for students who took the fourth-grade math test in their fourth-grade year, and who took the fifth-grade math test in their fifth-grade year. As a result, we are missing fifth-grade math test data for 257 (25%) students in Hanover County who took a different test than their peers in the fifth grade.

[19] The data sets provided by the districts are indexed with a unique student ID number, which is used to match student-level administrative data with the *Ocean Science Assessment* pre- and posttest data, as well as each student's intervention condition. Each observation includes the following demographic data: race, gender, date of birth, disability status, gifted and talented status, English language learner status, and free/reduced lunch status. We merged the individual-level data from the districts with the pre- and posttest questionnaire data using unique student ID numbers assigned by either the district or PRG.

baseline characteristics. Table 2 provides a description of the subsample of participants that were located in Frederick, Albemarle, and Powhatan Counties, which account for approximately 60% of the study sample. Average fourth-grade math achievement scores are available for students in all four districts. We report this statistic at the bottom of the table. Following the descriptive statistics of the combined sample, we present baseline balance statistics for treatment and control groups in the form of standardized differences.[20] Again, these balance statistics are only available for three districts for all but fourth-grade math SOL scores.

Table 2. Descriptive Characteristics of Participants

Characteristic	Number Reporting	Statistic
Age		
Mean age in years at baseline	1,510	10.7
Gender	(*n* = 1,510)	
Male	806	53.4%
Female	704	46.6%
Race	(*n* = 1,510)	
Black	89	5.9%
White	1,282	84.9%
Multiracial	75	5.0%
Other race[21]	64	4.2%
Disability status	(*n* = 1,514)	
Yes	203	13.4%
Free/reduced price lunch status	(*n* = 1,505)	
Yes	413	27.4%
English language learner status	(*n* = 1,512)	
Yes	55	3.6%
Gifted and talented status	(*n* = 1,514)	
Yes	57	3.8%
Math SOL score[22]		
Mean fourth-grade math	2,369	472.8

The pooled sample of randomized participants in Frederick, Albemarle, and Powhatan Counties include 1,537 students from 18 schools. Most students are reported to be White (85%), whereas a small number are reported to be Black/African American (6%). Just over one half of students are reported to be male (53%). On average, students were between 10 and 11 years old at the time they took their pretest assessment. Approximately one quarter (27%) of the students received free or reduced priced lunches, whereas a smaller proportion of students had some kind of disability (13%), received English language learner services (4%), or were considered gifted and talented (4%). PRG received fourth-grade math test

[20] Researchers are encouraged to assess baseline equivalence with standardized difference statistics rather than hypothesis tests, such as a t-test. Although there is no consensus on what value denotes balance, the *What Works Clearinghouse* specifies that differences less than or equal to 0.05 standard deviations requires no statistical adjustment to be considered equivalent. For differences between 0.05 and .025 standard deviations, an analysis must include an acceptable statistical adjustment for the baseline characteristic to meet equivalence standards. Differences above 0.25 standard deviations in value are considered to be nonequivalent.

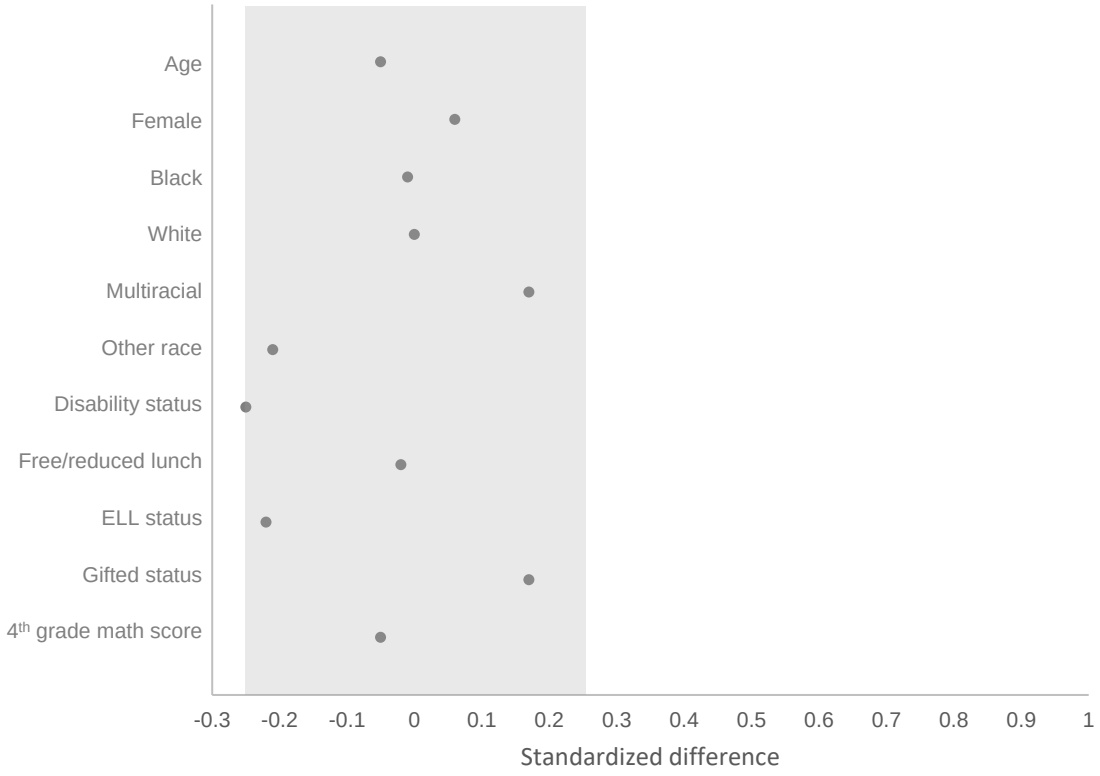
[21] *Other* includes participants who were reported to be *Asian, Hispanic/Latino*, or *Other.*

[22] Math SOL scores are on a scale from 0 to 600 points.

scores from all four school districts. Out of the 2,546 students who were randomized, we have fourth-grade math SOL scores for 2,369 (93%). On average, students scored 473 points (scale 0 to 600) on their fourth-grade SOL test, prior to entering their fifth-grade year.

Figure 2 presents the baseline balance diagnostics for the treatment and comparison groups in the form of standardized differences.[23] The gray shaded area within the figure indicates standardized differences that are equal to or less than 0.25 standard deviations, which represents a region of acceptable balance in education research.[24] Overall, the treatment and control groups appear to be balanced on characteristics for which we were able to obtain data. Standardized mean differences between the treatment and control groups are less than 0.25 for all but one characteristic, disability status, which has a standardized mean difference of –0.25. The proportion of participants in the control group who have a disability is about four percent higher than the proportion of participants in the treatment group with a disability. We include this information for descriptive purposes only. Because assignment was random and attrition from the study sample was extremely low, we can be confident that both groups are well balanced in observed as well as unobservable characteristics. The observed difference in disability status is attributable to random variation and is not likely indicative of a breakdown in randomization procedures.

Figure 2. Baseline Equivalence of Treatment and Control Groups



[23] Table B.1 in Appendix B provides the proportion of students in the treatment and control groups who meet each characteristic and the standardized difference for each.
[24] According to the *What Works Clearinghouse Standards Handbook Version 4.0* (2017). Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf May 1, 2018.

ANALYTIC SAMPLE

Table 3 describes the randomized and analytic sample for each of the five primary outcomes, including the overall attrition from the randomized sample, and the differential attrition between the treatment and control groups. Attrition occurs when outcome data are not available for all participants initially assigned to the intervention and comparison groups. A well-designed RCT may experience rates of attrition, or missing outcome data, that compromise the comparability of the intervention and comparison group and could lead to potentially biased estimates. Overall attrition refers the rate of missing data for the entire sample. Differential attrition represents the difference in missing data for the intervention and comparison groups. Each pose a different threat of bias of the estimated intervention effect.

Prior to implementation at each school (either during fall 2016, or spring 2017), 2,546 students in 123 classrooms from 30 schools were randomly assigned to participate in either the EngiLearn program or receive the standard science curriculum. The analytic sample varies slightly for each of the five outcomes, depending on item nonresponse on the *Ocean Science Assessment*. To be included in the noncognitive scale score calculations for a particular outcome, students must have responded to all of the scale items. The overall and differential attrition rates for all five outcomes are well below the cautious boundary for an acceptable threat of bias due to attrition, as outlined by the *What Works Clearinghouse.*[25]

Table 3. Randomized and Analytic Samples

Outcome	Number Randomized	Analytic Sample	Overall Attrition	Differential Attrition[26]
Ocean Science Test	2,546	2,484	2.4%	0.8%
Self-Efficacy in Science scale	2,546	2,433	4.4%	0.0%
STEM Intentions scale	2,546	2,324	8.7%	0.8%
Academic Engagement scale	2,546	2,349	7.7%	2.0%
Digital Literacy scale	2,546	2,424	4.8%	1.1%

ANALYTIC METHODS

In this section, we provide an overview of the analytic approaches for providing an empirical response to the research questions across the two outcome domains: academic achievement and noncognitive skills. The Model Specification section in Appendix A describes the analytic models used to generate the specific estimates of impact for the five primary and two secondary research questions.

The principal aim of this study is to determine whether offering the EngiLearn intervention to participants improves achievement in science, academic engagement, academic self-efficacy in science, STEM academic intentions, and digital literacy in comparison to instruction as usual. We do this within the intent-to-treat (ITT) framework, which means that the analysis aims to include all the students initially enrolled into the study, regardless of their actual exposure to the intervention. While this approach can seem obtuse because it fails to account for the variation in students' actual exposure to

[25] *What Works Clearinghouse Standards Handbook Version 4.0* (2017). Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf May 1, 2018.

[26] Differential attrition is the difference in attrition levels between the treatment and control groups. Higher levels of differential attrition indicate a potential threat of bias in the analytic sample because students in one group were more responsive to data collection than the other group.

the program, researchers adopt it because it provides an estimate of program impact that minimizes the potentially biased post-enrollment self-selection that motivates some people to attend more or less of the intervention. This estimate also has the added advantage of providing a more pragmatic estimate of the predicted impact of the program because it factors in the variation of exposure into the estimate, rather than controlling for it statistically or by design. We add to these confirmatory findings by conducting a number of exploratory analyses that are not necessarily causal in interpretation, but that help to provide additional contextual information about the primary study results.

PRIMARY RESEARCH QUESTIONS

OCEAN SCIENCE ACHIEVEMENT

To answer the first research question in this study, we construct an empirical model that estimates the effects of the intervention on ocean science test scores. We use a multilevel mixed effects model that estimates the treatment effect on the number of correct responses the student provided on his/her posttest. We regress the number of correct responses on the ocean science test on different predictors of interest (e.g., treatment or comparison group, the number of correct responses on the pretest) and a series of dummy variables that identify whether the student attended one of the 30 schools where the intervention is offered. We use a multilevel model as opposed to, say, an ordinary least squares (OLS) model because it is more appropriate for data that are collected at the individual level, but which are organized – and randomized – at a different level (e.g., classrooms).[27]

NONCOGNITIVE OUTCOMES

We estimate the effects of the intervention on four noncognitive outcomes (academic engagement, self-efficacy in science class, future STEM intentions, and digital literacy) with an empirical model that is similar to the one described above. In this case, we regress each noncognitive scale score on a pretest score, dummy school indicators, and the treatment indicator. The model is a three-level model where student-level data are nested within the classroom, and subsequently nested within the science teacher assigned to teach the classroom.

SECONDARY RESEARCH QUESTIONS

We estimate EngiLearn impacts on math achievement (Secondary RQ1) and environmental awareness (Secondary RQ2) using the same three-level model where student-level data are nested within the classroom, and classrooms are nested within the science teacher assigned to teach the classroom. The model regresses the outcomes on a highly predictive baseline measure, dummy school indicators, and the treatment indicator.[28]

RESULTS

PRIMARY RESEARCH QUESTIONS

1. OCEAN SCIENCE ACHIEVEMENT

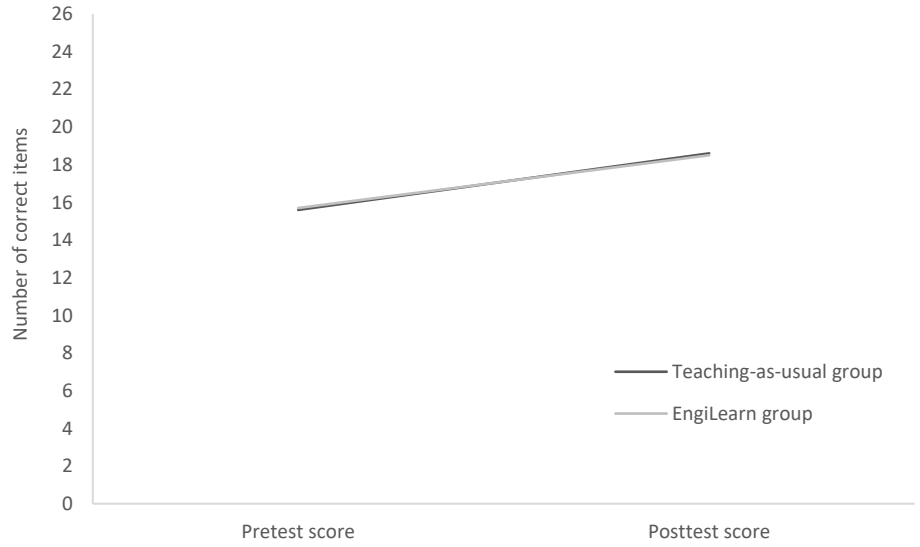
Statistical estimates indicate that the EngiLearn program had no effect on students' ocean science achievement. Students who were offered the EngiLearn intervention had virtually identical post-

[27] For this study, we use a multilevel model to adjust for the potential for the standard error to be too small, were we to use an OLS model. We use a three-level model where student-level data are nested within the classroom, which is then nested within the science teacher assigned to teach the classroom. We add the third level (science teacher) because one science teacher may have taught more than one classroom of students.

[28] For the assessment of the program's effect on math achievement, we use a student's fourth-grade math assessment score as our pretest measure.

intervention scores on the ocean science posttest assessment as students who were offered the teaching-as-usual curriculum. Figure 3 presents the unadjusted pre- and posttest ocean science test scores for both the EngiLearn group and the TAU group.[29] As illustrated by the graphic, the pretest, posttest, and improvement (slopes) of both sets of students is virtually identical.

Figure 3. Mean Number of Correct Responses on Ocean Science Pre- and Posttest, by Treatment Group



Although both groups improved over time, students in the control group improved at the same rate as EngiLearn students, suggesting that the EngiLearn experience, as implemented, provided no additional academic gains compared with the TAU experience. In other words, students in the EngiLearn group appeared to learn ocean science concepts, on average, just as well as those who were taught by conventional means.

2-5. NONCOGNITIVE OUTCOMES

Findings also suggest that the EngiLearn program had no effect on noncognitive outcomes. Results from the statistical models indicate that students who were offered the intervention had virtually identical post-intervention scores on the noncognitive assessments as students who were offered the teaching-as-usual curriculum. For each of the noncognitive scales, the difference between the EngiLearn and teaching-as-usual posttest scale scores was not meaningfully different.

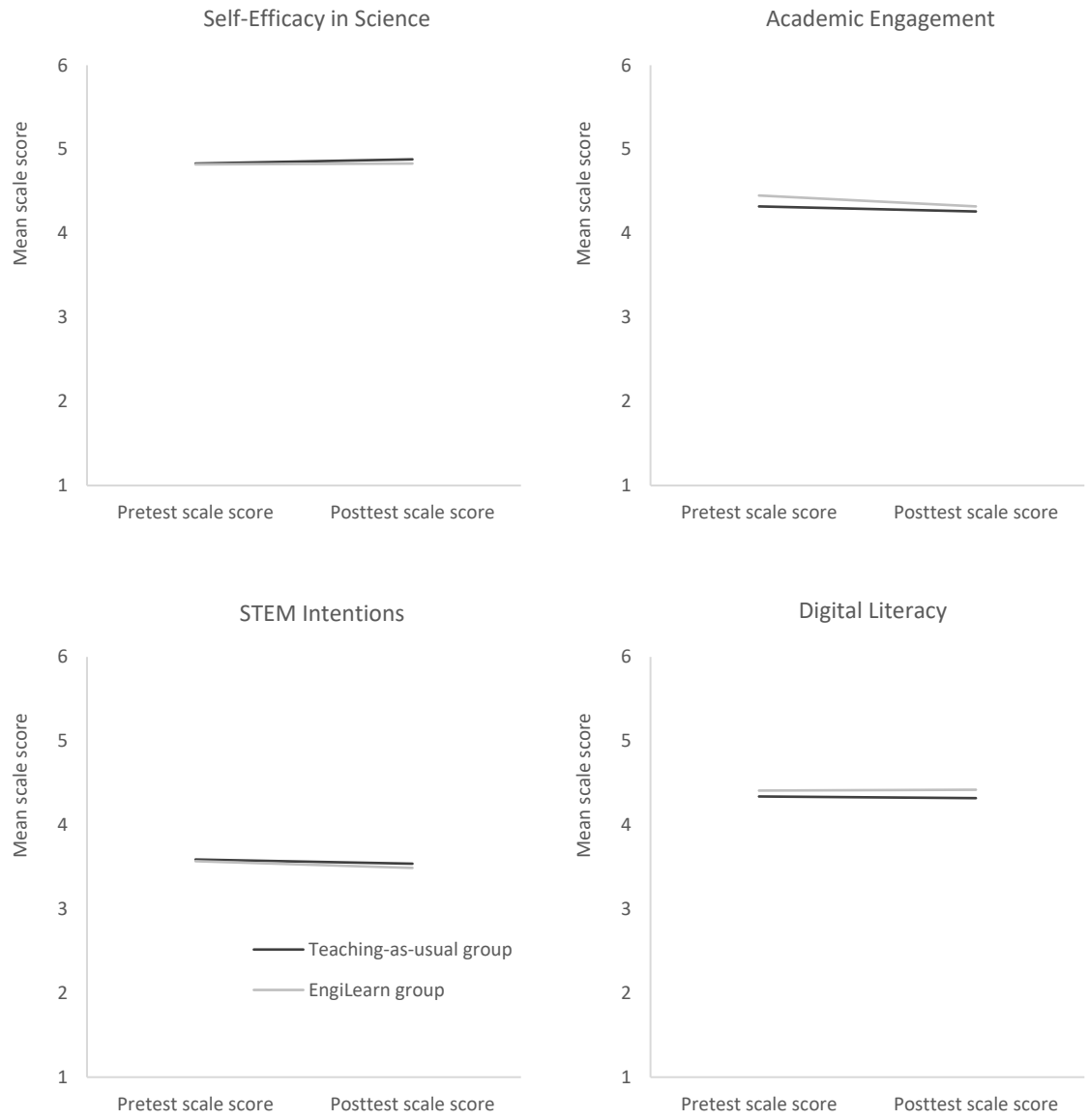
The average post-intervention *Self-Efficacy in Science scale* score is significantly lower for the EngiLearn group postintervention. We don't infer much from this; the estimated effect is so small in magnitude for the full sample of students that we have concluded that this is a statistical distinction without a difference, an artifact of the large sample size and small variation in responses to the self-efficacy scale.

Figure 4 presents the unadjusted mean scale scores at baseline and after the intervention for the EngiLearn and teaching-as-usual groups for all four noncognitive outcomes. Higher scale scores on these

[29] All of the graphical presentation of results in Figures 3 through 6 depict the unadjusted means for the treatment and control groups at pre- and posttest, and do not use the statistically adjusted means produced by the analytic models used to infer program impact. We do this for reasons of clarity for the reader in interpreting the benchmark analytic results. The unadjusted means are practically identical to the benchmark results provided in Appendix B and are more readily interpretable.

scales indicate more desirable outcomes for the noncognitive constructs. Overall, the graphics in the figure illustrate the statistical findings – little change over time and negligible differences between the two groups after the intervention.

Figure 4. *Mean Noncognitive Outcome Scale Scores, by Treatment Group*



SELF-EFFICACY IN SCIENCE

The first panel of the graphic (top left) shows that students in the EngiLearn and the teaching-as-usual groups report almost identical self-efficacy at pretest and again at posttest. Although TAU students

report slightly higher scores postintervention, both groups report relatively high and unchanging levels of self-efficacy (reported scores are closest to *agree* on the response scale).

ACADEMIC ENGAGEMENT

The second panel of the graphic (top right) shows a similar pattern: EngiLearn and TAU students report indistinguishable scores postintervention and neither changes meaningfully from baseline. Students in the EngiLearn group appear to regress slightly but the change is inconsequential. Students in both conditions at both time points report moderate agreement (*slightly agree* to *agree)* on the response scale.

STEM INTENTIONS

As is illustrated in the bottom left panel, students in both the EngiLearn and TAU groups report virtually identical STEM intentions at baseline and postintervention. Stem intentions are neutral at both time points (between *sort of agree* and *sort of disagree*), though both decrease slightly at posttest.

DIGITAL LITERACY

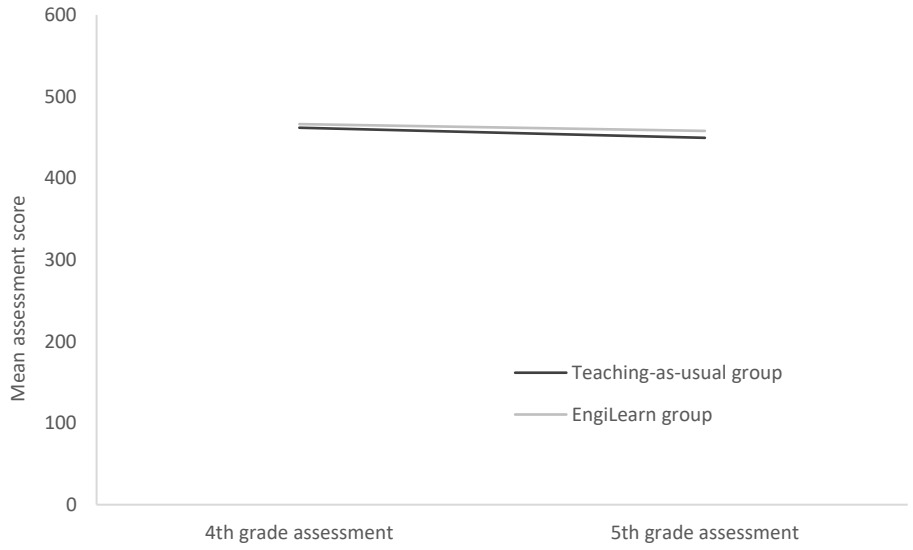
The bottom right panel shows that both the EngiLearn and TAU groups report reasonably high digital literacy (between *slightly agree* to *agree)* before and after the intervention. The EngiLearn group improved slightly more, but the difference at posttest is insignificant.

SECONDARY RESEARCH QUESTIONS

1. MATH ACHIEVEMENT

Model estimates indicate that the EngiLearn program had no effect on students' math achievement. Students who were offered the EngiLearn intervention had virtually identical scores on their fifth-grade math assessment as students who were offered the teaching-as-usual curriculum. Figure 5 presents the fourth- and fifth-grade math assessment scores for both the EngiLearn and teaching-as-usual groups. As illustrated by the graphic, students in the EngiLearn and teaching-as-usual groups had average scores of 458 and 449 (on a scale of 0 to 600), respectively, on their fifth-grade math assessment. This nine-point difference in scores was not statistically significantly different.

Figure 5. Mean Fourth- and Fifth-Grade Math Assessment Scores, by Treatment Group



2. ENVIRONMENTAL AWARENESS

Model estimates indicate that the EngiLearn program had no detectable effect on students' environmental awareness. Students in both groups had statistically indistinguishable post-intervention scores on the *Environmental Awareness scale*. Figure 6 illustrates that although both groups were similar at baseline (as would be expected) they also remain equivalent after the intervention.

Figure 6. Mean Environmental Awareness Scale Scores, by Treatment Group

The initial narrow partition of the lines illustrates the finding that both groups of students had similar, moderate, self-reported environmental awareness at baseline. The flat, parallel slope of the two lines demonstrates that this self-reported sense of environmental awareness did not change over the duration of the intervention.

DISCUSSION

The EngiLearn intervention did not have the hypothesized impact on primary or secondary outcomes. In terms of the mean effect on the full analytic samples, there is little variation to these findings. Although null results are not ever desirable, they are inevitable in applied research. They can also be very useful in the development of a program. Null findings can help a development team adjust the program or modify the theory of change. To assist the team at Challenger Center in their future development of the program, PRG conducted a range of exploratory analyses to determine whether the program was more or less effective for some subgroups (e.g., male or female, high- or low-achieving students).[30]

Most of the analyses we conducted were statistically insignificant or substantively meaningless and did not reveal anything new. Here we report on the findings that were significant and interesting. In no particular order, these findings were:

- High-achieving students at baseline (i.e., students who scored above the sample median on the baseline ocean science knowledge test) who participated in EngiLearn scored lower on their ocean science knowledge posttest compared with similar high-achieving students who received the TAU condition.
- Female students in Frederick, Albemarle, and Powhatan County School Districts who participated in EngiLearn scored lower on the *Self-Efficacy in Science scale* at posttest than female students who received the TAU condition.
- Looking further, the negative effect on self-efficacy is magnified when we constrict the analytic sample to female, high-achieving students.

We discuss each of these in turn, drawing upon literature and other qualitative data.

HIGH-ACHIEVING STUDENTS

When we constrict the analytic sample to just students who scored above the sample median on their ocean science pretest, we find that the students in the EngiLearn group scored significantly lower on their ocean science posttest assessment ($p = 0.05$). The mean difference between the EngiLearn and TAU students' scores is very small (EngiLearn students scored 0.5 points lower on their posttest compared with teaching-as-usual students). This difference of one half of a point on a 26-point scale is relatively inconsequential.

This finding, however, is surprising. Several explanations are worth considering. First, it is possible, based on the feedback received from teachers that allowing them more time to get through the curriculum may have enhanced student comprehension of the academic concepts and learning experience. The feedback that PRG and Challenger Center received from the teachers who implemented EngiLearn indicated that, although the curriculum was rigorous, and students enjoyed the intervention

[30] It's worth pointing out that some of these analyses are not causal in interpretation because they don't involve a comparison of the randomly assigned sample (but rather an endogenous subgrouping). There is also a risk in running a multitude of statistical tests and drawing inferences from those tests. Probability sampling dictates that significant findings will appear by chance alone and the risk of a spurious finding increases with multiple tests on an analytic sample.

experience, there was not enough time built into the curriculum to get through all the components.[31] In addition, the intervention is designed to be provided to students within one week, which is a relatively short length of time. Similar hands-on, collaborative, and immersive STEM curriculums that have proven to increase student science achievement when compared with the standard science curriculum offered in schools are typically longer in length. For instance, the *What Works Clearinghouse* identifies two similar technology-based curriculums that have shown to increase science achievement; one was delivered over two weeks and the other was delivered over 24 sessions.[32, 33]

While this may explain overall diminished impacts, it does little to clarify why the EngiLearn intervention appears to be disproportionately affecting high-performing students. In addition to science achievement, we see the same negative and statistically significant difference in self-efficacy in science for high-achieving students that we see in the benchmark results ($p = 0.02$). We do not see the same statistical difference among low-performing students (i.e., those who scored below the median at baseline). All else being equal, it seems reasonable to expect the high-performing students to better compensate (than the low-performing students) for diminished exposure to learning concepts. And in fact, this finding suggests that students who traditionally perform well under teaching-as-usual instruction are doing less well in the EngiLearn experience, however minimal the actual difference.

GENDER DISPARITIES IN STEM

The primary study findings indicate a potentially negative effect on students' self-efficacy in science, though the effect is not meaningful. Subgroup analyses indicate that the source of this negative effect is motivated by female students and primarily high-achieving female students. When we constrict the analytic sample to just females in the three districts where the data permit (i.e., Frederick, Albemarle, and Powhatan School Districts), we find that EngiLearn students reported significantly lower self-efficacy in science than the TAU group.[34, 35] We don't see a similar effect for males. Looking further, when we limit the analytic sample to high-achieving female students (students who score above the sample median on their ocean science assessment pretest), the effect becomes even more statistically significant ($p = 0.001$). This suggests that EngiLearn is *reducing* high-achieving female students' self-efficacy in science.[36] This finding is worth attending to for several reasons. First, the analytic sample of this subgroup analysis is smaller ($n = 347$), so a highly significant finding such as this is potentially a genuine effect. Second, the effect – if accurate – is precisely the opposite effect that would be desired. Social cognitive theory suggests that self-efficacy lays the groundwork for motivation and action and that without the belief that they can achieve a particular goal or perform well in a subject matter, students will lack the desire and self-motivation to attempt to do so.[37]

---

[31] A detailed overview of the feedback received from teachers who implemented EngiLearn can be found in PRG's *Teacher Feedback Memo*, which was distributed to Challenger Center in July 2017.

[32] Granger, E. M., Bevis, T. H., Saka, Y., & Southerland, S. A. (March 2010). Large scale, randomized cluster design study of the relative effectiveness of reform-based and traditional/verification curricula in supporting student science learning. Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, Philadelphia, PA.

[33] Zucker, A. A., Tinker, R., Staudt, C., Mansfield, A., & Metcalf, S. (2008). Learning science in grades 3–8 using probeware and computers: Findings from the TEEMSS II project. *Journal of Science Education and Technology*, *17*(1), 42–48.

[34] Subgroup analyses that constrict analytic sample to female or male students exclude students from the Hanover School District. We were unable to obtain any demographic characteristic data on students from this school district, and therefore cannot determine the gender of these students.

[35] Females in the EngiLearn group scored lower on the *Self-Efficacy in Science scale* ($p = 0.010$) compared with TAU females by approximately one tenth of a point.

[36] High-achieving females in the EngiLearn group scored lower on the *Self-Efficacy in Science scale* ($p = 0.001$) compared with TAU females by approximately two tenths of a point.

[37] Bandura, A. (1992). Social cognitive theory. In R. Vasta (Ed.), *Annals of child development. Vol. 6, Six theories of child development* (pp. 1–60). Greenwich, CT: JAI Press.

While it's beyond the scope of the available data to ascertain why female students felt less self-efficacious after participating in the EngiLearn intervention, previous research on gender disparities in the applied STEM field point us to a few possibilities. Previous research has suggested that high school female students do not report consistent gains in self-efficacy in science and math after applied STEM coursework compared with their male counterparts. In a study using data from the 2009 High School Longitudinal Study, Sublett and Plasman found that in general, earning credits in applied STEM career and technical education courses (e.g., hands-on, skills-based, real-world problem-solving learning) was predictive of increases in both math and science self-efficacy among high school students. [38] Their subgroup analyses indicated, however, that this was only true for male students, and that the correlation between applied STEM coursework and increased self-efficacy in science and math was not presented for female students.

EngiLearn is not a technical education program for high-school students; however, it is designed to be a hands-on, skills-building learning experience where students can apply the information they learn in class (abstract and theoretical) to real-world problem solving (applied learning). Many of the activities in the curriculum were designed to get students to collaborate in groups to complete different problem-solving tasks, rather than the teacher delivering the information through a lecture or more conventional format. The disparity in self efficacy observed in this evaluation and the study conducted by Sublett and Plasman suggests that applied exercises may not be equally efficacious for both boys and girls. Future developmental work on EngiLearn could explore this apparent disparity, verify whether it is an explanatory factor, and make necessary adjustments.

GENDER DISPARITIES IN GROUP WORK

Group work may also be a factor. Although the literature on STEM and group work in elementary school settings is limited, research on college-age students suggests that females experience collaborative group work differently than male counterparts. One survey of undergraduates found that female students were less likely to report taking on a leadership role in group projects, and more likely to take on roles that required less engagement (collaborator, listener, or recorder) within the group.[39] Male students, in other words, took the most engaging group roles, and therefore would theoretically get more out of the group exercises than females. Similarly, a second study found that, despite the overrepresentation of female students in the small group, peer-led activities, male students reported feeling more comfortable participating in the group and sharing their ideas compared with the female students.[40] They concluded that anxiety or reluctance to contribute to the group discussion may lead to female students losing out on opportunities for the high-level thinking that emerges from these discussions with peers.

The research literature is instructive but not definitive. EngiLearn is a group-centered and computer-mediated hands-on experience for fifth-grade students, not an applied course in high school technical education or a small group exercise in college. Challenger Center, nevertheless, has an opportunity as it continues to develop EngiLearn to explore these hypotheses and, if necessary, make adjustments in the theory or core components of change. One advantage at this stage is that the EngiLearn intervention is

[38] Sublett, C. & Plasman, J. S. (2017) How does applied STEM coursework relate to mathematics and science self-efficacy among high school students? Evidence from a national Sample. *Journal of Career and Technical Education*, *32*(1) 29-50.
[39] Eddy S. L., Brownell S. E., Thummaphan P., Lan M. C., and Wenderoth M. P. (2015). Caution, student experience may vary: social identities impact a student's experience in peer discussions, *CBE Life Sciences Education 14*(4), 1-17.
[40] Micari, M. & Drane, D. (2011). Intimidation in small learning groups: The roles of social-comparison concern, comfort, and individual characteristics in student academic outcomes. *Active Learning in Higher Education 12*(3), 175-187.

modest in size and scope. Future developmental research should include qualitative feedback from students and teachers as this may help identify specific components that need further attention.

CONCLUSIONS AND STUDY LIMITATIONS

Our primary impact study findings indicate that the EngiLearn program did not have any discernible impact on student outcomes, namely ocean science achievement, self-efficacy in science, academic engagement, STEM intentions, or digital literacy. The results indicate that the students who received the intervention experience were equally as knowledgeable, engaged, and confident in their science skills as the students who learned the same ocean science curriculum in the standard lecture format. There is some exploratory (e.g., not causal) evidence to suggest that female and male students experience the EngiLearn intervention differently, and that EngiLearn may impact students who have higher or lower science knowledge at baseline differently as well. These results provide Challenger Center with the opportunity to ask new questions about how they can make further improvements to the curriculum and student experience so that future iterations of the intervention may lead to positive student outcomes.

This study is not without its limitations. While PRG believes that the executed study design was the most rigorous one available, researchers who aim to measure intangible constructs such as knowledge, self-efficacy, engagement, educational aspirations, and computer literacy are tasked with the challenge of identifying the most effective measurement tool available to approximate those constructs. The scales that were used to measure the noncognitive outcomes were adapted from questionnaire scales that had been shown to be valid and reliable in previous research; however, it's possible there are alternative scales that may have more closely approximated the outcomes of interest. Alternatively, it's possible that there are mediating factors that the EngiLearn program may have impacted that are precursors to increased self-efficacy, engagement, educational aspirations, and computer literacy that we did not measure. Future research on the EngiLearn program may benefit from a more comprehensive focus on both short-term mediating factors as well as the more distal academic and noncognitive outcomes.

The aim of this study is to produce empirical, casual responses to the posed research questions, and is just one part of the comprehensive evaluation that PRG conducted on the intervention. The EngiLearn intervention was in its early phase of development at the time this evaluation took place. Thus, it is our hope that the combination of information gained from the EngiLearn impact study, the qualitative teacher feedback, and the EngiLearn implementation study provide Challenger Center with tools to ask the questions necessary to further develop the EngiLearn program and investigate how and why it might improve student outcomes in science learning.

APPENDIX A. METHODS

ASSIGNMENT PROCEDURES

This impact study is a cluster randomized control trial where the unit of assignment is the classroom, blocked by school. Class rosters were fixed before randomization occurred and any students who joined the class after randomization occurred were able to participate in classroom (treatment or control) activities but were excluded from the study sample. During the third week of the implementation semester and prior to randomization, schools provided PRG with classroom lists that identified the number of students enrolled in each class and the science teacher and homeroom teacher assigned to each class.

Within each school district, class lists at each school were arranged alphabetically by last name of science teacher, and, if a science teacher taught more than one classroom, we ordered classes for each science teacher by the last name of the homeroom teacher. For each school within the district, the first teacher on the list was assigned a random number from 0 to 1.0000 via a random-number generator. If the teacher was assigned a random number between 0 and 0.5000, his/her class was assigned to the treatment group. His/her class was assigned to the control group if the random number was between .5001 and 1.000. The rest of the classes in the school were assigned to treatment or control group based on the initial teacher's assignment. Control always followed treatment, and treatment always followed control. For example, if the first teacher was randomly assigned to the control group, then the second teacher on the list was assigned to the treatment group, the third teacher to the control group, the fourth teacher to the treatment group, and so on. This process was repeated for each school within the district. In other words, the first classroom in the school was randomly assigned to treatment or control, and the rest of the classrooms were assigned based on alternating assignment conditions. Assignment blocks for schools ranged in number from two to six classrooms.

In 4 school districts, a total of 123 classrooms within 30 schools were included in the study; 58 classes were assigned to the treatment condition and 65 were assigned to the control condition. A total of 2,546 students were randomized; 1,194 were assigned to the treatment condition and 1,352 were assigned to the control condition. On average, classrooms contained between 20 and 21 students, ranging from a minimum of 9 to a maximum of 33 students.

ANALYTIC METHODS

As detailed in the research questions, our impact study investigates whether offering the EngiLearn intervention to participants impacts their achievement in science, academic engagement, academic self-efficacy in science, STEM academic intentions, and digital literacy. We do this within the intent-to-treat (ITT) framework, which does not take into account participants' actual or measured exposure to the treatment itself, but, rather, the effect of the offer of the treatment (EngiLearn) relative to the offer of receiving the control condition (teaching as usual). This framework maintains the integrity of the experimental structure by including all participants who were randomized (except those who attrite) in the analytical sample, maintaining an exogenous assignment of participants to the experimental condition. Under this structure, we are able to produce an unbiased estimate of the treatment effect regardless of variation in program exposure. Bias can be insinuated, however, through self-selection if any participant who is randomized fails to provide posttest data (i.e., through unit or item nonresponse).

The analysis pools data across all schools where EngiLearn was offered and estimates effects using individual student-level data. The level of assignment and treatment is the classroom, and conditions

were randomly assigned, blocking by school. The level of inference is at the individual student level. We control for any variation that might occur at the individual, classroom, school, and school district level to improve the precision of our estimates.

This study examines whether the offer to participate in EngiLearn impacts the following outcomes: science achievement, academic engagement, STEM academic intentions, academic self-efficacy in science, and digital literacy. The following model is used for all contrasts. We estimate these impacts using a multilevel regression-estimated approach that models intervention effects while controlling for the baseline measure of the outcome variable. We use a model-based approach rather than a straight difference-of-means approach in order to increase the precision of our estimates and to correct for clustering. We estimate the empirical model with a multilevel model (using Stata).

We model all confirmatory outcomes identically, using the model below. Outcomes include science achievement, academic engagement, STEM academic intentions, academic self-efficacy in science, and digital literacy. For the sake of simplicity, we refer to a generic outcome score below.

MODEL SPECIFICATION

We use the same multilevel model to estimate the impact of EngiLearn on all outcomes:

Level 1: Students

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}(PreTest)_{ijk} + \varepsilon_{ijk}$$

Level 2: Classrooms

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}(TX)_{jk} + r_{0jk}$$
$$\pi_{1jk} = \beta_{10k}$$

Level 3: Teachers

$$\beta_{00k} = \gamma_{000} + \gamma_{001}(SchoolBlock)_k + \zeta_{00k}$$
$$\beta_{01k} = \gamma_{010}$$
$$\beta_{10k} = \gamma_{100}$$

Same Model in Mixed-Effects Format:

$$Y_{ijk} = \gamma_{000} + \gamma_{001}(SchoolBlock)_k + \gamma_{010}(TX)_{jk} + \gamma_{100}(PreTest)_{ijk} + \varepsilon_{ijk} + r_{0jk} + \zeta_{00k}$$

Where:

Y is the post-intervention score for the outcome for student *i* in class *j* with teacher *k*;

$SchoolBlock$ is a vector of *n* -1 school dummy variables (centered at grand mean for analysis);

$PreTest$ is the individual student's baseline score on the outcome measure (centered at grand mean for analysis);

TX is the treatment indicator variable;

γ_{000} is the intercept term, which estimates the mean outcome score for the comparison group;

γ_{010} is the estimated impact of the program – or the difference in post-intervention outcome scores between treatment and control students;

ε_{ijk} is the unmodeled variability of posttest scores at the individual student level;

r_{0jk} is the unmodeled variability of posttest scores between classrooms;

ζ_{00k} is the unmodeled variability of postest scores between teachers.

APPENDIX B. DETAILED ANALYTIC RESULTS

In this appendix, we present the detailed results of the baseline equivalence tests and the benchmark analytic models for both primary and secondary studies.

BASELINE EQUIVALENCE

Table B.1. Baseline Equivalence of Frederick, Albemarle, and Powhatan District Treatment and Control Groups

Characteristic	Treatment	Control	Standardized Mean Difference
Age	(*n* = 727)	(*n* = 783)	
Mean age in years at baseline	10.7	10.8	–0.05
Gender	(*n* = 727)	(*n* = 783)	
Female	47.9%	45.5%	0.06
Race	(*n* = 727)	(*n* = 783)	
Black	5.4%	6.4%	–0.01
White	86.2%	83.7%	0.00
Multiracial	5.4%	4.6%	0.17
Other race[41]	3.0%	5.4%	–0.21
Disability status	(*n* = 728)	(*n* = 786)	
Yes	11.1%	15.5%	–0.25
Free/reduced price lunch status	(*n* = 724)	(*n* = 781)	
Yes	25.4%	29.3%	–0.02
English language learner status	(*n* = 728)	(*n* = 784)	
Yes	3.0%	4.2%	0.22
Gifted and talented status	(*n* = 728)	(*n* = 786)	
Yes	4.4%	3.2%	0.17
Math SOL score[42]	(*n* = 1,103)	(*n* = 1,266)	
Fourth-Grade Math	472.15	473.34	–0.05

[41] *Other* includes participants who were reported to be *Asian, Hispanic/Latino*, or *Other*.
[42] Math SOL scores are on a scale of 0 to 600 points.

PRIMARY IMPACT STUDY

Table B.2. Multilevel Model Regression Results, Primary Study Outcomes

	Science Achievement		Self-Efficacy in Science		Academic Engagement		STEM Intentions		Digital Literacy	
Variable	***β***	**SE**	***β***	**SE**	***β***	**SE**	***β***	**SE**	***β***	**SE**
Treatment effect	-0.27	0.22	-0.06*	0.03	-0.04	0.03	-0.05	0.04	0.05	0.04
Number correct on pretest	0.67***	0.02	0.64***	0.02	0.76***	0.02	0.72***	0.02	0.74***	0.02
Indicator for imputed missing pretest score	0.19	1.21	-0.36***	0.11	-0.01	0.09	0.11	0.08	-0.34*	0.15
School block indicator[43]	-	-	-	-	-	-	-	-	-	-
Intercept	18.64***	0.13	4.89***	0.02	4.31***	0.02	3.54***	0.02	4.35***	0.02
Level 2: Classroom variance	0.59	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Level 3: Teacher variance	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Residual variance	10.38	0.46	0.36	0.02	0.43	0.08	0.55	0.02	0.86	1.81

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ~ $p < 0.10$

SECONDARY STUDY

Table B.3. Multilevel Model Regression Results, Secondary Study Outcomes

	Environmental Awareness		Math Achievement	
Variable	***β***	**SE**	***β***	**SE**
Treatment effect	0.01	0.04	3.44	3.05
Number correct on pretest	0.73***	0.02	0.74***	0.04
Indicator for imputed missing pretest score	0.08	0.14	-12.18~	6.87
School block indicator	-	-	-	-
Intercept	4.02***	0.02	452.14***	1.76
Level 2: Classroom variance	0.00	0.01	116.83	42.82
Level 3: Teacher variance	0.00	0.00	0.00	0.00
Residual variance	0.64	0.08	2,654.86	237.08

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ~ $p < 0.10$

[43] The benchmark analytic model included a series of dummy indicators to identify each school that participated in the study. We've omitted the detailed results of these dummy indicators in this table for the sake of parsimony as they do not provide readily interpretable information about particular schools.

APPENDIX C. VARIABLE OPERATIONALIZATION

All primary outcome data were obtained using the *Ocean Science Assessment* that study participants completed at the beginning and end of the fall 2016 or spring 2017 semester. In addition, fifth-grade standardized math test scores were obtained from the participating school districts in either spring or summer 2017.

OUTCOME VARIABLE OPERATIONALIZATION

OCEAN SCIENCE ACHIEVEMENT

Ocean science achievement is operationalized as the number of correct responses on the 26-item *Ocean Science Assessment* posttest. Scores range from 0 (no correct responses) to 26 (responded correctly to all test items).

Construction of this variable occurred in three stages. First, we create an indicator that counts the number of test times the student responded to on the posttest. If the student responded to at least one test item (count indicator >= 1), we then create an indicator variable for each of the 26 science test items, so that 0 indicates an incorrect response and 1 indicates the student selected the correct response to the test question. We code the response to a 0 if the student selected an incorrect response or he/she skipped the question altogether. Finally, we create an index of the number of correct responses by summing the 26 correct response indicators. Scores ranged from 0 to 26. If a student did not respond to any science test items on the posttest, he/she was excluded from the analytic sample for this outcome analysis.

SELF-EFFICACY IN SCIENCE [44]

Self-efficacy in science is operationalized as the mean scale score of the four items in the *Self-Efficacy in Science scale*. Students indicated how strongly they agreed or disagreed to four questionnaire items related to academic self-efficacy:

- I can get good grades in science class.
- I can always concentrate on my work during science class.
- I can remember facts I learn in science class.
- Even if the work in science class is hard, I can learn it.

Response options were on a 6-point scale that did not include a neutral option where 1 = *strongly disagree* and 6 = *strongly agree*. Mean scale scores were only constructed if the student responded to all four items in the questionnaire. If a student did not answer one or more items on the questionnaire, they were excluded from analysis.

Higher mean scale scores are indicative of greater levels of academic self-efficacy. A scale score of four or greater indicates that a student, on average, agrees that he/she has the ability to do well in science class. By contrast, mean scale scores that are less than four indicate that a student, on average, disagrees that he/she has the ability to do well in science class.

[44] Skinner, E. A., Chi, U., & The Learning-Gardens Educational Assessment Group. (2012). Intrinsic motivation and engagement as "active ingredients" in garden-based education: Examining models and measures derived from self-determination theory. *The Journal of Environmental Education*, 43(1), 16–36. Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (Vol. 5, pp. 307–337). Greenwich, CT: Information Age Publishing. Midgley, C., Maehr, M. L., Hruda, L. Z., Anderman, E., Anderman, L., Freeman, K. E., & Urdan, T. (2000). Manual for the patterns of adaptive learning scales. Ann Arbor, MI: University of Michigan.

STEM INTENTIONS [45]

STEM intentions is operationalized as the mean scale score of the six items in the *STEM Intentions scale*. Students indicated how strongly they agreed or disagreed to six questionnaire items related to STEM academic and career pathways:

- I plan to use science in my future career.
- I plan to use math in my future career.
- I plan to use technology in my future career.
- I want to be an engineer when I leave school.
- I want to be a computer programmer when I leave school.
- I want to be an environmental scientist when I leave school.

Response options were on a 6-point scale that did not include a neutral option where 1 = *strongly disagree* and 6 = *strongly agree*. Mean scale scores were only constructed if the student responded to all six items in the questionnaire. If a student did not answer one or more items on the questionnaire, they were excluded from analysis.

A scale score of four or greater indicates that a student, on average, agrees that he/she plans to use STEM in future careers. By contrast, mean scale scores that are less than four indicate that a student, on average, disagrees that he/she plans to use STEM in future careers.

ACADEMIC ENGAGEMENT [46]

Academic engagement is operationalized as the mean scale score of the five items in the *Academic Engagement scale*. Students indicated how strongly they agreed or disagreed to five questionnaire items related to academic engagement:

- I pay attention in class.
- I look forward to coming to school.
- I enjoy learning new things in school.
- I can't stand doing homework.*
- When we work on something in class, I feel bored.*

Response options were on a 6-point scale that did not include a neutral option where 1 = *strongly disagree* and 6 = *strongly agree*. Before calculating a mean scale score, responses to the two negatively worded questionnaire items (*I can't stand doing homework* and *When we work on something in class, I feel bored*), were first reverse coded so that higher scores for all questionnaire items indicated higher levels of academic engagement. For example, if a student indicated a score of 6 (strongly agree) for the statement *I can't stand doing homework*, we assigned a score of 1 to this question. If a student indicated a score of 5 on this item, we assigned a score of 2, and so on. Mean scale scores were only constructed if the student responded to all five items in the questionnaire. If a student did not answer one or more items on the questionnaire, they were excluded from analysis.

[45] Items were modeled after items in the Math/Science Intentions and Goals Scale, from Fouad, Smith, & Enochs (1997). Kier, M. W., Blanchard, M. R., Osborne, J. W., & Albert, J. L. (2014). The development of the STEM Career Interest Survey (STEM-CIS). *Research in Science Education, 44*(3), 461–481.

[46] Skinner, E. A., Kindermann, T. A., & Furrer, C. J. (2009). A motivational perspective on engagement and disaffection: Conceptualization and assessment of children's behavioral and emotional participation in academic activities in the classroom. *Educational and Psychological Measurement, 69*(3), 493-–525. Skinner, E. A., Chi, U., & The Learning-Gardens Educational Assessment Group. (2012). Intrinsic motivation and engagement as "active ingredients" in garden-based education: Examining models and measures derived from self-determination theory. *The Journal of Environmental Education, 43*(1), 16–36.

Higher mean scale scores are indicative of greater levels of academic engagement. A scale score of four or greater indicates that a student, on average, has greater levels of academic engagement. By contrast, mean scale scores that are less than four indicate that a student, on average, has lower levels of academic engagement.

DIGITAL LITERACY [47]

Digital literacy is operationalized as the mean scale score of the four items in the *Digital Literacy scale*. Students indicated how strongly they agreed or disagreed to four questionnaire items related to using technology, both in general and in the classroom:

- I learn better with computers.
- Computers make learning more interesting.
- Teachers should use computers more in teaching my classes.
- I can learn new computer skills easily.

Response options were on a 6-point scale that did not include a neutral option where 1 = *strongly disagree* and 6 = *strongly agree*. Mean scale scores were only constructed if the student responded to all four items in the questionnaire. If a student did not answer one or more items on the questionnaire, they were excluded from analysis.

Higher mean scale scores are indicative of more positive attitudes toward technology. A scale score of four or greater indicates that a student, on average, agrees that he/she has positive views on using technology in the classroom. By contrast, mean scale scores that are less than four indicate that a student, on average, disagrees that he/she has positive views on using technology in the classroom.

ENVIRONMENTAL AWARENESS [48]

Environmental awareness is operationalized as the mean scale score of the five items in the *Environmental Awareness scale*. Students indicated how strongly they agreed or disagreed to five questionnaire items related to reducing pollution and preserving the environment:

- To save water, I would be willing to use less water when I bathe.
- I would be willing to walk instead of ride in a car to reduce air pollution.
- I would give my own money to help the environment.
- I am interested in spending time working to help the environment, even though I realize this will cut into my free time.
- I talk to others about helping the environment.

Response options were on a 6-point scale that did not include a neutral option where 1 = *strongly disagree* and 6 = *strongly agree*. Mean scale scores were only constructed if the student responded to all five items in the questionnaire. If a student did not answer one or more items on the questionnaire, they were excluded from analysis.

---

[47] Ng, W. (2012). Can we teach digital natives digital literacy? *Computers & Education, 59*(3), 1065–1078.

[48] Erdogan, M., Ok, A., & Marcinkowski, T. J. (2012). Development and validation of children's responsible environmental behavior scale. *Environmental Education Research, 18*(4), 507–540. McBeth, B., Hungerford, H., Marcinkowski, T., Volk, T., Cifranick, K., Howell, J., & Meyers, R. (2011). National Environmental Literacy Assessment, phase two: Measuring the effectiveness of North American environmental education programs with respect to the parameters of environmental literacy. Final report. (Report to the U.S. Environmental Protection Agency, National Oceanic and Atmospheric Administration, and North American Association for Environmental Education). Thomson, G., Hoffman, J., & Staniforth, S. (2003). Measuring the success of environmental education programs. Ottawa: Canadian Parks and Wilderness Society and Sierra Club of Canada. Williams, H. (2011). Examining the effects of recycling education on the knowledge, attitudes, and behaviors of elementary school students. Outstanding Senior Seminar Papers. Paper 9. Retrieved September 7, 2017 from http://digitalcommons.iwu.edu/envstu_seminar/9.

Higher mean scale scores are indicative of more positive attitudes toward preserving the environment. A scale score of four or greater indicates that a student, on average, agrees that he/she would contribute to environmental sustainability. By contrast, mean scale scores that are less than four indicate that a student, on average, disagrees that he/she would contribute to environmental sustainability.

MATH ACHIEVEMENT

Math achievement is operationalized as the student's score on the Grade 5 Standards of Learning Mathematics Assessment, administered to most fifth-grade students in Virginia public schools at the end of the school year. Some students may take a different grade level assessment at the end of their fifth-grade year, depending on their academic standing (e.g., behind grade level or advanced). We requested end-of-grade standardized math test scores for all study students from the four participating school districts. Out of the 2,546 students randomized into the study, we obtained valid Grade 5 mathematics scores for 2,219 students, for a completion rate of 87%. We excluded students who took a different test other than the Grade 5 mathematics assessment at the end of their fifth-grade year (e.g., they took a Grade 7 math assessment, or a Grade 5 science assessment). Most students take the Grade 5 assessment using a computer-assisted testing mode; however, some of the study participants took the assessment on paper. In addition, some students took the plain English version of the Grade 5 SOL assessment, which contain adaptations on some test items in order to simplify the language and ensure test items are accessible to ELL students and to students with disabilities. The plain English versions of the tests have the same rigor, mathematical concepts, and item constructs as the regular SOL mathematics tests.[49] Math test scores range from 0 to 600.

COVARIATES

CLASSROOM IDENTIFIER

This variable is operationalized as a single categorical variable that groups students by the classroom they were enrolled in when they were enrolled in the study and randomly assigned to receive the EngiLearn or teaching-as-usual intervention. Students were assigned to 1 of 123 classroom groups to distinguish the 123 classrooms randomly allocated to EngiLearn or teaching-as-usual conditions. This variable was used as a level-two predictor in the multilevel model.

TEACHER IDENTIFIER

This variable is operationalized as a single categorical variable that groups individual students by the teacher who instructed their science class. Students were assigned to 1 of 84 science teacher groups to distinguish the number of teachers who taught the 123 classrooms of students.[50] In some cases, a science teacher taught a single classroom of students at a school, whereas in other cases, an individual teacher might have been the only science teacher at a school and therefore taught all sections of science classes at that school. This variable was used as a level-three predictor in the multilevel model.

[49] Information about the plain English assessment was retrieved from the Virginia Department of Education website: http://www.doe.virginia.gov/testing/sol/standards_docs/mathematics/plain_english_information.pdf March 1, 2018.

[50] Hanover School District did not provide PRG with a list of teacher names that could be directly linked to a student's intervention assignment, and instead provided a 4-digit numeric code to represent each classroom cluster. PRG examine the list of teachers at each study school within Hanover County, their intervention assignment, and classroom size to identify teachers who taught more than one science class at a particular school. PRG was able to decipher all but two teacher codes from the list provided by Hanover County. As a result, there are 84 teacher groupings used in our analytic model, though there were only 83 unique teachers who participated in the study.

APPENDIX D. DATA COLLECTION AND DATA MANAGEMENT

DATA COLLECTION

OCEAN SCIENCE ASSESSMENT

All primary outcome data were collected during the semester in which the EngiLearn intervention was delivered to students (i.e., fall 2016 or spring 2017). All *Ocean Science Assessment* data (pre- and posttests) were collected directly from students during the semester in which EngiLearn was delivered to their school. Student administrative data from the school districts were collected in the summer following the spring 2017 semester after students had taken their fifth-grade standardized math assessments. Table D.1 provides an overview of the *Ocean Science Assessment* data collection and EngiLearn intervention dates within the four school districts.

Table D.1. Ocean Science Assessment Data Collection Dates

School District	Pretest Administration Dates	EngiLearn Implementation Dates	Posttest Administration Dates
Frederick County	October 3–7, 2016	November 14–18, 2016	December 5–9, 2016
Hanover County	February 13–17, 2017	March 13–24, 2017[51]	March 27–31, 2017
Albemarle County	February 6–10, 2016	March 6–10, 2017[52]	March 20–24, 2017
Powhatan County	February 6–10, 2017	March 6–10, 2017	March 27–31, 2017

PRG mailed each school a package of questionnaires with prepopulated ID numbers for each classroom. Study ID numbers had previously been assigned to each student based on the classroom roster. Also included for each classroom were two questionnaires with blank ID numbers to be used in the event a questionnaire with a printed ID number was lost or damaged. Fifth-grade science teachers were trained by PRG to administer the questionnaires to classes that received the EngiLearn program and those that received the standard curriculum. Students completed the questionnaires using a paper instrument printed by Scantron. Teachers submitted completed questionnaires to a predetermined administrative contact at the school, who then shipped all questionnaires back to PRG.

STUDENT CHARACTERISTICS AND MATH ACHIEVEMENT DATA

PRG entered into a data-sharing agreement with the district office of each of the four counties in which EngiLearn was implemented to receive math SOL scores and demographic data for all students participating in the study at the end of the fourth- and fifth-grade years. We requested and received demographic data and math SOL scores from Frederick, Albemarle, and Powhatan School Districts. Hanover County School District did not agree to send PRG any demographic data for participating students but did provide fourth- and fifth-grade math SOL scores.

PRG requested the following data elements from Frederick, Albemarle, and Powhatan School Districts:

- Race/ethnicity
- ELL (English language learner) status
- Individualized education program/special education status
- Gifted and talented status

[51] Implementation at Beaverdam, Kersey Creek, Laurel Meadow, and South Anna Elementary Schools occurred during the week of March 13–17, 2017. Implementation at Cold Harbor, Cool Spring, Elmont, John Gandy, Mechanicsville, Pearson's Corner, Rural Point, and Washington-Henry Elementary Schools occurred during the week of March 20–24, 2017.

[52] Implementation at Hollymead Elementary School was delayed until March 13–17, 2017, due to a teacher's family emergency.

- Gender
- Free/reduced price lunch status
- Age
- School name
- Fifth-grade homeroom teacher full name
- Student grade
- Local student ID number
- Student-level mathematics SOL test scores
    - Fourth-grade test scores from spring 2016
    - Fifth-grade test scores from spring 2017

PRG requested the following data elements from Hanover School District:

- School name
- Fifth-grade homeroom teacher full name
- Student grade
- Unique study ID number
- Student-level mathematics SOL test scores
    - Fourth-grade test scores from spring 2016
    - Fifth-grade test scores from spring 2017[53]

DATA CLEANING PROCEDURES

To improve the validity and reliability of our estimates, prior to analysis, we followed several steps to prepare our data set and improve the quality of the data. After students completed the *Ocean Science Assessment* pre- and posttests, teachers returned the completed Scantron questionnaires to a predetermined administrative contact at the school after each administration. The contact packaged and shipped all questionnaires back to PRG using prepaid shipping labels. PRG reviewed the completed instruments and cleaned each questionnaire before submitting to Scantron for scanning. To clean each questionnaire, PRG staff ensured marked answers were dark enough, erased stray marks or notes near questions, and checked for clear patterns in responses. If a questionnaire was considered unreadable by either PRG or Scantron, it was entered into an electronic form by hand and checked by PRG staff. These steps were taken to improve the quality of data PRG received from Scantron and reduce the number of errors created by the scanning process.

Once all questionnaires from the two administrations were scanned, Scantron returned all paper questionnaires along with a district-specific electronic data set containing all scanned data. PRG conducted a 10% check of all data received from Scantron; 10% of all electronic entries were checked against the paper questionnaires to ensure there were no errors in the scanned data. If scanning errors were found in the 10% check of data, PRG's protocol required a 100% check of the batch of questionnaires in which the errors were found. This process was completed for both pre- and posttests in all four school districts to improve the validity and quality of our data.

[53] Frederick, Albemarle, and Powhatan School Districts provided PRG with the test name and test score for the test the student took to satisfy the math SOL requirement in Grades 4 and 5. For example, some students may have taken a higher- or lower-level test than the grade they were in. Hanover School District only provided PRG with test scores for students who took the fourth-grade math test in their fourth-grade year, and who took the fifth-grade math test in their fifth year. As a result, we are missing fifth-grade math test data for 257 (25%) of students in Hanover County who took a different test than their peers in the fifth grade.

MISSING DATA

We do not impute missing outcome data, as per *What Works Clearinghouse* evidence standards. Impact analysis samples include only those observations that have non-missing posttest data. Because the overall and differential attrition rates were minimal, missing covariate data, including missing pretest data, were handled according to the techniques outlined by the National Center for Education Evaluation.[54] With the assumption that data are missing at random (MAR), missing covariate data are treated using logical imputation according to guidance provided by Puma et al. (2009).

There were two reasons why a student may have been excluded from the analytic sample for a particular outcome. Table D.2 presents the number of students who were enrolled into the impact study and randomly assigned to either the intervention or teaching-as-usual group, the analytic sample size for each of the five primary outcomes, the number of students excluded from the analytic sample because they did not complete an *Ocean Science Assessment* posttest, and the number of students excluded from the analytic sample because they did not complete all of the items within a scale. Note, however, that for the ocean science test score, we re-coded missing responses to the ocean science questions as incorrect if the student answered at least 1 of the 26 items within that scale. We excluded one student who completed a posttest but did not answer any of the 26 ocean science knowledge questions.

Table D.2. Random Sample, Analytic Samples, and Exclusions

Outcome	Number Randomized	Analytic Sample	Missing Posttest	Missing Scale Items
Ocean science test	2,546	2,484	61	1
Self-efficacy in science scale	2,546	2,433	61	52
STEM intentions scale	2,546	2,324	61	161
Academic engagement scale	2,546	2,349	61	136
Digital literacy scale	2,546	2,424	61	61

Note: For the ocean science test, the value in the Missing Scale Items column identifies the number of students who completed an *Ocean Science Assessment* posttest but did not complete any ocean science knowledge questions.

[54] Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). *What to do when data are missing in group randomized controlled trials* (NCEE 2009-0049). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

APPENDIX E. IMPLEMENTATION STUDY

DATA COLLECTION

Challenger Center and participating teachers provided PRG with data related to the level of implementation that took place at each school. Data collected for the implementation study included: (1) treatment teacher professional development attendance provided to PRG by Challenger Center staff; and (2) *EngiLearn Fidelity Form* data completed by treatment teachers at the end of each lesson during implementation.

PROFESSIONAL DEVELOPMENT

Challenger Center collected treatment teacher professional development attendance data during the semester in which EngiLearn was being implemented in a particular school district. All treatment teachers were instructed by Challenger Center to complete a series of three online sessions focusing on general STEM content, training on use of the simulation software and embedded assessments, and instructions on implementing pre- and post-simulation activities. After completing the online sessions, teachers attended a two-day in-person training with Challenger Center staff prior to delivering the intervention. Challenger Center provided PRG with professional development attendance data in March 2017, after the spring implementation teachers had been trained on the intervention.

ENGILEARN FIDELITY FORM

During the week in which EngiLearn was delivered to treatment students, teachers completed the *EngiLearn Fidelity Form* to document whether they delivered each activity in the *Aquatic Investigators Teacher's Guide* curriculum completely, completed the activity but with changes, or did not complete the activity at all. All teacher fidelity forms were submitted to PRG with the *Ocean Science Assessment* posttests and entered into an electronic data set by PRG staff.

RESULTS

PROFESSIONAL DEVELOPMENT

Table E.1 presents the findings from the treatment teacher professional development attendance. Fifty-five science teachers taught a total of 58 treatment classrooms during fall 2016 and spring 2017 semesters. Any teacher who taught a classroom that was assigned to receive the EngiLearn intervention was required to complete the three online sessions to familiarize themselves with the EngiLearn simulation technology and science concepts, followed by a two-day in-person training provided by Challenger Center staff.

Table E.1. Treatment Teacher Professional Development Attendance (n = 55)

Professional Development Module	Attended Session	Did Not Attend Session
Online training		
Module A	52	3
Module B	52	3
Module C	52	3
In-person training		
Day 1	55	0
Day 2	55	0

Attendance data provided by the Challenger Center indicated that 52 of the 55 treatment teachers completed all 3 online modules prior to attending the in-person training. In addition, all 55 treatment teachers attended both in-person training days before delivering the intervention to their students.

ENGILEARN FIDELITY OF IMPLEMENTATION

A total of 58 classrooms were randomly assigned to receive the EngiLearn intervention. Teachers of each classroom were responsible for completing a *Fidelity Form* at the end of each day during the intervention week to document whether or not they completed each activity outlined in the curriculum. Table E.2 presents the results from the teacher self-reported fidelity forms.

Table E.2. Teacher-Reported Fidelity of Implementation (n = 58)

Curriculum Activity	Did Not Complete Activity	Completed Activity but With Changes	Completed Activity
Day 1			
Activity 1: Carbon Cycle	1	5	52
Activity 2: Earth's Sphere's Jigsaw	8	11	39
Day 2			
Activity 1: Water Cycle	3	3	52
Activity 2: Centers	1	4	53
Activity 3: Reflections	12	4	42
Day 3			
Phases 1–14 of mission simulation experience	1	2	55
Used embedded assessments during mission	8	N/A	50
Day 4			
Activity 1: Earthquake Activity	0	2	56
Activity 2: Power, Water, and Oxygen	3	6	49
Activity 3: Base Redesign	7	14	37
Day 5			
Activity 1: Writing	0	8	50
Activity 2: Publishing	4	9	45

Results from the fidelity forms indicate that on day one of intervention week, 52 classrooms completed the carbon cycle activity, but only 39 completed the Earth's spheres jigsaw puzzle activity. On day two, most classrooms completed the water cycle and centers activities (52 and 53 classrooms, respectively), whereas only 42 classrooms successfully completed the reflections activity. On day three, 55 classrooms completed the mission simulation experience, an additional 2 classrooms completed the mission experience, but with changes to the curriculum. The majority of classes (50) used embedded assessments during the mission simulation experience. Almost all classes (56) completed the earthquake activity on day four of intervention week. A smaller number completed the power, water, and oxygen activity and the base redesign activity (49 and 37, respectively). On the final day of intervention implementation, most classrooms completed the writing and publishing activities (50 and 45, respectively). A number of teachers indicated that they completed the activities, but with changes to the curriculum.

The implementation study findings corroborate the feedback that PRG and Challenger Center staff received from treatment teachers in the *EngiLearn Teacher Feedback Memo.* Teachers indicated in the qualitative interviews and on the *Fidelity Forms* that they tended to run out of time at the end of each day of implementation to complete all of the activities outlined in the *Aquatic Investigators Teacher's Guide*. The data in the above table reflect this feedback as more teachers reported they completed the first activity of each day, whereas fewer teachers reported being able to complete the last activity of each day.

APPENDIX F. LOGIC MODEL