Engineering is Elementary: An Evaluation of Year 6 Field Testing

Cathy P. Lachapelle
Christine M. Cunningham
Jennifer Jocz
Alison E. Kay
Preeya Phadnis
Stephen Sullivan

February, 2011

Engineering is Elementary
Museum of Science, 1 Science Park
Boston, MA 02114

Museum of Science®

This material is based upon work supported by the National Science Foundation under Grant No. 0454526. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Abstract

Results are presented from formative and summative evaluation of two *Engineering is Elementary* (EiE) units in development: *Designing Solar Ovens* and *Designing Parachutes*. These two units were field tested during the 2008-2009 school year, the sixth year of the EiE project, with results of field testing used to revise the units for publication. Teachers submitted pre- and post-assessments completed by students in field test and control classrooms. In field test classrooms, students participated in both a related science and an EiE unit, while in control classrooms, students only participated in related science instruction. Compared to control students, EiE students learned significantly more about engineering and technology. Where science knowledge was tested, EiE students also showed significantly more improvement than control students. In addition, in almost all cases, students from all demographic groups tested (gender, race/ethnicity, Individualized Education Program (IEP) status, and others) improved as much as other students on the assessments after participating in EiE.

1 Introduction

1.1 Engineering is Elementary: A Curriculum Development Project

Engineering is Elementary (EiE) is a research-based curriculum development project focused on creating engineering and technology curricular units to supplement core science instruction. Each EiE unit is designed to build on and apply science content through the design and development of a related technology (Cunningham & Hester, 2007). For example, in the *Designing Parachutes* unit, students design parachutes using the engineering design process (EDP): student designs are required to meet criteria for a mission travelling to a planet with an atmosphere thinner than Earth's—the parachutes students design must both pack tightly (because a parachute must fit on a spacecraft), and fall slowly (to prevent damage to a payload).

All EiE units have a four-lesson structure. In Lesson 1, students read a story involving a fictional child character who is faced with a problem that can be solved by creating a technology through the use of the engineering design process. The story serves to introduce the target field of engineering and also present a challenge similar to the one that students will undertake during the unit. Hands-on activities in Lesson 2 allow students to more broadly explore the unit's focal field of engineering. In the third lesson, students conduct controlled experiments which allow them to collect data and explore different materials, processes, or design elements. In the final lesson, using information gathered from experiments carried out in the Lesson 3, students use the engineering design process to plan, create, test, evaluate and improve their designs.

1.2 Theoretical Framework

The *Engineering is Elementary* curriculum is designed to (1) engage students and teachers in meaningful activity; (2) offer opportunities for both student and teachers to improve skills in and understanding of science, technology, engineering, and other areas; (3) provide scaffolding and support for new pedagogical styles and learning environments; and (4) build upon children's preexisting notions about the world. In addition, EiE units are (5) designed for the constraints that are typically faced by teachers working in elementary school classrooms.

Meaningful activity provides rich learning opportunities and motivates learning. A hundred years ago John Dewey (1938) advocated that schoolwork should be set in a meaningful learning environment that engages children in work that, while appropriate for children, nevertheless has its foundation in real adult work and makes interdisciplinary connections as well as connections to the child's lived experience. During the same time period, Piaget (Gruber & Vonèche, 1977) laid the foundations for our understanding of how children construct their own understanding of the world, working from concrete to abstract, while Vygotsky (1978) showed how children's skills and understanding grow through interaction with more competent others within the zone of proximal development.

Today, we have evidence that experience and practice help students learn both skills and how to apply concepts as they are immersed in rich contexts (Bruner, 2004; Lave & Wenger, 1991; Rogoff, 1990). Children's learning is more profound when they engage in realistic disciplinary practices (Sawyer, 2006), including the social and epistemic practices of a discipline, as students productively use key concepts (Duschl, 2008; Duschl & Grandy, 2008; Engle & Conant, 2002; Rosebery, Warren, & Conant, 1992). EiE therefore endeavors to construct a series of connections as the contextual basis of every unit: each EiE unit is set in a work and real-world context; children not only learn about the work of engineers in a particular field, but are also invited to "be engineers". Each EiE unit is interdisciplinary, with links to reading, social studies, science, and mathematics. In every unit, children are invited to use their experience with materials, inquiry, design, and their knowledge of the world around them to reflect upon and build their own understanding.

Because they incorporate activity and learning structures which are unusual in elementary school education in the United States, EiE lessons are heavily scaffolded for both teachers and students (Banilower, Smith, Weiss, & Pasley, 2006), and most students require significant scaffolding to participate successfully. A variety of activities are used, including large-group discussions, demonstrations, small-group hands-on activities, and individual brainstorming and reflection. Children are given the opportunity to share ideas and must, in fact, come to consensus about their group designs. Activities can easily be entered into collaboratively and cooperatively, and the roles within the group can be set by either the students or the teacher. Teacher tips on implementation are included in the EiE teacher guides, along with scripts for conducting guided reflection in group discussion and student handouts that structure activities as well as prediction and reflection. Teachers are encouraged to demonstrate and model good inquiry skills and proper experimental technique. Students discuss their ideas as they explore a new topic, share their findings, designs, and design improvements and are guided to reflect upon what they have learned, and to apply it. These pedagogical moves and learning structures—modeling, scaffolding, predicting, and reflecting—are important features of inquiry learning and project-based learning which evidence indicates are important for learning (Hmelo-Silver, Duncan, & Chinn, 2007; Kolodner et al., 2003).

Each EiE curricular unit is designed to build upon children's representations of and conceptions about natural phenomena and technology. In order to design an effective curriculum, it is important not only to know what notions children have, but also to more deeply understand why and how they hold those conceptions (Driver, 1989a, 1989b; Posner, Strike, Hewson, & Gertzog, 1982) since children's naïve conceptions provide a framework that they use to construct new understandings and to determine courses of action (Jones, 1997).

EiE units are design to mediate the conflicting demands often faced by teachers, such as the need to cover more content while working within a restricted time frame, or the obligation to meet standards while taking the time to engage students' interest, through connections and compromise, and above all, careful design to constraints. Any new curriculum designed for use in schools must take into account the wider political, social, cultural, and historical context of the school (Jones, 1997). Thus, the design of EiE curricular units takes into consideration the goals and requirements of state policymakers, school districts, and teachers. Additionally, the needs and realities of typical elementary school teachers and classrooms are also taken into consideration. The EiE Teacher Guides are created to be as easy to follow and as "teacher-friendly" as possible. Units are written to fit within a small time-frame of 6-8 lessons, and to take advantage of cross-disciplinary ties so that, for example, literacy, math or social studies time can be used to incorporate EiE and maximize limited class time. Furthermore, each unit is mapped to national standards and science content.

1.3 The EiE Research and Evaluation Program

EiE's research and evaluation program has four major goals: (1) to determine what students and teachers across the country know and believe about engineering, technology, and the engineering design process; (2) to improve the EiE curriculum by conducting classroom observations and incorporating feedback from teachers; (3) to develop assessment tools for that can be used by EiE, school districts, and others to evaluate the implementation of EiE; and (4) to evaluate the impact of the *Engineering is Elementary* curriculum on students' understanding of, attitudes toward, and interest in engineering, technology, and related science topics.

Each EiE unit undergoes a rigorous 2-3 year cycle of design, testing, and redesign. During the first year of EiE unit development (the "pilot" year), new units are developed and tested in classrooms across Massachusetts by EiE staff, with the help of pilot teachers. The EiE research and evaluation program has two tasks during the pilot year: developing assessments for the new units, and formative assessment of those same units. As part of the formative evaluation process, members of the EiE staff observe new units in classrooms as teachers implement the units for the first time. These observations, in combination with teacher and student feedback, are used to revise the units and assessments.

The second year of EiE unit development (the "field test" year) is devoted to testing the new units in a wider range of classrooms. In each of five states—during year 6 of the EiE project, these were California, Colorado, Florida, Massachusetts, and Minnesota—twelve teachers implement and test each revised unit in their classrooms. These field test teachers receive hands-on training in how to implement the units through attending professional development workshops in their home states along with materials they need to implement the unit in their classrooms. They collect pre- and post-assessments from their students as well as student demographics; these are sent to EiE where they are scanned and analyzed. Field test teachers also return information about their background, teaching practices, science teaching dates, and feedback on the EiE unit(s) they have taught. Results from the student pre-/post-assessments, as well as teacher feedback, are described in this report. Based on the results of field testing, each unit is revised again. If necessary, a third year of limited testing of revisions may also be carried out.

The EiE assessment development process has evolved and improved over time. For further details, see (Lachapelle & Cunningham, 2010).

Table 1. EiE Units Field Tested in Year 6

Year of EiE Project	EiE Unit Name	Design Challenge	Engineering Field	Science Applied
EiE Year 6 2008-2009	A Long Way Down	Designing Parachutes	Aerospace	Solar System; Atmospheres
	Now You're Cooking	Designing Solar Ovens	Green	Energy

In this paper, we describe findings from field test teacher feedback and EiE student assessments of science and engineering for two units: *A Long Way Down: Designing Parachutes* and *Now You're Cooking: Designing Solar Ovens*, field tested during the 2008-2009 academic year (Year 6 of EiE). These units are summarized in Table 1.

1.4 Research Questions

Questions to be addressed in this report include:

1. Does engaging in an engineering unit taught in combination with a science unit allow students to learn more science, engineering, and technology content as compared to when they are taught the science unit alone? If so, does teacher feedback provide some indications as to why?
2. Do students participating in EiE learn the science concepts that are typically taught in science curricula better than a control group learning science alone? As the evaluations reported here are not randomized controlled tests (RCT's), we cannot conclusively name EiE as the cause of any

differences we find. However, we are interested in determining if participation in EiE is a significant predictor of student gains as a promising indicator of the efficacy of EiE.

3. To what extent does participation in EiE predict student learning of science and engineering concepts – whether differentially or not – across student demographic groups? In this paper, we examine the following demographics: gender, students with and without Limited English Proficiency (LEP), students receiving or not receiving lunch from the National Free and Reduced-Price Lunch Program, students with or without an Individualized Education Program (IEP), and students from different racial groups.
4. How do teachers react after teaching an EiE unit for the first time? What effects do they see for their students, both academically and otherwise?

Although the goals of the EiE project are much larger than these four, time and available resources are limiting factors. In the future, we hope to more directly address questions about how engaging in EiE affects students' learning of content, their collaboration and communication skills, their ability to analyze and improve designs, and their technical skills such as engineering drawing and design of experiments.

2 Methodology

In this report, we detail the results of pre- and post-assessments of units in field test and control classrooms, as well as field test teacher feedback about their experiences implementing the EiE units. Field test classrooms are those where students were taught an EiE unit in combination with a related science unit, while control classrooms are those in which students were taught the related science, but not the EiE unit.

Results are to be considered indicative of the quality of the units reported here and of EiE as a curriculum, but not authoritative, for several reasons: (1) the units evaluated here were still in a preliminary form and were tested while still subject to extensive revisions—they have since been revised and improved; (2) the study is not a randomized controlled test—control and field test teachers were recruited at separate times; (3) the assessments for the units, as well as the assessment design process itself, were still under development at the time of testing—and the assessment versions used do not have high internal reliability; (4) some of the content measured in the assessments was taught in the field test, but not in control classrooms, leading to overalignment; (5) field tests suffered from substantial attrition.

2.1 Study Design

During the field test year for the EiE units, an attempt was made to recruit 12 field test teachers and 12 control teachers at each field site. EiE staff directly recruited Massachusetts field test and control teachers. Field test teachers in other areas were recruited and trained by local field site leaders who were EiE professional development providers local to each field test state. Professional development was provided for all field test teachers through a workshop at or near the beginning of the school year. Field test workshops introduced the EiE units that teachers would test. These workshops also addressed introductory topics in technology and engineering, including how to teach these concepts to students, as well as science and engineering background relevant to the unit. Teachers engaged in the hands-on activities from the unit, and discussed how these activities could be introduced to their students in an age-appropriate manner. Teachers were also provided with the materials needed for implementing an

EiE field test unit (teacher guide, a classroom pack of storybooks, and a materials kit), and were compensated with a small stipend after returning post-assessments and unit feedback.

In order to attract similar populations of teachers for both test and control groups, control teachers were recruited with the promise of receiving the same resources available to field test teachers, only after data collection. They were provided with admission to a local EiE workshop, EiE field test materials, and a stipend after returning post-assessments, just as field test teachers were. In some cases, teachers recruited and trained for field testing ended up collecting control data, for a variety of reasons – for example if their situation changed and they were unable to teach EiE, or they volunteered to collect control data instead of field test data (delaying their instruction of EiE) to even out the distribution of sample.

Unfortunately, control and field test teachers were neither recruited at the same time nor randomly assigned to a treatment group status. Because the original curriculum design grant did not include a significant evaluation component, and because grant money for control testing was received late in the evaluation cycle from secondary sources, control teachers were not recruited until after field testing had begun. Since this raises the possibility that there are unmeasured differences between these teachers or their students, our results must be considered indicative and not conclusive, and findings will be presented as preliminary.

Both field test and control teachers were recruited based on several criteria:

1. They must teach appropriate grade levels.
2. They must teach the science topic related to the EiE unit of interest, whether participating as a field test or control teacher.
3. They must teach in a formal classroom setting (not, for example, in an after-school program).
4. They must complete all surveys and assessments provided.
5. Field test teachers were also required to teach the EiE field test unit and provide feedback about their experience implementing the unit, while control teachers were required to refrain from teaching EiE until after the completion of data collection.

Grades 2-6 were recruited for both field test and control groups. Grades 3-6 completed the regular assessment, while a shorter, modified assessment was completed by students in grade 2. In addition, some grade 1 and kindergarten teachers participated in EiE field testing, but their students did not complete assessments. We did, however, collect feedback forms from kindergarten and grade 1 teachers. All grade 2-6 teachers were asked to collect pre- and post-assessments from their students. They were instructed to read the assessments aloud if possible, but not to provide any assistance with the assessments. Each student in a given classroom received the same assessments. Both field test and control students completed one general engineering assessment and/or an "engineering attitudes" survey in addition to a unit-specific assessment. The unit-specific assessments each included 9 or 10 science and engineering questions related to the engineering learning objectives of the unit and the target science concepts to be reinforced by the unit.

Whenever possible, students completed pre-assessments in October or November and post-assessments four to eight weeks after completion of the engineering and related science units (in the same academic year). However, due to varying circumstances for different teachers, the dates that pre-assessments were given to students vary. For example, some students are taught by science specialists who only see the students for a portion of the year. Other teachers first heard about the project and signed up as field test or control teachers in the spring of the academic year. No matter when teachers administered their assessments, they were instructed to administer pre-assessments to the students before the students were exposed to the engineering unit or any related science topic, and the post-assessments after they completed instruction in the science topic and (if a field test classroom) engineering lessons (Figure 1).

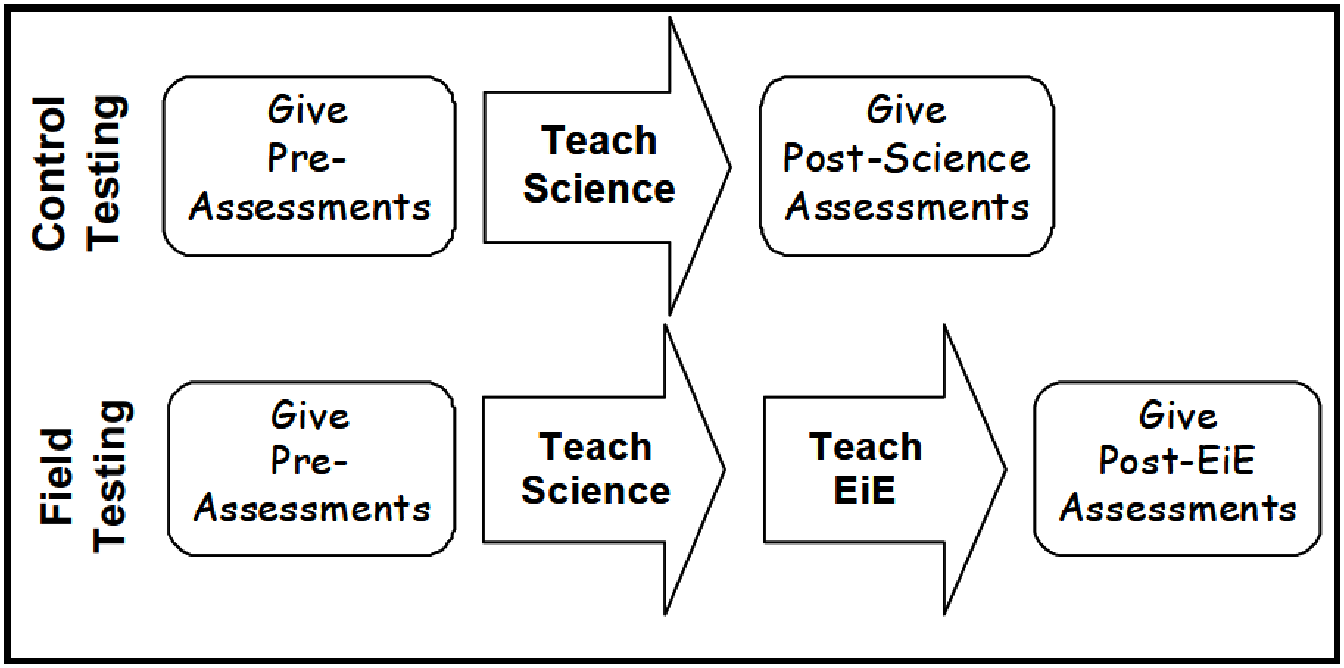


Figure 1. Timing of Assessments in Control and Field Classrooms

After completing instruction of the EiE unit they were testing, field test teachers filled out feedback forms. These feedback forms asked specific questions about implementation—time spent preparing for and teaching each lesson, which student handouts were used, and modifications made by the teacher—as well as suggestions for improvement of individual lessons, and ratings of the unit as a whole. Teacher responses were reviewed, compiled, and used to inform unit revisions.

2.2 *Assessment Instrument Development*

2.2.1 Development of Teacher Feedback Forms

Teacher Feedback Forms were developed in the third year of the curriculum development project and revised yearly to address the newest units being field tested. Teachers were first asked to rate and comment on aspects of the unit overall. Specifically, teachers were asked to discuss what they felt their students learned by taking part in the unit, to report the amount of time they spent prepping for and teaching each lesson, in both minutes and class periods—data that we used in our HLM models of student outcomes, and to explain if and how they deviated from the instructions given in the teacher guide and why. Finally, we asked teachers to rate and discuss each individual lesson in the unit. Feedback Forms for each unit were customized with specific questions about aspects of the lessons the curriculum writers were most concerned about.

2.2.2 Development of Student Assessments

Field test instruments for assessing students were developed using a two-step process. During the pilot (first) year of development for each unit, EiE staff devised a number of questions pertaining to the learning objectives of each unit, as well as to the related science content. Potential assessment items were also collected, when available, from standardized assessments such as the NAEP and state science assessments. The questions were cross-referenced with the unit's learning objectives, reviewed for correctness, and checked for ambiguity.

Each pilot assessment was distributed to approximately 100 pilot classroom students (students in classrooms using pilot versions of the EiE units) in grades 3-5 for testing. Based upon students' responses, questions that were too easy (greater than 90-95% correct) were removed. Other questions were removed at a later date if changes made to the EiE unit after pilot testing resulted in those questions being no longer relevant.

Validity evidence was gathered in the fall of 2008 for all questions from both assessments discussed in this report. EiE staff conducted cognitive interviews with children in grades 2-5 from Massachusetts classrooms, asking students to read each question aloud, explain what they thought the answer would be, read all answer choices aloud, and explain which answer they would choose and why. Based on the question read-alouds, minor changes were made to several diagrams as well as to the wording of some questions in order to improve clarity for later use of the assessments; however since data from the field test year had already been collected, the results reported here are based on the earlier, unimproved version.

Based upon our experiences in constructing assessments, EiE made significant changes to the assessment development process for the 2009-2010 school year. However, these changes do not apply to the evaluations reported here. For a complete description of the new process, see Lachapelle & Cunningham (2010).

2.3 Data Collection and Initial Cleaning

2.3.1 Data Collection of Teacher Feedback Forms

A Teacher Feedback Form was collected from each field test teacher after she/he completed the assigned EiE unit. Feedback Forms were then entered by hand into Microsoft Access® (Microsoft Corporation) for initial coding. Researchers ran queries to check for duplicate forms and missing entries, which were then checked against the originals. When duplicates were found, the most complete form was retained.

2.3.2 Data Collection of Student Assessments

Pre-assessments were mailed directly to participating teachers near the beginning of the school year. For a small subset of teachers who registered to field test mid-year, pre-assessments were sent immediately after registration. In a given classroom, each student received the same assessments. Each student who participated in the field test program as a field test student or a control student completed one general engineering assessment and/or an "engineering attitudes" survey. The results of the general engineering assessment and the engineering attitudes survey are not included in this paper; the results of the engineering attitudes survey are reported elsewhere (Cunningham & Lachapelle, 2010). In addition, each student completed a unit assessment that was related to the related science topic studied, and for field test students, to the EiE unit studied. The unit specific assessments included 9 or10 science and engineering questions related to the unit.

The timing of instruction and assessments in control and field test classrooms is shown in Figure 1 above. EiE students were tested at least twice—once before beginning the science curriculum and/or related *Engineering is Elementary* unit, and once after instruction was completed—allowing for a test-retest analysis.

Once all completed assessments were collected, they were digitized using an OMR scanner and then imported into a Microsoft Access database. This data was then exported to PASW Statistics version 17.0, together with student demographic data, for initial analysis. Student responses to each question were recoded as correct (1) or incorrect (0), and a preliminary scale item (AllScore), which was calculated as the sum of all questions (equivalent to the sum of all correct questions), was also computed for each assessment. Data collected from all classrooms were checked for any transcription errors and unusual or missing data by inspecting the means from the AllScore variable for the pre-assessment and post-assessment separately. Data entry and transcription errors were corrected before continuing.

2.4 Reliability Analysis and Scale Construction for Student Assessments

Reliability analysis for each unit was conducted on the responses of students who had returned both a pre- and a post-assessment before any further cleaning had taken place. The mean of all corrected items for a given unit assessment was calculated to create a score, the AllScore, as described above. Additionally, for each assessment with at least five science questions, a science sub-score (ScienceScore) was created in the same manner.

Initial scores were checked for internal reliability and factorability by means of reliability analysis and principal component analysis with direct oblimin rotation in PASW v17.0, on the pre- and the post-scores separately. As our scales had low internal reliability (Cronbach’s alpha less than .65) we used the aggregated score as a raw score to simply show the number of correctly answered questions on the assessment. Where 5 or more science questions were present on the assessment, we also calculated a science score as a sum of all correct answers on these items.

2.5 Methods of Analysis

2.5.1 Analysis of Teacher Feedback Forms

Open-ended responses were coded by examining the first 100-200 responses and classifying these responses into one or more mutually exclusive categories. Multiple codes could be applied to the same response if the teacher made comments that fit into more than one coding category. If codes were added or changed during the coding process, all responses were re-examined using this new coding scheme.

Open-ended responses, codes, and numeric answers for ratings and the amount of time spent teaching each unit were exported to Microsoft Excel® where the minimum, maximum, mean and standard deviation of ratings were calculated, and the most common codes were tabulated and representative responses marked for reporting. Teachers were able to report the amount of time they spent teaching in either minutes or class periods. If a teacher reported the amount of time teaching in periods but not in minutes, the number of periods was multiplied by 45 minutes in order to determine an approximate number of minutes spent teaching the lesson. Additionally, for lessons with multiple parts, the total number of minutes was calculated by adding up the number of minutes the teacher reported spending on each part of the lesson. As long as a teacher reported either minutes or periods taught for at least one part of the lesson, they were counted as having taught the lesson and their overall time was calculated.

2.5.2 Analysis of Student Assessments

Further data cleaning was completed after analysis of reliability of scores, but before beginning analysis. Any student missing three or more questions on either the pre- or post-assessment was removed from the dataset. All remaining missing responses were replaced with a zero – a coding of “incorrect”.

Further data cleaning was carried out through a multistep process using PASW Statistics. First, the number of students and classes missing level-1 demographic information was determined (one or more of gender, IEP status, LEP status, FRL status, and racial information) in order to decide if any of these variables needed to be dropped from the HLM analysis. We then examined the number of classrooms missing level-2 demographic information (number of minutes spent teaching, teacher position, and teacher experience). Generally, the threshold for excluding a variable depended on how many classes were in the dataset. If including the variable would have dropped more than three or four classes from the dataset, then the variable was not considered, except in the case where the final dataset still had more than 60 classes. We found that Free or Reduced-Price Lunch (FRL) was a problematic variable due to the fact that many school districts keep this information confidential. For the *Designing Solar Ovens* unit, the classrooms that reported FRL constituted too small a sample to analyze using HLM, thus FRL status was not considered in the analysis of that unit. Similarly, for *Designing Solar Ovens*, many teachers did not report the amount of time spent teaching the unit, thus that information was also not considered in the HLM analysis. Although a number of classrooms were also missing information about FRL status and the amount of time spent teaching for the *Designing Parachutes* unit, more than 60 classrooms remained after dropping classrooms missing this information, thus we were able to keep these variables in our analysis of that unit.

Several conditions must be met for the HLM analysis to be valid. These include multivariate normality, a linear relationship between the predictors and the outcome variable, and an absence of large outliers. Each assumption was tested using the HLM software to generate level-1 and level-2 residuals for the final model. Histograms and scatterplots of those residuals were subsequently inspected to determine normality and linearity and to detect outliers. For both units reported here, the residuals were found to be normally distributed at both levels, and were not correlated with any of the predictors, indicating linearity. However, the residuals did linearly increase with the observed post-assessment scores, but not with the pre-assessment scores, suggesting that the differences between students are not being fully captured by our models. However, enough variance at level-2 (the classroom level) was accounted for by our models to make the analysis informative. Major outliers were identified using both the residual scatterplots and a plot of the Mahalanobis distance for each class against the theoretical chi-squared distribution (assuming multivariate normality). The HLM software also provides a tool to test the assumption that the within-class variance (σ^2) is homogeneous. The homogeneity of σ^2 was tested, and where it was not homogeneous, the variance was modeled using HLM software.

Our primary question of interest is whether participation in EiE (the treatment) affects student understanding of the engineering learning objectives and related science concepts, as operationalized by scores on the post-assessment. To address this question, responses from students in EiE field test classrooms (test sample) were compared to responses from a control sample. While both the test sample and the control sample received science instruction after completing the pre-assessments and before completing post-assessments, the test sample also completed the EiE curriculum in conjunction with their regular science curriculum.

As stated previously, we were also interested in whether effects of the curriculum are modified by demographic variables such as gender, socio-economic status as measured by student participation in the National FRL Program, race, and status as an English language-learner with limited English Proficiency (LEP). Finally, we also wished to investigate how the amount of time a classroom spends on a unit moderates student outcomes. Again, the effects of FRL status and the amount of time spent implementing the unit could only be examined for *Designing Parachutes* as the sample for the

Designing Solar Ovens unit would have become too small had we dropped those students and classrooms missing this information.

A pre-test/post-test design was used to examine curriculum effects. Post-assessment scores served as the outcome variable, while pre-assessment scores were entered as a covariate. This strategy results in an ANCOVA model, where the pre-assessment is used to adjust treatment effect estimates by differences in pre-assessment scores between control and treatment samples. The ANCOVA model was implemented at the classroom level using hierarchical linear modeling (HLM) methodology that accounts for the clustering of students within classrooms. All HLM models were estimated using restricted maximum likelihood and deviance statistics were used to test the relative model fit from competing conditional models.

One of the assumptions of the HLM analysis is that the within-class variance is homogeneous between different classes. Violating this assumption can lead to biased significance estimates and Type I error. Fortunately, the HLM software (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2004) provides a module to easily test this assumption, and to model the level-1 variance with the model predictors. This allows us to understand what factors affect that variance, and allows HLM to take those factors into account when calculating significance estimates in the model. This leads to more accurate significance estimates, even when using the robust standard errors. For the *Designing Parachutes* unit, the level-1 within-class variance was found to be heterogeneous ($p<.05$). The level-1 variance was logarithmically modeled with level-1 and level-2 predictors using the HLM software. Though some were found to be significant predictors of the variance, they were not sufficient to explain all the heterogeneity.

A two step modeling process was performed:

Step 1: Student demographics were modeled at Level-1, the student level. The student level was modeled by first testing the significance of the fixed effects associated with student level demographics and a classroom-centered pre-assessment. Non-significant fixed effects were eliminated by order of least significance until only significant student level demographic fixed effects remained. Random effects corresponding to each fixed effect were then tested to determine if the fixed effects varied across classrooms.

Step 2: Because the curriculum was delivered to intact classrooms, the treatment effect and pre-assessment covariate were conceptualized as classroom level (Level-2) variables. The student level intercept was modeled at the classroom level using an ANCOVA strategy where classroom mean pre-assessment served as the covariate, and a dummy variable indicating treatment and control classes served as the treatment effect. For both units reported here, we found that significant variance existed among classrooms. Homogeneity of regression slopes (an ANCOVA assumption) was also tested and confirmed in all cases. The demographic coefficients from the finalized student level model were also predicted by the classroom level treatment dummy in order to examine if the treatment effects varied by demographic groups.

The pre-assessment scores were used as a covariate, and were group-mean centered. Group-mean centering, rather than grand-mean centering, was used because (1) the relationship between pre- and post-assessment scores becomes better specified since we are comparing each student to their own class (2) it allows us to interpret the grand intercept, G00, as the average class score on the post-assessment, rather than the average student score. Continuous level-2 predictors (class size, number of years teaching, the classroom means of pre-assessment scores, and in some cases, the number of minutes spent teaching EiE) were grand-mean centered, so that the treatment effect could be interpreted in the context of an average class, rather than in the context of a class with class size 0, classroom pre-assessment mean 0, and a novice teacher (Bryk & Raudenbush, 1992).

While these variables could be grand-mean centered in HLM, there were some instances where we needed to center them in PASW Statistics before exporting to HLM. The number of minutes spent teaching EiE (NumMinEiE) needed to be centered in PASW, since that number must be zero for control classrooms, but would have been negative had we used HLM to center it. This was also true for the variables used to measure the interaction effects of level-2 predictors with treatment (class size by treatment, number of years teaching by treatment, and classroom means of pre-assessment scores by treatment). Again, the control classrooms needed to have value 0 after being centered; however, had we set the control values to 0 and then centered them in HLM, the control values would have become negative, causing a misspecification of the treatment effect. Whenever one of these interaction effect variables was significant, it was necessary to re-center it (and the corresponding level-2 predictor) in PASW Statistics and then redo the HLM analysis. The easiest way to center a variable using the PASW software is to take its z-score, and since our variables were normally distributed, the z-score of the variable was just as valid to use in the HLM analysis as the variable itself. Therefore, we centered the variables using their z-scores and then set the control values to zero.

The only level-2 predictor that was not centered in any way was the science specialist dummy variable (and, correspondingly, the variable indicating its interaction with treatment), since we wanted to interpret our results in the context of a regular classroom teacher.

3 Results

In the following sections, the results of evaluation are reported separately for the two year 6 field test units.

3.1 Results for the "Designing Solar Ovens" Unit Evaluation

The EiE unit *Green Engineering: Designing Solar Ovens* focuses on green engineering, the science of energy and energy transfer, and how green engineers use their knowledge about energy, resources, and local consumers to create environmentally-friendly technologies. Lesson 1 of the unit explores the story of Lerato, a young girl in Botswana, who learns about solar ovens from an older girl from her village who is studying to be a green engineer. Lerato improves and uses a solar oven to cook food for her family so that she and her brothers and sister do not have to collect firewood to build cooking fires. This lesson introduces students to the basic concepts and vocabulary of green engineering, as well as heat energy, thermal conductors and insulators, and heat transfer. Lesson 2 teaches students about "life cycle assessments" used by engineers to evaluate the environmental impact of various technologies. Children follow the life cycle of a glass bottle and think about the differences for the environment when the bottle is reused, recycled, or simply thrown out. In Lesson 3, students learn about thermal conductors versus thermal insulators and conduct experiments to test the thermal properties of different materials. This leads them into Lesson 4, in which they use the results of those experiments to design and test their own solar ovens.

The *Designing Solar Ovens* assessment was created and piloted in Year 5 (2007-2008) of the EiE project. The assessment was revised, and the revision distributed to field test teachers, during the 2008-2009 school year (Year 6 of the EiE project). Revisions were again made following field testing, and all field test teachers (and control teachers who chose to implement the unit) were given copies of the publication version in the fall of 2009.

3.1.1 Formative Evaluation

Feedback forms for *Designing Solar Ovens* were completed by 33 teachers who taught the unit in their classrooms. These teachers were from five states: California, Colorado, Massachusetts, Minnesota, and New Hampshire. Classrooms ranged from grades 2-6. The majority of the forms were from grade 3 and 4 classrooms.

Table 2. *Designing Solar Ovens* – Classroom Grade and State Distribution

State	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	Grade Mixed*	Total
CA		3					3
CO		3	5	2			10
FL							0
MA	1	1	5	3	2	1	13
MN		1	4	1			6
NH				1			1
Total	1	8	14	7	2	1	33

***In MA there was one mixed-grade classroom with students from fifth and sixth grade.**

Teachers were asked to rate the unit overall, as well as provide specific feedback on the individual lessons. Table 3 shows the number of the responses (N), average response, standard deviation (SD), and minimum (Min) and maximum (Max) response to the overall unit, based on a seven point scale, with 1 representing “Not at all”, 3 representing “Slightly”, 5 representing “Moderate”, and 7 representing “Very”.

Table 3. *Designing Solar Ovens* Feedback – Unit Ratings

Question	N	Mean (1-7)	SD	Min	Max
Did this unit further your objectives for science in your classroom?	16	6.1	.85	5	7
Did this unit further your objectives for engineering?	28	6.5	.79	4	7
Did this unit further your objectives for another content area in your classroom?	26	5.3	1.54	1	7
Did this unit positively affect your students' motivation?	28	6.5	.74	5	7
Were the concepts presented in this unit age-appropriate for your classroom?	28	6.0	1.35	3	7
Were the materials and activities age-appropriate?	28	6.1	1.04	3	7
Did preparation for this unit require reasonable time, materials, and skill?	28	5.5	1.43	1	7
Were the Lesson Plans easy to understand?	26	5.9	1.14	3	7
Was this unit reasonably easy to implement and manage?	28	5.5	1.53	1	7
If you used a kit, were sufficient materials provided?	28	6.3	1.08	4	7
What is the likelihood that you will choose to teach this unit again in your classroom?	28	6.0	1.61	3	7

Teachers were asked whether this unit furthered their objectives for another content area (Table 3) beyond science and engineering; 21 teachers responded with math mentioned more than half the time (52%, n=11). One third of these teachers reported that the unit furthered their objectives for language arts (29%, n=6), while others noted social studies (19%, n=4).

Eighteen of the 33 teachers responded to the open-ended question, "How did your students benefit, academically or otherwise, from taking part in this unit?" Table 4 shows the three coding categories used to code teacher responses.

Table 4. SO – Categories of Teacher Responses to the Open-Ended Question, "How did your students benefit, academically or otherwise, from taking part in this unit?"

Coding Category	Number of teachers	Percentage of respondents (N=18)
Students had the opportunity to learn about the work of an engineer, engineering (including the EDP), and science	13	72.2
Students had the opportunity to improve in other areas (problem solving, critical thinking, literacy, planning, cultural awareness)	6	33.3
Students had fun, were engaged, and were positively challenged	10	50.0
Total Number of Comments*	**18**	**100.0**

***The total number of positive comments is less than the sum of individual coding categories because some teachers provided comments that fit into more than one coding category.**

As shown in Table 4, almost three-quarters of responding teachers (72%, n=13) noted that the unit enhanced their students' knowledge of science and engineering topics. Specifically, teachers noted that their students learned about the properties of materials (including thermal insulation and conduction) as well as solar energy and the engineering design process.

> *"They became environmentally conscious and saw correlations to the thermal conductor/insulator discussions in our science units. They were also particularly excited with the design and implementation of the solar oven."* – Colorado teacher, grade 3

> *"They learned about energy and energy transfer. They also thought more deeply about climate change."* – Massachusetts teacher, grade 4

In addition to science and engineering concepts, some teachers (33%, n=6) noted that the unit provided their students with opportunities to practice and improve other skills. All of these six teachers (100%) reported that their students engaged in problem solving, while half (50%, n=3) also mentioned that students benefited from practicing teamwork and cooperation.

> *"Students' interest in science soared. Creative and critical thinking grew by involvement in this process. Cooperation and group work also showed growth."* – Massachusetts teacher, grade 4

> *"Cooperation was always needed. Thinking through process and justifying use of materials was very thought provoking."* – Colorado teacher, grade 5

Additionally, half of respondents to this question (50%, n=9) noted that their students enjoyed the unit and were motivated during class. Of these nine teachers, 4 mentioned that students were especially engaged during the unit's hands-on activities.

> *"My students are always more engaged and seem to understand more when we do hands-on learning activities such as EiE."* – Colorado teacher, grade 4

> *"[This unit] allowed them to experience engineering and hands-on science."* – Colorado teacher, grade 4

"Great hands-on learning experience–always makes the learning better." – Colorado teacher, grade 4

Teachers were asked to provide a rating on a scale of 1-7 indicating how likely they were to teach the unit again. Twenty-eight teachers responded and the average score was 6, with the lowest score being 3 out of 7, indicating that teachers are likely to teach the unit again. Of the 14 teachers that responded when asked to explain their rating, nearly a third noted that students learned and were engaged (29%, n=4), however, the same number expressed concerns about time constraints (29%, n=4).

3.1.1.1 Lesson 1 of "Designing Solar Ovens"

In Lesson 1, students read *Lerato Cooks Up a Plan*, a story set in Botswana. The book tells the story of Lerato, a young girl who works hard to look after her younger siblings while completing daily chores, like gathering firewood for cooking. After a friend returns from university and introduces Lerato to the concept of green engineering, she begins to wonder about the resources people consume during everyday activities. With her friend's help, Lerato uses the engineering design process and her understanding of thermal insulators and conductors to design a solar oven that will eliminate the need to gather wood so that she has more time to weave baskets.

Of the 33 teachers, 26 reported teaching this lesson in their classrooms. As shown in Table 5, teachers spent an average of 105 minutes teaching this lesson. When asked to rate the quality of the lesson on a scale of 1-7, 24 teachers gave Lesson 1 an average score of 5.6.

Table 5. *Designing Solar Ovens* Feedback – Lesson 1

Question	N	Mean	SD	Min	Max
How much time did you spend teaching this lesson? (Minutes)	26	125.2	64.41	45	300
How would you rate the quality of this lesson, overall? (Scale: 1-7)	24	6.1	1.06	3	7

Table 6. SO – Categories of Teacher Responses to the Open-Ended Question, "Please explain your rating [of Lesson 1]"

Coding Category		Number of teachers	Percentage of respondents (N=13)
Positive	High-quality activities / supporting materials	8	61.5
	Provided good opportunities for learning	5	38.4
	Other positive comment	0	0.0
	Total Positive Comments*	**10**	**76.9**
Negative	Criticism of activities / supporting materials	4	30.8
	Time constraints	4	30.8
	Other negative comments	0	0.0
	Total Negative Comments*	**7**	**53.8**

***The total number of comments is less than the sums of individual and grouped coding categories because some teachers provided comments that fit into more than one coding category.**

Of the 26 teachers who taught Lesson 1, 13 provided an explanation for their rating of lesson quality. Table 6 shows coding categories used to code this question. Although more than three-quarters of teachers (77%, n=10) provided at least one positive comment, about one-half (54%, n=7) provided at least one negative comment.

Teachers were most concerned with the amount of time it took to complete the lesson. Four teachers (31%) mentioned this problem.

"I think given more time and following all lesson instructions it would have been better." – Massachusetts teacher, grade 4

"The challenge with EiE is fitting it into our existing science curriculum. We have more to teach than time to teach it in. Given that, the story book was longer than would be ideal given the age of my students (3rd grade) and our time restrictions." – Massachusetts teacher, grade 3

A number of teachers (31%, n=4) also criticized the lesson activities and supporting materials.

"I gave this a 4 because the vocabulary work was good but the other parts were only somewhat helpful for students." – Minnesota teacher, grade 4

"It illustrated the concept well but the students were not as highly engaged as I would hope. – Colorado teacher, grade 3

In spite of these concerns, Table 6 shows that more than half of the respondents (62%, n=8) reported that their students thought the story was interesting and engaging.

"The story really drew the kids in. There was also a news article I found that discussed the use of solar ovens in war-torn Africa – how it has protected women and children caught in the war when they go out to collect firewood. The kids were intrigued by the story and how relevant it was." – California teacher, grade 3

"I thought that the book was very good." – Colorado teacher, grade 5

"Great, engaging story!" – Minnesota teacher, grade 5

Additionally, more than one-third of teachers (38%, n=5) mentioned that the lesson provided opportunities for their students to learn. Of these five teachers, four (80%) mentioned that the lesson stimulated classroom discussion; while three teachers (60%) appreciated that the storybook introduced students to another culture.

"The story is very exciting and interesting – allowed for thoughtful discussion." – California teacher, grade 3

"Learning about another culture - how enriching for them!" – Minnesota teacher, grade 4

3.1.1.2 Lesson 2 of "Designing Solar Ovens"

In Lesson 2, students identify and discuss the resources and relative environmental impacts of using glass versus plastic bottles and then recycling or reusing them. Students talk about the life cycle of a bottle, and discuss ways to reduce the environmental impacts of using drink bottles, including reducing, reusing, and recycling.

Of the 33 teachers, 21 reported teaching Lesson 2. As shown in Table 7, teachers spent an average of 75 minutes teaching the lesson and 16 teachers gave it an overall rating of 6.0 on a scale of 1-7.

Table 7. *Designing Solar Ovens* Feedback – Lesson 2

Question	N	Mean	SD	Min	Max
How much time did you spend teaching this lesson? (Minutes)	21	75	26.3	30	120
How would you rate the quality of this lesson, overall? (Scale: 1-7)	16	6.0	1.10	4	7

Only 8 teachers provided an explanation for their rating of lesson quality. Table 8 shows coding categories used to code this question. In general, teachers were positive about the unit with three-quarters of responding teachers (73%, n=6) providing at least one positive comment, while only three teachers (38%) responded negatively.

As shown in Table 8, half (50%, n=4) of those commenting reported that students were able to understand the science and engineering content and improved in their ability to make real-life engineering connections. Additionally, three teachers (38%) also mentioned that their students found this lesson to be fun, interesting, and challenging.

> *"The students were able to talk, at the end, about the things which surprised them – the amount of transportation used, the length of time it takes for a bottle to break down, and the amount of energy it takes for everything. I think the lesson was thought-provoking."* – Minnesota teacher, grade 4

> *"They really had a lot of 'ah-ha' moments going through this process, and began taking other materials through the process on their own."* – California teacher, grade 3

> *"I think it was enlightening for students and myself to see all the steps of the bottle life cycle. It encouraged students to think more seriously about recycling and waste."* – Colorado teacher, grade 3

Table 8. SO – Categories of Teacher Responses to the Open-Ended Question, "Please explain your rating [of Lesson 2]"

Coding Category		**Number of teachers**	**Percentage of respondents (N=8)**
Positive	Provided opportunities to learn engineering and science content	4	50.0
	Promoted conversation, discussion, and teamwork	1	12.5
	Lesson was fun, engaging, and challenging	3	37.5
	Hands-on experimentation	0	0.0
	Other positive comment	1	12.5
	Total Positive Comments*	**6**	**75.0**
Negative	Not engaging/too easy	1	12.5
	Difficult or confusing	2	25.0
	Time constraints	0	0.0
	Problems with setup/execution of lesson	0	0.0
	Other negative comments	0	0.0
	Total Negative Comments*	**3**	**37.5**

***The total number of comments is less than the sum of individual coding categories because some teachers provided comments that fit into more than one coding category.**

3.1.1.3 Lesson 3 of "Designing Solar Ovens"

In Lesson 3, students conduct a series of hands-on experiments to determine how well different materials perform as thermal insulators. Students also explore the materials in terms of their potential environmental impacts. Finally, they combine their data and analysis to discuss which materials would be best to use in their solar ovens.

Of the 33 teachers, 24 reported that they taught Lesson 3. As shown in Table 9, teachers spent an average of 127 minutes teaching this lesson and 18 teachers gave it an overall rating of 5.7 on a scale of 1-7.

Table 9. *Designing Solar Ovens* Feedback – Lesson 3

Question	N	Mean	SD	Min	Max
How much time did you spend teaching this lesson? (Minutes)	24	127.0	44.28	40	210
How would you rate the quality of this lesson, overall? (Scale: 1-7)	18	5.7	1.27	3	7

Thirteen teachers commented on their rating of lesson quality. The coding categories used to code this question are shown in Table 10. Although two-thirds of the teachers (67%, n=8) provided at least one positive comment, a little more than half (58%, n=7) gave a negative response.

Table 10. SO – Categories of Teacher Responses to the Open-Ended Question, "Please explain your rating [of Lesson 3]"

Coding Category		Number of teachers	Percentage of respondents (N=13)
Positive	Provided opportunities to learn engineering and science content	2	16.7
	Provided opportunities to improve other skills (discussion, problem solving, teamwork)	0	0.0
	Lesson was fun and engaging	3	25.0
	Good teacher resources	1	8.3
	Good hands-on experimentation	5	41.7
	Other positive comment	1	8.3
	Total Positive Comments*	**8**	**66.7**
Negative	Time constraints/difficult to manage in class	2	16.7
	Content was difficult to teach/learn	3	25.0
	Not interesting/less hands-on	1	8.3
	Other negative comments	1	8.3
	Total Negative Comments*	**7**	**58.3**

***The total number of comments is less than the sums of individual and grouped coding categories because some teachers provided comments that fit into more than one coding category.**

As shown in Table 10, the major concern among a few respondents (25%, n=3) was that this lesson was somewhat difficult, both for teachers to implement and for students to understand.

> *"Excellent goals – difficult to carry out."* – New Hampshire teacher, grade 5

> *"I omitted some of this but I also feel that it was beyond fourth grade."* – Massachusetts teacher, grade 4

> *"To do it, you have to be a very organized and experienced teacher. If you have classroom management problems it could hamper the learning." – Minnesota teacher, grade 4*

However, nearly half of commenting teachers reported that students enjoyed the hands-on activities and experiments in this lesson (56%, n=5).

> *"We enjoyed the hands-on activities and being able to test the materials. We also enjoyed comparing materials that were better insulators versus those that are greener."* – Colorado teacher, grade 3

> *"Fantastic – hands-on, data collection, graphing analysis. Excellent!"* – Colorado teacher, grade 4

> *"The kids loved the measurement piece."* – Massachusetts teacher, grade 4

A concern among a few respondents was that this lesson was somewhat difficult, both for teachers to implement (22%, n=2) and for students to understand (22%, n=2).

> *"Excellent goals – difficult to carry out."* – New Hampshire teacher, grade 5

> *"I omitted some of this but I also feel that it was beyond fourth grade."* – Massachusetts teacher, grade 4

> *"To do it, you have to be a very organized and experienced teacher. If you have classroom management problems it could hamper the learning."* – Minnesota teacher, grade 4

3.1.1.4 Lesson 4 of "Designing Solar Ovens"

In Lesson 4, students use the engineering design process to create their own solar ovens, applying what they learned about the thermal properties of materials from Lesson 3.

Out of 33 teachers, 25 reported that they taught this lesson in their classroom. As shown in Table 11, on average, teachers spent 180 minutes teaching this lesson. More than half of the teachers gave an overall rating, with the average response being 5.9 on a scale of 1-7.

Table 11. *Designing Solar Ovens* Feedback – Lesson 4

Question	N	Mean	SD	Min	Max
How much time did you spend teaching this lesson? (Minutes)	25	180.2	98.83	60	540
How would you rate the quality of this lesson, overall? (Scale: 1-7)	15	5.9	1.19	3	7

Table 12. SO – Categories of Teacher Responses to the Open-Ended Question, "Please explain your rating [of Lesson 4]"

Coding Category		Number of teachers	Percentage of respondents (N=11)
Positive	Fun, motivating, and enjoyable	6	54.5
	Provided opportunities to engage in and improve other skills (teamwork, communication, problem solving, critical thinking)	2	18.2
	Lesson was hands-on	0	0.0
	Students made connections to the real world	0	0.0
	Other positive comment	1	9.1
	Total Positive Comments*	7	63.6
Negative	Time constraints/too long	1	9.1
	Lesson was difficult/confusing for students and teacher	3	27.3
	Too easy/not engaging	0	0.0
	Other negative comments	0	0.0
	Total Negative Comments*	6	54.5

***The total number of comments is less than the sum of individual coding categories because some teachers provided comments that fit into more than one coding category.**

Eleven teachers commented on their rating of lesson quality. Table 12 shows coding categories used to code this question. Although nearly two third of teachers (64%, n=7) provided at least one positive comment, about half of teachers (55%, n=6) commented negatively.

As shown in Table 12, about half (55%, n=6) of these respondents indicated that students enjoyed and were motivated and engaged by this lesson.

“The lesson gave children the chance to show all they had understood in a fun and concrete way and gave them excellent use of the EDP. They were very motivated throughout.” – Massachusetts teacher, grade 2

“Students really enjoyed the process of improving their solar ovens.” – Minnesota teacher, grade 5

“Very organized and pacing was excellent. They could quickly see that their design needed changes. During observation time, they began discussing what thought went wrong with their oven and what changes to make.” – Minnesota teacher, grade 4

“It was fun and motivating. We had a great discussion about environmental impact and trade-offs.” – Minnesota teacher, grade 4

Three respondents (27%), however, noted that certain aspects of the lesson were difficult and time-consuming for both students and teachers.

“Students excited and understood goals, but it was a complex understanding.” – New Hampshire teacher, grade 5

“I like the idea of teaching about green engineering and solar ovens. I just think the lessons provided are too lengthy and time-consuming.” – Massachusetts teacher, grade 4

“My only concern is about the timing and measuring [activities], which seems a little excessive for third graders. Otherwise, this is a fabulous lesson.” – Massachusetts teacher, grade 3

3.1.1.5 Summary: “Designing Solar Ovens” Formative Evaluation

Overall, teachers found this unit to be motivating and engaging for their students. They noted that their students learned about both science and engineering and also improved their problem solving and teamwork skills. In addition to successfully integrating this unit with science topics, teachers mentioned that they also used this unit to further objectives in math, language arts, and social studies. The most common concern raised by teachers was in regards to the amount of time necessary to implement this unit. Despite concerns about time, teachers reported that they are very likely to teach the *Designing Solar Ovens* unit again because it provided a wonderful opportunity to illustrate connections between engineering and everyday life. Teachers were also very satisfied with the extent to which this unit raised students’ environmental awareness.

3.1.2 Summative Evaluation

3.1.2.1 Assessment Design: “Designing Solar Ovens”

The *Designing Solar Ovens* assessment was designed and first tested during the 2007-2008 school year. Twelve multiple-choice questions and four choose-all-that-apply questions were chosen for the pilot assessment. During the summer of 2008, a number of questions were dropped or revised to create a new field test assessment. In 2008-2009 it was used for field testing in five states; the assessments were given to students in field test and control classrooms. The field test version included nine multiple-choice questions.

On the identical pre- and post-assessments students were asked science questions about heat energy and insulation. They were also asked engineering questions about green engineering and how solar ovens work. Table 13 describes the text for the questions with the correct answer shown in brackets.

Table 13. *Designing Solar Ovens* Assessment: Questions (Text)

Question #	Category	Question Text
1	Green Engineering	Sophie practices green engineering. She is helping to design water bottles in a factory. Which would she be MOST concerned with? [How to make the water bottles easy to reuse or recycle.]
2	Green Engineering	Too much trash is a problem around the world. How could the trash problem be improved in a way that is MOST friendly to the environment? [Find ways to make less trash.]
3	Green Engineering	Which of the following are resources that are needed to make cookies? [All of the above (flour, sugar, and other cookie ingredients / people who grow the wheat in the cookies / energy used to bake the cookies)]
4	Science / Engineering (Insulation)	Ken has an insulated lunch bag. He is using it to keep his yogurt cool. How does it work? [It allows heat to pass through it very slowly.]
5	Science / Engineering (Insulation)	The next day, Ken wants to bring hot soup to work. Can he use the same lunch bag to keep his soup warm? [Yes, an insulated lunch bag can keep hot foods warm and cold foods cool.]
6	Science (Heat Energy)	If you put a wooden spoon and a metal spoon into a pot of boiling water the metal spoon will feel much warmer. Why? [Heat moves more easily through metal than through wood.]
7	Science (Heat Energy)	Juan poured hot soup into a cool bowl. Soon his bowl was too hot to carry. Which of the following also happened? [The soup got cooler than it had been.]
8	Engineering a Solar Oven	It is a sunny day but Raj's solar oven is not getting hot enough. What could he do to make it work better? [All of these ideas would work. (Insulate the box better. / Turn the oven so it catches more sunlight. / Tape up all the holes in the box so no air can get in or out.)]
9	Engineering a Solar Oven	Which of the following BEST describes how a solar oven works? [It reflects light energy from the Sun onto a cooking pot.]

3.1.2.2 Scale Construction: "Designing Solar Ovens"

Scales were constructed after completion of the post-assessments. It was anticipated that reliable scales would align with the science and engineering objectives of the unit. Reliability analysis was conducted on the 1253 students who had returned both a pre- and a post-assessment.

There were insufficient questions to construct engineering scales, however we constructed a science scale to ascertain whether EiE students learned science content better after participation in the program. This scale had very low internal reliability, so was not used in any further analysis (PreScience: Cronbach's alpha of .066, (n=1232); PostScience: Cronbach's alpha of .121 (n=1229)). The overall scores also had low internal reliability (PreAll: Cronbach's alpha of .310 (n=1210); PostAll: alpha=.352 (n=1216)). Consequently, the raw scores (number correct) are used as outcome variables.

As there were a total of 9 questions on this instrument, we calculated an overall score if students had answered seven of them. If students had answered fewer than seven questions on either the pre- or post-assessment, they were dropped from further analysis. This reduced our sample size by two students.

3.1.2.3 Sample: "Designing Solar Ovens"

Designing Solar Ovens (SO) assessments were collected from students in grades 2 through 6. Grade 2 students received a different version of the assessment than grades 3-6, which is not analyzed in this report. There were not enough grade 6 classrooms to include them for comparative purposes.

The full sample of collected data for SO included 1815 students in 86 classrooms, grades 3-5. Of the 589 students in the control group, 43.1% were excluded due to a missing pre- or post-assessment, and of the 1226 students in the test group, 42.5% were excluded due to a missing pre- or post-assessment. Additionally, 24 individual students were dropped due to missing demographic information. Eleven classes were missing information on which students received FRL, so that variable was not considered in the analysis. Two students answered fewer than seven questions on either the pre-assessment or the post-assessment, and were dropped from the dataset. One class was identified as a statistical outlier using methods described in the methodology section above and dropped.

The final dataset used for analysis included 836 students (Level-1 units) in 46 classrooms (Level-2 units), with an average cluster size of 22.7 and a standard deviation of 5.2. The majority of the sample consisted of students in grade 5—see Table 14.

Table 14. SO Treatment * Grade Crosstabulation

	Grade (# Students)				Grade (# Classrooms)			
	3	4	5	Total	3	4	5	Total
Control	127	7	160	294	8	1	8	17
Treatment	131	126	285	542	8	7	14	29
Total	258	133	445	836	16	8	22	46

Males made up exactly half the dataset (see Table 15). The control and test groups had slightly different racial compositions. Asian students made up 6% of the control group and 13% of the test group, while Hispanics made up 21% of the control group but only 10% of the test group.

Table 15. SO Proportions for Level-1 Dichotomous Variables

Treatment		Gender (male)	Limited English Proficiency	Individualized Education Program
Control		.49	.09	.13
Treatment		.51	.13	.07
Total	Proportion	.50	.11	.09
	N	418	93	78

Table 16. SO Proportions for Level-1 Variables – Race

Treatment		White	Asian	Black	Hispanic	Other	Total
Control	Proportion	.57	.06	.10	.21	.06	1.0
	N	167	19	29	61	18	294
Treatment	Proportion	.66	.13	.07	.10	.04	1.0
	N	359	71	36	53	23	542
Total	Proportion	.63	.11	.08	.14	.05	1.0
	N	526	90	65	114	41	836

Of the 34 teachers participating in field testing for the *Designing Solar Ovens* unit, 22 were in the test group and 13 in the control group. These 34 teachers taught in 28 schools. The control group included 2 science specialists, while the test group had 5—all other teachers were regular classroom teachers. Teachers had 15 years of experience on average, with test group teachers averaging 3 more years of experience than control group teachers. Both the test group and the control group included relatively inexperienced as well as veteran teachers. Class sizes were similar for both treatment groups, at approximately 23 students.

Table 17. SO Teacher and Class Demographics

	Number of Teachers	Number of Schools	Number of Science Specialists	Experience (Number of Years Teaching)				Class Size			
				Mean	SD	Min	Max	Mean	SD	Min	Max
Control	13	11	2	13.6	8.98	4	39	23.4	4.12	16	31
Test	22	19	5	16.6	12.59	2	43	22.2	5.82	6	30
Total	34*	28**	6*	15.5	11.34	2	43	22.7	5.24	6	31

***One teacher taught both a control and a test class. **Two schools contained both control and test classes.**

Schools with field test and control classrooms had similar proportions of Black and Asian students; field test schools on average had fewer Hispanic students and more White students (Table 18). The mean proportion of FRL students at schools with participating classrooms was slightly more than a third for both groups (Table 18). Schools with control participants tended to be slightly larger, on average, than schools with field test participants (Table 19). Only 2 rural schools were represented in the field test group, with the rest distributed nearly half-and-half between suburban and urban settings; urban schools represented nearly half of the control group sample, with the other half evenly distributed between rural and suburban settings (Table 20).

Table 18. SO School Demographics: Proportions by Race, FRL

	Mean		St. Dev.		Minimum		Maximum	
	Control	Test	Control	Test	Control	Test	Control	Test
Black	0.10	0.10	0.07	0.11	0.01	0	0.22	0.47
Hispanic	0.40	0.25	0.32	0.30	0.02	0.01	0.83	0.94
Asian/PI	0.06	0.09	0.06	0.11	0	0	0.18	0.46
White	0.42	0.54	0.35	0.34	0	0	0.97	0.94
Native American/ Mixed/Other	0.01	0.02	0.02	0.02	0	0	0.04	0.09
FRL	0.38	0.36	0.33	0.29	0	0	0.97	0.97

Table 19. School Demographics: Enrollment and # Classrooms per School

	Mean		St. Dev.		Minimum		Maximum	
	Control	Test	Control	Test	Control	Test	Control	Test
Enrollment	580	492	399	223	88	104	1427	919
# Classrooms participating per school*	1.9	1.6	1.1	0.8	1	1	4	4

***Two schools had both test and control classrooms.**

Table 20. SO School Demographics: Setting

	Number of Schools		Total
	Control	Test	
Rural	3 / 27%	2 / 11%	5
Suburban	3 / 27%	9 / 47%	12
Urban	5 / 45%	8 / 42%	11*
Total	11	19	28*

***Two urban schools had both test and control classrooms,
and were classified with both the test and the control groups.**

3.1.2.4 Results: "Designing Solar Ovens"

One outcome variable, PostAll, was used to test for a treatment effect on the performance of students grades 3 to 5 on the EL assessment. It was computed by adding together the number of correct answers for each of the questions listed in Table 13 above. This score had a possible range from 0 to 9. Mean PreAll scores were similar between treatment groups, while the test group mean PostAll score was a point higher than that of the control group. There were not enough questions to compute a science sub-score.

Table 21. SO Descriptive Statistics: PreAll and PostAll Scores by Treatment

	PreAll				**PostAll**			
	Mean	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**
Control	3.7	1.65	0	8	4.1	1.50	1	8
Test	3.8	1.60	0	8	5.2	1.55	1	9

For *Designing Solar Ovens*, the residuals were found to be normally distributed at both levels. The residuals were not correlated with any of the predictors, indicating linearity, but the model overestimated the lowest PostAll scores and underestimated the highest PostAll scores, indicating the model is not fully specified. Major outliers were identified using both the residual scatterplots and a plot of the Mahalanobis distance for each class against the theoretical chi-squared distribution (assuming multivariate normality). In this case, only one major outlier was identified, and was removed. It was also confirmed that the within-class variance (σ^2) was homogeneous between classes ($p > 0.5$).

For this unit, we tested the student level (level-1) variables of IEP and LEP status, as well as gender and race (Asian, Black, Hispanic, and Other). We also included the PreAll (PreScience) score as a covariate. Next, we tested whether each level-1 variable was best modeled as a fixed or random coefficient, the default being a fixed coefficient, unless the random variance was found to be significant ($p < .05$). At the classroom level (level-2) we tested the effect of the treatment on the level-2 intercept, including the classroom PreAll Means (PreScience Means), class size, and number of years teaching as covariates. We also tested the effect of grade and of teacher specialty (Science Specialist) here. For each of these variables, we simultaneously tested the variable's interaction with the treatment (*e.g.* PreAll Means by Treatment, Grade 3 by Treatment, Number of Years Teaching by Treatment, etc.), which were calculated by multiplying the variable by the treatment dummy variable, *i.e.* setting the variable to zero for control classrooms. This allowed us to test if the effect of these variables differed for the test and control groups. Finally, we tested the interaction of the treatment with any level-1 variables that were found to be significant.

The final two-level conditional model for the *Designing Solar Ovens* (SO) assessment includes all variables and random variance coefficients which were found to be statistically significant ($p<.05$) using the given procedure (see Figure 2). This final model shows evidence of a main treatment effect (γ_{02} p= .000) with a Cohen's *d* effect size of 0.689 ($0.930 / \sqrt{1.737 + .0828}$ – see Table 22 and Table 23). For comparison, gender (male) has an effect size of −0.230 (−0.303 / 1.318) and IEP status has an effect size of −0.344 (−0.453 / 1.318), half or less than half the effect size of the treatment. The conditional model explains approximately 63% of the classroom-level variance (proportional reduction in variance statistic: (1 – (0.288 / 0.788) – see Table 23 and Table 24).

Level-1 Model:

$$\text{PostAll} = \beta_0 + \beta_1(\text{Gender}) + \beta_2(\text{IEP}) + \beta_3(\textbf{PreAll}) + \beta_4(\text{Is\_Black}) + \beta_5(\text{Is\_Other}) + \text{r}$$

Level-2 Model:

$$\beta_0 = \gamma_{00} + \gamma_{01}(\textit{PreAllMean}) + \gamma_{02}(\text{Treatment}) + \gamma_{03}(\textit{NumYrsTeaching\_by\_Treatment}) + \text{u}_0$$

$$\beta_1 = \gamma_{10}$$

$$\beta_2 = \gamma_{20}$$

$$\beta_3 = \gamma_{30}$$

$$\beta_4 = \gamma_{40}$$

$$\beta_5 = \gamma_{50}$$

Bold indicates group mean centered.
Italicized indicates grand-mean centered.

Figure 2. *Designing Solar Ovens* PostAll Score – Conditional Model

Table 22. SO PostAll Score Conditional Model – Final Estimation of Fixed Effects (with robust standard errors)

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. df	P-value
For Intercept 1, β_0					
Intercept 2, γ_{00}	4.398	0.104	42.471	42	0.000
PreAll Classroom Mean, γ_{01}	0.570	0.072	7.951	42	0.000
Treatment, γ_{02}	0.930	0.118	7.914	42	0.000
Number of Years by Treatment, γ_{03}	0.186	0.068	2.756	42	0.009
For Gender slope, β_1					
Intercept 2, γ_{10}	-0.303	0.091	-3.337	827	0.001
For IEP slope, β_2					
Intercept 2, γ_{20}	-0.454	0.120	-3.792	827	0.000
For PreAll (Group mean centered) slope, β_3					
Intercept 2, γ_{30}	0.291	0.039	7.472	827	0.000
For Is_Black slope, β_4					
Intercept 2, γ_{40}	-0.416	0.197	-2.109	827	0.035
For Is_Other slope, β_5					
Intercept 2, γ_{50}	-0.587	0.281	-2.086	827	0.037

Table 23. SO PostAll Score Conditional Model – Final Estimation of Variance Components

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
Intercept 1, u_0	0.288	0.083	42	79.74	0.000
level-1, r	1.318	1.738			

Table 24. SO PostAll Score Unconditional Model – Final Estimation of Variance Components

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
Intercept 1, u_0	0.788	0.622	45	310.83	0.000
level-1, r	1.402	1.965			

3.1.2.5 Summary: "Designing Solar Ovens" Summative Evaluation

Our analysis suggests that students learning through EiE's *Designing Solar Ovens* engineering design unit in conjunction with their science unit about heat and energy learn about engineering and science concepts, and that this learning is significantly greater than that of control students who spend time only on their heat and energy science unit. The effect of the EiE *Designing Solar Ovens* field test appears to be moderately high on the PostAll score (d= 0.689). This difference was consistent across gender, race, LEP status, and IEP status, indicating that the treatment has the same effect across these demographic variations. Though boys, black students, and those with IEP's tended to have lower scores, these effects were the same for both EiE students and the control group.

3.2 Results for the "Designing Parachutes" Unit Evaluation

In the EiE unit *Designing Parachutes*, students are introduced to the field of aerospace engineering and how atmospheric properties and the concept of drag affect the design and operation of parachutes. The unit begins with the discussion of a fictional story read as part of Lesson 1. In the next lesson, students discuss and brainstorm the work of aerospace engineers and the properties of the planets and other celestial bodies as they determine how a fictional spacecraft will navigate the solar system. In Lesson 3, they explore the concept of drag by observing and analyzing a model that demonstrates how the thickness of an atmosphere affects how fast objects fall. Students also collect and analyze their own data relating to three variables of a parachute: suspension line length, canopy size, and canopy material. Finally, in Lesson 4, students utilize their data to inform the design of their own parachutes. Using the engineering design process they plan, create, test, and improve their parachute designs to meet packing and speed criterion for a mission travelling to a planet with an atmosphere thinner than Earth's.

3.2.1 Formative Evaluation

Feedback forms for *Designing Parachutes* were completed by 42 teachers who taught the unit in their classrooms. These teachers were from five states: California, Colorado, Massachusetts, Florida, and Minnesota; classrooms ranged from grades 3-6. Not surprisingly since this unit is designed for advanced grade levels, the majority of the forms were from grade 4 and 5 classrooms.

Teachers were asked to rate the unit overall, as well as provide specific feedback on the individual lessons. Table 26 shows the number of the responses (N), average response, standard deviation (SD), and minimum (Min) and maximum (Max) response for questions about the unit overall, based on a seven point scale, with 1 representing "Not at all", 3 representing "Slightly", 5 representing "Moderate", and 7 representing "Very".

Table 25. *Designing Parachutes* – Classroom Grade and State Distribution

State	Grade 3	Grade 4	Grade 5	Grade 6	Grade Mixed*	Total
CA			3			3
CO	1	6	1			8
FL	1	1	6			8
MA	4	5	3	2		14
MN	3	3	2		1	9
NH						0
Total	9	15	15	2	1	42

*** In MN there was one homeschooled class of 10-13 year-olds.**

Table 26. *Designing Parachutes* Feedback – Unit Ratings

Question	N	Mean (1-7)	SD	Min	Max
Did this unit further your objectives for science in your classroom?	40	6.4	.74	4	7
Did this unit further your objectives for engineering?	40	6.5	.75	4	7
Did this unit further your objectives for another content area in your classroom?	39	5.2	1.39	2	7
Did this unit positively affect your students' motivation?	41	6.7	.52	5	7
Were the concepts presented in this unit age-appropriate for your classroom?	42	6.3	.82	4	7
Were the materials and activities age-appropriate?	42	6.5	.67	4	7
Did preparation for this unit require reasonable time, materials, and skill?	41	5.8	.95	4	7
Were the Lesson Plans easy to understand?	40	6.4	.70	5	7
Was this unit reasonably easy to implement and manage?	41	6.0	.95	4	7
If you used a kit, were sufficient materials provided?	40	6.5	1.08	3	7
What is the likelihood that you will choose to teach this unit again in your classroom?	40	6.7	.62	5	7

Teachers were asked whether this unit furthered their objectives for another content area beyond science and engineering (Table 26); 27 teachers responded, with math being mentioned most often (56%, n=15). Multiple teachers also mentioned that the unit furthered their objectives for language arts (37%, n=10) and a few repeated the connection with science (15%, n=4).

Of the 42 teachers, 38 responded to the question, "How did your students benefit, academically or otherwise, from taking part in this unit?" As shown in Table 27, about three-quarters of the respondents (76%, n=29) noted that students had the opportunity to learn about science and engineering topics such as the properties of parachutes (including the forces of gravity and drag) and atmospheres. Of these 29 teachers, nearly one-third (31%, n=9) specifically noted that their students practiced and / or gained skills in using the engineering design process.

> *"Students learned a great deal about atmosphere and drag. Students learned how to experiment and use actual scientific data to make choices about the canopy size, canopy material, and suspension line length." –* Massachusetts teacher, grade 4

"The students utilized the Scientific Method to design different types of parachutes and make hypotheses and predictions concerning them. They used measuring skills." – Florida teacher, grade 5

Table 27. PA – Categories of Teacher Responses to the Open-Ended Question, "How did your students benefit, academically or otherwise, from taking part in this unit?"

Coding Category	Number of teachers	Percentage of respondents (N=38)
Students had the opportunity to learn about the work of an engineer, engineering (including the EDP), and science	29	76.3
Students had the opportunity to improve in other areas (problem solving, critical thinking, literacy, planning, cultural awareness)	23	60.5
Students had fun, were engaged, and were positively challenged	16	42.1
Total Number of Comments*	38	100.0

***The total number of positive comments is less than the sum of individual coding categories because some teachers provided comments that fit into more than one coding category.**

Six of these sixteen teachers specifically noted that their students were introduced to the work of an engineer and were able to make real-life connections to engineering.

"It increased their understanding of engineering. It gave them authentic practice in reading and math. It provided the opportunity for problem solving and creative thinking and reinforced understanding of science concepts such as resistance, atmosphere, gravity." – Minnesota teacher, grade 5

"It enhanced their knowledge of technology and how it affects our lives. The students were also introduced to various fields of engineering, and how to use creativity and critical thinking skills to solve problems." – Florida teacher, grade 5

In addition to science and engineering concepts, many teachers (61%, n=23) reported that their students improved in other areas as well. Of these 23 teachers, 13 noted that the unit also helped their students improve skills such as teamwork and 12 said the units helped them to improve their problem solving skills (52%).

"My students were able to think about the Design Process and put it into use. They were challenged to work as a team and problem solve together to be successful." – Massachusetts teacher, grade 5

Again as shown in Table 27, 16 respondents (42%) noted that their students enjoyed the unit and were motivated during class; 3 of these 16 teachers explicitly stated that the students were especially engaged during the unit's hands-on activities.

"My students really enjoyed the hands-on experience and being able to determine which types of material would work better for their parachutes to float." – Florida teacher, grade 3

"The hands-on learning was a huge bonus. The students loved to actually build and test what they built." – Colorado teacher, grade 4

"My students were very excited during the parachute program. It was wonderful watching them research the conditions of the planet they were going to build a parachute for. They said they really liked building something to solve a problem." – Massachusetts teacher, grade 3

> *"My three different third grade classes all enjoyed working in pairs or teams to create and test their parachutes."* – Minnesota teacher, grade 3

> *"Their motivation to experiment and try out new ideas significantly increased. Their scientific knowledge of drag, solar system, and engineering became very real to them. They see a need for these areas, and believe they are capable, better problem solvers and team players."* – Minnesota teacher, grade 4

Teachers were also asked to provide a rating on a scale of 1-7 indicating how likely they were to teach the unit again. The average score was 6.7, with the lowest score being 5 out of 7, indicating that teachers are highly likely to teach the unit again. Of the 26 teachers who explained their rating, the most common reasons cited were that students enjoyed the unit (35%, n=9) and that it enhanced the curriculum (31%, n=8).

3.2.1.1 Lesson 1 of "Designing Parachutes"

In Lesson 1, students read *Paulo's Parachute Mission,* a story set in Brazil. The book tells the story of Paulo, a twelve-year-old boy who decides to build parachutes that will help him safely transport fruit from trees to the ground. With the help of a new friend, Paulo uses the engineering design process and the concepts of atmosphere and drag to design parachutes that will help prevent the fruit he picks from smashing open when he drops them to his friend.

Of the 42 teachers completing this form, 35 teachers reported teaching this lesson in their classroom. As shown in Table 28, teachers spent an average of 86.7 minutes teaching this lesson. When asked to rate the quality of the lesson on a scale from 1-7, teachers gave Lesson 1 an average score of 6.3.

Table 28. *Designing Parachutes* Feedback – Lesson 1

Question	N	Mean	SD	Min	Max
How much time did you spend teaching this lesson? (Minutes)	35	86.7	42.20	30	180
How would you rate the quality of this lesson, overall? (Scale: 1-7)	35	6.3	.96	4	7

Of the 39 teachers who taught Lesson 1, 33 teachers provided an explanation for their rating of lesson quality. Table 29 shows coding categories used to code this question. Overall, teachers responded positively to Lesson 1 with the majority of teachers (82%, n=27) providing at least one positive comment. Fewer than one-third of the responding teachers (27%, n=9) provided a negative comment.

Table 29. PA – Categories of Teacher Responses to the Open-Ended Question, "Please explain your rating [of Lesson 1]"

Coding Category		Number of teachers	Percentage of respondents (N=33)
Positive	High-quality activities / supporting materials	25	75.8
	Provided good opportunities for learning	7	21.2
	Other positive comments	0	0.0
	Total Positive Comments*	**27**	**81.8**
Negative	Criticism of activities/supporting materials	9	27.3
	Time constraints	5	15.2
	Other negative comments	0	0.0
	Total Negative Comments*	**9**	**27.3**

***The total number of comments is less than the sums of individual and grouped coding categories because some teachers provided comments that fit into more than one coding category**

As shown in Table 29, 25 teachers (76%) commented on the high quality of the lesson activities. Twelve of these 25 teachers specifically reported that their students found the story interesting and engaging.

> *"The students enjoyed reading 'Paulo's Parachute Mission' and were very anxious to learn more about aerospace engineers."* – Florida teacher, grade 5

> *"The students related to Paulo's feelings of moving and meeting new people."* – Minnesota teacher, grade 3

Additionally, some teachers (21%, n=7) mentioned that the lesson provided good opportunities for students to learn. Of these seven teachers, 3 mentioned that the lesson inspired discussion and critical thinking.

> *"I feel it was a great way to introduce the lesson."* – Colorado teacher, grade 5

> *"Great for building background knowledge."* – Colorado teacher, grade 4

> *"The book gave students a starting point and then we worked in a large group to get the answers."* – Minnesota teacher, grade 3

3.2.1.2 Lesson 2 of "Designing Parachutes"

In Lesson 2, students gain a broader understanding of what the job of an aerospace engineer involves by discussing how the features of the Solar System influence the design of spacecraft. Students draw and label a fictional spacecraft designed to explore a specific celestial body in our Solar System, and explain how its features match the spacecraft's destination and mission goals. Through this activity, students learn that aerospace engineers must address many constraints and criteria as they design spacecraft to explore the unique features of celestial bodies.

Of the 42 teachers, 37 reported teaching Lesson 2. As shown in Table 30, teachers spent an average of 77.7 minutes teaching the lesson and gave it an overall rating of 6.3 on a scale of 1-7.

Table 30. *Designing Parachutes* Feedback – Lesson 2

Question	N	Mean	SD	Min	Max
How much time did you spend teaching this lesson? (Minutes)	37	77.7	33.49	40	180
How would you rate the quality of this lesson, overall? (Scale: 1-7)	38	6.3	.93	4	7

Twenty-eight teachers provided an explanation for their rating of lesson quality. Table 31 shows coding categories used to code this question. In general, teachers were positive about Lesson 2 with more than three-quarters of responding teachers (79%, n=22) providing at least one positive comment. Only a few teachers (21%, n=6) contributed a negative comment.

As shown in Table 31, 12 of the commenting teachers (43%) mentioned that their students found Lesson 2 to be fun, interesting, and engaging. Of these 12 teachers, 4 specifically mentioned that the hands-on activities engaged their students.

> *"'This activity really got students thinking about Aerospace Engineering and outer space crafts."* – Minnesota teacher, grade 3

> *"It challenged the students to be creative."* – Florida teacher, grade 5

> *"Students loved designing their own spacecraft!"* – Florida teacher, grade 5

Table 31. PA – Categories of Teacher Responses to the Open-Ended Question, “Please explain your rating [of Lesson 2]”

Coding Category		Number of teachers	Percentage of respondents (N=28)
Positive	Provided opportunities to learn engineering and science content	7	25.0
	Promoted conversation, discussion, and teamwork	2	7.1
	Lesson was fun, engaging, and challenging	12	42.9
	Hands-on experimentation	3	10.7
	Other positive comments	3	10.7
	Total Positive Comments*	22	78.6
Negative	Not engaging/too easy	2	7.1
	Difficult or confusing	2	7.1
	Time constraints	1	3.6
	Problems with setup/execution of lesson	0	0.0
	Other negative comments	1	3.6
	Total Negative Comments*	6	21.4

***The total number of comments is less than the sums of individual and grouped coding categories because some teachers provided comments that fit into more than one coding category.**

3.2.1.3 Lesson 3 of "Designing Parachutes"

In Lesson 3, students watch a demonstration of how atmospheres affect falling objects. They then experiment with different variables (canopy size, suspension line length, canopy material) to explore how each affects the rate at which a parachute drops. Students then analyze their data and determine the properties of a parachute that make it fall more slowly.

Of the 42 teachers, 38 reported that they taught Lesson 3. Teachers spent an average of 128.9 minutes teaching this lesson and gave it an overall rating of 6.6 on a scale of 1-7.

Table 32. *Designing Parachutes* Feedback – Lesson 3

Question	N	Mean	SD	Min	Max
How much time did you spend teaching this lesson? (Minutes)	38	128.9	52.20	30	240
How would you rate the quality of this lesson, overall? (Scale: 1-7)	37	6.6	.64	5	7

Fifteen teachers commented on their rating of lesson quality. The coding categories used to code this question are shown in Table 33. In general, teachers were very positive about Lesson 3. All 15 commenting teachers (100%) provided at least one positive comment, while only two of these teachers (13%) also gave a negative response.

As shown in Table 33, about half of the responding teachers (53%, n=8) mentioned that they appreciated the lesson's hands-on activities. Furthermore, 6 teachers (40%) mentioned that their students enjoyed the lesson and were engaged, with 3 of the 6 specifically mentioning that the hands-on activities were especially engaging.

> *"Students really enjoyed making their parachutes and being able to test them out!"* – Florida teacher, grade 3

> *"The students loved it!"* – Colorado teacher, grade 4

"I enjoyed it, the students enjoyed it and learned from it, and it appears that they will be able to take it with them (understanding for life)." – Massachusetts teacher, grade 5

"Wow!! They loved this lesson. The hands-on activities were great and everyone was involved." – Florida teacher, grade 4

Table 33. PA – Categories of Teacher Responses to the Open-Ended Question, "Please explain your rating [of Lesson 3]"

Coding Category		Number of teachers	Percentage of respondents (N=15)
Positive	Provided opportunities to learn engineering and science content	3	20.0
	Provided opportunities to improve other skills (discussion, problem solving, teamwork)	2	13.3
	Lesson was fun and engaging	6	40.0
	Good teacher resources	1	6.7
	Good hands-on experimentation	8	53.3
	Other positive comments	0	0.0
	Total Positive Comments*	**15**	**100.0**
Negative	Time constraints/difficult to manage in class	2	13.3
	Content was difficult to teach/learn	0	0.0
	Not interesting/less hands-on	0	0.0
	Other negative comments	0	0.0
	Total Negative Comments*	**2**	**13.3**

***The total number of comments is less than the sums of individual and grouped coding categories because some teachers provided comments that fit into more than one coding category.**

3.2.1.4 Lesson 4 of "Designing Parachutes"

In Lesson 4, students use the engineering design process and their new knowledge of the properties of parachutes based on experimentation in Lesson 3 to design and create a parachute that is able to fall within a certain range of speed.

Of the 42 teachers, 36 reported that they taught this lesson in their classroom. On average, teachers spent 130.6 minutes teaching this lesson and gave it an overall rating of 6.4 on a scale of 1-7.

Table 34. *Designing Parachutes* Feedback – Lesson 4

Question	N	Mean	SD	Min	Max
How much time did you spend teaching this lesson? (Minutes)	36	130.6	57.08	30	270
How would you rate the quality of this lesson, overall? (Scale: 1-7)	34	6.4	.78	4	7

When asked to comment on their rating of lesson quality, 30 teachers responded. Table 35 shows coding categories used to code this question. Overall, teachers responded positively about Lesson 4 with about three-quarters (77%, n=23) providing at least one positive comment. Six teachers (20%) contributed negative comments.

Table 35 shows that 18 teachers (60% of those responding) said that Lesson 4 was fun and engaging for their students. Five of these 18 teachers noted that the hands-on activities were especially engaging and motivating.

"The lesson was a great ending to the unit. Very hands-on, active, motivating and age-appropriate. The students also had to use their prior knowledge from lessons 1-3." – Minnesota teacher, grade 3

"The students enjoyed the challenge of creating a parachute that was mission ready and deciding on their own materials." – Florida teacher, grade 5

Table 35. PA – Categories of Teacher Responses to the Open-Ended Question, "Please explain your rating [of Lesson 4]"

Coding Category		Number of teachers	Percentage of respondents (N=30)
Positive	Fun, motivating, and enjoyable	18	60.0
	Provided opportunities to engage in and improve other skills (teamwork, communication, problem solving, critical thinking)	4	13.3
	Lesson was hands-on	5	16.7
	Students made connections to the real world	1	3.3
	Other positive comments	3	10.0
	Total Positive Comments*	23	76.7
Negative	Time constraints/too long	1	3.3
	Lesson was difficult/confusing for students and teacher	1	3.3
	Too easy/not engaging	3	10.0
	Other negative comments	1	3.3
	Total Negative Comments*	6	20.0

***The total number of comments is less than the sums of individual and grouped coding categories because some teachers provided comments that fit into more than one coding category.**

Eight teachers (27%) contributed their observations that students were especially successful in this lesson because they used prior knowledge gained in the three previous lessons.

"Students could make a successful parachute because of the previous lesson. They really liked showing how their parachute worked. I can see some of the kids continuing this at home." – Minnesota teacher, grade 4

"The class was able to use their prior knowledge of the previous lessons and discussions to complete their parachute. The unit was very hands-on, motivating, and engaging!" – Minnesota teacher, grade 5

"The kids were really able to apply what they had learned to the construction of their parachutes. The kids also knew what they needed to do to make it better." – Minnesota teacher, grade 4

3.2.1.5 Summary: "Designing Parachutes" Formative Evaluation

Overall, teachers found this unit to be fun and engaging for their students. They noted that their students learned about both science and engineering and also improved their critical thinking, problem solving, and teamwork skills. In addition to successfully integrating this unit with science topics, teachers mentioned that they also used this unit to further objectives in language arts, social studies and math. The majority of concerns regarding implementation of this unit appeared in Lesson 4, with teachers having difficulty finding the space necessary to effectively prepare and test the students' parachute designs. Despite concerns about available time and space, teachers reported that they are very likely to teach the *Designing Parachutes* unit again because of their students' level of enjoyment and its enhancement of student understanding across multiple subject areas.

3.2.2 *Designing Parachutes:* Summative Evaluation

3.2.2.1 Assessment Design: "Designing Parachutes"

The *Designing Parachutes* assessment was designed and first tested during the 2007-2008 school year. Seventeen multiple-choice questions, 6 choose-all-that-apply questions, and 11 true/false questions were chosen for the pilot assessment. During the summer of 2008, a number of questions were dropped or revised to create a new field test assessment. In 2008-2009 the new instrument was used for field testing in five states; the assessments were given to students in field test and control classrooms. The field test version included six multiple-choice questions and four True/False questions.

On the identical pre- and post-assessments students were asked science questions about atmospheric properties and the effects of drag on falling objects. They were also asked engineering questions about the design and operation of parachutes and the work of aerospace engineers. Table 36 describes the text for the questions with the correct answer shown in brackets.

Table 36. *Designing Parachutes* Assessment: Questions (Text)

Question #	Category	Question Text
1	Engineering	The atmosphere on Mars is thinner than the atmosphere on Earth. How can Mr. Ino change a parachute that works well on Earth to help a robot land safely on Mars? [Make the parachute larger]
2	Engineering	Wilson constructed a model rocket. It didn't fly for as long as he wanted it to. He decided to change the shape of the fins on the rocket. Based on his testing data (shown in the table), what should Wilson do to make his rocket fly longer? [Use curved fins] (See Figure 3)
3	Engineering	At work, an aerospace engineer might: [Figure out ways to help airplanes land safely.]
4	Engineering	Sven designed a parachute. When he created and tested it, he found that it fell too slowly. What would work BEST to make his parachute fall more quickly? [Make the canopy smaller.]
5	Science	On a planet the same size as Earth, where the atmosphere is thinner than it is on Earth, objects will: [Fall faster]
6	Science	David and Dipa took identical sheets of paper. They crumpled one into a ball and left one flat. They dropped them both at the same time. The crumpled paper ball dropped quickly straight down. The flat sheet of paper fell slowly and drifted to the ground. Why? [More air got in the way of the flat sheet of paper and slowed it down.]
7	Engineering	A parachute will work on a planet with no atmosphere. [False]
8	Science	Mars takes 365 days to circle the Sun. [False]
9	Science	Only Earth has an atmosphere. [False]
10	Science	Air can affect how things fall. [True]

Wilson constructed a model rocket. It didn't fly for as long as he wanted it to. He decided to change the shape of the fins on the rocket.

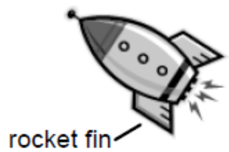
Based on his testing data (shown in the table), what should Wilson do to make his rocket fly longer?

Ⓐ Use rectangle-shaped fins.

Ⓑ Use curved fins.

Ⓒ Use triangle-shaped fins.

Ⓓ It is impossible to tell from the data.



Try #	Rocket Fin Shape	Length of Flight
1	Rectangle	7 seconds
2	Rectangle	8 Seconds
3	Rectangle	6 Seconds
4	Curved	7 Seconds
5	Curved	8 Seconds
6	Curved	7 seconds
7	Triangle	7 Seconds
8	Triangle	6 Seconds
9	Triangle	6 seconds

Figure 3. *Designing Parachutes* Question 2

3.2.2.2 Score Construction: "Designing Parachutes"

For the grade 3-5 assessment we started with a sample of 2113 students who returned a pre- and a post-assessment. On this assessment we only had three questions that were solely assessing engineering objectives, so we did not construct a score for this assessment. However, we constructed a science sub-score to ascertain whether EiE students learned science content better after participation in the program. Science scores were constructed by summing the number of items correct on the science questions. The science sub-scores had fairly low internal reliability (PreScience: Cronbach's alpha of .310 (n=1975); PostScience: Cronbach's alpha=.503 (n=2016)), so we did not investigate component structure, instead choosing to use the raw science scores as outcome and covariate. The overall scores for the pre-assessment (PreAll) and the post-assessment (PostAll) also had fairly low reliability (PreAll: Cronbach's alpha= .340 (n=1933); PostAll alpha=.574 (n=1991)). Again, we chose to forgo further investigation into component structure, instead using the raw scores as outcome and covariate.

As there are a total of ten questions on this instrument, we calculated an overall score if students had answered eight of them. If students had answered fewer than eight questions on either the pre- or the post-assessment they were dropped from further analysis. This reduced our sample size by 85 students.

3.2.2.3 Sample: "Designing Parachutes"

Designing Parachutes (PA) assessments were collected from students in grades 2 through 6. Grade 2 students received a different version of the assessment which is not discussed here. There were not enough grade 6 classrooms to include them in the sample.

The grades 3-5 sample of collected data for PA included 3344 students in 142 classrooms. Out of the 1081 students initially in the control group, 49.3% were dropped for missing a pre- or post-assessment, along with 30.8% of the 2263 students in the test group. Fourteen classes were excluded due to missing demographic information. Three classes were excluded due to missing information on how many minutes they had spent on the EiE curriculum. Additionally, eighty-five individual students were excluded because they answered fewer than 8 questions on either the pre- or post-assessment. Three classes were identified as statistical outliers using the methods described in the Methodology section and dropped. The final dataset used for analysis included 1200 students (Level-1 units) in 66 classrooms (Level-2 units), with an average cluster size of 18.18 and a standard deviation of 5.0.

The test group was about 60% larger than the control group. Both groups contained about 17% 3rd grade students and classrooms and a larger proportion of 5th grade students and classrooms.

Table 37. PA Treatment * Grade Crosstabulation (N of Students)

	Grade			Total
	3	4	5	
Control	86	101	273	460
Treatment	125	228	387	740
Total	211	329	660	1200

Table 38. PP Treatment * Grade Crosstabulation (N of Classrooms)

	Grade			Total
	3	4	5	
Control	5	6	15	26
Treatment	7	11	22	40
Total	12	17	37	69

Boys and girls made up approximately half of the sample for both the treatment and control groups (see Table 39). Both groups were comprised of approximately 8% students with LEP, 28% students receiving FRL, and 10% students with an IEP. The control group also had a slightly higher proportion of Hispanic students and a slightly lower percentage of white students (see Table 40).

Table 39. PA Proportions for Level-1 Dichotomous Variables

Treatment		Gender (male)	LEP	FRL	IEP
Control		.52	.06	.27	.07
Treatment		.53	.08	.29	.12
Total	Proportion	.53	.07	.28	.10
	N	631	84	338	121

Table 40. PA Proportions for Level-1 Variables – Race

Treatment		White	Asian	Black	Hispanic	Other	Total
Control	Proportion	.59	.06	.05	.26	.03	1.0
	N	273	29	25	118	15	460
Treatment	Proportion	.72	.06	.08	.10	.04	1.0
	N	531	47	57	73	32	740
Total	Proportion	.67	.06	.07	.16	.04	1.0
	N	804	76	82	191	47	1200

Thirty-seven teachers contributed to student assessments from their classrooms which were used in this analysis (see Table 41). Each teacher taught either control or test classrooms, but not both. Most schools included only one teacher participating in the field test of the *Designing Parachutes* unit, but some schools included more than one. The majority of teachers in both treatment groups were regular classroom teachers, not science specialists. On average, teachers had about 10 years of experience, though both novice and veteran teachers were included in both treatment groups. The mean class size was 18 students, quite similar across both treatment groups.

Table 41. PA Teacher and Class Demographics

	Number of Teachers	Number of Schools	Number of Science Specialists	Experience (Number of Years Teaching)				Class Size			
				Mean	SD	Min	Max	Mean	SD	Min	Max
Control	16	12	1	10.8	9.99	0	29	17.7	4.48	9	26
Test	21	19	4	8.6	8.50	0	32	18.5	5.31	8	31
Total	37*	29**	5	9.6	9.24	0	32	18.2	4.98	8	31

Table 39 shows the mean proportion of students in participating schools of each racial group, compared by test and control. The mean proportion of FRL students in schools is also shown. Schools with test students and schools with control students had similar proportions of Black and Asian students. The test group schools had a smaller mean proportion of Hispanic students and larger mean proportion of White students. Students receiving free or reduced-price lunch (FRL students) made up 40% on average of the schools with control participants, and 36% of schools with field test participants.

Table 42 shows the mean enrollment of schools where teachers and students participated in either control or test groups. Schools with participants in the control group tended to be larger than field test schools (Table 43). Table 44 shows that schools had similar distributions across settings, with approximately a third of schools classified as suburban (64% of control, 59% of test) and the rest distributed nearly equally between rural and urban settings.

Table 42. PA School Demographics: Proportions by Race, FRL

	Mean		St. Dev.		Minimum		Maximum	
	Control	Test	Control	Test	Control	Test	Control	Test
Black	0.09	0.09	0.07	0.08	0	0	0.23	0.30
Hispanic	0.30	0.18	0.29	0.21	0.04	0	0.89	0.62
Asian/Pacific Islander	0.04	0.07	0.05	0.11	0	0	0.17	0.54
White	0.52	0.63	0.35	0.32	0.01	0.03	0.93	1
Mixed/Other	0.04	0.03	0.04	0.03	0	0	0.10	0.10
FRL	0.40	0.36	0.36	0.30	0.01	0	0.96	0.90

Table 43. PA School Demographics: Enrollment and # Classrooms per School

	Mean		St. Dev.		Minimum		Maximum	
	Control	Test	Control	Test	Control	Test	Control	Test
Enrollment	850	638	369	233	433	230	1568	1078
# Classrooms participating per school*	2.6	2.2	2.2	1.4	1	1	8	6

***Three schools had both test and control classrooms.**

Table 44. PA School Demographics: Setting

	Number / % of Schools		Total
	Control	Test	
Rural	2 / 18%	4 / 18%	6
Suburban	7 / 64%	13 / 59%	18*
Urban	2 / 18%	5 / 23%	6*
Total	11	22	30*

***Two suburban schools and one urban school had both test and control classrooms, and were classified with both the test and the control groups.**

3.2.2.4 Results: "Designing Parachutes"

For the *Designing Parachutes* (PA) student assessment, researchers computed an overall score by counting the number of correct answers on all questions listed in Table 36, as described above. This score has a possible value range of zero to 10. The overall score on the post-assessment (PostAll) was used as the dependent variable, and the overall score on the pre-assessment (PreAll) as covariate, in an ANCOVA model modeled at two levels (student and classroom) using HLM (see Figure 4). There was only a 0.1 difference between control and test treatment groups on the mean PreAll scores; however, the test group had a mean PostAll score which was 1.8 points higher that the control group (see Table 45).

Table 45. PA Descriptive Statistics: PreAll and PostAll Scores by Treatment

	PreAll				**PostAll**			
	Mean	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**
Control	5.3	1.71	1	10	5.8	1.69	0	10
Test	5.4	1.65	1	10	7.6	1.91	1	10

Initial examination of intraclass correlation found that significant variance existed among classrooms (ρ=.412). Thus, researchers obtained sufficient evidence in favor of the utilization of HLM methodology.

For this unit, we tested the student level (level-1) variables of IEP, LEP, and Free or Reduced Price Lunch status, as well as gender and race (Asian, Black, Hispanic, and Other). We also included the PreAll (PreScience) score as a covariate. Next, we tested whether each level-1 variable was best modeled as a fixed or random coefficient, the default being a fixed coefficient, unless the random variance was found to be significant ($p < .05$). At the classroom level (level-2) we tested the effect of the treatment on the level-2 intercept, including the amount of instruction (EiE Minutes, EiE Minutes squared, and EiE Minutes cubed), classroom PreAll Means (PreScience Means), class size, and number of years teaching as covariates. We also tested the effect of grade and of teacher specialty (Science Specialist) here. For each of these variables, we simultaneously tested the variable's interaction with the treatment (*e.g.* PreAll Means by Treatment, Grade 3 by Treatment, Number of Years Teaching by Treatment, etc.), which were calculated by multiplying the variable by the treatment dummy variable, *i.e.* setting the variable to zero for control classrooms. This allowed us to test if the effect of these variables differed for the test and control groups. Finally, we tested the interaction of the treatment with any level-1 variables that were found to be statistically significant.

The two-level conditional model for the PA assessment (see Figure 4) includes variables and random variance coefficients which were found to be statistically significant ($p<.01$) using the above procedure. The final model shows evidence of a treatment effect (γ_{02} p=<.001) for non-inclusions students given the average numbers of minutes spent teaching EiE with a Cohen's *d* effect size of 1.292 (1.982 / $\sqrt{2.042 + 0.030 + 0.282}$ – see Table 46 and Table 47). To put the treatment's effect size in context, Individualized Education Program status (IEP) has an effect size of −0.483 (−0.689 / 1.429), Free or Reduced Price Lunch has an effect size of −0.260 (−0.372 / 1.429), Gender (male) has an effect size of 0.279 (0.399 / 1.429), and Is_Black has an effect size of −0.251 (−0.359 / 1.429).

When demographic coefficients from the finalized student level model were predicted by the classroom level treatment dummy, the treatment effects were significant for IEP students, indicating that the treatment effect is moderated by IEP status: here, IEP students appear to benefit less from the treatment than non-IEP students. This effect may, however, be associated with the very small sample size of IEP students. Predictors were not correlated with residuals, but PostAll and PreAll scores at the lowest and highest ranges were overestimated and underestimated, respectively. This suggests that the model is not fully capturing the differences between students.

The conditional model explains approximately 83% of the classroom-level variance (proportional reduction in variance statistic: 1 – (0.282 / 1.716) – see Table 47 and Table 49).

Level-1 Model:

$$\text{PostAll} = \beta_0 + \beta_1(\text{Gender}) + \beta_2(\text{IEP}) + \beta_3(\text{FRLunch}) + \beta_4(\textbf{PreAll}) + \beta_5(\text{Is\_Black}) + \text{r}$$

$$\text{Var}(r) = \sigma^2$$

and

$$\ln \sigma^2 = \alpha_0 + \alpha_1(\text{PreAllMean}) + \alpha_2(NumYrsTeaching\_by\_Treatment) + \alpha_3(\text{IEP}) + \alpha_4(\text{PreAll})$$

Level-2 Model

$$\beta_0 = \gamma_{00} + \gamma_{01}(PreAllMean) + \gamma_{02}(\text{Treatment}) + \gamma_{03}(EiE\_Minutes\_Taught) + \gamma_{04}(EiE\_Minutes\_Taught\_Squared) + \text{u}_0$$

$$\beta_1 = \gamma_{10}$$

$$\beta_2 = \gamma_{21}(\text{Treatment})$$

$$\beta_3 = \gamma_{30}$$

$$\beta_4 = \gamma_{40} + \text{u}_4$$

$$\beta_5 = \gamma_{50}$$

Figure 4. PA PostAll Score – Conditional Model

Table 46. PA PostAll Score Conditional Model – Final Estimation of Fixed Effects (with robust standard errors)

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. df	P-value
For Intercept 1, β_0					
Intercept 2, γ_{00}	5.707	0.111	51.498	61	< 0.001
PreAll Classroom Mean, γ_{01}	0.845	0.102	8.300	61	< 0.001
Treatment, γ_{02}	1.982	0.174	11.406	61	< 0.001
Minutes of EiE Instruction, γ_{03}	0.361	0.114	3.160	61	0.003
Minutes of EiE Instruction Squared, γ_{04}	-0.212	0.077	-2.743	61	0.008
For Gender slope, β_1					
Intercept 2, γ_{10}	0.399	0.080	4.962	1190	< 0.001
For IEP slope, β_2					
Treatment, γ_{21}	-0.689	0.187	-3.682	1190	< 0.001
For FR Lunch slope, β_3					
Intercept 2, γ_{30}	-0.372	0.129	-2.883	1190	0.004
For PreAll (Group mean centered) slope, β_4					
Intercept 2, γ_{20}	0.295	0.035	8.489	65	< 0.001
For Is Black slope, β_5					
Intercept 2, γ_{30}	-0.359	0.165	-2.168	1190	0.030

Table 47. PA PostAll Score Conditional Model – Final Estimation of Variance Components

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
Intercept 1, u_0	0.531	0.282	61	236.43	< 0.001
PreAll (Group mean Centered) slope, u_1	0.174	0.030	65	109.79	< 0.001
level-1, r	1.429	2.042			

Table 48. PA PostAll Score Conditional Model – Model for Level-1 Variance

Parameter	Coefficient	Standard Error	Z-ratio	P-value
Intercept 1, A_0	1.411	0.337	4.188	< 0.001
PreAll Classroom Mean, α_1	-0.162	0.061	-2.643	0.009
Number of Years Teaching by Treatment, α_2	0.015	0.005	3.265	0.001
IEP, α_3	0.374	0.143	2.604	0.010
PreAll, A_4	-0.089	0.030	-2.987	0.003

Table 49. PA PostAll Score Unconditional Model – Final Estimation of Variance Components

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
Intercept 1, u_0	1.31006	1.71627	65	875.08216	< 0.001
level-1, r	1.56482	2.44865			

Figure 5 shows a graph of the amount of time spent on EiE instruction (EiE_Minutes_Taught) against PostAll scores for the *Designing Parachutes* unit. The PostAll scores of the students in the test group are graphed against the z-score of the number of minutes of EiE instruction that they received. The black line represents the LOESS trend curve, while the red line represents the best-fit curve from the HLM analysis, and the minima and maxima of these curves are labeled accordingly. The equation for the curve from the HLM analysis is ($\text{PostAll} = .141(EiE\_Minutes) - .097(EiE\_Minutes)^2$). The recommended range of minutes (370-460) is represented with vertical lines.

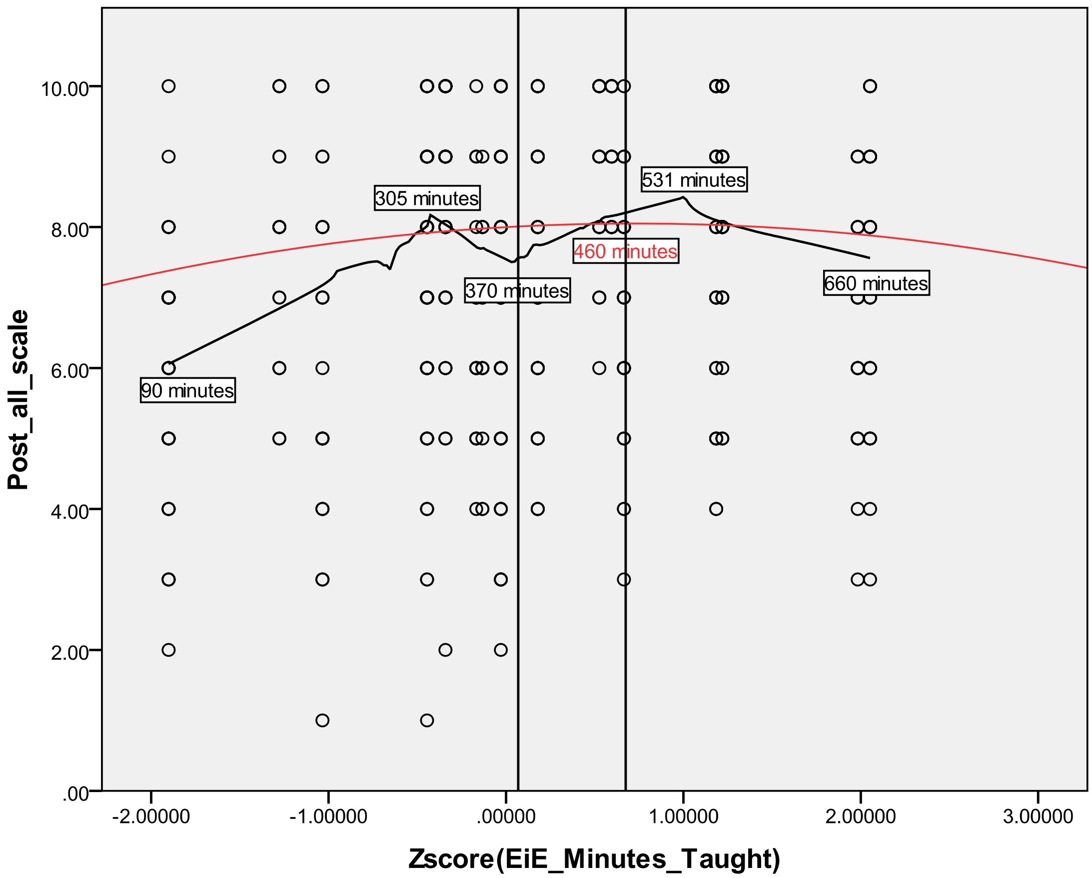


Figure 5. The Relationship Between Minutes of EiE Instruction and Students' PostAll Scores for *Designing Parachutes*

Figure 5 shows that the best scores on the *Designing Parachutes* assessment were attained by students who received between 460 and 531 minutes of instruction. This is just above the recommended range (460 minutes is exactly the highest recommended amount of instruction). The LOESS curve shows a slight peak at 305 minutes, and again a dip at exactly the lowest recommended amount of instruction, 370 minutes. This suggests that the relationship between the scores and the amount of instruction may be significantly influenced by confounding factors not accounted for in our model, such as teaching style and school environment. However, it is evident that additional instruction is not helpful after a certain point is reached.

Researchers also computed a Science Score by counting the number of correct answers on the science questions listed in Table 36 above. This score has a possible value range of zero to 5. The Science Score on the post-assessment was used as the dependent variable, and the Science Score on the pre-assessment as covariate, in an ANCOVA model modeled at two levels (student and classroom) using HLM (see Figure 6). Mean PreScience scores were similar between treatment, while mean PostScience scores were greater for the test group compared to the control group (see Table 50).

Table 50. PA Descriptive Statistics: PreScience and PostScience Scores by Treatment

	PreScience				PostScience			
	Mean	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**
Control	2.5	1.11	0	5	2.8	1.14	0	5
Test	2.6	1.11	0	5	3.9	1.11	0	5

When demographic coefficients from the finalized student level model were predicted by the classroom level treatment dummy, the treatment effects were not significant for any demographic group, unlike the PostAll score model. However, the treatment was found to be significantly more effective for those students with pre-test scores that were lower than their class average.

Level-1 Model:

$$\text{PostScience} = \beta_0 + \beta_1(\text{Gender}) + \beta_2(\text{IEP}) + \beta_3(\text{FRLunch}) + \beta_4(\mathbf{PreScience}) + \beta_5(\text{Is\_Black}) + r$$

$\text{Var}(r) = \sigma^2$ and

$$\ln \sigma^2 = \alpha_0 + \alpha_1(\text{Grade4}) + \alpha_2(\text{PreScienceMean}) + \alpha_3(\text{PreScience}) + \alpha_4(\text{Is\_Hispanic})$$

Level-2 Model:

$$\beta_0 = \gamma_{00} + \gamma_{01}(\textit{PreScienceMean}) + \gamma_{02}(\text{Treatment}) + \gamma_{03}(\textit{EiE\_Minutes}) + \gamma_{04}(\textit{EiE\_Minutes\_Squared}) +$$
$$\gamma_{05}(\textit{Class\_Size\_Squared}) + \gamma_{06}(\textit{Class\_Size\_Squared\_by\_Treatment}) + u_0$$

$$\beta_1 = \gamma_{10}$$

$$\beta_2 = \gamma_{20}$$

$$\beta_3 = \gamma_{30}$$

$$\beta_4 = \gamma_{40} + \gamma_{41}(\text{Treatment})$$

$$\beta_5 = \gamma_{50}$$

BOLD indicates group mean centered.
Italicized indicates grand mean centered.

Figure 6. PA PostScience Score – Conditional Model

The two-level final PostScience conditional model (see Figure 6) includes variables which were tested using the same procedure as described for the PostAll model, except that PreScience and PreScienceMean were used in place of PreAll and PreAllMean. This model explains approximately 91% of the classroom-level variance (proportional reduction in variance statistic: 1 – (0.048 / 0.516) – see Table 52 and Table 54). The conditional model for the PostScience score is slightly changed from that of the PostAll score. When inspecting the residuals of the model it became clear that class size was quadratically associated with the outcome variable. This was also true for the PostAll score, but the relationship was not quite statistically significant ($p = .06$). For the PostScience score, the relationship is significant (γ_{05}, $p < .001$). The class size interaction with the treatment is also significant (γ_{06}, $p = .001$), and indicates that while very small and very large classes had large science scores overall, very small and very large classes in the test group had better science scores overall.

The two-level PostScience conditional model shows evidence of a treatment effect (γ_{02} p<.001) for non-IEP students given the average numbers of minutes spent teaching EiE with a Cohen's *d* effect size of 1.046 ($1.045 / \sqrt{0.048+0.950}$ – see Table 52). To put the treatment's effect size in context, IEP status has an effect size of −0.338 (−0.329 / 0.974), free or reduced-price lunch status (FRL) has an effect size of −0.245 (−0.238 / 0.974), Gender (male) has an effect size of 0.272 (0.265 / 0.974), and Is_Black has an effect size of −0.272 (−0.266 / 0.974) – none are even half the magnitude of the treatment's effect size. For the PostScience score, the relationship is significant (γ_{05}, p<.001). The class size interaction with the treatment is also significant (γ_{06}, p= .001), and indicates that while very small and very large classes had large science scores overall, very small and very large classes in the test group had better science scores overall.

Table 51. PA PostScience Score Conditional Model – Final Estimation of Fixed Effects (with robust standard errors)

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. df	P-value
For Intercept 1, β_0					
Intercept 2, γ_{00}	2.954	0.072	41.055	59	< 0.001
PreScience Classroom Mean, γ_{01}	0.456	0.090	5.086	59	< 0.001
Treatment, γ_{02}	1.045	0.105	9.947	59	< 0.001
Minutes of EiE Instruction, γ_{03}	0.141	0.059	2.377	59	0.021
Minutes of EiE Instruction Squared, γ_{04}	-0.097	0.041	-5.220	59	0.020
Class Size Squared, γ_{05}	-0.280	0.054	-5.220	59	< 0.001
Class Size Squared by Treatment, γ_{06}	0.236	0.062	3.795	59	0.001
For Gender slope, β_1					
Intercept 2, γ_{10}	0.265	0.054	4.893	1187	< 0.001
For IEP slope, β_2					
Intercept 2, γ_{20}	-0.329	0.094	-3.494	1187	< 0.001
For FR Lunch slope, β_3					
Intercept 2, γ_{30}	-0.238	0.095	-2.519	1187	0.012
For PreScience (Group mean centered) slope, β_4					
Intercept 2, γ_{40}	0.309	0.053	5.837	1187	< 0.001
Treatment, γ_{41}	-0.159	0.062	-2.557	1187	0.011
For Is Black slope, β_5					
Intercept 2, γ_{50}	-0.266	0.128	-2.080	1187	0.037

Table 52. PA PostScience Score Conditional Model – Final Estimation of Variance Components

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
Intercept 1, u_0	0.219	0.048	59	130.55	< 0.001
level-1, r	0.974	0.950			

Table 53. PA PostScience Score Conditional Model – Model for Level-1 Variance

Parameter	Coefficient	Standard Error	Z-ratio	P-value
Intercept 1, α_0	0.408	0.261	1.568	0.117
Grade 4, α_1	0.320	0.095	3.377	0.001
PreScience Classroom Mean, α_2	-0.237	0.098	2.433	0.015
PreScience, α_3	-0.104	0.041	-2.525	0.012
Is Hispanic, α_4	0.253	0.117	2.157	0.031

Table 54. PA PostScience Score Unconditional Model – Final Estimation of Variance Components

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
Intercept 1, u_0	0.718	0.516	65	631.17	< 0.001
level-1, r	1.022	1.044			

3.2.2.5 Summary: "Designing Parachutes" Summative Evaluation

Our analysis suggests that students learning through a combination of EiE's *Designing Parachutes* engineering design unit and their astronomy science unit learn about engineering and science concepts, and that this learning is significantly greater than that of control students who participate in only the astronomy science unit. The effect of the EiE *Designing Parachutes* field test appears to be high on both the PostAll Score (*d*= 1.292) and the PostScience score (*d*= 1.046); however effect sizes are likely to be inflated due to heterogeneity of within-group variance at level 1. Positive effects of treatment were consistent across gender, race, and LEP status, indicating that the treatment has the same effect across these demographic variations. Though girls, black students, and those receiving FRL tended to have lower scores, these effects were the same for both EiE students and the control group. IEP students may have benefitted less from the treatment than other demographic groups as shown by analysis of the overall score (no such difference was found in analysis of the science score); however this finding may be an artifact attributable to the very small number of IEP students in this sample.

4 Conclusions

For both field test units we have evaluated in this report, surveys of implementing teachers revealed that teachers, overall, found the units worthwhile and indicated that they would teach them again in the future. Teachers most often commented on students' enjoyment of the unit, and especially noted the role of hands-on activities in increasing student engagement. Teachers also reported that the units helped to promote discussion and increase students' skills in critical thinking and problem solving. Additionally, teachers credit the unit with helping students learn about engineering design, as well as improved understanding of science content.

In our summative assessment, we have found effect sizes for both overall scores and science sub-scores associated with the test treatment (EiE). These effect sizes ranged in size from moderate to large. Demographic variables including whether a student received free or reduced-price lunch (FRL), has an Individualized Education Program (IEP), or is from an underrepresented minority group (Is_Black) were associated with poorer performance on both pre- and post-assessments, but this relationship was not moderated by treatment: the difference was the same for both the control and test groups.

4.1 Discussion

4.1.1 Formative Assessment

Results from teacher feedback forms indicate that teachers feel the EiE units help to further their students' understanding of both science and engineering content. For both units, teachers reported that their students learned unit-specific science and engineering content and were better able to recognize engineering in everyday life. Additionally, for the *Designing Parachutes* unit, teachers noted that their students gained a deeper understanding of the work of an engineer. It was also reported that EiE helps students learn other important skills. For both units, teachers described how their students' problem solving, critical thinking, and communication skills increased. For *Designing Parachutes*, teachers also mentioned that students' teamwork skills were strengthened (see Table 55).

Additionally, it was noted for both units that students enjoyed the unit and were engaged. For *Designing Parachutes*, teachers specifically mentioned that the hands-on activities in Lessons 2 and 4 were very engaging for students. This suggests that engaging students in hands-on activities may help to promote learning.

Finally, teacher feedback results indicate that teachers utilized the interdisciplinary nature of the curriculum by using an EiE unit to further objectives in other content areas. For both units, teachers rated highly EiE's reinforcement of science learning objectives (mean ratings of 6.1 to 6.4 out of 7 for *Designing Solar Ovens* and *Designing Parachutes*, respectively). Integration with language arts and math was also mentioned for both units. The vast majority of teachers reported that they would teach the unit again, indicating that this was due in large part to the fact that EiE enhanced the curriculum and connected to other content areas. This suggests that EiE is succeeding in the goal of maximizing cross-disciplinary ties in order to make units accessible to typical classrooms and engaging for students.

Teacher feedback highlights that the units reported here exhibit many of the characteristics described by Dewey (1938, 1998). Teachers appreciated the interdisciplinary nature of the units and reported that their students were fully engaged in meaningful hands-on activities and made connections to engineering in their every day lives. In addition, summative assessment results suggest that students learning through a combination of EiE and a traditional science curriculum learn both engineering and science concepts better than students taught the science curriculum alone. The results reported here reinforce the notion that providing a learning environment that engages students in a realistic disciplinary context can help to foster learning and understanding of the concepts being taught.

Table 55. Summary of Coding of Feedback Form Responses—Comparative by Unit

		SO	PA	# Units	Overall # Units
Fun/engaging	Unit	X	X	2	2
	L1	X	X	2	
	L2	X	X	2	
	L3		X	1	
	L4	X	X	2	
Increased science/engineering knowledge	Unit	X	X	2	2
	L1			0	
	L2			0	
	L3		X	1	
	L4			0	
Recognized engineering in everyday life	Unit		X	1	2
	L1			0	
	L2	X		1	
	L3			0	
	L4			0	
Problem solving and critical thinking skills increased	Unit	X	X	2	2
	L1		X	1	
	L2			0	
	L3			0	
	L4			0	
Hands-on activities increased engagement	Unit	X	X	2	2
	L1			0	
	L2		X	1	
	L3			0	
	L4		X	1	
Promoted teamwork	Unit		X	1	1
	L1			0	
	L2			0	
	L3			0	
	L4			0	
Promoted good discussions	Unit			0	2
	L1	X	X	2	
	L2			0	
	L3			0	
	L4			0	
Learned about the work of an engineer	Unit		X	1	1
	L1			0	
	L2			0	
	L3			0	
	L4			0	
Connected to other content area	LA	X	X	2	2
	Sci	X	X	2	
	SS			0	
	M	X	X	2	

Key: L1 = Lesson 1, etc.; LA = Language Arts, Sci = Science, SS = Social Studies, M = Math

4.1.2 Summative Assessment

In general, we found that the *Engineering is Elementary* curriculum appears to have a moderate to high effect on both the overall (engineering plus science) scores and the science-only scores of our assessments (Table 56 and Table 57). Some of this is likely to be due to over-alignment between the assessments and the treatment, because of a lack of alternative engineering instruction for the control group. However, this is unlikely to account for all of the difference, since teachers from both groups reported that they taught the required science instruction, and science questions comprised approximately half of the assessment questions. It is also possible that control teachers for some reason related to our non-random methods of recruitment provided less or poorer-quality science instruction, a possibility for which we do not have the data necessary to refute. What we do know is that our analysis of the *Designing Parachutes* unit shows that the treatment appears to have a similar effect on science scores as it does on overall scores, suggesting (but not proving) that EiE does cause students to learn science concepts better when compared to a traditional science curriculum.

The differences between treatments also may have been due to independent, unmeasured differences between teachers that followed from our use of convenience samples, even though we attempted to keep the groups similar by using some field test teachers for control sampling (by having them complete data collection before beginning the unit) and by offering similar incentives to both treatment groups, with the only difference being timing of receipt of professional development and materials (after data collection for the control group). Therefore, we treat the effects we found with caution.

There were some instances in which the effect of the treatment was moderated by demographic variables. For the *Designing Parachutes* unit, the treatment was less effective for IEP students' overall scores, though it still affected their scores positively. Additionally, the science scores in *Designing Parachutes* were boosted more heavily for students who had lower pre-assessment scores (relative to their class). It is unclear if these effects are due to sampling bias, or if they will persist in future randomized studies.

There were also instances in which the treatment interacted with other classroom-level predictors. For *Designing Parachutes*, class size was negatively associated with the outcome variable, but within the test group, class size was positively associated with the outcome variable. We have no theory as to why these results are so variable, since we would have expected class size to be negatively associated with performance (with smaller classes showing better outcomes than larger classes).

The treatment also had interactions with teacher experience. For *Designing Solar Ovens*, teacher experience correlated positively with post-assessment scores within the test group. Again, it is unclear if this effect is an artifact of sampling.

Table 56. Effect Sizes: Student-Level Variables—Comparative Across Units

Unit	Outcome Variable	Treatment (L2 variable for comparison)	IEP	LEP	Race				Gender (Male)	FRL*	Notes
					Black	Asian	Hispanic	Other			
SO	PostAll	0.689	−0.344	*	−0.316	*	*	−0.445	−0.230	NA	
PA	PostAll	1.292	−1.298	*	−0.251	*	*	*	0.279	−0.260	Level-1 variance modeled, but within-class variance remained heterogeneous (p= .006)
	PostScience	1.046	−0.338	*	−0.272	*	*	*	0.272	−0.245	Level-1 variance modeled, but within-class variance remained heterogeneous (p<.001)

***Variable tested but not found to be significant.**
****NA = not applicable. Due to small sample sizes FRL was dropped for SO.**

Table 57. Effect Sizes: Classroom-Level Variables and Interactions

Unit	Outcome	Treatment Effect	Pre-assessment Classroom Mean	Teacher Demographics	Grade	Class Size	Minutes of EiE Instruction*
SO	PostAll	0.689	PreAllMean = 0.570	Number of years teaching by treatment = 0.138	*	*	NA
PA	PostAll	1.292	PreAllMean = 0.551	*	*	*	Minutes = 0.235 Minutes_squared = −0.138
	PostScience	1.045	PreScienceMean = 0.456	*	*	Class size squared = −0.280 Class size squared by treatment = 0.236	Minutes = 0.141 Minutes_squared = −0.0968

***Variable(s) tested but not found to be significant.**
****NA = not applicable. Due to small sample sizes, minutes was dropped for SO.**

For both units, we found that our model both overestimated the scores of those who did worst on the post-assessment, and underestimated the scores of those who did best (see Figure 7). However, the residuals of the post-assessment scores were not correlated with the pre-assessment scores (see Figure 8). While this figure shows a slight upward trend at both ends, overall there is very little pattern in the residuals, indicating that the relationship between the PreAll and PostAll scores (and, where tested, the PreScience and PostScience scores) have been accurately modeled. As there was no correlation with any of our other predictors, this indicates that something is missing from our model: we have not adequately modeled how the rate and depth of learning differ between students. While this does not change our results regarding the overall efficacy of *Engineering is Elementary*, it does show that there is room to increase our understanding of what factors are at play when students are learning the material. This can also be seen in the models themselves—often the relationship between pre- and post-assessment scores is best modeled with a randomly varying coefficient rather than a fixed one, indicating that that relationship is very different between classrooms.

We also investigated whether the amount of time spent teaching the *Engineering is Elementary* curriculum to students affected their scores. Unfortunately, we could only include this variable in the analysis for the *Designing Parachutes* unit as we did not have this data for enough classrooms to include it in the analysis of *Designing Solar Ovens*. For *Designing Parachutes*, we found that the relationship between the amount of EiE instruction and the students' scores was curvilinear. This was first found by visually inspecting the graphs before running the HLM analysis, and including square and cubic terms in the analysis as a result. Indeed, the HLM analysis showed that the best model was not linear for either unit.

The highest PostAll scores were correlated with a median amount of instruction very roughly corresponding to the recommended time of instruction for the unit. For *Designing Parachutes*, the highest PostAll scores were associated with EiE instruction of between 460 and 531 minutes—see Figure 5. EiE's recommended instructional time for the *Designing Parachutes* unit is 370 to 460 minutes, which is slightly less than what was found. These findings suggest that the relationship between student scores and the amount of instructional time spent on EiE may be significantly influenced by confounding factors not accounted for in our model, such as teaching style, amount of science instructional time, and school environment. Also, it appears that additional instruction may not be helpful after a certain point is reached.

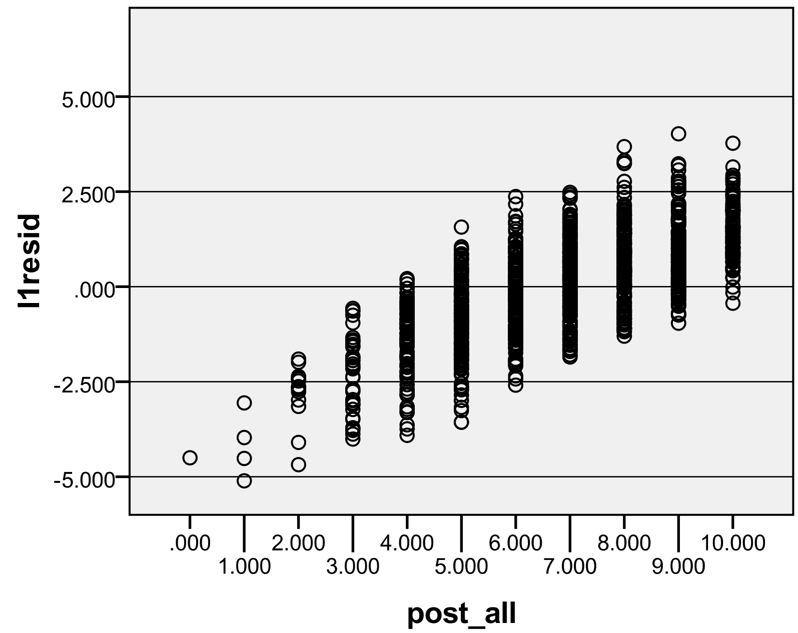


Figure 7. Level-1 (Student) Residuals vs. the PostAll score for *Designing Parachutes*.

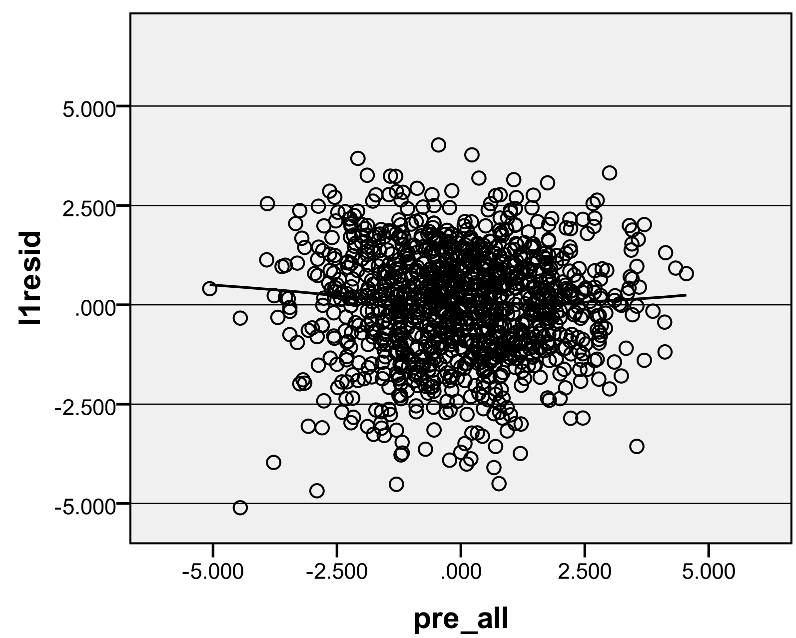


Figure 8. Level-1 (Student) Residuals vs. the PreAll Score for *Designing Parachutes*

4.2 Recommendations

Because the evaluations of student learning we conducted for both of the units analyzed in this report were conducted with available samples and not subject to randomized assignment, we cannot make causal attributions to our findings. However, we feel that the findings, both formative and summative, are strong enough to be suggestive that EiE impacts student learning of both science and engineering. Teachers, without being prompted to mention science or engineering, consistently listed these as academic benefits of the units. For most units, they also offered their beliefs that the units helped students to learn problem solving skills and critical thinking skills, promoted discussions and teamwork, and increased engagement.

Analysis of student data consistently showed that EiE was associated with statistically significant effects on assessments of students' science and engineering knowledge, and effect sizes that ranged from moderate to large. Though the short assessments, with their low internal reliability, were a definite liability for this evaluation, still the consistency observed implies that the effects are likely to be persistent. EiE appears, for the most part, to work well for different demographic groups equally, including students with limited English proficiency, students receiving free or reduced-price lunch, students of both genders, and underrepresented minorities. In addition, EiE appears to work slightly better in the hands of experienced teachers.

Further study is required to establish EiE as causing improvements in students understanding of science and engineering, and to better determine the effects of moderating variables, the reasons for those effects, and the features of the program that are most important for leading to increased student learning in engineering and science.

5 References

Banilower, E., Smith, P., Weiss, I., & Pasley, J. (2006). The status of K-12 science teaching in the United States. In D.W. Sunal & E. Wright (Eds.), *The impact of state and national standards on K-12 science teaching* (pp. 83-122). United States of America: Information Age Publishing, Inc.

Bruner, J. (2004). The psychology of learning: A short history. *Daedalus*, *133*(1), 13-20.

Bryk, A., & Raudenbush, S. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Cunningham, C. M., & Hester, K. (2007). Engineering is Elementary: An Engineering and Technology Curriculum for Children. Presented at the American Society for Engineering Education Annual Conference & Exposition, Honolulu, HI: American Society for Engineering Education.

Cunningham, C. M., & Lachapelle, C. P. (2010). The impact of Engineering is Elementary (EiE) on students' attitudes toward engineering and science. Presented at the ASEE Annual Conference and Exposition, Louisville, KY.

Dewey, J. (1938). *Experience and education*. Indianapolis: Kappa Delta Pi.

Driver, R. (1989a). Students' conceptions and the learning of science. *International Journal of Science Education*, *11*(5), 481-490.

Driver, R. (1989b). The construction of scientific knowledge in school classrooms. In R. Millar (Ed.), *Doing science: Images of science in science education* (pp. 83-106+). The Falmer Press.

Duschl, R. (2008). Science education in three-part harmony: Balancing conceptual, epistemic, and social learning goals. *Review of Research in Education*, *32*(1), 268-291.

Duschl, R., & Grandy, R. (2008). *Teaching scientific inquiry: Recommendations for research and implementation*. Rotterdam, The Netherlands: Sense Publishers.

Engle, R. A., & Conant, F. R. (2002). Guiding principles for fostering productive disciplinary engagement: Explaining an emergent argument in a Community of Learners classroom. *Cognition and Instruction*, *20*(4), 399-483.

Gruber, H., & Vonèche, J. (1977). *The essential Piaget*. Routledge & Kegan Paul.

Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist*, *42*(2), 99-107.

Jones, A. (1997). Recent research in learning technological concepts and processes. *International Journal of Technology and Design Education*, *7*(1), 83-96.

Kolodner, J. L., Camp, P. J., Crismond, D., Fasse, B., Gray, J., Holbrook, J., Puntambekar, S., et al. (2003). Problem-Based Learning meets Case-Based Reasoning in the middle-school science classroom: Putting Learning by Design™ into practice. *The Journal of the Learning Sciences*, *12*(4), 495-547.

Lachapelle, C. P., & Cunningham, C. M. (2010). Assessing elementary students' understanding of engineering and technology concepts. Presented at the ASEE Annual Conference and Exposition, Louisville, KY.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge, UK: Cambridge University Press.

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, *66*, 211-227.

Raudenbush, S., Bryk, A., Cheong, Y., Congdon, R., & du Toit, M. (2004). *HLM 6: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.

Rogoff, B. (1990). *Apprenticeship in thinking: Cognitive development in social context*. New York, NY: Oxford University Press.

Rosebery, A. S., Warren, B., & Conant, F. R. (1992). Appropriating scientific discourse: Findings from language minority classrooms. *Journal of the Learning Sciences*, *2*(1), 61-94.

Sawyer, R. (2006). *The Cambridge handbook of the learning sciences*. Cambridge, UK: Cambridge University Press.

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.