International Journal of Engineering Education Vol. 35, No. 2, pp. 685–697, 2019
Printed in Great Britain

0949-149X/91 $3.00+0.00

Impact of Flipping a First-Year Course on Students' Ability to Complete Difficult Tasks in the Engineering Design Process*

ANN SATERBAK**
Department of Biomedical Engineering, 101 Science Drive, Box 90281, Duke University, Durham, NC 27708, USA.
E-mail: ann.saterbak@duke.edu

TRACY M. VOLZ
School of Engineering, 6100 Main St., Rice University, Houston, TX 77005, USA. E-mail: tmvolz@rice.edu

MATTHEW A. WETTERGREEN
Oshman Engineering Design Kitchen, 6100 Main St., Rice University, Houston, TX 77005, USA.
E-mail: mwettergreen@rice.edu

The impact of flipping an engineering course on student learning outcomes remains an open question because the literature presents a perplexing array of results. In this multi-year study, we evaluate the impact of flipping a first-year project-based engineering design course. We pose two questions: (1) Does a flipped project-based engineering design course produce different levels of mastery of engineering design process knowledge compared to the lecture version of the course? (2) Do students in the lecture class and flipped class achieve the same levels of design process knowledge when given tasks that range in difficulty? Three strands of data were used to assess design process knowledge: first drafts of team technical memos, pre- and post-critiques of a Gantt chart of a proposed design process, and an exam. Teams in the flipped classroom performed significantly better on the technical memos that evaluated solution ideas and established a testing plan, both of which required high-level cognitive capacities. However, no differences were found between lecture and flipped models for the exam or technical memo that established design criteria, which required low-level cognitive activities. The Gantt chart assessment produced mixed results. This study concludes that flipping a project-based design course can significantly improve student learning on more difficult tasks. We urge researchers and educators to consider the tasks and assignments given to students when drawing conclusions about the effectiveness of flipped pedagogy.

Keywords: flipped classroom; engineering design; technical memos; task difficulty

1. Background

1.1 The impact of flipped pedagogy remains unclear despite growing body of literature

Despite the numerous papers that have been published about the use of flipped pedagogy in engineering education in the past five years, engineering educators are still struggling to understand the impact of this teaching method on student learning. There are two reasons for this—the absence of an established theoretical framework and the variable methods and evidence used to determine its effectiveness. As a result, there is a persistence of mixed outcomes [1].

First, there is a lack of consensus regarding which theoretical frameworks best support the use of the flipped model in engineering education [2, 3]. Karabulut-Iglu et al.'s review lists theoretical frameworks that have been used in flipped implementations, the most common being active learning [2], which is not surprising given its widely accepted pedagogical benefits [4]. Talbert's recent book *Flipped Learning* also emphasizes the value of active learning but does not include it among the list of theoretical frameworks that explain and validate the use of flipped instruction: self-determination, cognitive load, and self-regulated learning [5]. Others have adopted constructivism, problem-based learning, and cooperative learning [6] as their theoretical frameworks.

Second, attempts to assess the effectiveness of flipped engineering courses rely primarily on results from student attitudinal surveys, focus groups, interviews, and course evaluations, not direct measures of student learning [7–11]. One recent study of a mechanical engineering course extended beyond course evaluations and noted an increase in student motivation and self-efficacy [12]. While student attitudes about teaching methods are important, attitudes alone should not determine which methods of instruction ought to be deployed in engineering courses.

While an increasing number of papers about flipped implementations report on student learning outcomes, those findings are deeply bound up in the context of the course, which includes the type of

** Corresponding author.

* Accepted 4 January 2019.

engineering course, the level of the course, learner characteristics, the type and timing of the assessments, and the difficulty of the task used to measure student learning [2, 7]. Unfortunately, most papers do not provide detailed descriptions of these study attributes. It's quite common for researchers to use generic labels such as "exam" or "project" to describe the assignments used to measure learning outcomes with only an occasional mention of their timing or difficulty. There are a few exceptions [13–15]; however, the general lack of contextual specificity in the literature makes it hard to identify features that contribute to the success or failure of a teaching intervention and to discern trends across multiple studies [16].

1.2 Flipped engineering courses produce mixed learning outcomes

Surveys of the literature on student learning outcomes in flipped engineering courses reveal mixed findings [1, 7], with some evidence of a slightly positive effect overall [2, 3, 17]. In many cases, conclusions about student performance are based solely on summative measures (e.g., exam grades, course grades, pre-/post-concept inventories) [2, 17]. Even within this subset of studies, the outcomes range from positive to negative to neutral. Gross and Musselman reported improved student performance on exams in flipped upper-level structural design courses compared to non-flipped courses [11]. Olson, on the other hand, found that students in a flipped fluid mechanics course did not perform as well on the final exam as students in the traditional lecture course [18]. Another researcher who flipped two courses reported a negative effect on exam grades in an introductory course and a positive effect in an upper-level course [19]. We previously found that partially flipping a first-year, project-based design course had no impact on student performance based on summative assessments [20]. Similarly, in a first year Environmental Engineering course, Velegol et al. assessed student performance on summative exam scores, comparing two different flipped versions (one incorporated small group work and another used reviews and quizzes) and traditional lecture. They showed that summative assessment was not impacted regardless of teaching intervention [8].

Interestingly, studies that combine summative and formative assessments in project-based engineering courses appear to capture different stages of students' acquisition, application, and mastery of knowledge. Formative assessments typically include interim project grades, homework problems, or quizzes [6, 14, 21]. Once again, the learning outcomes are mixed, sometimes within the context of a single course. For example, Mason et al. flipped a control systems course, compared student performance on quizzes and exams, and reported uneven results [13]. Students in the flipped course performed better on problems related to three out of five course concepts and on an open-ended design problem ($p = 0.001$), but all other problems produced non-significant results. Day and Foley compared student performance in a flipped, senior-level project-based human-computer interaction course to an unflipped control [15]. They analyzed homework grades, interim project grades, exam grades, and final course grades. Students performed significantly better on homework assignments and projects in the flipped section. However, students' exam grades in the flipped section, while slightly higher, were only marginally significant. Their work suggests that studies of blended learning that use only final exams as performance metrics may not fully reflect the impact of flipped instruction.

1.3 Task difficulty and assessment

To better understand how flipped instruction may influence student learning, it is useful to look at studies that have investigated how pedagogical techniques such as active learning affect students' ability to complete tasks that vary in difficulty and thus elicit different levels of critical thinking. Seminal work by Menekse et al. in an introductory materials science course teases out the effects of different active learning strategies on students' cognitive gains [16]. For their investigation, they designed questions on materials science concepts and classified them by level of difficulty: verbatim, integration or inference. Then, Menekse et al. assessed student performance on these questions and correlated the outcomes to four learning environments, ranging from interactive to constructive to active to passive. They found that students in classrooms that use interactive and constructive active learning techniques outperformed students in active or passive classrooms on inference questions, which are the most challenging. Inference questions involve multiple ideas, implicit information, and the construction of new knowledge [16]. By differentiating the type and difficulty of questions, the authors were able to see the impact of different active learning methods on performance.

Bloom's Taxonomy [22–24] has been used to differentiate the levels of critical thinking associated with the engineering design process [25]. For example, Safoutin et al. classified short design challenges and concluded that most involve low- to mid-level cognitive processing such as knowledge, comprehension or application [26]. Subsequent studies by other researchers took this a step further and assessed student performance on design tasks that require different levels of critical thinking. Atman et

al. compared how students and professional engineers approach the engineering design process and found that professionals engage in the evaluation and the creation of new knowledge more readily than novices [27]. In the context of flipped instruction, Yelamarthi, Drake and Prewitt's introductory project-based digital circuits course was also informed by Bloom's Taxonomy [14]. They reported statistically significant improvements in exam scores as well as higher mean scores on course learning outcomes, especially those associated with design tasks that involve higher order thinking.

1.4 Contributions from this study

In summary, understanding the impact of a flipped classroom model on student learning, specifically how to capture and interpret its effects, remains an open area of research. We are encouraged by the increasing number of studies documenting the successes and failures of this method of instruction. However, the outcomes they have produced are highly variable and seldom statistically significant as reported in reviews of the literature [1–3, 7]. This situation is perplexing and undermines the legitimacy of the approach.

Our multi-year study of a flipped, project-based engineering design course uses a combination of formative and summative assessments to investigate whether task difficulty affects learning outcomes. This study is unique because it features consistent course instructors, content and assignments, formative and summative assessments, and a high number of participants for statistical power.

2. Study design

2.1 Course description

Introduction to Engineering Design is a semester-long, first-year design course. In constructing the course, we implemented a combination of active learning techniques, including cooperative (or team-based) learning and project-based learning. In this elective course, first-year students work in multidisciplinary teams on an open-ended, client-sponsored design project. Students are expected to achieve three learning outcomes upon completion of the course:

1. Successfully solve a client-based design challenge by following steps in the engineering design process (Fig. 1).
2. Effectively communicate the progress of their design project through written, oral, and visual communication.
3. Develop project management skills and function effectively on a high-performance team.

During the first half of the semester, students complete the first five steps of the engineering design process. During the second half of the semester, students construct their design (physical object or computer program) and iteratively prototype and test until a final solution is reached. Previous work describes the course implementation, including community partners, projects, assignments, and logistics [28, 29].

2.2 Course assignments

There are nine technical memos assignments, which constitute the largest percentage of the course grade. The 1–4 page technical memos document the team's work and justify the team's decisions at each step in the engineering design process (Fig. 1). The instructors use technical memos to track each team's performance and to provide feedback as the team progresses through its design project. Teams are also evaluated on their final prototype, two oral presentations, their contributions to their team, and class participation. Other assignments include an exam as well as a Gantt chart assignment [20].

2.3 Flipping the course

From the inception of the course, class time has been divided between instruction and team meetings. From 2011 through spring 2014, during the first 30–45 min of class, the faculty typically lectured on a step in the engineering design process or a professional skill. During the remainder of the 75-min class, students worked in teams to apply that new knowledge to solve their design challenge. Based on instructor observations and student feedback, students struggled to jump from passively listening to a lecture to applying that information to their design project within the class period. This problem motivated us to transform the lecture component of the course to a flipped model [20].

By fall 2015, we had created all of the materials to

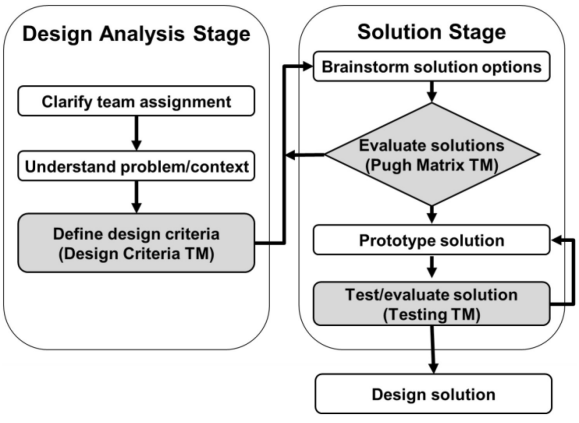

Fig. 1. Engineering design process, including technical memo assignments used for assessment (adapted from [20]).

fully implement a flipped classroom for engineering design. In this model, students view video lectures [30] prior to class and take a short, multiple-choice quiz to assess their understanding of the material. The videos cover all steps in the engineering design process (Fig. 1) as well as professional skills (e.g., teaming, communication). During the instruction time, students complete in-class exercises (ICEs) in small groups facilitated by the faculty; these ICEs were designed to reinforce key concepts and illustrate common misconceptions [31]. This strategy facilitates the scaffolding of knowledge [32] and ensures the deliberate practice of important tasks prior to students tackling their own project. Following this 30 min of practice, teams apply those learnings to their design project. Note that the content and topics covered in the course did not change upon adoption of the flipped model. The flipped classroom materials (videos, quizzes, and ICEs) and its implementation were previously described [20].

2.4 Current study and research questions

As previously noted, few studies have used multiple assessments of student learning to measure the effects of flipped instruction in a project-based design course. We build upon our previous work in which we analyzed an exam and Gantt chart evaluation [20]. Here, we additionally analyze changes in student performance on technical memo writing assignments to assess the effects of fully flipping the class on student understanding and application of the design process. Unlike other studies [10], we used the same assessment methods for both fully flipped and traditional lecture environments, allowing us to directly compare the effects of flipping on students' design process knowledge.

This study examines the following research questions (RQ):

RQ#1. Does a flipped project-based engineering design course produce different levels of mastery of engineering design process knowledge compared to the lecture version of the course?

RQ#2. Do students in the lecture class and flipped class achieve the same levels of design process knowledge when given tasks that range in difficulty?

3. Research methods

3.1 Participants

Participants in this study were first-year students at a highly-selective, STEM-focused private university in the United States. The characteristics (e.g., SAT scores, AP scores) of the incoming students across this study were unchanged. All students either intended to major in engineering or were strongly considering it. We obtained IRB approval from our institution as well as written consent from all participants whose coursework is included in the study.

3.2 Overview of assessment methods

While engineering researchers have provided overviews of engineering design instruction and assessment methods [33–35], no standard method exists to quantitatively assess student learning after students engage in design activities. Therefore, we used the course work, as it reflects the diversity of tasks the teams were asked to complete. Guided by Bloom's Taxonomy, Bailey and Szabo recommend using writing assignments to discern higher order thinking [25]. To specifically investigate the impact of fully flipping the classroom on student performance, we used three strands of assessment data taken from student assignments:

1. Performance on first drafts of team technical memos assessing students' ability to:
   (a) Develop design criteria based on user-defined needs and constraints.
   (b) Create a Pugh Matrix to evaluate solution options and select a design.
   (c) Describe tests to measure the extent to which the prototype fulfills the design criteria.
2. Performance at the beginning (pre-test) and end (post-test) of the semester when students evaluated a 14-week Gantt chart of the design process.
3. Performance on an end-of-term exam that covers the application of steps in the engineering design process and descriptions of best practices of professional skills.

Data from three academic years (2013–2016) were included in the study.

3.3 Technical memos (TMs)

3.3.1 Selection of the TMs

For the purpose of this study, we chose to analyze the three TMs that emphasize design criteria at different steps in the design process. The Design Criteria TM requires teams to establish and justify design criteria that conform to a client's needs. The Pugh Matrix TM asks teams to evaluate possible design solutions against their established design criteria using both Screening and Scoring Pugh Matrices; the outcome of this effort is a selected design solution. In the Testing TM, teams are asked to provide detailed descriptions of the tests they plan to perform on their design prototype to determine whether it satisfies the established design

criteria. While all three of these TM assignments share a common focus on design criteria, the assignments vary in difficulty.

3.3.2 Development of scoring rubrics and training materials

To analyze the memos, we developed scoring rubrics to evaluate specific technical attributes related to the application of design knowledge. Note that these rubrics are different from the grading rubrics used in the class. The assessment rubrics focus on technical application (and did not evaluate usage, clarity, and other topics present in the grading rubrics). The number of tracked features varies across TMs from three (Design Criteria TM) to six (Testing TM) to 11 (Pugh Matrix TM). Some features yield one result per team, thus sample size ranges from 60 to 74. For features with multiple results, such as each team's list of design criteria, data was collected from alternating criteria; sample size per team was two to four, and thus total sample size ranged up to 287.

We also produced detailed rater training materials to ensure accurate and consistent evaluations. The scoring rubrics required raters to apply numerical rating scales (e.g., 0, 1, 2, 3) or qualitative rating scales (e.g., yes, no) to evaluate student performance. For example, the scoring rubric for the Design Criteria TM required raters to determine if quantitative design criteria were stated.

3.3.3 Raters and scoring TMs

We recruited 14 recent engineering alumni to participate in the assessment of the TMs and compensated them for their time. Alumni had either completed Introduction to Engineering Design as first-year students or had served as a teaching assistant for the course. For each of the three TM types, 60–74 memos were scored by one or two pair of trained raters. Raters did not score TMs they wrote when enrolled in the course.

3.4 Gantt chart assignment

A Gantt chart assignment was administered to the students in the first and last weeks of the course to evaluate their knowledge of the design process. Both the pre- and post-tests were administered as take-home, individual assignments as described previously [20]. Briefly, students were given a short explanation of Gantt charts and were then asked to critique a 14-week design project schedule. Student responses varied in length from 0.5–2 pages.

Trained raters used an adapted rubric from Bailey and Szabo to rate six levels (i.e., design topics) on a 3-point scale (0, 1, 2) [25, 29]. As before, responses were included in the analysis only if both pre- and post-tests were completed [36]. Data includes 78 student responses from the lecture course and 92 responses from the fully flipped course. Each level of each Gantt chart was scored by a pair of raters, who were upper-class writing mentors and teaching assistants for the course.

3.5 Exam assignment

An exam given toward the end of the course assessed design process knowledge and general professional skills. Data from the exam include 132 responses from lecture and 120 from flipped courses. The instructors and senior teaching assistants graded the exams following a well-specified rubric.

3.6 Data management and statistical analysis

All identifying information was stripped (e.g., name, date, and pre- vs post-, if necessary) and assignments were randomized prior to evaluation. For all statistical analysis, a p-value < 0.05 indicates a statistical difference between measurements.

3.6.1 Bhapkar test for Gantt chart (pre- versus post-test improvement in each level)

For paired, ordered categorical data, a Bhapkar test was used [37] to compare gains in the pre- versus post-test for each level of the Gantt chart. We first calculated the average rater score for each student submission and then summarized the data in a 5×5 contingency table (each table side as 0, 0.5, 1, 1.5, 2). This analysis directly incorporates the pairing of the pre- and post-test data, as each student's pre-test and post-test value results in one value scored. Thus, trends for pre-test and post-test responses can be seen, as well as changes in student responses. The test statistic and p-value were computed from the 5×5 table.

3.6.2 Permutation test for TMs and Gantt chart (flipped versus lecture improvement)

For non-paired, nominal data, a permutation test was used to compare gains in student performance for flipped versus lecture models for all features in the TMs and the Gantt chart assessment. Briefly, all data for a specific assignment (mean ratings) were randomized, and a series of 1,000,000 permutations of the data was run to evaluate if data from the flipped model fell under the same distribution as the lecture model [29].

3.6.3 Method to calculate difference between exam scores

For the exam, we used a 2-sided t-test to compare student grades in the flipped versus the lecture mode.

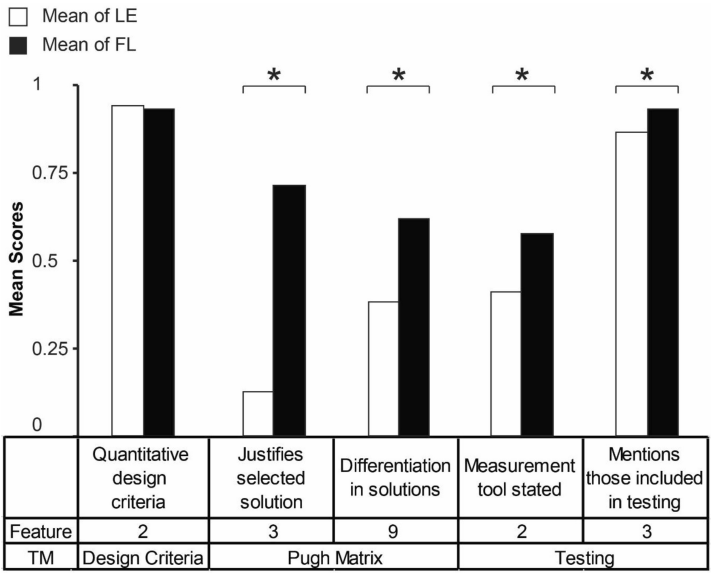

Fig. 2. Representative summary of the Design Criteria TM, Pugh Matrix TM, and Testing TM that compare mean scores of lecture (LE) to fully flipped (FL) models. Asterisk (*) indicates significant difference between mean FL and mean LE ($p < 0.05$).

3.6.4 Rater consistency

Fleiss' kappa (κ) measurement was used to evaluate inter-rater reliability for a fixed pair of raters [38, 39]. This method can skew when analyzing binary (e.g., yes/no) data, and a single disagreement can skew the kappa values dramatically, resulting in a very low or negative agreement. Because of this issue, we also ran an agreement proportion test [40]. Average kappa and percent agreement are calculated for the technical memo and Gantt chart assignments.

3.7 Evaluation of task difficulty

In the absence of a widely-accepted model that describes the degree of difficulty of tasks associated with the engineering design process, we borrowed from Safoutin et al.'s [26] design attribute framework, which is based on Bloom's Taxonomy [22]. The course instructors, together with other experts, categorized the cognitive capabilities necessary to complete the assignments tasks, especially as they relate to the features (TM rubrics) and levels (Gantt chart assessment). For example, a task that requires students to apply a rule (e.g., each criterion needs a standard in a Pugh Matrix) was categorized as "Apply." In contrast, a task that required describing that a Pugh Matrix can be used to evaluate solutions was categorized as "Remember."

4. Results

4.1 Results from formative TM assessments

We assessed the mean gains in student performance by comparing mean flipped (FL) and lecture (LE) ratings for each TM feature. Fig. 2 shows representative features that we assessed from the three TMs.

4.1.1 Design criteria memo

We found that students performed equally well on the Design Criteria TM in the lecture and flipped models; in other words, students had the same mean ratings on three of three features (Table 1). For example, Feature #2 indicates whether a stated design criterion is quantitative (0.95 for LE and a 0.94 for FL on a 0/1 binary scale). Thus, the flipped model did not affect student performance on assess-

Table 1. Rated features of Design Criteria TM. Differences between the lecture (LE) and fully flipped (FL) models are significant when $p < 0.05$

#	Feature of Design Criteria TM	Rating scale*	Mean LE	Mean FL	*p*-value	Difference LE vs FL?
1	Describes attribute specifically	1, 2, or 3	2.45	2.42	0.664	No
2	Design criterion is quantitative	Y or N	0.95	0.94	0.846	No
3	Design criterion is justified	0, 1, 2, or 3	1.60	1.65	0.501	No

* Abbreviations: Y = Yes, N = No; Point scales: (Y/N = 1/0), all others correspond to their numerical values. Sample size ranges from 285 to 287 taken from the TMs of 74 teams.

ment strand *1(a): Develop design criteria based on user-defined needs and constraints.*

4.1.2 Pugh matrix memo

In contrast to the Design Criteria TM, students demonstrated significant improvement in the flipped model relative to the lecture model for the Pugh Matrix TM (Table 2). Specifically, there was a greater mean rating in eight of 11 features, relating to strand *1(b): Create a Pugh Matrix to evaluate solution options and select a design.*

Statistically significant improvements were seen across the majority of queries regarding the quality of the Pugh Matrices. For example, students' ability to justify their selected solutions by addressing design criteria (Feature #3) increased from 0.13 in the LE to 0.73 in the FL model on a 0/1 binary scale. The means for differentiation of solutions (Feature #9) were 0.39 for LE and 0.63 for FL. We conclude that the flipped model had a positive effect on student performance related to developing Pugh Matrices.

4.1.3 Testing memo

Students also demonstrated significant improvement in the flipped model relative to the lecture model for the Testing TM (Table 3). Specifically, there was a greater mean rating in five of six features of the Testing TM, relating to strand *1(c): Describe tests to measure the extent to which the prototype fulfilled the design criteria.*

For example, the measurement tool used for testing (Feature #2) was included more frequently in the FL model (0.59 on a 0/1 binary scale) compared to the LE model (0.42). Similarly, teams mention the test participants (Feature #3) more often for the FL model (0.94) versus the LE model (0.88). Thus, we conclude that the fully flipped model also had a positive effect on student performance for the learning outcome related to developing a testing plan.

4.2 Results for summative Gantt chart assessment

Students in the fully flipped model gained knowledge ($p < 0.001$) of the engineering design process

Table 2. Rated features of Pugh Matrix TM. Differences between the lecture (LE) and fully flipped (FL) models are significant when $p < 0.05$

#	Feature of Pugh Matrix TM	Rating scale*	Mean LE	Mean FL	*p*-value	Difference LE vs FL?
Overall Features of Pugh Matrix TM						
1	# of Screening Pugh Matrices shown as a table	Count from 0	0.96	2.71	< 0.001	Yes
2	# of Scoring Pugh Matrices shown as a table	Count from 0	1.09	1.31	0.061	No
3	Justifies selected solution by addressing design criteria	Y or N	0.13	0.73	< 0.001	Yes
Errors in Scoring Pugh Matrix						
4	Row missing the standard value of 3	0, 1, 2, 3+ errors	0.48	0.48	0.918	No
5	Only 3 point range used, given 5 point scale	0, 1, 2, 3+ errors	2.15	1.29	< 0.001	Yes
6	Design criteria weighted < 5%	0, 1, 2, 3+ errors	0.96	0.35	< 0.001	Yes
Characteristics of Scoring Pugh Matrix						
7	Appropriate # of design criteria (e.g., 4–7)	Y or N	0.83	0.91	0.083	No
8	Standard value of 3 floats across solutions	Y or N	0.76	0.98	< 0.001	Yes
9	Differentiation in solutions (i.e., not clustered)	Y or N	0.39	0.63	0.009	Yes
10	# of levels (1–5) that are specified in detail	Count from 0	1.40	3.86	< 0.001	Yes
11	Levels (1–5) are justified	0, 1, 2, or 3	0.18	0.41	0.002	Yes

* Abbreviations: Y = Yes, N = No; Point scales: (Y/N = 1/0), all others correspond to their numerical values. Sample sizes range from 60 to 221 taken from the TMs of 60–67 teams.

Table 3. Rated features of Testing TM. Differences between the lecture (LE) and fully flipped (FL) models are significant when $p < 0.05$

#	Feature of Testing TM	Rating scale*	Mean LE	Mean FL	*p*-value	Difference LE vs FL?
1	Detailed test description	0, 1, 2, or 3	1.80	2.06	< 0.001	Yes
2	Measurement tool stated	Y or N	0.42	0.59	< 0.001	Yes
3	Includes people involved in testing	Y or N	0.88	0.94	0.021	Yes
4	# tests / trials listed	Y or N	0.55	0.73	< 0.001	Yes
5	# of Likert or user-defined scales created	Count from 0	0.45	0.84	0.031	Yes
6	Target value or range established on scale	Y or N	0.67	0.68	0.808	No

* Abbreviations: Y = Yes, N = No; Point scales: (Y/N = 1/0), all others correspond to their numerical values. Sample size ranged from 41 to 253 taken from the TMs of 65 teams.

Table 4. Comparison of pre-test values versus post-test values for the Gantt chart assessment for the lecture and fully flipped model

#	Level Description	Pre-test (mean ± stdev*)	Post-test (mean ± stdev*)	*p*-value
1	Needs assessment and establishing design criteria	0.49 ± 0.60	1.40 ± 0.74	< 0.001
2	Design context review	0.52 ± 0.72	1.37 ± 0.76	< 0.001
3	Idea generation/brainstorming	1.30 ± 0.58	1.71 ± 0.47	< 0.001
4	Analysis and decision-making	0.89 ± 0.48	1.84 ± 0.43	< 0.001
5	Building and testing	1.57 ± 0.51	1.80 ± 0.41	< 0.001
6	Overall layout of a design process and iteration	1.15 ± 0.55	1.74 ± 0.42	< 0.001

* Each level in the pre- and post-tests was scored on a 3-point scale (0, 1, 2), with 0 being low and 2 being high. Sample size is 92.

Table 5. Results comparing performance gain in Gantt chart exercise for lecture (LE) and fully flipped (FL) models. Mean gains are shown as well as *p*-values, which capture the difference between the two models

#	Level Description	Mean Gain for LE* [20]	Mean Gain for FL†	*p*-value	Difference LE vs FL?
1	Needs assessment and establishing design criteria	0.97	0.91	0.617	No
2	Design context review	0.81	0.85	0.762	No
3	Idea generation/brainstorming	0.39	0.41	0.833	No
4	Analysis and decision-making	1.10	0.95	0.141	No
5	Building and testing	0.10	0.23	0.093	No
6	Overall layout of a design process and iteration	0.31	0.59	0.004	Yes

* Mean gain for lecture = post-test LE mean – pre-test LE mean; † Mean gain for flipped = post-test FL mean – pre-test FL mean. Sample size is 170.

Table 6. Results comparing exam performance for lecture (LE) and flipped (FL) models

	Mean ± Stdev (%) LE*	Mean ± Stdev (%) FL*	*p*-value	Difference LE vs FL?
Exam Grade	83.5 ± 9.2	81.5 ± 9.7	0.091	No

* Grade percentage, reported as mean ± standard deviation. Sample size is 252.

by the end of the semester for all levels of the Gantt chart assignment (Table 4). To compare the lecture [20] and flipped models, we used a permutation test with the paired pre- and post-test values for six levels in the Gantt chart exercise. Mean gains are shown as well as *p*-values, which capture the difference between the two didactic models (Table 5). For Levels 1–5, the gains were not statistically significantly different between the lecture and flipped models. In contrast, there was a significant gain in understanding of Level 6: *Overall layout of a design process and iteration* ($p < 0.005$) for the fully flipped model relative to the lecture model.

4.3 Results for summative exam assessment

The results for the end-of-term exam scores show no significant change between the lecture model compared to the flipped model ($p > 0.05$) (Table 6).

4.4 Verification of inter-rater reliability

Overall, the inter-rater reliability and percent agreement calculations demonstrate that our methods were reliable and robust. For the TM analysis, the mean kappa value was 0.70, and the percent agreement was 0.85. For the Gantt chart analysis, the average kappa value was 0.64 and percent agreement was 0.79.

5. Discussion

5.1 Impact of flipped learning—results of RQ#1

This study evaluated the impact of flipping a first-year engineering design course through five separate assignments that measured different aspects of the student learning outcomes. Overall, we see mixed results in response to RQ#1 (Table 7). Specifically, we saw improvements from lecture to flipped for the Pugh Matrix and Testing TMs as well as Level 6 of the Gantt chart. We did not see changes for the Design Criteria TM, Levels 1–5 of the Gantt chart, or the exam. In no case was the performance of the students in the fully flipped model lower than in the lecture model. Thus, we found that flipping does not harm students' understanding of the engineering design process, which is consistent with our previous work [20]. Importantly, we found that some assessments showed improvements upon flipping, suggesting that students benefitted from this change. Because our mixed results are within the context of one class, we have a unique opportunity to offer an explanation of these differences.

Table 7. Summary of assessment results for lecture (LE) and flipped (FL) models

	Assessment Method	Difference LE vs FL?
Summative	Gantt chart assessment, Levels 1–5	No
	Gantt chart assessment, Level 6	Yes
	Exam	No
Formative	Design Criteria TM	No
	Pugh Matrix TM	Yes
	Testing TM	Yes

5.2 Task difficulty and complexity as a moderator

In the work described earlier by Menekse et al. [16], more cooperative forms of active learning enabled students' ability to answer more difficult questions. Specifically, when asked to generate ideas beyond the information presented, students in classrooms that use interactive activities (i.e., pairs of students who construct, build and resolve knowledge) outperformed students in less active or passive classrooms. However, had the authors looked only at the overall quiz score rather than analyzing by question type, no difference in student performance would have been detected among the teaching methods.

Building on the idea that a clear differentiation in question type may be needed to observe the impact of teaching methods, we discuss our mixed results in light of the difficulty of the task. Given the structure and expectations of the ENGI 120 class, we posit that the difficulty of assignments is the main moderator of our mixed assessment results (RQ#2).

5.3 Technical memos assignments vary in difficulty

In a professional setting, teams gather research, customer interviews, and product surveys and then synthesize that information into design criteria. Thus, establishing design criteria can be a complicated and difficult task. In an academic environment, Atman et al. found that both freshman and senior engineering design students struggled to identify design criteria [27]. In the study, the project prompt (given in the article) was open-ended and, as a result, difficult for inexperienced students.

In contrast, the task of defining design criteria was relatively straightforward in ENGI 120. Students had access to the project prompt (written by an instructor) that clearly stated some of the important design criteria (with associated target values) prior to completing an interview with their client to clarify any uncertainty. When preparing the Design Criteria TM, teams needed to simply recall and define their quantitative design criteria; these tasks are low on Bloom's Taxonomy. The low-level of difficulty of the assignment was evident in the lecture model values for Features #1 and #2, which were near the upper end of the scale (Table 1). The students were already performing well on this assignment in the lecture model, and no change was seen in the flipped model. Feature #3 assesses the justification of selected design criteria, which was done poorly by students in the lecture and flipped models.

As noted by Dym and colleagues in a seminal paper on design thinking, the cycles of divergent and convergent thinking necessary for successful design are very difficult cognitive tasks [34]. Arguably the most difficult assignment in our course, the Pugh Matrix TM challenged students to evaluate potential design concepts against their established design criteria, converge on a solution, and justify their chosen solution. Students began by applying a Screening Pugh Matrix to the 30+ design concepts generated during brainstorming. Design concepts were compared to a "standard" for each design criterion, rated as better, worse, or same. This screening process narrowed the field to 5–10 viable designs. Then, students created a Scoring Pugh Matrix wherein each design criterion was assigned a weight based on the client's priorities. Students then used engineering analysis and research to rate how well each newly-created design concept satisfied each design criterion using a scale of 1–5. For this assignment, students not only learn two new tables and how to create them, but they must also evaluate large numbers of ideas using a rigorous and structured method and then justify those scores.

Considering these tasks through the lens of Bloom's Taxonomy, students must understand and apply new heuristics to construct their Pugh Matrix. Correctly constructing and applying a Pugh Matrix is captured in Features #5, #8, #9, and #10. To construct a technically sound Scoring Pugh Matrix, teams must analyze many potential solutions and critically evaluate them. Steps associated with these higher levels of difficulty include Features #6, #11, #12 and #13. Overall, the flipped model enabled students to perform significantly better on this more challenging assignment as compared to the control.

Another difficult task in the course was the development of a testing plan. In the Testing TM, students provided detailed descriptions of the tests they planned to perform on their design prototype

to determine whether it satisfied the established design criteria. For each test, students had to identify the criterion being measured, the method used, the number of testing trials, and the people involved in conducting the tests. This involved students applying their knowledge to a new situation, and teams performed significantly better in the flipped model than in the lecture model.

As the major formative assessment tool in this study, these technical memo results reveal how students perform when presented with assignments requiring different cognitive capabilities. We see that for the Design Criteria TM assignment requiring recall and understanding, there was no difference upon flipping the class. But, we see that for the Pugh Matrix and Testing TMs requiring application and evaluation, there was a significant, positive impact upon flipping the class. Similar to Menekse et al. [16] and Yelamarthi et al. [14], the impact of an alternate teaching method was only measured when difficult tasks were presented. Stated another way, the impact of a flipped classroom model may be evident only if the assessment tool captures students completing difficult or challenging tasks.

5.4 Gantt chart exercise requires a mix of cognitive processes

The Gantt chart exercise was established to measure acquisition and application of knowledge about the design process. Bailey and Szabo [25] argue that to complete the Gantt chart exercise, students need to elaborate (requiring remembering and understanding) as well as identify the pros and cons (requiring evaluation and critique). As a summative form of assessement in ENGI 120, this exercise captured what students learned from the didatic material and by completing their design project.

In this exercise students critique a poorly constructed, 14-week Gantt chart. The assessment rubric for Levels 1-5 focused on simple explanations [36] and can be summarized:

0—No mention of step or topic.
1—Mentions that this step should be completed.
2—Mentions how or why a step should be completed.

To evaluate Levels 1–5, the raters evaluated the depth of discussion for each step in the design process. Raters were instructed to allocate points for a level based on the discussion of ''what'' and ''how''—tasks that require recall, understanding and some application. It should be noted that Levels 1–5 tracked disrete steps in the engineering design process, each of which was associated with a lecture or a video/quiz/ICE, as well as completing the project and writing a technical memo. In comparing differences between the lecture and fully flipped models, no statistically signficant differences were seen for Levels 1–5. This result is consistent with our previous work that compares the lecture and partially flipped models [20].

Level 6 of the Gantt chart exercise was different in that it was evaluated holistically by reading the full paper. Raters were instructed to determine whether the student demonstrated a grasp of the overall design process and understood that design is not strictly linear. As noted by Atman et al., iteration during the engineering design process is a complex, high-level task, and experienced engineers more fluidly move among steps as the unique design problem dictates [27]. In order to earn a high score of 2 on Level 6, students had to demonstrate an understanding of iteration, overall allocation of time, and/or concurrent or sequential activities. In contrast to Levels 1–5, Level 6 drew upon a student's schema of the design process and its high-level features and attributes. In fact, there was not an explicit video module or lecture on this topic; instead, students had to construct this knowledge from the repeated mention of the design process in the other course lectures or videos and completion of the project.

Because Level 6 reflects an integrated and holistic understanding of the design process, it is notable that students in the fully flipped model scored statistically signficantly higher than the students in the lecture model. These results are in contrast to our previous work that compares the lecture and partially flipped models [20]. We attribute this to the partial implementation of the flipped model in the previous study.

5.5 Summative exam designed to reveal general student understanding of design

Administered toward the end of the course, the exam tested design process knowledge and related professional skills. The summative exam was written based on content taught in the lecture or video formats, depending on the year. Students were asked questions such as:

- State four quantitative design criteria for a posed design problem.
- Name and elaborate on three rules of brainstorming.
- Complete a pairwise comparison chart given ranking information.

As designed, these questions required students to recall, understand and apply knowledge from the class. Students were allowed to prepare a one-page ''cheat sheet'' with key words, definitions, and examples. With mean exam scores in the 80's, there was no statistically significant difference between the fully flipped and lecture models. This

result is consistent with our previous work comparing the lecture to a partially flipped class [20]. The exam was not challenging for most students because it did not require them to synthesize or evaluate their knowledge of the engineering design process or apply it in a new way. Others have questioned the validity of using exams to measure proficiency in design practice [34, 35].

5.6 Recommendations for assessing flipped instruction

Our results prompt us to make recommendations for instructors and educational researchers. First, we strongly encourage instructors experimenting with the flipped model to consider the way in which task difficulty may affect measured learning outcomes. In the design of assessments, instructors should use this knowledge as a lens to contextualize results. At a minimum, we recommend that details about the assessments and their difficulty be published alongside evaluation of learning outcomes to inform the interpretation of results and to allow for comparisons across studies.

We also recommend using a combination of formative and summative assessments because summative assessments may mask the effects of flipped instruction. Summative measures reflect explicit didactic instruction (lecture, video, coaching) as well as the accumulation of knowledge that occurs throughout the completion of a course, whereas formative assessment measures are more likely to capture students' first attempt at application. Therefore, it's important to consider the timing of the assessments in study design, particularly in project-based courses.

6. Conclusions

This study focused on a first-year, project-based engineering design course. Based on observations and student feedback, we saw students struggle to jump from passively listening to a lecture to applying that information to their own design project within the same class period. To respond to this need, we flipped the course and measured student learning outcomes in the lecture and fully flipped models. This study included a sample size of >200 pieces of student work, from six semesters representing 12 sections of the course.

Our assessment methods represent some advances over previous approaches. We are one of the few studies to have used multiple assessments of student learning to measure the effects of flipped instruction on engineering students' ability to apply their knowledge of the design process in a fully flipped project-based design course. Specifically, three strands of data were used to measure the application of engineering design process knowledge. We used the same assessment methods for both fully flipped and traditional lecture environments, allowing direct comparison of classroom models.

Our mixed findings underscore the complexity of answering the simple question: "Do learning outcomes improve in a flipped classroom?" Our study produced mixed results, which indicate that flipping a project-based course can significantly improve student learning on tasks of higher difficulty or that require higher levels of critical thinking. We urge researchers and educators to carefully consider the tasks and assignments given to students when drawing conclusions about the effectiveness of flipped pedagogy.

Acknowledgments—We would like to thank Dr. Qiwei Li for performing the statistical analysis, Dr. Jordan Trachtenberg for contributing to the manuscript, and Dr. Margaret Beier for commenting on the manuscript. We would also like to acknowledge the efforts of all of the raters. This work was supported by an NSF DUE grant (#1244928).

References

1. L. Abeysekera and P. Dawson, Motivation and cognitive load in the flipped classroom: Definition, rationale and a call for research, *Higher Education. Research & Development*, **34**(1), pp. 1–14, 2015.
2. A. Karabulut-Ilgu, N. J. Cherrez and C. T. Jahren, A systematic review of research on the flipped learning method in engineering education, *Br. J. Educat. Tech.,* **49**(3), pp. 398–411, 2018.
3. J. O'Flaherty and C. Phillips, The use of flipped classrooms in higher education: A scoping review, *The Internet and Higher Education*, **25**, pp. 85–95, 2015.
4. S. Freeman, S. Eddy, M. McDonough, M. Smith, N. Okoroafor, H. Jordt and M. P. Wenderoth, Active learning increases student performance in science, engineering, and mathematics, *PNAS*, **111**(23), pp. 8410–8415, 2014.
5. R. Talbert, *Flipped learning: A Guide for Higher Education Faculty*, Sterling, VA: Stylus Publishing, 2017.
6. E. M. Choi, Applying inverted classroom to software engineering education, *International Journal of e-Education, e-Business, e-Management and e-Learning*, **3**(2), pp. 121–125, 2013.
7. J. Bishop and M. Verleger, The flipped classroom: A survey of the research, in *Proceedings of the 120th Annual ASEE Annual Conference & Exposition*, Atlanta, GA, 2013.
8. S. B. Velegol, S. E. Zappe and E. Mahoney, The evolution of a flipped classroom: Evidence-based recommendations, *Advances in Engineering Education*, **4**(3), pp. 1–37, 2015.
9. J. Everett, J. Morgan, K. Mallouk and J. Stanzione, A hybrid flipped first year engineering course, in *Proceedings of the 6th First Year Engineering Experience Conference*, College Station, TX, 2014.
10. K. M. Calabro, Flipping the classroom on an established introduction to engineering design course, in *Proceedings of the 5th First Year Engineering Experience Conference*, Pittsburgh, PA, 2013.
11. S. Gross and E. Musselman, Observations from three years of implementing an inverted (flipped) classroom approach in structural design courses, in *Proceedings of the 122nd ASEE Annual Conference & Exposition*, Seattle, WA, 2015.
12. J. Yan, L. Li, J. Yin and Y. Nie, A comparison of flipped and traditional classroom learning: A case study in mechanical engineering, *International Journal of Engineering Education*, **34**(6), pp. 1876–1887, 2018.

13. G. S. Mason, T. R. Shuman and K. E. Cook, Comparing the effectiveness of an inverted classroom to a traditional classroom in an upper-division engineering course, *IEEE Transactions on Education*, **56**(4), pp. 430–435, 2013.
14. K. Yelamarthi, E. Drake and M. Prewett, An instructional design framework to improve student learning in a first-year engineering class, *Journal of Information Technology Education: Innovations in Practice*, **15**, pp. 195–222, 2016.
15. J. A. Day and J. D. Foley, Evaluating a web lecture intervention in a human-computer interaction course, *IEEE Transactions on Education*, **49**(4), pp. 420–431, 2006.
16. M. Menekse, G. S. Stump, S. Krause and M. Chi, Differentiated overt learning activities for effective instruction in engineering classrooms, *Journal of Engineering Education*, **102**(3), pp. 346–374, 2013.
17. G. V. Oddsson and R. Unnthorsson, Flipped classroom improves the student's exam performance in a first year engineering course, *International Journal of Engineering Education,* **33**(6A), pp. 1776–1785, 2017.
18. R. Olson, Flipping engineering probability and statistics – lessons learned for faculty considering the switch, in *Proceedings of the 121st ASEE Annual Conference & Exposition*, Indianapolis, IN, 2014.
19. J. Maarek and B. Kay, Assessment of performance and student feedback in the flipped classroom, in *Proceedings of the 122nd ASEE Annual Conference & Exposition*, Seattle, WA, 2015.
20. A. Saterbak, T. Volz and M. Wettergreen, Implementing and assessing a flipped classroom model for first-year engineering design, *Advances in Engineering Education*, **5**(3), pp. 1–29, 2016.
21. C. A. Reidsema, L. Kavanagh and J. E. McCredden, Project design and scaffolding for realising practitioner learning in a large first year flipped classroom course, in *Proceedings of the 27th Australasian Association for Engineering Education Conference*, Coffs Harbour, NSW, Australia, 2016.
22. B. S. Bloom, M. D. Engelhart, E. J. Furst, W. H. Hill and D. R. Krathwohl, *Taxonomy of Educational Objectives, Handook 1: Cognitive Domain*, New York: McKay, 1956.
23. R. Irish, Engineering thinking: Using Benjamin Bloom and William Perry to design assignments, *Language and Learning Across the Disciplines*, **3**(2), pp. 83–102, 1999.
24. D. Krathwohl, A revision of Bloom's taxonomy: An overview, *Theory into Practice*, **41**(4), pp. 212–218, 2002.
25. R. Bailey and Z. Szabo, Assessing engineering design process knowledge, *International Journal of Engineering Education*, **22**(3), pp. 508–518, 2006.
26. M. J. Safoutin, C. J. Atman, R. S. Adams, T. Rutar, J. C. Kramlich and J. L. Fridley, A design attribute framework for course planning and learning assessment, *IEEE Transactions on Education,* **43**(2), pp. 188–199, 2000.
27. C. J. Atman, R. S. Adams, M. E. Cardella, J. Turns, S. Mosborg and J. Saleem, Engineering design processes: A comparison of students and expert practitioners, *Journal of Engineering Education*, **96**(4), pp. 359–379, 2007.
28. A. Saterbak, M. Embree and M. Oden, Client-based projects in freshman design, in *Proceedings of the 119th ASEE Annual Conference & Exposition*, San Antonio, TX, 2012.
29. A. Saterbak and T. Volz, Assessing design capabilities following a client-based freshman design course, in *Proceedings of the 4th First Year Engineering Experience Conference*, Pittsburgh, PA, 2012.
30. M. Wettergreen and A. Saterbak, Learn Engineering Design, YouTube, [Online]. Available: http://goo.gl/dPMdaO. [Accessed 10 December 2018].
31. A. Saterbak and M. Wettergreen, Engineering Design Materials, [Online]. Available: https://goo.gl/A7cK4S. [Accessed 10 December 2018].
32. J. M. Le Doux and A. A. Waller, The problem solving studio: An apprenticeship environment for aspiring engineers, *Advances in Engineering Education*, **5**(3), pp. 1–27, 2016.
33. S. Sheppard and R. Jenison, Freshman engineering design experiences: An organizational framework, *International Journal of Engineering Education,* **13**(3), pp. 190–197, 1997.
34. C. L. Dym, A. M. Agogino, O. Eris, D. D. Frey and L. J. Leifer, Engineering design thinking, teaching, and learning, *Journal of Engineering Education*, **94**(1), pp. 103–120, 2005.
35. C. Atman, R. S. Adams and J. Turner, Using multiple methods to evaluate a freshman course, in *Proceedings of the 30th ASEE/IEEE Frontiers in Education Conference*, Kansas City, MO, 2000.
36. A. Saterbak and T. Volz, Assessing knowledge and application of the design process, in *Proceedings of the 121st ASEE Annual Conference & Exposition*, Indianapolis, IN, 2014.
37. V. P. Bhapkar, A note on the equivalence of two test criteria for hypotheses in categorical data, *Journal of the American Statistical Association*, **61**(313), pp. 228–235, 1966.
38. J. L. Fleiss, Measuring nominal scale agreement among many raters, *Psychological Bulletin*,**76**(5), pp. 378–382, 1971.
39. K. L. Gwet, Testing the difference of correlated agreement coefficients for statistical significance, *Educational and Psychological Measurement*, **76**(4), pp. 609–637, 2016.
40. A. R. Feinstein and D. V. Cicchetti, High agreement but low kappa: I. The problems of two paradoxes, *J. Clin. Epidemiol.*, **43**(6), pp. 543–549, 1990.

Dr. Ann Saterbak is Professor of the Practice in Biomedical Engineering and Director of the First-Year Engineering Program. Since joining Duke in June 2017, she launched the new *Engineering Design and Communication* course. In this course, first-year students work in teams to solve community-based, client-driven problems and build physical prototypes. Prior to Duke, she taught a similar course at Rice University, where she was on the faculty since 1999. Saterbak is the lead author of the textbook, *Bioengineering Fundamentals*. At Rice, Saterbak's outstanding teaching was recognized through four university-wide teaching awards. In 2013, Saterbak received the ASEE Biomedical Engineering Division Theo C. Pilkington Outstanding Educator Award. For her contribution to education within biomedical engineering, she was elected Fellow in the Biomedical Engineering Society and the American Society of Engineering Education.

Dr. Tracy Volz has been collaborating with engineering faculty at Rice University to embed communication instruction in engineering curricula for two decades. She was recently named the director of engineering communication in the School of Engineering. Prior to this role, she spent five years leading Rice's Program in Writing and Communication where she oversaw the First-Year Writing-Intensive Seminar program, the Center for Written, Oral and Visual Communication, ESL programming for international students, and communication in the disciplines projects. Her research interests include technical presentations, posters and engineering education.

Dr. Matthew Wettergreen is Associate Teaching Professor at the Oshman Engineering Design Kitchen at Rice University. He teaches engineering design courses, including first-year engineering design and the follow-on engineering design courses. Additionally he teaches students how to manufacture prototypes using low fidelity prototyping, iterative design, and the use of advanced manufacturing tools. In 2017 the engineering design courses at the OEDK were combined into one of the first engineering design minors in the country, credentialing students for a course of study in engineering design,

teamwork, and client-based projects. Dr. Wettergreen is the faculty mentor for Rice's Design for America chapter, for which he has been given the Hudspeth Award for excellence in student club mentoring. He is the designer of a number of academic products that help students improve their prototyping techniques, including a low fidelity prototyping cart and the Laser Cutter Prototyping Library.