Proceedings of the 52nd Hawaii International Conference on System Sciences | 2019

Introducing Low-Stakes Just-in-Time Assessments to a Flipped Software Engineering Course*

Hakan Erdogmus
Electrical & Computer Engineering
Carnegie Mellon University
hakan.erdogmus@sv.cmu.edu

Soniya Gadgil
Eberly Center
Carnegie Mellon University
soniyag@andrew.cmu.edu

Cécile Péraire
Electrical & Computer Engineering
Carnegie Mellon University
cecile.peraire@sv.cmu.edu

Abstract

***Objective**: We present a Teaching-as-Research project that implements a new intervention in a flipped software engineering course over two semesters. The short-term objective of the intervention was to improve students' preparedness for live sessions. The long-term objective was to improve their knowledge retention evaluated in time-separated high-stakes assessments. **Intervention**: The intervention involved adding weekly low-stakes just-in-time assessments to course modules to motivate students to review assigned instructional materials in a timely manner. The assessments consisted of, per course module, two preparatory quizzes embedded within off-class instructional materials and a non-embedded in-class quiz. **Method:** Embedded assessments were deployed to two subgroups of students in an alternating manner. In-class assessments were deployed to all students. The impact of embedded assessments on in-class assessments and on final exam performance was measured. **Results:** Embedded assessments improved students' preparedness for live sessions. The effect was statistically significant, but variable. Embedded assessments did not impact long-term knowledge retention assessed on final exam. We have decided to keep the intervention and deploy it to all students in the future.*

1. Introduction

Foundations of Software Engineering (FSE) is a graduate-level course [1] in Carnegie Mellon University's Master in Software Engineering program offered on the Silicon Valley campus. The instructors designed the course in 2014 as a gateway to the program, and it has since been taught to 50-80 students each semester. FSE has a *flipped classroom* format, a technology-assisted pedagogical method that inverts the traditional delivery of theory and application [2].

In a traditional classroom, theory is provided during live plenary lectures. Students apply the theory outside the classroom through take-home assignments. In a flipped classroom, the locations are reversed. Students self-learn the theory outside the class through pre-recorded lectures and other supporting materials. During in-class, or *live*, sessions, they apply the theory through supervised activities. The advantages and disadvantages of this increasingly popular format is well-understood in both lower and higher educational settings. Disadvantages in particular include difficulty in sustaining student motivation to perform both preparatory work before live sessions and reinforcing work after them [3].

1.1. Motivation

Although over time we have evolved FSE to a blended-flipped format [4] to address some of the shortcomings of flipped classroom, the course has essentially maintained its flipped character. Like any flipped course, FSE relies heavily on off-class instructional materials—videos and readings—that students are required to review each week before attending live sessions. During live sessions, we normally perform a team activity applying the off-class content, often complemented by a mini lecture to improve preparedness. While the addition of mini lectures has helped, student motivation has remained a primary issue, as exemplified by this excerpt from [1]: "In a flipped classroom, students must be strongly and repeatedly encouraged to prepare for live sessions by viewing the assigned videos and reading the assigned materials, as well as ask clarification and probing questions. We are still looking for effective non-grade-based strategies for incentivizing students to come to class better prepared."

We aimed for "non-grade based" strategies because the course already had many moving parts with rigorously graded high-stakes components. Our student

* This work was facilitated by the Teaching as Research service, provided by CMU's Eberly Center for Teaching Excellence and Educational Innovation.

URI: https://hdl.handle.net/10125/60201
ISBN: 978-0-9981331-2-6

population exhibits strong grade-oriented behavior, and in FSE especially, they have historically reported high stress levels and higher-than-average work loads in faculty course evaluations. Daniels et al. argue that students can be motivated without complicated grade-oriented or rubric-based assessments [5].

As an additional longer-term objective and for lasting effects, it was also desirable for any new teaching intervention to not only improve students' preparedness in immediate course components, but also their knowledge retention as measured by performance on more rigorous, time-separated assessments.

1.2. Approach

To follow up, we sought guidance from the university's center for teaching excellence. Based on that guidance, in the Fall 2017 offering, we decided to introduce *low-stakes*, *just-in-time assessments* to each course module. The assessments are *low-stakes* in that they are quick, frequent, and formative in nature [6]. Although students get a score from each assessment, the score does not contribute to their final grade. Instead, they receive participation points for completing the assessments as an incentive. The assessments are *just-in-time* in that they are either attached to off-class instructional materials (embedded), or performed in-class right before an activity that applies theory and concepts conveyed in the most recent off-class content.

To be able to simultaneously evaluate the effectiveness of this teaching intervention while introducing it, we did not deploy it all at once, but instead we took an experimental approach and formulated the initiative as a *Teaching-as-Research* (TAR) project [7]. The *Center for the Integration of Research, Teaching and Learning* defines TAR [8] as "the deliberate, systematic, and reflective use of research methods to develop and implement teaching practices that advance the learning experiences and outcomes of students and teachers."

This paper thus presents a TAR project for introducing a new teaching intervention to a flipped course, FSE, with the main goal of motivating students to optimally prepare. The paper explains the design, implementation, and outcomes of the project, focusing on how we introduced, measured, and evaluated the intervention to make an informed permanent decision. The TAR project ran from September 2017 to May 2018 over two offerings of FSE.

2. Related Work

Several studies have examined the benefits of flipped classroom in higher STEM and computer science education, and reported that the format had a positive impact on student outcomes and syllabus design [9, 10, 11].

However flipped instruction is not without challenges, which include high initial cost for instructors, poor reception by students, and weak student motivation [12]. A failure case is recounted by Towey [13], in which the instructor had great difficulty motivating students to be active in the classroom, noting also the role of insufficient preparation as a hindrance.

Köppe et al. presents several flipped classroom patterns [14] to increase the odds of success. Chief among them is the importance of timely preparation to avoid lagging students from sabotaging their peers' learning. They recommend controlling the pace of the students through explicit planning and concrete preparation activities. Giannakos et al.'s [12] lists low-stakes quizzes outside and inside the classroom among successfully employed strategies. Horton et al. [11] emphasize the effectiveness of low-stakes prep exercises with a "small-grade reward" leading to higher participation.

Enfield [15] successfully employed low-stakes quizzes, both in-class and off-class, to motivate students to do the prep work and encourage them to keep up with the course's pace. The in-class quizzes were created from a subset of prep quiz questions to encourage students to review the assigned instructional materials. Students then conducted an in-class activity that pertained to the quiz's topic. His approach is very similar to our intervention. A marked reduction in engagement and attendance was noticed after the prep quizzes ceased. Over 80% of the students stated that they were more likely to watch the videos because there were quizzes. Gehringer et al. [16] also used both prep and in-classes quizzes as an incentive with positive results. They report that pre-quizzes forced students to keep up with the material and the students who took the pre-quizzes performed better. Carvalho and Machado [17] reported that in-class quizzes motivated the students to prepare and improved class attendance.

The use of low-stakes embedded assessment, such as quizzes embedded to videos, are common in flipped courses and have been reported to improve interactivity and engagement outside the class and incentivize students to review the instructional materials. Examples are provided by Cummins et al. [18] and Campblell et al. [19]. The latter authors report that in their flipped course far more students completed the prep quizzes

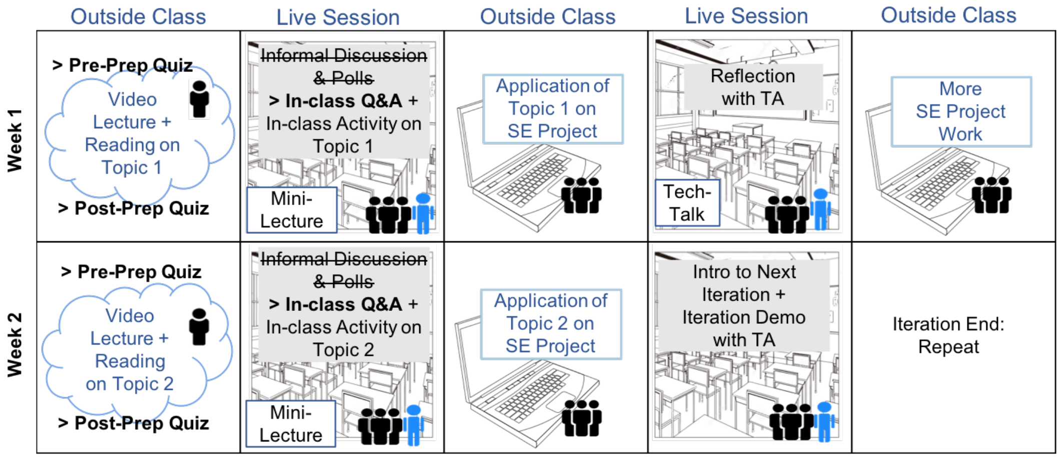
Outside Class
Live Session
Outside Class
Live Session
Outside Class
Week 1
> Pre-Prep Quiz
Video Lecture + Reading on Topic 1
> Post-Prep Quiz
~~Informal Discussion & Polls~~
> In-class Q&A +
In-class Activity on Topic 1
Mini-Lecture
Application of Topic 1 on SE Project
Reflection with TA
Tech-Talk
More SE Project Work
Week 2
> Pre-Prep Quiz
Video Lecture + Reading on Topic 2
> Post-Prep Quiz
~~Informal Discussion & Polls~~
> In-class Q&A +
In-class Activity on Topic 2
Mini-Lecture
Application of Topic 2 on SE Project
Intro to Next Iteration + Iteration Demo with TA
Iteration End: Repeat

Figure 1. Iteration structure before and after introducing low-stakes assessments.

than attended class. This observation suggests that prep quizzes are not enough, and in-class quizzes would add value by increasing attendance.

Iwasaki's survey on flipped instruction [20] conducted with the faculty supports use of both prep and in-class assessment. The tendency of the faculty was to augment off-class materials with quizzes or follow-up with in-class quizzes. Only 27% used only lecture videos without any assessment.

We haven't encountered any work that reports employing a preparatory pre-assessment in addition to a post-assessment with off-class instructional materials to evaluate students' existing knowledge of a topic before the materials are reviewed. In our intervention, we included such an assessment as an integral part of the students' prep work, to benchmark their knowledge, promote self-awareness, increase feedback, and evaluate their improvement with instructional materials.

3. Pre-intervention course structure

FSE is a breadth course spread over 14 weeks, with twice weekly 110-minute live sessions. It is designed to require on average 12 weekly hours of student effort, including preparation for live sessions, contact hours, team project components, and various assessments. Course modules align with project iterations, each of which emphasizes a distinct software engineering discipline, starting with *Teamwork & Technology* (Iteration 0), during which students also receive collaboration training, followed by *Architecture & Design* (Iteration 1), *Construction* (Iteration 2), *Testing & Quality* (Iteration 3), *Requirements* (Iteration 4), and *Integration, Packaging & Presentation* (Iteration 5).

Each iteration typically contains one to three theory modules and lasts two weeks. It is structured with the following components: (1) brief presentation of iteration objectives; (2) concept-based live sessions where students learn by applying the underlying theory necessary to reach an iterations goal; and (3) project-based live sessions where students reflect on their teamwork or demonstrate their results. Figure 1 shows this structure. Before the new intervention, bold components were missing and struck-through components were present. Live sessions had kicked off with an informal discussion driven by ungraded, anonymous online polls to gate-check students' knowledge and address questions and potential misunderstandings of off-class content.

In both the old and new structure, an iteration starts with a brief introduction by the instructor, explaining the iterations theme, requirements, practices, and deliverables. Live sessions are typically dedicated to a team activity related to iteration objectives. Students prepare for live session by watching the assigned video lectures and reviewing other supporting resources on Canvas, the Learning Management System (LMS) used by the university. Video lectures cover modules that last 10-25 minutes. Mini lectures or short *tech talks* are occasionally added to live sessions to illustrate concrete applications of theory using code examples or pertinent technologies or tools.

4. Post-intervention course structure

The first change to the course was the *embedding* of short, online quizzes with each weekly module. Students take these quizzes on their own outside the class immediately before and immediately after reviewing the instructional content of a module. We call these paired assessments the *pre-prep quiz* and the *post-prep quiz*, respectively (or collectively, *embedded assessments*). The quizzes typically contain 5-12 automatically-scored questions that assess the same knowledge items right before ("how much do I know?") and right after ("how much have I learned?") the student's self-study of the instructional content. They are untimed, and students can complete them at their own pace in multiple attempts.

The second change was the replacement of the informal discussion and polls that had typically started a live session with a more formal online assessment, called the *in-class Q&A*. The in-class Q&A involves a deeper assessment compared to the embedded ones. It also consists of 5-12 automatically-scored questions and covers the same knowledge area as the preceding embedded assessments, but using a different set of more complex questions. The in-class Q&A grounds the live session activity that follows it. It is timed and lasts up to 12 minutes. Solutions are discussed immediately afterwards in a plenary Question-and-Answer period (hence the name).

All three kinds of assessments are administered on Canvas, the LMS that hosts all course materials. Students receive immediate feedback after an attempt. The assessments are mandatory: students who skip them lose participation points, although they are not penalized for incorrect answers.

Table 1 illustrates the distribution of course modules within the two-week iteration structure. For each module, it gives the type and number of off-class instructional materials as well as the number of low-stakes assessment sets, where each set is a triplet consisting of a pre-prep quiz, a post-prep quiz, and an in-class Q&A.

5. Research questions and hypotheses

Executing a TAR project involves both "creating objectives for student learning" and "developing hypotheses for practices to achieve the learning objectives." [8]. Below we convert our objectives to research questions and provide the corresponding hypotheses to be tested. We only declare the alternative hypotheses and omit the underlying null hypotheses. In each case, the null hypothesis implicitly stipulates equal performance between the groups tested.

Our first objective is to ascertain that embedded assessments make sense relative to the off-class instructional content that they assess. We expect instructional materials to give students fresh concepts and theory and embedded assessments to reveal this effect. If not, there is a problem with either the materials, assessments, or alignment between them. This expectation leads to the first research question and corresponding hypothesis.

RQ1: *Do embedded assessments show that students learn new theory and concepts after reviewing the associated instructional content?*

H1: *Students' post-prep quiz scores on average will be higher than their pre-prep quiz scores.*

Answering RQ1 affirmatively is a pre-requisite for our two-part grand objective: to incentivize students to better absorb theory and concepts with timely preparation. This objective in turn leads to our top-level research question that is the project's primary driver:

RQ2: *How does receiving an embedded assessment before and after reviewing instructional content impact students' uptake of theory and concepts targeted in that content in the short term and in the long term?*

We break down RQ2 into two hypotheses to address short- and long-term implications separately.

Our most central interest with the teaching intervention was to push students to better prepare for live sessions. If students are ill-prepared, they forgo the benefits of in-class activities, which are a central part of flipped instruction. If live sessions fail, the whole purpose of flipped classroom is ultimately defeated. Moreover, since class activities are team-based, ill-prepared students also hamper their teammates' learning. The first hypothesis captures this pivotal, immediate implication:

H2.1: *Students who receive the pre- and post-prep quizzes before and after reviewing instructional content will score higher on a following in-class Q&A on the same topic compared to students who do not receive the pre- and post-quizzes.*

Since the in-class Q&A evaluates theory and concepts required in the subsequent class activity, we assume a better Q&A score correlates with better preparedness for the live session.

The second set of hypotheses concerns a somewhat less urgent matter: longer-term retention of theory and concepts. Are the benefits of timely preparation persistent, as measured by students' performance on a distant high-stakes assessment? If they are, then the following should hold:

H2.2: *On the final exam, students will perform better on questions based on topics for which they had received prep quizzes, compared to questions based on topics for which they had not received prep quizzes.*

6. Study design

To answer RQ1, we simply compared (1) the average gain from pre-prep to post-prep quiz scores for each student, aggregating over all modules, and (2) on a

Table 1. Distribution of low-stakes assessment sets to course modules within the two-week iteration structure.

Iteration	Week #	Module	#Videos	#Readings	# Low-Stakes Assessment Sets
N/A	1	M1-Software Engineering Methods	2	0	1
0	2	M2-Agile and Lean	2	0	1
		M3-Planning & Estimation	2	1	1
	3	M4-Object Technology & UML	1	1	1
1	4	M5-Object-Oriented Analysis & Design	2	1	1
	5	M6-Design Patterns	2	1	1
		M7.1-Architecture	1	1	1
		M7.2-Object-Orientation in JavaScript	0	2	
2	6	M8-Technical Practices	2	2	1
	7	M9-Testing	2	1	1
3	8	M10-Technical Debt	0	1	1
	9	M11-Achieving High Quality	2	0	1
4	10	M12-Requirements	1	1	1
	11	M13-Quality Attributes	1	1	1
5	12-14	Integration, Final Presentations, Exam	0	0	0
		Total	20	13	13

module-by-module basis aggregating over all students.

To answer RQ2, we deployed the embedded assessments using a single-factor, multi-object, repeated-measures design [21, 22]. We divided the cohort of a course offering to two fixed, groups based on students' course section. At the beginning of the semester, we randomly designated one group as the *treatment* (T) and the other group as the *control* (C). The T group received the intervention, the pair of embedded assessments, corresponding to the module covered in that week, and the C group did not. The groups were then switched in alternating weeks. During the first week, all students received the intervention to get used to it. After the first week, each student received the intervention in every other week only, six times out of a total of the remaining 12 opportunities. As an exception, week 5 had two sets of embedded assessments, which we also deployed in an alternating manner.

Unlike embedded assessments, we deployed in-class Q&As to all students at once as a common *post-test* to evaluate the short-term effectiveness of the embedded assessments (H2.1). To do this, we compared the average in-class Q&A performance of each student with and without prior embedded assessments by using group assignment (T or C) as an independent variable and in-class Q&A score as a dependent variable. We evaluated long-term effectiveness (H2.2) using final exam scores in the same manner. We sectioned the exam questions depending on their originating module so that a student's average score for questions for which the student had received a prior embedded assessment could be compared to the same student's average score for questions for which the student had not received a prior embedded assessment.

6.1. Differences between course offerings

The course's Fall 2017 (F17) offering had a total of 53 students divided into two sections of 26 and 27 students and the Spring 2018 (S18) offering had 61 students divided, again, into two sections of 28 and 33 students. The same instructors taught all four sections.

There were no demographic differences between the two sections of the F17 offering: all students were incoming Master's level students, and both sections were local. No particular patterns were present with respect to differences in gender, skill level, past academic performance, and ethnicity. However, there were notable demographic differences between the sections of the S18 offering with respect to academic background, ethnicity, and skill level: one section was remote and involved two different remote locations while the other section was local.

We effected two changes from F17 to S18. First, we incorporated better feedback mechanisms to post-prep quizzes so that, after completing a quiz, a student could not only see the wrong answers, but also the correct answers. This had not been available in F17. Second, we incorporated additional low-cognitive-level questions, called *rapid fire questions*, to the final exam resembling the questions used in the low-stakes assessments. We did this in the hopes of pinpointing whether time separation or question complexity dominated, or erased, long-term effects.

6.2. IRB protocols and ethical considerations

Since the new teaching intervention was an integral part of normal course work, the TAR project relied on data sources that were readily available. We guaranteed anonymity and confidentiality, and students only needed to consent to their data being used in the analysis. All students received instruction of equal value: no student was disadvantaged by being assigned to a treatment or control group more times than any other student. These characteristics allowed us to use an exempt IRB protocol applied to low-risk human studies in an educational context.

7. Results

We first comment on drop-out rates and relate it to incentivization. Then we test the stated hypotheses and answer the underlying research questions. In each case, the independent variable is group affiliation (T or C), indicating whether a subject has participated in the teaching intervention, and the dependent variable is the subject's average performance in a follow-up assessment (post-prep quiz, in-class Q&A, or final exam).

Sample sizes were large enough ($>=$ 50) in all cases. Therefore we did not have to worry about normality. To test the underlying null hypotheses, we used a paired-samples t-test to match the straightforward two-factor, repeated-measures design. To remain conservative, all tests were two-tailed regardless of the direction of the alternative hypotheses, with an alpha level of 0.01. Cohen's d measured effect size for paired samples [23].

Mortality was low in all instances (see Table 2), and reduced the sample size only slightly in some analyses. Overall, participation was high, and reported levels suggest low-stakes incentivization worked as expected.

7.1. Pre-quiz to post-quiz gain

Embedded assessments were available for 12 of 13 course modules. Overall all students who took these assessments improved significantly from the pre-prep to the post-prep quiz. Table 3 shows average scores, standard deviations, and t-test results separately for both cohorts. The results were significant and consistent across the cohorts and sections within them, allowing us to reject the null underlying hypothesis and support H1. For each cohort, the difference between mean gains for the two sections were small and statistically insignificant, suggesting the sections had similar characteristics. Therefore, the data from both sections could be pooled within each cohort. The effect sizes were large in each case, being above one standard deviation. The averages for pre-prep quiz scores were over 50%; hence students had some prior familiarity with the topics covered. The relative percentage improvement from pre- to post-prep scores in F17 and S18 were on average 33% and 22%, respectively—evidence that instructional materials also contained new information.

Module-by-module comparisons are charted in Figures 2 and 3. Learning gains from pre- to post-prep quizzes were not uniform across all modules. Students did not improve markedly on five out of 12 modules analyzed in one cohort, and two of these topics—*Architecture & Object Orientation in JS* and *Testing*—were problematic in both cohorts. Thus these modules and their associated assessments deserve further investigation.

RQ1: *Embedded assessments reveal that, overall, students learn new theory and concepts after reviewing assigned instructional content.*

7.2. Performance on in-class Q&As

We compared the average in-class Q&A scores of students who completed both pre- and post-prep quizzes (treatment group T) to those of students who did not receive the prep quizzes at all (control group C). The results are shown in Table 4. The relative percentage improvement with embedded assessments in F17 was 6%, and in S18, it was 16%. The effect size was small-medium (slightly above one-third of a standard deviation) in F17, but quite large in S18 (over 1.5 times a standard deviation). Both results were significant. Thus we reject the underlying null hypothesis, and supportive of H2.1. Speculatively, the much larger effect in S18 could be due to the improved feedback mechanism added in S18 to post-prep quizzes.

Module-by-module analysis is shown in Figures 4 and 5. Again, the improvements were non-uniform across cohorts and modules within each cohort. Some modules (*Planning & Estimation*, *Object Technology & UML*, *Architecture & Object Orientation in JS*, *Technical Practices*, and *Requirements*) registered an improvement in one cohort, but not the other. *Object-Oriented Analysis & Design* (OOAD) curiously exhibited a noticeable reverse effect in F17, although it registered an improvement in S18. OOAD was one of the most abstract and difficult topics in the syllabus, and unlike other modules, the majority of the questions on its in-class Q&A addressed content that the prep quizzes did not specifically target or targeted in a more

Table 2. Participation and mortality in terms of percentage of students who missed a certain number of assessments of a given type for each semester.

Sem.	Missed Embedded Assessments (out of 24 assessments)				Missed In-class Q&As (out of 13 assessments)			
	None	1-4	> 4	> 9	None	1-3	> 3	> 6
F17	48%	46%	6%	0%	46%	52%	2%	0%
S18	39%	52%	11%	0%	51%	48%	2%	0%

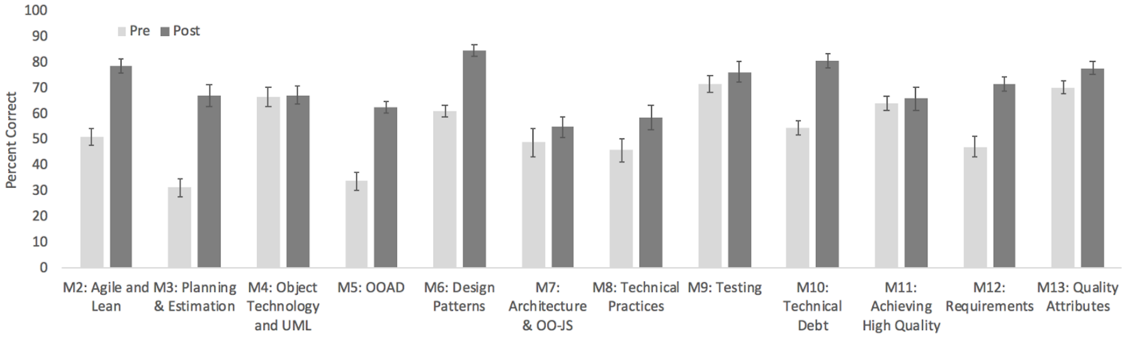

Figure 2. Module-by-module comparison of pre- to post-prep quiz performance for F17 cohort.

Table 3. Pre-quiz to post-quiz gain.

Semester	Timing (N)	Avg. (Std. Dev.)
F17	Pre (52)	53.45% (8.97%)
	Post (52)	70.65% (9.08%)
	$t = 10.99$; $p < 0.001$; $d = 1.22$	
S18	Pre (58)	51.61% (10.23%)
	Post (58)	65.18% (12.70%)
	$t = 9.32$; $p < 0.001$; $d = 1.16$	

superficial way. The observed reversal effect in F17 could be spurious. However, to be sure, we will revisit instructional materials and assessment questions for all problematic modules, upgrading their embedded assessment questions when warranted.

RQ2 (short-term): *Receiving low-stakes embedded assessments impacts students' learning in the short-term, but the impact's magnitude vary from cohort to cohort and module to module.*

Table 4. In-class Q&A performance of control (C) and treatment (T) groups.

Semester	Group (N)	Avg. (Std. Dev.)
F17	T (50)	64.6% (9.54%)
	C (50)	61.2% (9.95%)
	$t = 3.13$; $p = 0.003$; $d = 0.37$	
S18	T (58)	61.07% (12.41%)
	C (58)	51.05% (11.09%)
	$t = 6.95$; $p < 0.001$; $d = 1.58$	

7.3. Final exam performance

For the final exam effects, we mapped each exam question to the module of the related content. We then tagged each student's score on that question with a code depending on whether or not the student was in the control group C or the treatment group T for that question. Computation and comparison of average scores for C-coded questions and T-coded questions followed next.

In F17, the final exam included 24 mixed-format questions that included both manually-graded and automatically-graded ones. For S18, we also added 13 automatically-graded *rapid-fire* questions—one for each module—for a total of 46 questions. Final exam questions of the two cohorts did not overlap. We classified the *rapid-fire* questions as low-cognitive (knowledge and comprehension in Bloom's taxonomy [24]), while the remaining questions targeted higher cognitive levels (application, analysis, and synthesis in Bloom's taxonomy).

The results are given in Table 5. The differences were not significant for either cohort: we cannot reject the underlying null hypothesis. For S18, we also analyzed the *rapid-fire* questions separately, but did not observe a different behavior. Thus embedded assessments did not have any discernible effect on students' final exam performance regardless of question complexity. Hypothesis H2.2 was therefore not supported.

RQ2 (long-term): Receiving a low-stakes embedded

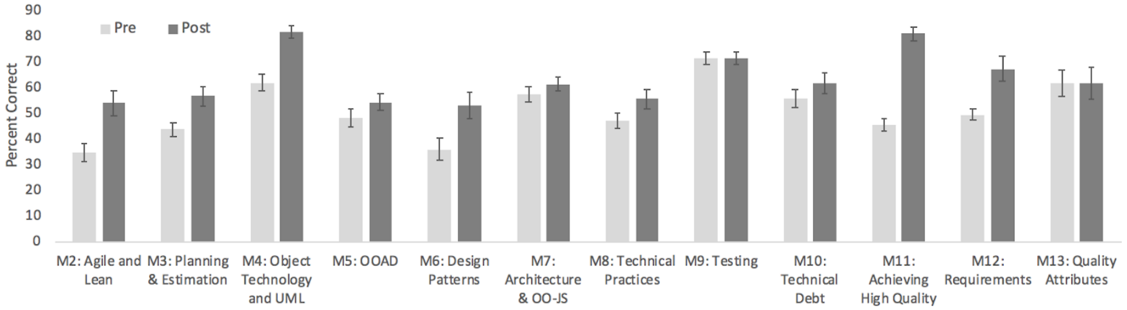

Figure 3. Module-by-module comparison of pre- to post-prep quiz performance for S18 cohort.

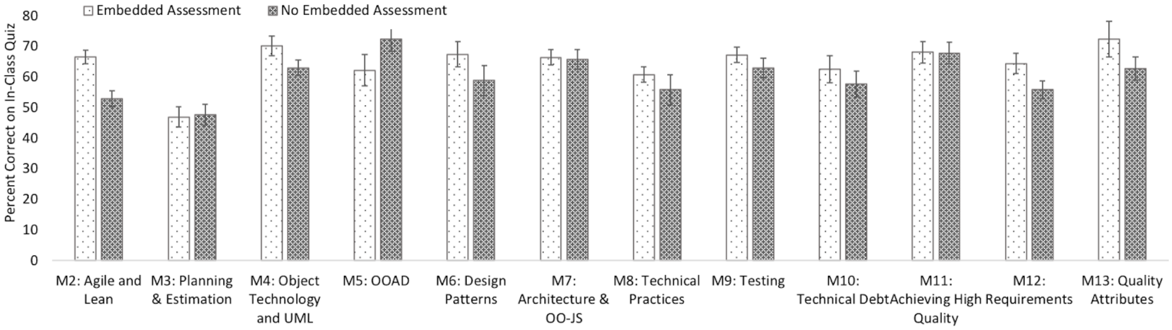

Figure 4. Module-by-module comparison of in-class Q&A performance for F17 cohort.

assessments does *not* impact students' learning in the long-term.

That all students performed in-class Q&As following embedded assessments might have masked embedded assessments' impact in the long run. This phenomenon could have further been amplified by students' continued application—throughout the course, during in-class activities and course project—of the concepts and theory targeted in embedded assessments from the time they performed the embedded assessments to the time they took the final exam. Because of significant time separation, students had ample opportunity to learn, with or without embedded assessments, until the final exam. At this point, we do not have a further theory to modify our hypotheses regarding this point, and accept this lack of long-term effect.

Table 5. Final exam performance of control (C) and treatment (T) groups.

Semester	Group (N)	Avg. (Std. Dev.)
F17	T (52)	58.84% (16.57%)
	C (52)	59.24% (15.06%)
	$t = 0.142$; $p = .88$ (n.s.)	
S18 (Overall)	T (58)	67.48% (16.12%)
	C (58)	63.64% (15.51%)
	$t = 1.65$; $p = 0.104$ (n.s.)	
S18 (Rapid Fire)	T (58)	80.76 % (19.23%)
	C (58)	78.09 % (17.24%)
	$t = .939$; $p = 0.352$ (n.s.)	

8. Limitations

Internal and conclusions validity. The study groups for embedded assessments were naturally formed based on course sections. Thus the formation of the groups was not random. Demographic differences between groups were notable in the S18 cohort. However the repeated-measures, within-subjects design overcame the potentially confounding effect of any such differences by comparing a student's performance under the control condition to the same student's performance under the treatment condition. Thus the analysis used only paired-sample tests. Since the repeated measures used different objects (course modules), the resulting multi-object design potentially introduced a different kind of bias related to object complexity. Testing each student on several objects of varying complexity, six modules for each study condition for a total of 12 objects, alleviated this threat. Averaging-out affects should kick in with this degree of object diversity, with each student receiving the intervention for a mix of topics ranging from easy to difficult. The differences in

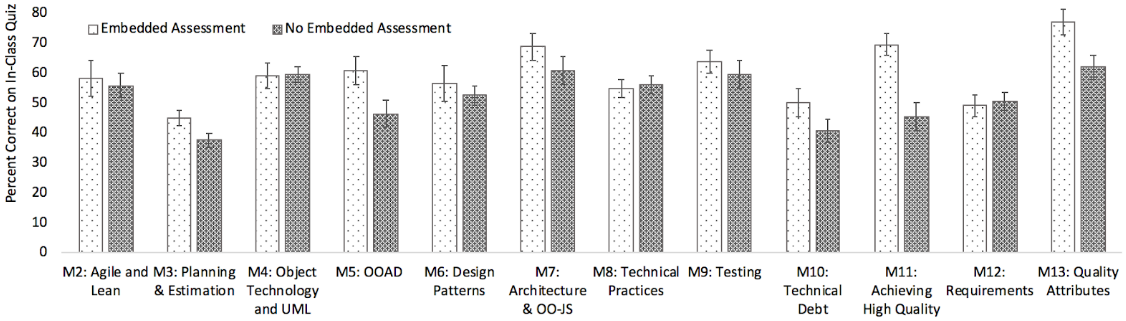

Figure 5. Module-by-module comparison of in-class Q&A performance for S18 cohort.

module-by-module results were not great or pervasive enough to warrant concern, and possibly represent natural variation due to object differences. Mortality was low in all instances and did not pose a threat.

External validity. While our study was conducted with a specific population (masters-level computer-science students at a highly selective university), our results suggest that embedded assessments have the potential to improve student preparedness for face-to-face sessions. These findings need to be replicated in other flipped format courses and across different institutions and settings, to test for generalizability.

Construct validity. Our dependent constructs were "immediate preparedness for live sessions" and "long-term retention of theory and concepts." Proxies, namely the in-class Q&A performance in the former case and final exam performance in the latter case, measured both. The proxies are objective, but, imperfectly capture the underlying constructs. The low-stakes assessments sampled only a small proportion of the knowledge covered in each module, so could not be considered comprehensive. On the other hand, there were a total of 12 sets, again constituting a sufficiently diverse sample. It was not feasible to increase coverage by using more questions due to workload concerns and other practical reasons: the assessments were not simply artifacts of a laboratory experiment, but an integral part of a real course as per the TAR approach. The possibility of students taking the embedded assessments in an unintended order poses a process conformance threat. However, since the assessments were low-stakes, the students would not gain anything by such behavior, which Canvas access logs confirmed.

9. Conclusions and future actions

The weekly embedded assessments in general aligned well with instructional materials and showed improvement in students' knowledge of course topics with high significance and large effect size. However the effect varied from module to module, with half of the modules falling short of expectations in at least one cohort. Incidentally, these are the modules with which the students historically have struggled most.

The embedded assessments overall improved students' preparedness for live sessions. The improvement was consistent and statistically significant in both cohorts, although the effect size varied greatly from one cohort to the other. The improvement was absent in five out of 12 modules in one cohort, but not the other, and one module registered a reverse effect in one cohort, warranting further investigation.

Action: Review all modules—including associated instructional materials and low-stakes assessments—(a) for which embedded assessments showed little improvement in student learning or (b) in which embedded assessments made little or no difference in class preparedness. During their review, make sure the instructional materials and assessments are well aligned and the assessments target the right knowledge elements. Give priority to modules that failed to meet expectations for both cohorts and in both dimensions (a) and (b).

While we met our short-term objective reasonably well, time separation appeared to erase any immediate positive impact of embedded assessments. In final exam scores, we could not observe any significant improvements attributable to embedded assessments. Thus we could not meet our long-term objective. The lack of impact on final exam scores can be explained by (a) the exam's high-stakes nature, forcing the students to study more; and (b) the shadowing effect of students' advancement over time through various other course components that reinforced the theory and concepts imparted in off-class content.

Separately, a significant majority of the students found the low-stakes assessments helpful in an independent course workload study, which anecdotally supports their added value from the students' perspective. Given that the assessments are naturally incorporated to the course, their impact on student work load is marginal, and their grading is automatic, full deployment makes sense.

Action: Fully deploy low-stakes just-in-time assessments in course's future offerings.

This paper demonstrated the application of a TAR project to improve a flipped, software engineering course using alternating, partial introduction of a new teaching intervention and measuring the effects. Educators can apply the approach and design in a variety of software engineering and other STEM courses.

References

[1] H. Erdogmus and C. Péraire, "Flipping a graduate-level software engineering foundations course," in *Proceedings of the 39th International Conference on Software Engineering: Software Engineering and Education Track*, ICSE-SEET '17, pp. 23–32, IEEE Press, 2017.

[2] P. N. Kiat and Y. T. Kwong, "The flipped classroom experience," in *2014 IEEE 27th Conference on Software Engineering Education and Training (CSEE T)*, pp. 39–43, April 2014.

[3] E. M. D., R. Ingram, and J. C. Liu, "A review of flipped classroom research, practice, and technologies," *International HETL Review*, vol. 4, no. 7, 2014. [Online at: https://www.hetl.org/feature-articles/a-review-of-flipped-classroom-research-practice-and-technologies; accessed 12 June, 2018].

[4] R. M. Clark and A. Kaw, "How do you like your course - blended or flipped?: A preliminary comparison," in *Proceedings of the American Society for Engineering Education*, ASEE, 2015.

[5] M. Daniels, A. Berglund, A. Pears, and S. Fincher, "Five myths of assessment," in *Proceedings of the Sixth Australasian Conference on Computing Education - Volume 30*, ACE '04, (Darlinghurst, Australia, Australia), pp. 57–61, Australian Computer Society, Inc., 2004.

[6] V. J. Shute and Y. J. Kim, "Formative and stealth assessment," in *Handbook of Research on Educational Communications and Technology*, pp. 311–321, Springer.

[7] A. J. Blanford and J. B. Scott, eds. The University of Colorado Boulder Graduate Teacher Program, 2017. [Online at: https://www.colorado.edu/gtp/sites/default/files/attached-files/tiger_tar_edited_volume.pdf; accessed 12 June, 2018].

[8] CIRTL Network, "Teaching-as-research." Center for the Integration of Research, Teaching and Learning, [Online at: https://www.cirtl.net/about/core_ideas/teaching_as_research; Accessed: 12 June, 2018].

[9] 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*, vol. 56, pp. 430–435, Nov 2013.

[10] C. Papadopoulos and A. S. Roman, "Implementing an inverted classroom model in engineering statics: Initial results," in *Proceedings of the American Society for Engineering Education*, ASEE, 2010.

[11] D. Horton and J. Campbell, "Impact of reward structures in an inverted course," in *Proceedings of the 2014 Conference on Innovation &#38; Technology in Computer Science Education*, ITiCSE '14, (New York, NY, USA), pp. 341–341, ACM, 2014.

[12] M. N. Giannakos, J. Krogstie, and N. Chrisochoides, "Reviewing the flipped classroom research: Reflections for computer science education," in *Proceedings of the Computer Science Education Research Conference*, CSERC '14, (New York, NY, USA), pp. 23–29, ACM, 2014.

[13] D. Towey, "Lessons from a failed flipped classroom: The hacked computer science teacher," in *2015 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, pp. 11–15, Dec 2015.

[14] C. Köppe, R. Niels, R. Holwerda, L. Tijsma, N. Van Diepen, K. Van Turnhout, and R. Bakker, "Flipped classroom patterns: Designing valuable in-class meetings," in *Proceedings of the 20th European Conference on Pattern Languages of Programs*, EuroPLoP '15, (New York, NY, USA), pp. 26:1–26:17, ACM, 2015.

[15] J. Enfield, "Looking at the impact of the flipped classroom model of instruction on undergraduate multimedia students at csun," *TechTrends*, vol. 57, pp. 14–27, Nov 2013.

[16] E. F. Gehringer and B. W. Peddycord, III, "The inverted-lecture model: A case study in computer architecture," in *Proceeding of the 44th ACM Technical Symposium on Computer Science Education*, SIGCSE '13, (New York, NY, USA), pp. 489–494, ACM, 2013.

[17] A. A. Carvalho and C. T. Machado, "Flipped classroom and quizzes to motivate learning: Students perspectives," in *2017 12th Iberian Conference on Information Systems and Technologies (CISTI)*, pp. 1–6, June 2017. [in Portuguese].

[18] S. Cummins, A. R. Beresford, and A. Rice, "Investigating engagement with in-video quiz questions in a programming course," *IEEE Transactions on Learning Technologies*, vol. 9, pp. 57–66, Jan 2016.

[19] J. Campbell, D. Horton, M. Craig, and P. Gries, "Evaluating an inverted cs1," in *Proceedings of the 45th ACM Technical Symposium on Computer Science Education*, SIGCSE '14, (New York, NY, USA), pp. 307–312, ACM, 2014.

[20] C. Iwasaki, "Faculty support for effective flipped classrooms in higher education," in *2017 International Symposium on Educational Technology (ISET)*, pp. 261–267, June 2017.

[21] D. C. Montgomery, ed., *Design and Analysis of Experiments*. Wiley, 8 ed., 2012.

[22] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wessln, *Experimentation in Software Engineering*. Springer, 2012.

[23] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. L. Erlbaum Associates, 2 ed., 1988.

[24] B. S. Bloom, M. D. Engelhart, E. J. Furst, W. H. Hill, and D. R. Krathwohl, *Taxonomy of Educational Objectives: The Classification of Educational Goals - Handbook I: The Cognitive Domain*. David McKay Company, 1956.