J Sci Educ Technol (2017) 26:207–222
DOI 10.1007/s10956-016-9673-x

The Impact of Design-Based STEM Integration Curricula on Student Achievement in Engineering, Science, and Mathematics

S. Selcen Guzey[1] · Michael Harwell[2] · Mario Moreno[2] · Yadira Peralta[2] · Tamara J. Moore[3]

Published online: 29 November 2016

Abstract The new science education reform documents call for integration of engineering into K-12 science classes. Engineering design and practices are new to most science teachers, meaning that implementing effective engineering instruction is likely to be challenging. This quasi-experimental study explored the influence of teacher-developed, engineering design-based science curriculum units on learning and achievement among grade 4–8 students of different races, gender, special education status, and limited English proficiency (LEP) status. Treatment and control students (n = 4450) completed pretest and posttest assessments in science, engineering, and mathematics as well as a state-mandated mathematics test. Single-level regression results for science outcomes favored the treatment for one science assessment (physical science, heat transfer), but multilevel analyses showed no significant treatment effect. We also found that engineering integration had different effects across race and gender and that teacher gender can reduce or exacerbate the gap in engineering achievement for student subgroups depending on the outcome. Other teacher factors such as the quality of engineering-focused science units and engineering instruction were predictive of student achievement in engineering. Implications for practice are discussed.

✉ S. Selcen Guzey
sguzey@purdue.edu

[1] Department of Curriculum and Instruction and Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA

[2] Department of Educational Psychology, University of Minnesota, Minneapolis, MN 55455, USA

[3] School of Engineering Education, Purdue University, West Lafayette, IN 47907, USA

Keywords Engineering curriculum · Engineering integration · STEM · Student learning

Reports from the National Academy of Engineering (NAE 2014) and the National Research Council (NRC 2010, 2011) have emphasized the importance of improving K-12 science and mathematics education in the USA to motivate more students to pursue science, technology, engineering, and mathematics (STEM) fields in college. An especially promising model is the integrated STEM education, defined by the merging of the disciplines of science, technology, engineering, and mathematics in order to (a) deepen student understanding by contextualizing concepts, (b) broaden student understanding through exposure to socially and culturally relevant STEM contexts, and (c) increase interest in STEM disciplines and expand the pathways for students to enter STEM fields (Guzey et al. 2014; Moore et al. 2014). The premise is that integrated STEM education will play a critical role in increasing US competitiveness in the global economy.

In K-12 science classrooms, integrated STEM education most often refers to the use of engineering design and practices as a vehicle to teach science and mathematics. Recent national documents such as the Framework for K-12 Science Education (NRC 2012) and the Next Generation Science Standards (NGSS Lead States 2013) recognize engineering as an important element in the new vision for science education. Throughout grades K-12, students are expected to "actively engage in scientific and engineering practices and apply cross-cutting concepts to deepen their understanding of the core ideas in these fields" (NRC 2012, p. 8–9). The intent is to encourage more K-12 students to explore engineering design, learn about interconnections of science and engineering, and apply science knowledge and skills to solve engineering challenges in their science classes.

However, this new vision of science education is challenging to implement for most science teachers. Lack of knowledge about engineering, time constraints, lack of quality teaching materials to teach engineering, unavailable resources, and an unsupportive school structure frequently limit the teachers' efforts to successfully integrate engineering into their science teaching (Dare et al. 2014; Guzey et al. 2014; Nelson et al. 2015; Riskowski et al. 2009; Wang et al. 2011). Another limiting factor is the lack of clarity about approaches to engineering integration and the effectiveness of these approaches on student learning and achievement. From a practical perspective, science teachers need more guidance and support in effectively using a range of engineering and science practices. From a research perspective, rigorous studies are needed to show if engineering integration supports student learning and achievement in science and if so, in what ways. Findings of the effectiveness of engineering in supporting science learning is mixed at the elementary (Lachapelle et al. 2011; Wendell and Rogers 2013), middle (Cantrell et al. 2006; Mehalik et al. 2008; Riskowski et al. 2009; Schnittka and Bell 2011), and secondary school levels (Apedoe et al. 2008; Berland et al. 2014; Tran and Nathan 2010; Valtorta and Berland 2015). These studies also tend to vary substantially in quality.

This study reports results from a large-scale National Science Foundation (NSF), Mathematics and Science Partnership (MSP) project whose purpose is to increase student learning of engineering, science, and mathematics (data analysis and measurement) concepts in grades 4–8 using an engineering design-based approach to the teachers' professional development and curricular development. Over 200 teachers and 15,000 students from three partner school districts in a Midwest state are involved in this ongoing 5-year project. The current study presents findings from the first year of the project in which 59 teachers and 4450 students participated. Our main research question asked "In what ways does participation in the engineering design-based science curriculum affect the students' content knowledge in the STEM disciplines?" A second important question was "Does teacher participation in professional development and curricular development reduce gaps in achievement among students of different races, gender, special education status, and limited English proficiency status?"

Background

Quality science and engineering integration focuses on designing effective learning experiences that allow students to actively engage in their learning, participate in collaborative problem-solving using real-world problems or situations, develop disciplinary knowledge and skills, and make connections across disciplines (NRC 2012). Thus, the integration of engineering design and practices in K-12 science classrooms fits well with the theory of situated learning, which views learning as the development and use of knowledge and practices in an authentic activity that involves social interaction and collaboration among learners (Brown et al. 1989; Cobb and Bowers 1999). In the case of engineering and science integration, the success of student learning in situated learning experiences relies on purposeful and meaningful design-based science activities and social interactions.

At the K-12 level, engineering education focuses on engineering design processes, application of science and mathematics in engineering, and engineering habits of mind (NRC 2009). Engineering design is a critical aspect of engineering whose goal is to identify and solve problems. Engineering design is "(1) highly iterative; (2) open to the idea that a problem may have many possible solutions; (3) a meaningful context for learning scientific, mathematics, and technological concepts; and (4) a stimulus to systems thinking, modeling, and analysis" (NRC 2009, p. 4). Engineering uses knowledge from different disciplines, such as science and mathematics and applications of science and mathematics concepts to solve engineering problems and supports instructional efforts to help students make connections across disciplines and engage in practices from different disciplines. Habits of mind refer to twenty-first century skills such as systems thinking, creativity, collaboration, and communication. Using these three elements of engineering education (design, application of science and mathematics, and habits of mind) effectively is important in efforts to improve student learning of STEM subjects.

Several studies have shown the benefits of engineering education and the positive effects of student participation in engineering design in K-12 science classrooms (e.g., Berland et al. 2014; Brophy et al. 2008; Carlson and Sullivan 2004; Lachapelle and Cunningham 2014; Wendell and Rogers 2013). For example, Wendell and Rogers (2013) found that students who engaged in engineering-based science units demonstrated greater science content knowledge compared to students who did not participate in engineering units in science classrooms. Similarly, in an efficacy study of elementary engineering curricular units, Oh et al. (2016) tested the influence of engineering design units on student learning of science. Students in treatment classrooms engaged in an engineering unit in conjunction with the related science unit, whereas control students participated in a related science unit. All students completed pre and post content assessments. These authors found that treatment students learned significantly more about science concepts than control students. However, students with special needs, low-income students and students with limited English proficiency (LEP), had lower scores on the tests than mainstream students. In addition, qualitative studies that focused on student design discourse in engineering-based science units provided promising results related to improved student decision-making and scientific

thinking as a result of participating in engineering design activities (Azevedo et al. 2015).

Other studies have raised questions about the effectiveness of engineering on student learning of science and mathematics. For example, Tran and Nathan (2010) reported no measurable advantage on science assessments for students who attended classes focused on engineering. Related research has also demonstrated that students have high motivation to engage in the qualitative aspects of engineering (e.g., identifying the problem) but little motivation to explore or apply the science and mathematics content necessary to solve engineering challenges (Berland et al. 2014). It is important to point out that a common thread in research of the impact of engineering integration in science classrooms is the scarcity of large-scale studies that examine the relationship between engineering integration and student outcomes. A few large-scale studies examining outcomes related to engineering are available (e.g., Lachapelle et al. 2011) but more are needed to replicate and generalize findings.

It is also important to note that much of the existing research does not distinguish between different approaches to engineering integration and curriculum approaches (NAE and NRC 2014). There are several ways to integrate engineering; however, little is known about the effectiveness of each strategy on student learning. For example, engineering can be used as a context to teach science content or engineering content that can be taught as part of a science unit (Moore et al. 2014). The difference between these two strategies lies in the goals and objectives of engineering integration. Furthermore, integration strategies can be grouped based on when the engineering concepts are presented in a science course. For example, various aspects of engineering can be infused into a single science unit. Students may learn about force and motion and then design a car or roller coaster under constraints (e.g., budget, materials). In another strategy, a teacher may embed different elements of engineering in each science unit so students can explore engineering throughout a science course. Regardless of the type of engineering approach used if engineering integration is not made explicit, then students often do not make the connections between engineering and science. Without these connections, the motivations for learning science to aid with the engineering design (and vice versa) are easily lost. In general, teachers need to provide opportunities that make engineering and science connections explicit to help students increase their knowledge in both disciplines (NAE and NRC 2014).

Because many teachers do not have adequate experience and knowledge in different approaches to engineering integration and curriculum design, providing opportunities for teachers to engage in engineering design and to learn to use effective strategies to teach engineering in science classrooms is critical. Previous research has demonstrated the relationship between participation in professional development and student achievement growth (Desimone 2011; Desimone et al. 2013; Garet, Porter, Desimone, Brimna, Yoon 2001; Wilson 2013). It is also well known that long-term teacher professional development specifically linked to classroom lessons or particular instructional approaches is more likely improve teacher practices and foster student learning. However, as Desimone and Garet (2015) argue "teachers vary considerably in their response to the same [professional development]" (p. 255) which can lead to variation in student outcomes. For example, in the case of engineering education, teachers come to professional development with varying levels of experience and knowledge in engineering design and practices. These factors, in addition to classroom factors such as classroom context, influence what teachers learn from the professional development activities and how effectively they transfer new knowledge and practices into their instruction.

The Study

At the onset, teachers from three partner school districts participating in the project received intensive 3-week-long summer professional development during which they designed engineering-focused STEM integration curricular materials corresponding to their science teaching area (e.g., seventh grade life science). These materials were linked to and guided by state and national standards. Teachers who developed curricular materials through the professional development comprised the "treatment" condition and subsequently taught using the curricular materials they developed. Teachers who agreed to participate in the study but did not participate in the professional development served as a "business as usual" control condition that taught curricula using state and national standards but did not use engineering as a vehicle to teach science and related mathematics. Available evidence suggests that in the business as usual approach science and mathematics are taught as separate disciplines, and, thus, connections among engineering, mathematics, and science are not emphasized; in the treatment condition, these connections are emphasized.

Population and Sampling

The target population consisted of upper elementary and middle school students and classrooms in the USA in public schools, whereas the sampled populations consisted of students and classrooms in grades 4–8 in a Midwest state. Our sample included three school districts, one of which was urban with a diverse student and teacher population in terms of race and socioeconomic status, one was a boundary district straddling both urban and suburban neighborhoods with less but still substantial diversity, and one was a suburban district with decidedly less diversity. Data from 42 treatment teachers, 17 control teachers, and 4450 students were available.

The sampling of school districts was purposive in the sense that districts varied in ways we think that enhance generalizability to the target populations. Shadish et al. (2002) pointed out that purposive sampling does not ensure generalizability, but we argue that our samples represents a cross section of schools, teachers, and students similar to many in the USA. To provide general evidence of this assertion, we turned to several educational indicators that allowed us to make broad comparisons between the characteristics of the upper Midwest state where our samples were taken from (blinded for review) and those of the USA as a whole.

For example, the average pupil/teacher ratio in the state (blinded for review) in 2007 (15.8) was quite similar to that for the USA as a whole (15.5) (U.S. Department of Education 2010), as was per pupil spending in 2012 for elementary and secondary students in the state (US$10,796) versus the USA as a whole (US$10, 208) (U.S. Bureau of the Census 2014). Scores on the National Assessment of Educational Progress (NAEP) for fourth and eighth grade students in the state are on average among the highest in the USA (U.S. Department of Education 2014a, b). However, high school graduation rates for the state are slightly lower (77%) than for the USA (79%) (U.S. Department of Education 2014d). Demographically, the state is less diverse than the USA as a whole. For example, the state has a smaller percentage of children 5–17 years old living in poverty (12.9%) than the USA (21%) and a higher percentage of whites (86.2%) and a smaller percentage of blacks (5.7%) compared to the USA (77.7, 13.2%) (U.S. Department of Education 2014c). This provides some evidence for generalizing our findings to states and school districts that (a) resemble the USA as a whole on key educational indicators like per pupil spending and average pupil/teacher ratio and (b) show higher than average achievement and are less racially diverse with fewer children in poverty than the USA as a whole.

Research Design

The study employed a prospective cohort (nonequivalent, pretest-posttest, quasi-experimental) cluster design for cross-sectional data in which classrooms (teachers) represented clusters and were self-selected into the treatment condition. Following What Works Clearinghouse (2014) guidelines for quasi-experimental designs, we used control variables in our statistical models to account for preexisting differences between the treatment and control conditions (see below) as well as matching. This helped to ensure, but does not guarantee, credible inferences.

Variables

Outcomes The most important outcome variables were project-constructed assessments capturing achievement in engineering, mathematics (data analysis and measurement), and science that were designed to be sensitive to the engineering design-based science curricula that teachers developed and taught. These assessments were developed, scaled, and validated following the process described in the Standards for Educational and Psychological Testing (American Educational Research Association 1999; Harwell et al. 2015). Briefly, a team of classroom teachers, school curriculum specialists, and academic researchers with collective expertise in engineering, science, and mathematics developed the assessments. The team clarified the purpose of each test and carefully described the knowledge domain to be sampled. A curricular map of topics consistent with state standards embedded in the teacher-constructed curricula was then developed. All science and mathematics items on each content test (e.g., middle school ecosystems and environments; elementary school plate tectonics) were obtained from public item banks linked to the Trends in International Mathematics and Science Study (TIMMS), the NAEP, and the American Association for Advancement of Science (AAAS). Central to selecting items was the requirement that they assess topics and knowledge within a contain domain consistent with the curricular map developed earlier. Items in some cases were modified slightly to be consistent with state standards.

All engineering items were developed by the authors following the Standards for Educational and Psychological Testing (American Educational Research Association 1999). These items were initially mapped to the K-12 engineering education framework (Moore et al. 2014) and focused on the students' knowledge about engineering design processes, conceptions of engineering and engineers, and engineering habits of mind. Engineering items were piloted and then calibrated using Rasch IRT modeling as described by Harwell and colleagues (2015).

All items were multiple choice and for middle school, the assessment consisted of 45 items with 15 items dedicated to each content area of engineering, science, and mathematics; for elementary school, the assessment included 30 items with 10 items dedicated to each content area of engineering, science, and mathematics. Rasch item response theory model was used to perform an equating study to ensure that the elementary and middle school student's engineering and mathematics/data analysis test scores reflected the same proficiency (Kolen and Brennan 2004); science tests were not equated because none included anchor items and were analyzed separately for elementary and middle school students.

Student test scores were reported in logits which are widely used in Rasch analyses of test data and estimate a student's proficiency on an outcome. In general, a logit represents the natural log of the ratio of the number of items answered correctly divided by the number of items answered incorrectly (Ludlow and Haley 1995). Thus, a logit of zero can generally be interpreted to mean that a student answered half of the items correctly, a positive logit to mean that a student

answered more items correctly than incorrectly, and a negative logit to mean that they answered fewer items correctly than incorrectly. Logits have a monotonic relationship with raw total correct scores.

Because of the project's focus on integrated STEM instruction and assessment, all students took the engineering and mathematics assessments as a single test that was administered at the beginning and end of the engineering design-based science unit in which these topics are covered. Several science-oriented assessments were also developed. Three of these assessments were similar in content for elementary and middle school students (i.e., heat transfer, plate tectonics, ecosystems) and one was different (erosion for elementary school students, particle theory for middle school students). These tests were given to students before and after the new curricular materials. Thus, both pretest and posttest data were available for these assessments, with posttest data serving as the outcome.

Scores obtained at the end of the school year for a state-mandated assessment were also available and served as a general measure of student achievement in mathematics. However, comparisons of state-mandated test scores among elementary and middle school students were not equated and thus, student grade level differences (elementary vs. middle school) must be taken into account. Altogether, there were three outcome variables taken by all sampled students (project-constructed assessments in engineering and mathematics/data analysis, state-mandated mathematics test). For the science portion, project-constructed assessments in science were taken by subsets of elementary or middle school students, depending on the topic covered in their curricular unit.

Predictors Student predictors consisted of gender (0 = male, 1 = female), race (black, Asian, Hispanic, with white students serving as the reference group), and a pretest for the content tests (e.g., engineering pretest). Classroom (teacher) predictors were treatment (treatment, control), years of teaching experience, years in current position, years in current school, and percentage of special education and LEP students. The discontinuous and coarse nature of the distributions of the latter two variables prompted us to rescale them to better capture the patterns in these data. Accordingly, we rescaled the percentage of special education and LEP students into quartiles and used these values (1, 2, 3, 4) in the modeling. Level (0 = elementary, 1 = middle school) was also used as a classroom predictor to capture any remaining differences due to grade level for project-constructed assessments after equating. For the state-mandated mathematics test, we relied on level to take grade differences into account. The presence of a modest amount of missing teacher data (<12%) prompted us to use data imputation under the assumption these values were missing at random (Little and Rubin 2002)—otherwise, teachers with any missing data and their students would be omitted from the multilevel analyses.

We also included the following three classroom/teacher variables in our analyses: quality of teaching, quality of curriculum unit, and type of engineering integration. These variables helped to take into account the fact that each treatment teacher developed a different curriculum for the same content, a process that was central to the project rationale and consistent with project goals. Although the curricula developed by the teachers used the same principles and guidelines, the net effect is that the treatment varied somewhat across treatment teachers and thus represented a possible confounder. To take this into account, we developed teacher-level predictors that were used in our main analyses. Specifically, we developed quality of teaching, quality of curriculum unit, and type of engineering integration variables to help capture differences among the curricula developed by treatment teachers.

For the quality of teaching variable, we observed each treatment teacher teaching their engineering design-based science units using a revised version of the Reformed Teaching Observation Protocol (RTOP; Sawada et al. 2002) to capture STEM integration teaching practices. Implementation of the curriculum units took 3–4 weeks, and each teacher observed 12–20 class periods. RTOP was used to evaluate each classroom observation, and average RTOP score was used to group teachers based on their RTOP score into one of the three quality of teaching categories (0 = low, 1 = medium, and 2 = high). Control teachers were also observed while teaching the targeted science units using the revised RTOP. Classroom observers completed training sessions for the revised RTOP prior to observing teachers to ensure the collection of reliable and valid data.

To assess the quality of the curriculum units developed by the teachers, we used a curriculum evaluation tool (Guzey et al. 2016) to assign a score to each teacher (treatment, control) based on the presence of key points in their teaching (0 = not present, 1 = weak, 2 = adequate, 3 = good, and 4 = excellent). Finally, we grouped the teachers based on the engineering integration strategy they used: 0 = add-on, 1 = implicit, and 2 = explicit. During each observation, we collected data on *when* teachers introduced and taught engineering and *how much time* they spent on engineering in their science unit. In explicit integration, teachers introduce engineering challenges early in the unit and integrate engineering into every science lesson, and students complete an engineering challenge at the end of the science unit. In implicit integration, teachers introduce engineering early in the unit but do not connect engineering and science in every lesson. Students again complete an engineering challenge at the end of the science unit. The add-on category represents a strategy in which teachers do not introduce the engineering challenge or the design until the end of the science unit. We argue that these variables collectively should help to minimize the potential confounding

effect of different treatment teachers developing different curricula.

Data Analyses

Descriptive analyses and *t* tests were initially used to examine differences in scores between the pretests and posttests in engineering, mathematics/data analysis, and science content areas and between treatment and control conditions. We also examined achievement differences among gender and racial groups. The main analyses used single-level regression models (Neter et al. 1996) for the science outcomes and multilevel models (students nested within teachers; Raudenbush and Bryk 2002) for the engineering, mathematics (data analysis and measurement), and state-mandated mathematics test outcomes. Data for each outcome variable were analyzed separately.

Single-level regression models were fitted to the science outcomes separately for elementary school students (heat transfer, ecosystems, plate tectonics, erosion) and middle school students (heat transfer, ecosystems, plate tectonics, particle theory)—there were too few classrooms to use multilevel modeling. The predictors in these models were gender, race, and whether a student was in the treatment or control condition.

The multilevel models fitted to the engineering, mathematics (data analyses and measurement), and state-mandated mathematics outcomes were

$$Y_{ij}\beta_{0j} + \sum \beta_{qj} X_{iq} + e_{ij} \quad \text{(student model)} \tag{1}$$

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \sum W_{0p}\gamma_{0p} + u_{0j} \\ \beta_{qj} &= \gamma_{00} + \sum W_{pj}\gamma_{pq} + u_{qj} \end{aligned} \quad \text{(classroom models)} \tag{2}$$

In Eqs. (1) and (2), Y_{ij} is the outcome of the ith student in the jth classroom, β_{0j} is the intercept of the jth classroom, β_{qj} is the slope capturing the impact of the qth student level predictor X_{iq} for the jth classroom, e_{ij} is the student-level residual, γ_{00} is the intercept for the classroom intercept model, γ_{0p} is a slope capturing the impact of the pth classroom-level predictor W_{pj}, u_{0j} is the residual for the classroom intercept model, γ_{q0} is the intercept of the classroom slope model for the qth predictor, γ_{pq} is the classroom slope capturing the impact of W_{pj}, and u_{qj} is the classroom residual for the slope model for the qth predictor (Raudenbush and Bryk 2002). The HLM7 software (Bryk et al. 2011) was used to perform the multilevel modeling with $\alpha = 0.05$ used for each statistical test.

A priori sample size and power analyses via the Optimal Design software (Spybrook et al. 2011) indicated that $J = 65$ classrooms with an average of 30 students per classroom and an intraclass correlation (ICC) of 0.18 for an unconditional model (Hedges and Hedberg 2007) produce a statistical test of the treatment effect with a power of 0.94 to detect an effect of 0.40 SDs. Because our sample contained 59 classrooms, our actual statistical power to detect a treatment effect of 0.40 SDs was somewhat lower than 0.94 but still exceeded 0.90. Model checking suggested that underlying assumptions were generally satisfied (e.g., normality, linearity, and no multicollinearity). For significant predictors, we estimated the variance in the outcome attributable to this variable.

Results

The descriptive analyses and *t* tests provided evidence of patterns in the data for all variables. These findings are reported in Tables 1 and 2 for elementary and middle school students using raw total correct scores for study outcomes and suggest (a) differences in prior achievement among students in treatment and control conditions especially among elementary school students; (b) consistent treatment-control differences on the engineering, mathematics (data analyses/measurement), and state-mandated mathematics tests with fewer differences on the science tests; (c) with the exception of a few science tests for middle school students, that the explained variance statistics for the significant *t* tests were generally small (i . e., ≤0.03).

Tables 3 and 4 report bivariate correlations among study variables for elementary and middle school students, and the nonnegligible values suggest the potential value of fitting regression models to posttest data. Correlations among teacher/classroom variables are not presented but also showed nonnegligible relationships with classroom achievement. The teacher-level predictors, years of teaching experience, years in current position, and years in current school were each coded as 1 ≤ 5 years, 2 = 6–10 years, 3 = 11–15 years, and 4 ≥ 15 years, with bivariate correlations ranging from 0.43 to 0.62.

Tables 5 and 6 report single-level regression results for science outcomes expressed in logits for elementary and middle school students. Treatment was not a significant predictor of the three science outcomes for elementary school students, although the overall *F* tests for these outcomes were statistically significant; i.e., the statistical null hypothesis Ho : $\boldsymbol{\beta} = 0$ was rejected where $\boldsymbol{\beta}$ is a $(q + 1) \times 1$ vector of slopes and $\mathbf{0}$ a $(q + 1) \times 1$ vector of zeros. The only significant results for elementary students were for the plate tectonics outcome, which produced a significant result favoring females ($\beta = 0.29$) and white students compared to Asian students ($\beta = -0.44$). The R^2 values for these models ranged from 0.21 to 0.49, with most explained variance attributable to the pretest predictor.

Several significant results emerged for the science outcomes among middle school students. Treatment was a significant predictor of every science outcome but only favored

Table 1 Descriptive statistics for elementary school students

	Treatment-control variable	Number	Mean	Std. deviation	*p* value
Engineering pre score	Control	234	5.35	2.57	.099
	Treatment	628	5.03	2.60	
Engineering post score	Control	238	6.56	2.09	.0*
	Treatment	559	5.71	2.79	
Math pre score	Control	234	5.87	2.31	.001*
	Treatment	584	5.24	2.47	
Math post score	Control	237	6.52	2.36	.004*
	Treatment	531	5.98	2.58	
Erosion pre score	Control	69	3.11	1.64	.077
	Treatment	181	3.55	2.00	
Erosion post score	Control	69	4.15	1.59	.55
	Treatment	159	4.00	2.23	
Heat T. pre score	Control	24	2.50	1.38	.675
	Treatment	194	2.62	1.36	
Heat T. post score	Control	22	3.13	1.42	.007*
	Treatment	195	4.12	2.19	
Ecosystems pre score	Control	50	4.96	2.39	.65
	Treatment	100	4.77	2.41	
Ecosystems post score	Control	57	4.66	2.60	.323
	Treatment	96	5.10	2.65	
Plate T. pre score	Control	91	4.91	1.71	.75
	Treatment	158	4.82	2.33	
Plate T. post score	Control	90	5.21	1.61	.633
	Treatment	107	5.33	2.06	
State math test score	Control	247	542.0	39.65	
	Treatment	712	519.0	65.66	

*Statistically significant difference between treatment and control groups

the treatment condition for heat transfer ($\beta = 0.50$); for the remaining outcomes, this effect favored the control condition. Race differences favoring whites emerged on every science outcome, and the variance explained by the models for middle school were substantial ranging from 0.31 to 0.58.

Table 2 Descriptive statistics for middle school students

	Treatment-control variable	Number	Mean	Std. deviation	*p* value
Engineering pre score	Control	567	10.12	3.37	.00*
	Treatment	2469	8.96	3.83	
Engineering post score	Control	486	11.17	3.41	.00*
	Treatment	2364	9.33	3.88	
Math pre score	Control	511	6.44	3.54	.310
	Treatment	2331	6.27	3.43	
Math post score	Control	463	7.31	3.94	.008*
	Treatment	2267	6.79	3.54	
Particle T. pre score	Control	407	6.98	3.66	.393
	Treatment	244	7.23	3.61	
Particle T. post score	Control	325	10.41	3.94	.00*
	Treatment	205	8.44	3.93	
Heat T. pre score	Control	50	2.88	1.49	.00*
	Treatment	622	6.82	2.67	
Heat T. post score	Control	68	3.08	1.50	.00*
	Treatment	622	7.54	3.13	
Ecosystem pre score	Control	41	8.02	3.19	.00*
	Treatment	838	5.31	3.59	
Ecosystem post core	Control	27	9.33	3.49	.00*
	Treatment	796	5.72	3.79	
Plate T. pre score	Control	61	6.01	2.79	.197
	Treatment	729	5.46	3.22	
Plate T. post score	Control	53	7.81	2.99	.00*
	Treatment	720	5.87	3.70	
State math test score	Control	580	697.23	112.01	
	Treatment	2540	724.81	140.04	

*Statistically significant difference between treatment and control groups

Table 3 Correlations of study variables for elementary school students

	1	2	3	4	5	6	7	8	9	10	11	12
1. Eng. pre	1	0.60[a]	0.42[a]	0.41[a]	0.17[b]	0.13	0.78[a]	0.67[a]	0.37[a]	0.46[a]	0.21[a]	−0.05
2. Math pre	0.60[a]	1	0.45[a]	0.36[a]	0.18[a]	0.08	0.65[a]	0.66[a]	0.47[a]	0.44[a]	0.23[a]	−0.11[a]
3. Ero. pre	0.42[a]	0.45[a]	1	0.42[a]	–	–	–	–	–	–	0.19[a]	0.10
4. Ero. post	0.41[a]	0.36[a]	0.42[a]	1	–	–	–	–	–	–	0.14[*]	−0.03
5. Heat T. pre	0.17[b]	0.18[a]	–	–	1	0.05	–	–	–	–	0.10	0.02
6. Heat T. post	0.13	0.08	–	–	0.05	1	–	–	–	–	0.12	0.13[b]
7. Eco. pre	0.78[a]	0.65[a]	–	–	–	–	1	0.69[a]	–	–	0.09	−0.03
8. Eco. post	0.67[a]	0.66[a]	–	–	–	–	0.69[a]	1	0.51[a]	–	0.16[b]	0.08
9. Plate T. pre	0.37[a]	0.47[a]	–	–	–	–	–	0.51[a]	1	0.45[a]	0.27[a]	−0.02
10. Plate T. post	0.46[a]	0.44[a]	–	–	–	–	–	–	0.45[a]	1	0.61[a]	0.03
11. State math	0.21[a]	0.23[a]	0.19[a]	0.14[b]	0.10	0.12	0.09	0.16[b]	0.27[a]	0.61[a]	1	−0.16[a]
12. Treatment-control variable	−0.05	−0.11[a]	0.103	−0.03	0.02	0.13[b]	−.03	0.08	−0.02	0.03	−0.16[a]	1

Correlations based on $N > 180$

[a] Correlation at 0.01 (two-tailed)

[b] Correlation at 0.05 (two-tailed)

For the multilevel analyses, we began by fitting an unconditional model (no predictors) and estimating the ICC. The ICC for the unconditional model for the engineering posttest (expressed in logits) was 0.43, meaning that 43% of the variation in these scores was between teachers (classrooms) which justifies a multilevel approach; equivalently, the average dependency of these scores within the classrooms was 0.43. Preliminary analyses showed that slopes for the engineering pretest and whether a student was Hispanic or black varied significantly across classrooms, and variation in these slopes was modeled.

Results for the conditional model for the engineering posttest outcome appear in Table 7. On average, classrooms with higher concentrations of special education students and more experienced teachers were associated with lower performances on the engineering posttest ($\gamma_{04} = -0.19, \gamma_{06} = -0.36$), with the variance attributable to these effects of 2.6 and 4.3%. On the other hand, classrooms where engineering was *explicitly* integrated into the curriculum were associated with higher average performances ($\gamma_{010} = 0.37, 1.5\%$). Female students outperformed male students on average ($\gamma_{20} = 0.10$), and Asian students on average scored lower than White students ($\gamma_{30} = -0.17$).

Table 4 Correlations of study variables for middle school students

	1	2	3	4	5	6	7	8	9	10	11	12
1. Eng. pre	1	0.56[a]	0.55[a]	0.39[a]	0.46[a]	0.44[a]	0.79[a]	0.79[a]	0.37[a]	0.40[a]	0.22[a]	−0.11[a]
2. Math pre	0.56[a]	1	0.56[a]	0.44[a]	0.37[a]	0.38[a]	0.64[a]	0.67[a]	0.45[a]	0.45[a]	0.24[a]	−0.01
3. Particle T pre	0.55[a]	0.56[a]	1	0.45[a]	–	–	–	–	–	–	0.20[a]	0.03
4. Particle T post	0.39[a]	0.44[a]	0.45[a]	1	–	–	–	–	–	–	0.24[a]	−0.23[a]
5. Heat T. pre	0.46[a]	0.37[a]	–	–	1	0.64[a]	–	–	–	–	0.01	0.37[a]
6. Heat T. post	0.44[a]	0.38[a]	–	–	0.64[a]	1	–	–	–	–	−.10[a]	0.41[a]
7. Eco. pre	0.79[a]	0.64[a]	–	–	–	–	1	0.82[a]	–	–	0.19[a]	−0.15[a]
8. Eco. post	0.79[a]	0.67[a]	–	–	–	–	0.82[a]	1	–	–	0.15[a]	−0.16[a]
9. Plate T. pre	0.37[a]	0.45[a]	–	–	–	–	–	–	1	0.75[a]	0.17[a]	−0.04
10. Plate T. post	0.40[a]	0.45[a]	–	–	–	–	–	–	0.75[a]	1	0.18[a]	−0.13[a]
11. State math	0.22[a]	0.24[a]	0.20[a]	0.24[a]	0.017	−0.10[a]	0.19[a]	0.15[a]	0.17[a]	0.18[a]	1	0.07[**]
12. Treatment control variable	−0.11[a]	−0.01	0.03	−0.23[a]	0.37[a]	0.41[a]	−0.15[a]	−0.16[a]	−0.04	−0.13[a]	0.07[a]	1

Correlations based on $N > 180$

[a] Correlation at 0.01(two-tailed)

[b] Correlation at 0.05(two-tailed)

Table 5 Single level science regression results for elementary school students

	β	Std. error	β
Ecology post score[a]			
Constant	−0.38	0.21	
Treatment-control variable	0.19	0.23	0.06
Gender of students	0.34	0.22	0.11
Asian/Pacific Islander stud.	0.51	0.49	0.07
Hispanic students	−0.65	0.36	−0.12
Black stud.	0.47	0.38	0.08
Eco. pre score	0.71	0.07	0.68
F	16.4		
R^2	.49		
N	109		
Plate T. post score[a]			
Constant	0.20	0.14	
Treatment-control variable	0.16	0.14	0.07
Gender of the student	0.29	0.14	0.13*
Asian/Pacific Islander students	−0.44	0.18	−0.18*
Hispanic students	−0.40	0.25	−0.11
Black students	−0.37	0.19	−0.13
Plate T. pre score	0.30	0.05	0.36
F	07.71		
R^2	.22		
N	172		
Ero. post score[a]			
Constant	0.14	0.19	
Treatment-control variable	−0.07	0.17	−0.03
Gender of the student	−0.09	0.14	−0.04
Asian/Pacific Islander students	−0.29	0.19	−0.13
Hispanic students	0.07	0.20	0.03
Black students	−0.14	0.21	−0.05
Ero. pre score score	0.48	0.07	0.43
F	7.5		
R^2	.21		
N	174		

Heat results not reported because the F test = 1.9 ($p > .05$ was not significant)

*Statistically significant ($p < .05$).

[a] Dependent variable

Table 6 Single level science regression results for middle school students

	β	Std. error	β
Heat post score[a]			
Constant	−0.14	0.17	
Treatment-control variable	0.50	0.17	0.10*
Gender of students	0.08	0.07	0.03
Asian/Pacific Islander stud.	−0.31	0.09	−0.11
Hispanic students	−0.30	0.13	−0.07*
Black stud.	−0.31	0.11	−0.10*
Heat pre score	0.67	0.04	0.57
F	74.6		
R^2	.45		
N	549		
Ecology post score[a]			
Constant	0.65	0.20	
Treatment-control variable	−0.55	0.20	−0.06*
Gender of the student	0.021	0.06	0.00
Asian/Pacific Islander students	−0.17	0.08	−0.05*
Hispanic students	−0.33	0.15	−0.05*
Black students	−0.28	0.10	−0.07*
Eco. pre score	0.76	0.02	0.72
F	162.5		
R^2	.58		
N	696		
Plate T. post score[a]			
Constant	0.68	0.15	
Treatment-control variable	−0.38	0.14	−0.07*
Gender of the student	0.04	0.07	0.01
Asian/Pacific Islander students	−0.49	0.09	−0.17*
Hispanic students	−0.53	0.14	−0.10*
Black students	−0.63	0.12	−0.15*
Plate T. pre score	0.75	0.03	0.62
F	125.6		
R^2	.54		
N	642		
Particle T. post score[a]			
Constant	1.55	0.14	
Treatment-control variable	−0.96	0.14	−0.28*
Gender of the student	0.20	0.13	0.06
Asian/Pacific Islander students	−0.57	0.21	−0.11*
Hispanic students	−0.52	0.18	−0.12*
Black students	−0.40	0.18	−0.10*
Plate T. pre score	0.51	0.05	0.41
F	32.6		
R^2	.31		
N	439		

*Statistically significant ($p < .05$)

[a] Dependent variable

There was also evidence that higher scores on the engineering pretest were associated with higher scores on the posttest ($\gamma_{10} = 0.54$), and that this relationship was stronger in classrooms with higher quality curricula ($\gamma_{19} = 0.08$). The relationship between whether a student was Hispanic and their score on the engineering posttest was on average weaker in classrooms with a female teacher ($\gamma_{48} = -0.39$). Increasing concentrations of LEP students in a classroom were associated

Table 7 Multilevel results for the engineering posttest outcome

Fixed effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p* value
For intercept level 1, β_0					
Intercept level 2, γ_{00}	0.79	0.42	1.87	35	0.06
Treatment, γ_{01}	−0.80	0.46	−1.73	35	0.09
Level, γ_{02}	0.36	0.22	1.65	35	0.10
LEP, γ_{03}	0.11	0.08	1.30	35	0.20
Special education, γ_{04}	−0.19	0.08	−2.25	35	0.03*
Years of teaching experience, γ_{05}	0.12	0.11	1.06	35	0.29
Years in current position, γ_{06}	−0.36	0.12	−2.84	35	0.00*
Years in current school, γ_{07}	0.27	0.14	1.85	35	0.07
Gender of teacher, γ_{08}	−0.12	0.19	−0.61	35	0.54
Quality of curriculum unit, γ_{09}	−0.05	0.13	−0.40	35	0.69
Type of eng. integration, γ_{010}	0.37	0.16	2.3	35	0.02*
RTOP, γ_{011}	0.02	0.21	0.12	35	0.89
For engineering pre score slope, β_1					
Intercept level 2, γ_{10}	0.54	0.14	3.84	35	<0.001*
Treatment, γ_{11}	0.11	0.16	0.68	35	0.49
Level, γ_{12}	0.08	0.07	1.04	35	0.30
LEP, γ_{13}	−0.04	0.02	−1.59	35	0.12
Special education, γ_{14}	−0.03	0.02	−1.44	35	0.15
Years of teaching experience, γ_{15}	0.03	0.04	0.72	35	0.47
Years in current position, γ_{16}	0.04	0.04	0.90	35	0.37
Years in current school, γ_{17}	−0.07	0.05	−1.37	35	0.17
Gender of teacher, γ_{18}	−0.06	0.06	−1.01	35	0.31
Quality of curriculum unit, γ_{19}	0.08	0.04	2.08	35	0.04*
Type of eng. integration, γ_{110}	0.09	0.05	1.82	35	0.07
RTOP, γ_{111}	−0.07	0.07	−1.04	35	0.30
For gender of student slope, β_2					
Intercept level 2, γ_{20}	0.10	0.04	2.62	2316	0.009*
For Asian slope, β					
Intercept level 2, γ_{30}	−0.17	0.06	−2.87	2316	0.004*
For Hispanic slope, β_4					
Intercept level 2, γ_{40}	0.25	0.43	0.59	35	0.55
Treatment, γ_{41}	0.74	0.48	1.54	35	0.13
Level, γ_{42}	0.06	0.23	0.28	35	0.77
LEP, γ_{43}	−0.06	0.08	−0.70	35	0.48
Special education, γ_{44}	−0.07	0.09	−0.78	35	0.43
Years of teaching experience, γ_{45}	−0.09	0.14	−0.63	35	0.53
Years in current position, γ_{46}	0.15	0.14	1.08	35	0.28
Years in current school, γ_{47}	−0.13	0.18	−0.74	35	0.46
Gender of teacher, γ_{48}	−0.39	0.17	−2.23	35	0.03*
Quality of curriculum unit, γ_{49}	0.09	0.12	0.72	35	0.47
Type of eng. integration, γ_{410}	−0.10	0.14	−0.68	35	0.49
RTOP, γ_{411}	−0.32	0.21	−1.53	35	0.13
For black slope, β_5					
Intercept level 2, γ_{50}	−0.26	0.33	−0.77	35	0.44
Treatment, γ_{51}	0.18	0.40	0.46	35	0.64
Level, γ_{52}	0.22	0.19	1.15	35	0.25
LEP, γ_{53}	−0.16	0.07	−2.27	35	0.02*
Special education, γ_{54}	0.03	0.07	0.54	35	0.58
Years of teaching experience, γ_{55}	−0.06	0.12	−0.52	35	0.60
Years in current position, γ_{56}	−0.10	0.12	−0.81	35	0.42
Years in current school, γ_{57}	0.08	0.15	0.52	35	0.60
Gender of teacher, γ_{58}	0.31	0.14	2.15	35	0.03*
Quality of curriculum unit, γ_{59}	0.16	0.12	1.31	35	0.19
Type of eng. integration, γ_{510}	−0.03	0.11	−0.26	35	0.79
RTOP, γ_{511}	−0.25	0.17	−1.44	35	0.15

Gender is coded 1 = female and 0 = male; quality of curriculum unit is coded 0 = not present, 1 = weak, 2 = adequate, 3 = good, and 4 = excellent; engineering integration is coded 0 = add-on, 1 = implicit, and 2 = explicit; RTOP is coded 0 = low, 1 = medium, and 2 = high; and level is coded 0 = elementary and 1 = middle school

*Statistically significant ($p < 0.05$)

with a larger black/white gap on the engineering posttest that favored white students ($\gamma_{53} = -0.16$), but this gap was on average smaller in classrooms with female teachers ($\gamma_{58} = 0.31$).

The ICC for the unconditional multilevel model for the project-constructed mathematics (data analysis and measurement) posttest was 0.34, meaning that 34% of the variation in these scores was between the teachers (classrooms), justifying a multilevel approach. Analyses showed that slopes for the mathematics (data analysis and measurement) pretest and whether a student was black varied significantly across classrooms, and variation in these slopes was modeled. Results for the conditional model for the mathematics posttest outcome appear in Table 8.

On average, teachers who had been in their current position longer were associated with lower classroom performances on the mathematics/data analysis posttest ($\gamma_{06} = -0.35, 5\%$) but were associated with higher performances if they had been at their school longer ($\gamma_{07} = 0.24, 3\%$). Higher scores on the engineering pretest were associated with higher scores on the posttest ($\gamma_{10} = 0.57$), and Asian students scored on average lower than white students ($\gamma_{30} = -0.15$). Teachers in their current position longer were associated with classrooms with a weaker relationship between whether a student was black and the mathematics/data analysis posttest ($\gamma_{56} = -0.34$), but this relationship was stronger when teachers had been at a school longer ($\gamma_{57} = 0.40$).

The ICC for the unconditional multilevel model for the state-mandated mathematics test was 0.58, meaning that 58% of the variation in these scores was between the teachers (classrooms), again justifying a multilevel approach. Preliminary analyses showed that none of the slopes for the student predictors varied significantly across classrooms, and variation in these slopes was constrained to zero. Results for the conditional model for the state mathematics outcome appear in Table 9. Because this test was not constructed to support comparisons of scores across grades, the significant level effect ($\gamma_{02} = 231.4, 50\%$) favoring middle school students is not surprising (mean of this test is about 600, standard deviation is about 130). The only other significant effect showed that Hispanic students and black students on average scored lower than white students ($\gamma_{30} = -14.6,\ \gamma_{40} = -22.2$).

In sum, the most important finding for the single-level analyses of science outcomes is that treatment was either not significant (elementary school students) or was a significant predictor (middle school students) which favored treatment over control for only one outcome (heat transfer). Treatment was not a significant predictor in any of the multilevel analyses, but there was evidence that higher quality curricula based on engineering produced higher achievement on average. Teacher gender and how long teachers had been in their position and school moderated differences in achievement among some student subgroups.

As noted earlier, we used several predictors to control for selection bias among treatment and control teachers to help ensure unbiased (or almost unbiased) treatment effect estimates (WWC 2014). We also performed additional analyses to assess the sensitivity of our findings to selection bias by matching teachers via propensity scores (Dehejia and Wahba 2002). The sample of matched teachers was used to repeat key analyses where possible to assess the impact of selection bias. Similar patterns of findings for results relying on predictors to adjust for preexisting treatment-control differences and results based on matching would provide additional evidence of the success in controlling for selection bias.

We began the matching process by fitting a logistic regression to the treatment variable data (1 = treatment, 0 = control) using teacher background variables and characteristics of their classes such as the percentage of special education and LEP students. The resulting propensity scores were then used to match the 42 treatment and 17 control teachers using the R package "Matching" (Sekhon 2011). Unequal number of treatment and control teachers means that the traditional "one-to-one" matching would produce 17 matched pairs. While "many-to-one" matching methods ideally use all treatment and control teachers who are available (Guo and Rosenbaum 1993), their complexity and underlying assumptions led us to perform and report matching based on one-to-one matching. The "nearest neighbor" method was used with the propensity scores to produce 17 matched pairs (one treatment teacher and one control teacher per pair).

Once matching was complete, we refitted the multilevel models to the 34 teachers that had been matched. Following the recommendations of Gelman and Hill (2007), these analyses included variables used in the matching as predictors to ensure that correlations induced by matching were taken into account and standard errors were adjusted accordingly. These results showed no treatment effect for the engineering, mathematics/data analysis, or state-mandated mathematics outcomes and generally similar patterns of findings for other variables to those in Tables 7, 8, and 9. For example, for the engineering posttest teacher gender was a significant moderator of Hispanic/White and black/white differences on this outcome.

Discussion and Conclusion

This study explored the influence of engineering design-based STEM curriculum units on the learning and achievement of students. Single-level analyses of science outcomes produced a significant treatment effect for curriculum units focused on heat transfer but only appeared for middle school students. Previous research has produced mixed results of the impact of engineering design-based science units in supporting the students' science content knowledge (e.g., Lachapelle et al. 2011; Wendell and Rogers 2013). For example, Wendell and Rogers (2013) reported significant science content gains in life science and physical science domains by elementary school students who participated in engineering design-based science units. However, in a study of a life science-

Table 8 Multilevel results for the mathematics/data analysis posttest outcome

Fixed effect	Coefficient	Standard error	*t* ratio	Approx. *d.f.*	*p* value
For intercept Level 1, β_0					
Intercept level 2, γ_{00}	0.42	0.28	1.51	35	0.13
Treatment, γ_{01}	−0.07	0.32	−0.22	35	0.82
Level, γ_{02}	0.02	0.14	0.15	35	0.88
LEP, γ_{03}	0.013	0.05	0.22	35	0.82
Special education, γ_{04}	−0.11	0.05	−2.01	35	0.05
Years of teaching experience, γ_{05}	0.13	0.08	1.59	35	0.11
Years in current position, γ_{06}	−0.35	0.09	−3.59	35	<0.001*
Years in current school, γ_{07}	0.24	0.11	2.21	35	0.03*
Gender of teacher, γ_{08}	−0.06	0.13	−0.50	35	0.62
Quality of curriculum unit, γ_{09}	−0.01	0.08	−0.21	35	0.82
Type of eng. integration, γ_{010}	0.15	0.10	1.46	35	0.15
RTOP, γ_{011}	−0.17	0.14	−1.22	35	0.23
For mathematics pre score slope, β_1					
Intercept level 2, γ_{10}	0.57	0.16	3.54	35	0.001*
Treatment, γ_{11}	0.08	0.18	0.48	35	0.62
Level, γ_{12}	0.14	0.08	1.66	35	0.10
LEP, γ_{13}	−0.04	0.03	−1.39	35	0.17
Special education, γ_{14}	−0.04	0.03	−1.32	35	0.19
Years of teaching experience, γ_{15}	−0.00	0.04	−0.09	35	0.92
Years in current position, γ_{16}	−0.00	0.05	−0.08	35	0.93
Years in current school, γ_{17}	0.02	0.06	0.33	35	0.74
Gender of teacher, γ_{18}	0.01	0.07	0.23	35	0.81
Quality of curriculum unit, γ_{19}	0.04	0.04	0.85	35	0.39
Type of eng. integration, γ_{110}	0.08	0.05	1.40	35	0.17
RTOP, γ_{111}	−0.10	0.08	−1.30	35	0.20
For gender of student slope, β_2					
Intercept level 2, γ_{20}	0.00	0.04	0.04	2176	0.96
For Asian slope, β_3					
Intercept level 2, γ_{30}	−0.15	0.05	−2.64	2176	0.008*
For Hispanic slope, β_4					
Intercept level 2, γ_{40}	−0.14	0.07	−1.93	2176	0.05
For black slope, β_5					
Intercept level 2, γ_{50}	0.09	0.35	0.27	35	0.78
Treatment, γ_{51}	−0.72	0.43	−1.66	35	0.10
Level, γ_{52}	0.11	0.20	0.56	35	0.57
LEP, γ_{53}	−0.08	0.07	−1.06	35	0.29
Special education, γ_{54}	0.10	0.07	1.42	35	0.16
Years of teaching experience, γ_{55}	−0.12	0.13	−0.90	35	0.36
Years in current position, γ_{56}	−0.31	0.14	−2.19	35	0.03*
Years in current school, γ_{57}	0.40	0.16	2.40	35	0.02*
Gender of teacher, γ_{58}	0.00	0.15	0.02	35	0.98
Quality of curriculum unit, γ_{59}	−0.00	0.13	−0.01	35	0.99
Type of eng. Integration, γ_{510}	−0.19	0.12	−1.54	35	0.13
RTOP, γ_{511}	0.23	0.18	1.23	35	0.22

Gender is coded 1 = female and 0 = male; quality of curriculum unit is coded 0 = not present, 1 = weak, 2 = adequate, 3 = good, and 4 = excellent; engineering integration is coded 0 = add-on, 1 = implicit, and 2 = explicit; RTOP is coded 0 = low, 1 = medium, and 2 = high; and level is coded 0 = elementary and 1 = middle school

*Statistically significant ($p < 0.05$)

Table 9 multilevel results for the state-mandated mathematics outcome

Fixed effect	Coefficient	Standard error	*t* ratio	Approx. *d.f.*	*p* value
For intercept level 1, β_0					
Intercept level 2, γ_{00}	570.63	42.28	13.49	36	<0.001*
Treatment, γ_{01}	53.37	45.98	1.16	36	0.25
Level, γ_{02}	213.43	21.64	9.86	36	<0.001*
LEP, γ_{03}	−12.63	8.95	−1.41	36	0.16
Special education, γ_{04}	−15.62	8.53	−1.82	36	0.07
Years of teaching experience, γ_{05}	−4.82	11.79	−0.40	36	0.68
Years in current position, γ_{06}	24.12	12.68	1.90	36	0.06
Years in current school, γ_{07}	−5.31	14.63	−0.36	36	0.71
Gender of teacher, γ_{08}	13.78	19.83	0.69	36	0.49
Quality of curriculum unit, γ_{09}	3.03	13.43	0.22	36	0.82
Type of eng. integration, γ_{010}	0.05	16.08	0.00	36	0.99
RTOP, γ_{011}	−42.28	21.59	−1.95	36	0.05
For Gender of student slope, β_1					
Intercept level 2, γ_{10}	4.08	3.55	1.15	3278	0.25
For Asian slope, β_2					
Intercept level 2, γ_{20}	−1.32	5.06	−0.26	3278	0.79
For Hispanic slope, β_3					
Intercept level 2, γ_{30}	−14.68	6.36	−2.30	3278	0.02*
For black slope, β_4					
Intercept level 2, γ_{40}	−22.20	5.23	−4.24	3278	<0.001*

Gender is coded 1 = female and 0 = male; quality of curriculum unit is coded 0 = not present, 1 = weak, 2 = adequate, 3 = good, and 4 = excellent; engineering integration is coded 0 = add-on, 1 = implicit, and 2 = explicit; RTOP is coded 0 = low, 1 = medium, and 2 = high; and level is coded 0 = elementary and 1 = middle school

*Statistically significant ($p < 0.05$)

focused engineering unit on designing model membranes for elementary school students, Lachapelle et al. (2011) did not find statistically significant differences on posttest science outcomes between the treatment condition who completed a science unit about organisms and engineering unit and the control condition who only completed a science unit on organisms. These authors found significant gains by students who completed physical science units, as was the case in the current study. Our findings add to this body of literature, supporting the positive impact of engineering on student learning only in physical science, particularly the heat transfer concept.

On the other hand, our multilevel analyses did not show a significant treatment effect for project-constructed assessments in engineering, mathematics/data analysis, or scores on a state-mandated mathematics test. However, two of the teacher variables, quality of curriculum units and type of engineering integration, were associated with student achievement. For the former, the evidence is that quality engineering design-based curriculum units are associated with higher achievement in engineering, and for the latter, that more effective integration of engineering into curriculum units strengthens the relationship between the engineering pretest and posttest. This novel contribution of the current study highlights the importance of the features or types of engineering integration approaches in supporting student learning. We argue that simply adding engineering into science instruction is not necessarily supportive of better student learning—teaching high-quality curriculum units that purposefully and meaningfully connect science concepts and the practices of those of engineering is essential to produce positive student outcomes.

Teachers come to professional development programs with varying levels of knowledge and experiences. To accommodate and address the teachers' needs, it is critical to provide a variety of opportunities for teachers and differentiate professional development activities (Desimone and Garet 2015; Garet et al. 2001). To do that, our program provided teachers with opportunities to learn and explore engineering and increase content knowledge of their desired science discipline. Several engineering design-based science units developed by the authors were implemented in the program. Teachers were not required to follow a specific engineering integration strategy; instead, they were asked to use an engineering integration approach that would fit to their school context and structure, appropriate for the science content they choose and right for their students. We believe that teachers benefited from the approach of providing collective experiences that targeted groups of teachers with similar needs. However, the classroom

observation data reflected that not all teachers practiced or implemented what was learned in the professional development program. Our data provide evidence that teachers changed their instruction at varying levels by including engineering design and practices, but that changed instruction failed to improve student learning in many areas of science content and engineering and mathematics. Thus, we believe that the primary reason for variation in student outcomes was because many teachers did not achieve the desired or required change in engineering integration.

As noted earlier, the approaches used by the teachers in this study were grouped in three categories: explicit, implicit, and add-on. The vast majority of the teachers in this study used the "add-on" strategy in their curriculum unit design. In this strategy, engineering is simply considered as an addition to a science unit or seen as an end of science unit project. As previous studies show, design activities or engineering activities should be explicitly integrated with science activities so that students can see the connections among different disciplines and increase disciplinary knowledge (Guzey et al. 2014; Moore et al. 2014; NAE and NRC 2014; Wendell and Rogers 2013). However, considering the fact that engineering was a new concept for the majority of the treatment teachers, the use of an add-on strategy was mainly preferred. It is not easy for teachers to effectively transfer everything they learn in professional development programs into classroom practices (Desimone and Garet 2015). It often requires time and a long-term commitment to adapt new practices.

Portions of the multilevel results indicated that engineering integration had different effects across race and gender. White students had on average higher science, engineering, and mathematics scores on the posttests compared to others. Hispanic and black students had lower scores on the state-mandated mathematics test. Thus, there was not compelling evidence that engineering integration practices reduced achievement gaps. These gaps could be related to the difficulty most teachers seemed to have in effectively integrating engineering into their teaching. The performance gaps in the study may also be related to the context used in engineering science units and classroom instruction. Previous research has demonstrated the importance of using engaging and motivating context for improving student learning (Berland et al. 2014; Brown et al. 1989; Brophy et al. 2008). These findings provide empirical support to arguments that the design of engineering-based science units should take into consideration the interests, abilities, needs, and backgrounds of students within every classroom to reduce or eliminate gender and culture differences or provide opportunities for all students to enter into the challenge from their perspective. Furthermore, teachers need to embed different types of support in science and engineering instruction for diverse learners. By ensuring that all students engage with the engineering challenges, teachers can promote better learning in students.

The multilevel analyses also revealed teacher predictors that interacted with student outcomes. We found that teacher gender plays a factor in decreasing or increasing achievement gaps in engineering among students, signaling the need to identify those characteristics that impact this gap. For example, female teachers were associated with a weaker relationship between whether a student was Hispanic and their posttest engineering scores. The finding that teacher experience correlated negatively with the engineering and mathematics posttests was not surprising because it is generally more challenging for experienced teachers to adopt new classroom practices; changing classroom practices are established over the years, and replacing a traditional science curriculum with an engineering-focused curriculum may not be easy for many experienced science teachers. This finding suggests the need for new strategies to better support the development of the experienced science teachers' engineering instruction.

We note that the majority of year one treatment teachers decided to come back for the second year of the program, which raises several important questions for future work: What is the trajectory of teacher learning of engineering design and engineering practices? What instructional approaches are most likely to help students increase their disciplinary knowledge and make connections between and among science and engineering? What science concepts can be learned better through engineering integration approaches? These questions in many ways emphasize the need to research teacher practices of science and engineering in classrooms. Research on teacher practices may help to explain why and how certain strategies of engineering integration better support student learning. It seems clear that as growing numbers of science teachers deliver engineering instruction, there is much more need to be learned about the nature and outcomes of integrated science and engineering experiences.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999) Standards for educational and psychological testing. American Educational Research Association, Washington, DC

Apedoe XS, Reynolds B, Ellefson MR, Schunn CD (2008) Bringing engineering design into high school science classrooms: the heating/cooling unit. J Sci Educ Technol 17(5):454–465

Azevedo FS, Martalock PL, Keser T (2015) The discourse of design-based science classroom activities. Cult Stud Sci Educ 10(2):285–315

Berland, L., Steingut, R., & Ko, P. (2014). High school student perceptions of the utility of the engineering design process: creating opportunities to engage in engineering practices and apply math and

science content. *Journal of Science Education and Technology*, 705–720.
Brophy S, Klein S, Portsmore M, Rogers C (2008) Advancing engineering education in K-12 classrooms. J Eng Educ:369–387
Brown JS, Collins A, Duguid P (1989) Situated cognition and the culture of learning. Educ Res 18(1):32–42
Bryk AS, Raudenbush SW, Congdon R (2011) Hierarchical linear and nonlinear modeling [computer software manual]. Scientific Software International, Lincolnwood, IL
Cobb P, Bowers J (1999) Cognitive and situated learning perspective in theory and practice. Educ Res 28(2):4–15
Cantrell P, Pekcan G, Itani A, Velasquez-Bryant N (2006) The effects of engineering modules on student learning in middle school science classrooms. J Eng Educ 95:301–309
Carlson LE, Sullivan JF (2004) Exploiting design to inspire interest in engineering across K-16 curriculum. Int J Eng Educ 20(3):372–380
Dare E, Ellis J, Roehrig GH (2014) Driven by beliefs: understanding challenges physical science teachers face when integrating engineering and physics. J Pre-College Eng Educ Res 4(2)
Dehejia RH, Wahba S (2002) Propensity score-matching methods for nonexperimental causal studies. Revi Econ Stat 84(1):151–161
Desimone LM (2011) A primer on effective professional development. Phi Delta Kappan 92(6):68–71
Desimone LM, Smith TM, Phillips KJR (2013) Linking student achievement growth to professional development participation and changes in instruction: a longitudinal study of elementary students and teachers in Title I schools. Teach Coll Rec 115:1–46
Desimone L, Garet M (2015) Best practices in teachers' professional development in the United States. Psychol Soc Educ 7(3):252–263
Gelman A, Hill J (2007) Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, Cambridge, U.K.
Garet MS, Porter AC, Desimone L, Birman BF, Yoon KS (2001) What makes professional development effective? Results from a national sample of teachers. Am Educ Res J 38(4):915–945
Guo XS, Rosenbaum PR (1993) Comparison of multivariate matching methods: structures, distances, and algorithms. J Comput Graph Stat 2:405–420
Guzey SS, Moore T, Harwell M (2014) Development of an instrument to measure students' attitudes toward STEM. Sch Sci Math 114(6): 271–279
Guzey SS, Moore T, Harwell M (2016) Building up STEM: an analysis of teacher-developed engineering design-based STEM integration curricular materials. Journal of Pre-College Engineering Education Research (JPEER) 6(1). doi:10.7771/2157-9288.1129
Harwell M, Philips A, Mareno M, Guzey SS, Moore T (2015) A study of STEM assessments in Engineering, Science, and Mathematics Assessments for elementary and middle school students. Sch Sci Math 115(2):66–74
Hedges LV, Hedberg EC (2007) Intraclass correlation values for planning group-randomized trials in education. Educ Eval Policy Anal 29:60–87
Kolen MJ, Brennan RL (2004) Test equating, scaling, and linking. Springer, New York, NY
Lachapelle C, Cunningham C (2014) Engineering in elementary schools. In: Purzer S, Strobel J, Cardella M (eds) Engineering in pre-college settings: synthesizing research, policy, and practices. Purdue University Press, West Lafayette: IN, pp. 61–88
Lachapelle CP, Cunningham CM, Jocz J, Kay AE, Phadnis P, Wertheimer J, Arteaga R (2011) Engineering is elementary: an evaluation of years 4 through 6 field testing. Museum of Science, Boston, MA
Little RJA, Rubin DB (2002) Statistical analysis with missing data (2nd Ed). Wiley, New York, NY
Ludlow LH, Haley SM (1995) Rasch model logits: interpretation, use, and transformation. Educ Psychol Meas 55:967–975
Mehalik MM, Doppelt Y, Schunn CD (2008) Middle-school science through design-based learning versus scripted inquiry: better overall science concept learning and equity gap reduction. J Eng Educ 97(71–85)
Moore TJ, Stohlmann MS, Wang H, Tank KM, Glancy AW, Roehrig GH (2014) Implementation and integration of engineering in K-12 STEM education. In: Purzer S, Strobel J, Cardella M (eds) Engineering in pre-college settings: Research into practice. Purdue University Press, West Lafayette, pp 35–60
National Academy of Engineering and National Research Council (2014) STEM integration in K-12 education: status, prospects, and an agenda for research. The National Academies Press, Washington, DC
National Research Council (2009) Engineering in K-12 education: understanding the status and improving the prospects. The National Academies Press, Washington, DC
National Research Council (2010) Standards for K-12 engineering education? The National Academies Press, Washington, DC
National Research Council (2011) Successful K-12 STEM education: identifying effective approaches in science, technology, engineering, and mathematics. National Academies Press, Washington, DC
National Research Council (2012) A framework for K–12 science education. Retrieved from www.nap.edu/catalog.php?record_id=13165
Nelson T, Lesseig K, Slavit D, Kennedy C, Seidel R (2015) Supporting middle school teachers implementation of STEM design challenges. Paper presented at NARST conference. IL, Chicago
Neter J, Kutner MH, Nachtsheim CJ, Wasserman W (1996) Applied linear statistical models (4th ed.). Irwin, Chicago, IL
NGSS Lead States (2013) Next generation science standards: for states, by states. The National Academic Press, Washington, DC
Oh Y, Lachapelle C, Shams M, Hertel J, Cunnigham C (2016) Evaluating the efficacy of engineering is elementary for student learning of engineering and science concepts. Poster presented at the American Educatonal Research Association Annual Meeting. Washington, DC Retrieved from http://www.eie.org/sites/default/files/research_article/research_file/aera_oh_evaluating_the_efficacy_poster.pdf
Raudenbush SW, Bryk AS (2002) Hierarchical linear models: applications and data analysis methods (2nd Ed). Sage, Newbury Park, CA
Riskowski JL, Todd CD, Wee B, Dark M, Harbor J (2009) Exploring the effectiveness of an interdisciplinary water resources engineering module in an eighth grade science course. Int J Eng Educ 25(1): 181–195
Sawada D, Piburn MD, Judson E, Turley J, Falconer K, Benford R, Bloom I (2002) Measuring reform practices in science and mathematics classrooms: the reformed teaching observation protocol. Sch Sci Math 102(6):245–253
Schnittka CG, Bell RL (2011) Engineering design and conceptual change in the middle school science classroom. Int J Sci Educ 33:1861–1887
Sekhon JS (2011) Multivariate and propensity score matching software with automated balance optimization: the matching package for R. J Stat Softw 42(7):1–52
Shadish WR, Cook TD, Campbell DT (2002) Experimental and quasi-experimental designs for generalized causal inference. Houghton-Mifflin, Boston
Spybrook, L., Bloom, H., Congdon, R., Hill, C., Martinez, A., & Raudenbush, S. (2011). Optimal design plus empirical evidence: documentation for the "Optimal Design" software (version 3.0) [computer software]. Retrieved from http://sitemaker.umich.edu/group-based
Tran NA, Nathan MJ (2010) Pre-college engineering studies: an investigation of the relationship between pre-college engineering studies and student achievement in science and mathematics. J Eng Educ 99(2):143–157
U. S. Department of Education. (2010). Digest of education statistics. Retrieved from http://nces.ed.gov/programs/digest/d09/tables/dt09_066.asp

U. S. Department of Education. (2014a). Digest of education statistics. Retrieved from http://nces.ed.gov/programs/digest/d13/tables/dt13_221.40.asp

U. S. Department of Education. (2014b). Digest of education statistics. Retrieved from http://nces.ed.gov/programs/digest/d13/tables/dt13_222.50.asp

U. S. Department of Education. (2014c). Digest of education statistics. Retrieved from http://nces.ed.gov/programs/digest/d13/tables/dt13_102.40.asp

U. S. Department of Education. (2014d). Public high school four-year on-time graduation rates and event dropout rates: school years 2010–11 and 2011–12. Retrieved from http://nces.ed.gov/pubs2014/2014391.pdf

U. S. Bureau of the Census. (2014). Public education finances 2012. Retrieved from http://www2.census.gov/govs/school/12f33pub.pdf

Valtorta CG, Berland LK (2015) Math, science, and engineering integration in a high school engineering course: a qualitative study. J Pre-College Eng Educ 5(1):15–29

Wang HH, Moore T, Roehrig G, Park MS (2011) STEM integration: teacher perceptions and practice. Journal of Pre-College Engineering Education Research (J-PEER) 1(2). doi:10.5703/1288284314636

Wendell K, Rogers C (2013) Engineering design-based science, science content performance, and science attitudes in elementary school. J Eng Educ 102(4):513–540

What Works Clearinghouse (2014). Procedures and standards handbook. Retrieved from http://ies.ed.gov/ncee/wwc/

Wilson SM (2013) Professional development for science teachers. Science 340:310–313