FEAL: Fine-Grained Evaluation of Active Learning in Collaborative Learning Spaces

Sixing Lu, Loukas Lazos, Roman Lysecky
Department of Electrical and Computer Engineering
University of Arizona, Tucson, AZ

Abstract

Numerous studies have shown the effectiveness of collaborative active-learning pedagogies compared to traditional lectures across STEM fields. However, incorporating active learning in large classes presents unique challenges in stimulating student engagement and developing quality activities. A growing trend at universities is to create collaborative learning spaces (CLSs) that are purposefully designed and equipped to facilitate active learning. While some research has identified some effective learning strategies for CLS environments based on learning psychology, the success of individual activities is neither defined nor measured. This gap in knowledge is often met with a trial and error approach over numerous semesters. Activity adjustments are made solely based on the instructors' partial perceptions, whereas activity effectiveness is neither directly evaluated nor correlated to student performance.

We present a novel measurement instrument called Fine-grained Evaluation of Active Learning (FEAL) for large CLS-based classes. FEAL can be quickly administered by preceptors to record key measures of activity success such as student engagement, activity difficulty, activity time, and associated lecture time. Other relevant information such as the concepts covered by the activity and the activity type are also recorded to be later cross-analyzed. FEAL can be applied to code exam questions and to assess student performance for the same concepts. We applied FEAL to a large freshman-level computer programming course with an enrollment of 200 students over the course of one semester. We present an overview of FEAL, its administration process within the CLS, and a detailed account of our evaluation methodology. We also highlight key lessons learned on the engagement and success achieved by individual activities, and outline planned improvements to in-class activities based on the obtained results.

Assessment of Collaborative Learning

Numerous studies have demonstrated the effectiveness of collaborative active-learning pedagogies compared to traditional lectures across STEM fields [1][2][3][4] and computer science education in particular [5][6][7]. Active-learning techniques include think-pair-share exercises [8][9], peer instruction [10], group problem solving, activities in CLS environments and extensive discussions, among others. Incorporating active learning in large classes (greater than 100 students) presents unique challenges in classroom management, teaching assistant and preceptor training, and student engagement. Additionally, most active-learning techniques involve extensive student interaction. However, the predominant design of classroom spaces has focused on lecture-style pedagogies, which further impedes the effective adoption of active learning.

Toward alleviating some of these challenges, a growing trend at universities is to create collaborative learning spaces (CLSs) that are purposefully designed and equipped to facilitate active learning [11]. While some research has identified learning activities that are effective in CLS environments [12] building upon learning psychology and general activity structure, the success of individual activities is neither defined nor measured. As such, instructors are left with little guidance on the quality and effectiveness of the activities that they design. This gap in knowledge is often met with a trial and error approach over numerous semesters and adjustments are made based on instructors' partial perceptions. Most importantly, activity effectiveness is not directly evaluated or correlated to student performance.

Several classroom measurement instruments have been proposed for evaluating teaching practices. However, none can directly assess individual activities. The Teaching Dimensions Observation Protocol (TDOP) [13] documents classroom behaviors by periodically marking which of 46 behaviors (or codes) were observed in the classroom. Code categories are divided into teaching methods, teacher-student dialog, instructional technology, pedagogical strategies, student engagement, and other groupings. Observations are recorded periodically (i.e., every two minutes), and provide a high-level overview of teaching, whereas a more fine-grained observation is needed to assess activities. Additionally, the TDOP requires three days of training to ensure inter-rater reliability. The Classroom Observation Protocol for Undergraduate STEM (COPUS) [14] was designed to code how both instructors and students spend class time and requires minimal training (less than 2 hours). COPUS enables instructors to understand what percentage of time they lecture, pose questions, write on a whiteboard, answer student questions with entire class involved, hold one-on-one discussions, etc. Moreover, COPUS records the percentage of time students are listening, working individually, engaged in group discussions, etc. COPUS can be very effective for assessing how different teaching practices are being used within the classroom. It provides a measure of the extent that active-learning pedagogies are applied, but the quality of the activities performed by students is not assessed. Importantly, most observation tools are meant to be used by non-experts in the subject matter, requiring very little training. In contrast, providing observations on the quality of individual activities requires expertise to gauge the difficulty or appropriateness of a given activity.

To address these limitations, we present an activity quality measurement instrument called Fine-grained Evaluation of Active Learning (FEAL). FEAL can be quickly administered by preceptors to record key measures of activity success such as student engagement, student success, activity difficulty, activity time, and associated lecture time. The instrument is designed to require minimal training and minimal effort within the classroom for recording observations. Quick administration of the instrument is critical because the preceptors recording with FEAL are primarily tasked with engaging the students with the given activities. A key difference of FEAL with other tools is that the preceptors must have expertise on the subject matter, as the intent is to evaluate activity quality. Moreover, relevant information such as the concepts covered by the activity are recorded and analyzed. The instrument is further applied to code exam questions accordingly and is used to correlate student performance for the same concepts. We applied FEAL to a large freshman-level computer programming course with an enrollment of 200 students over the course of one semester. The course was taught within a CLS by a team of two instructors and eight preceptors.

The application of FEAL on this course demonstrates a positive correlation between the student engagement and exam performance, and the student engagement is higher when the introduced concepts are more difficult. Moreover, the time spent on activities does not necessarily yield better student performance in the exam. FEAL also enables an instructor to quickly identify outliers from the expectations (e.g., recursion activities) that may need to be redesigned, such as by adding more activity difficulty to enhance student engagement and to match difficulty level with the questions in the exam.

Observing and Assessing In-class Collaborative Learning Activities Using FEAL

Class Information and Operation

We applied FEAL at a freshman-level programming CS1 course at a four-year university, using the *C* programming language. The typical course enrollment for one section is about 200 students. The course primarily serves the College of Engineering, but students from other colleges are frequently enrolled. Although the course is intended for freshmen, it is equally attended by sophomores, and juniors. Some senior and graduate students (primarily outside the College of Engineering) are also enrolled.

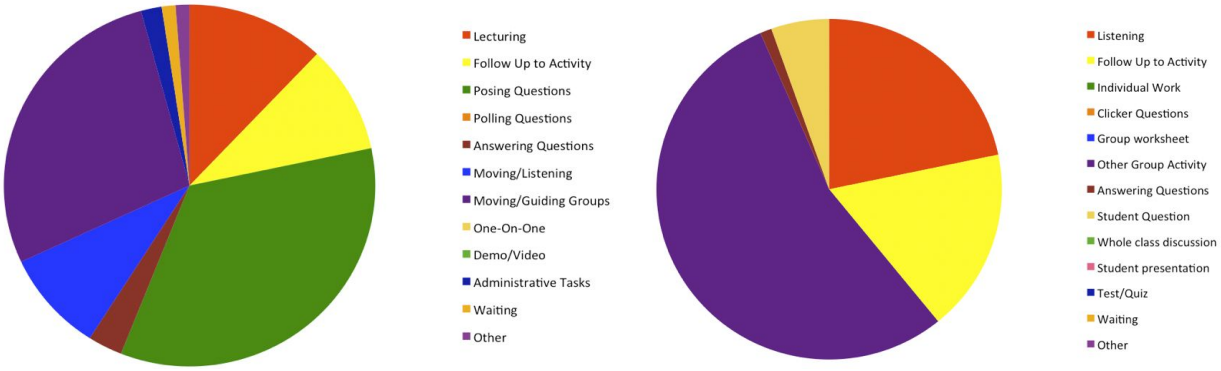
The class is taught in a CLS with a maximum capacity of 260 students. Students are organized in round tables of up to six persons. Each table is equipped with 1-2 whiteboards and a table number. An A-type whiteboard is also available per three tables. The space is further equipped with over 20 screens placed around the room so that projected material is visible from every table and angle. The CLS layout, as it is seen from the instructor's station point of view, is shown in Figure 1.

Fig. 1. Instructor view of the Collaborative Learning Space.

The class consists of three main components: (a) reading assignments using the zyBooks online interactive book platform [15], (b) 75 minutes in-class sessions held twice a week, and (c) a 3-hour lab held weekly. Students are requested to complete a set of participation and challenge questions before every in-class session. These are automatically graded through the zyBooks platform. The in-class time is structured as a sequence of active-learning tasks, and lecturing/ demonstration periods. The administration of the activities is assisted by preceptors (teaching assistants and undergraduate learning assistants that have previously taken the course). A typical distribution of the instructors' and students' activities during a 75 minutes class session, as it is

recorded with the COPUS tool, is shown in Figure 2. The COPUS tool verifies that the majority of the in-class time is spent on group activities, polling questions, demonstrations, group guiding, and student questions, whereas little time is devoted to traditional lecturing. However, the COPUS graphs cannot be used to infer the activity quality.

Fig. 2. Overview of instructor's and student's time in a single class session using COPUS.



(a) Instructor's time (b) Student's time

The FEAL Instrument

The primary challenge in observing the activity quality for a large class is to collect accurate and comprehensive data with minimal interruption on the student-instructor team engagement and overall class flow. A single dedicated recorder is not able to observe the student activities in the entire class. To overcome this challenge, we designed a simple form, shown in Table 1, for collecting lecture time on specific concepts, activity time, level of easiness, level of student engagement, student success on each activity, and student attendance. The form is meant to be administered in a distributed fashion by each preceptor who is typically responsible for facilitating students in four or five tables (around 20-25 students).

Table 1. The FEAL data collection form for one activity.

Class Date:	**Preceptor Name:**	**Attending Students:**		
Activity 1				
Simple Description:				Notes
Lecture Time:	**Activity Time:**			
	<50%	50% - 70%	70% - 90%	>90%
Engagement (students)				
Engagement (time)				
Correctness				
Easiness	Very difficult	Slightly difficult	Average	Very easy

The form consists of the following fields:

Class Date, Preceptor Name, and Attending Students: Records the date of the class, the name of the preceptor administrating the form, and the number of students assigned to the preceptor.

Simple Description: A concise description of the activity and the associated concept. One activity could cover multiple concepts.

Lecture Time (min): The time spent by the instructor covering the concepts related to the activity. This time includes lecturing, demonstrations, activity description and instructions, and any follow-up.

Activity Time (min): The time spent on the activity.

Engagement (% students): The percentage of students that are engaged during the activity. The preceptors are trained to capture behaviors of students to determine if they are engaged in the activity. A student who reads the activity requirements, refers to the relevant concepts, writes code on paper or whiteboard, and discusses with other group members is considered to be engaged. Each preceptor is assigned to evaluate a predefined set of students (usually on the order of 25 students).

Engagement (% time): The percent of time that students remained engaged in a particular activity, as a fraction of the activity time.

Correctness (%): An estimated percent of the activity that is completed by students. The preceptors, who are subject experts, evaluate correctness based on the product created by the student groups and the discussions held between the student and the preceptor.

Easiness (numeric): An estimation of the easiness level for the activity. The preceptors evaluate the easiness of the activity based on their subject expertise and interactions with the students.

Notes: Quick notes made by the preceptors on the activity delivery and execution.

To facilitate the fast administration of the instrument, the easiness, engagement, and correctness are fuzzified to four levels, as shown in Table 1. The fuzzification mitigates the challenge of precisely determining a percentage for these observations, improves inter-rater reliability (although we do not directly assess the latter in this paper), and reduces the time needed to record observations.

Administration of the FEAL Instrument in the CLS

The FEAL instrument was administered by a team of eight preceptors, who were responsible for facilitating engagement and student understanding in the CLS. Preceptors' training for the FEAL instrument included an initial one hour training along with ad hoc training in the first few weeks of classes, totaling less than two hours. Within the CLS, 34 tables were divided into eight regions of four or five tables each. Tables accommodated up to six students each. Each preceptor is responsible for 20-25 students.

During an activity, preceptors are directed to work with students for the entire activity time. Once an activity is finished and control is regained by the instructor for post-activity discussions or introduction of the following topic, preceptors have the opportunity to complete the form based on their observations. To correlate the FEAL data with exam performance, graders record the score of individual exam questions.

Analysis of Collected Observations

In this section, we describe how the data collected through the FEAL instrument is analyzed to determine the overall engagement, easiness, and correctness for each activity. First, the values recorded in the forms are aggregated across all preceptors. Table 2 shows an example of the aggregated data for a single activity. The median percentage of each level is used for calculating the aggregated value for each metric. For instance, the aggregated engagement (E) is calculated as:

$$E = \sum_{i=1}^{4} p_i \, n_i / N \times \sum_{i=1}^{4} t_i \, n_i / N,$$

where $p_i \in \{0.25, 0.6, 0.8, 0.95\}$ denotes the median percentage of each level in the number of students that were engaged, $t_i \in \{0.25, 0.6, 0.8, 0.95\}$ denotes the average fraction of time that students were engaged, n_i denotes the number of students engaged in each level and for each metric (engagement (students) or engagement (time)), and N denotes the total number of students attending. For correctness and easiness, the aggregated value is the average of the values recorded across each level, weighted by the median percentages. For easiness, a scale from 1 to 4, with 1 being the most difficult and 4 being the easiest, is used.

Table 2. Example of aggregated observations for a single activity.

	<50% (level 1)	**50%-70% (level 2)**	**70-90% (level 3)**	**>90% (level 4)**	**Student Attendance**	**Aggregate**
Engagement (students)	0	35	49	48	132	80.2% × 80.5% = 64.5%
Engagement (time)	0	47	18	67	132	
Correctness	0	27	93	12	132	77.3%
Easiness	0	8	112	12	132	3.03

Table 3. Concept categorization for introductory C programming course.

Concept	**Sub-concepts**
Basics	variables, ASCII encoding, operators, scanf, printf
Loops	for loops, while loops, nested loops
Conditions	if-else, switch
Functions	functions, arguments, return values
Pointers	single pointers, double pointers, dereferencing
Arrays	integer arrays, floating point arrays, char arrays
2D Arrays	numeric matrices, char matrices
Strings	strings
Structures	structures
Linked Lists	single linked lists, doubly linked lists
Recursion	recursive functions

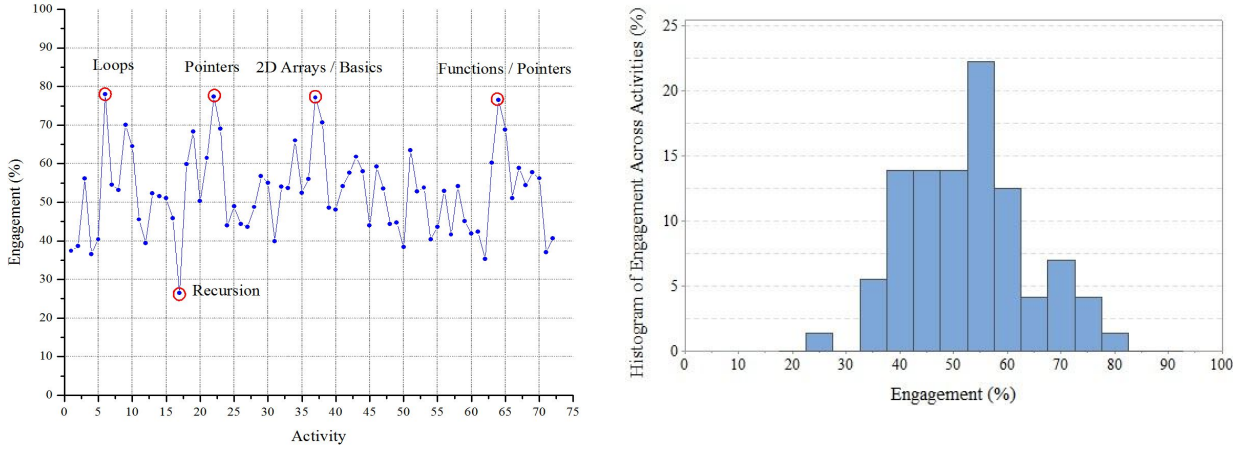
Concept Categorization

Each activity and exam question is categorized by concept at the granularity presented in Table 3. These concepts are specific to each course and are defined by the instructor. Categorization could be also assisted by using existing concept inventories. The concepts are used to classify activities and aggregate the engagement, correctness, and easiness metrics per concept. This categorization also enables the cross-correlation of the activity scores with exam scores.

What Drives Student Engagement

The first question that we wanted to answer using FEAL is what drives student engagement. In particular, we want to identify those activity traits that yield a high level of student engagement. Engagement is a key component of active learning and has been shown to positively correlate with learning. Figure 3(a) shows the engagement level for each of the 72 activities that were observed over the course of the semester. The engagement does not increase or decrease monotonically throughout the semester, but rather depends on the concepts and design of the activities. Peaks in engagement are primarily observed when a new concept is introduced with the exception of the concept of recursion, which achieved the lowest engagement. This particular activity asked students to revisit a recursive function previously discussed in class and ask at least one question to clarify anything they did not understand. The goal of this activity was to facilitate discussion and help determine any misconceptions that students may have had. However, the notes made by the preceptors attributed the low engagement to the vagueness of the activity. As students were not asked to produce a specific product, the majority did not attempt to ask any question. The FEAL instrument enables an instructor to quantitatively notice the low engagement, and redesign the activity. Figure 3(b) shows the normalized histogram of student engagement across activities during the semester. We observe that most of the time engagement was between 40% and 60%, and only for 15% the engagement exceeded 70%. This indicates that a significant portion of the time devoted to activities is not used as efficiently as possible.

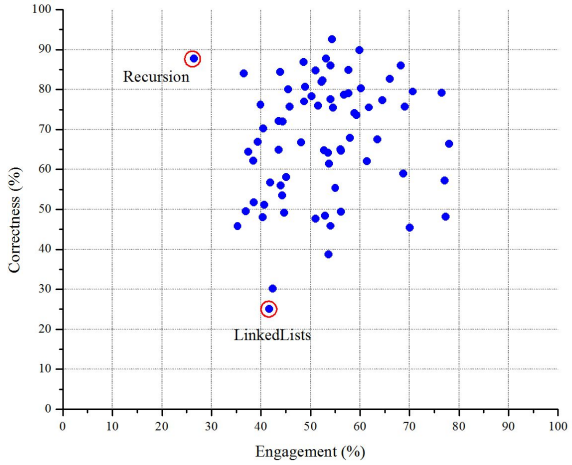
Fig. 3. Engagement for individual activities during the semester.



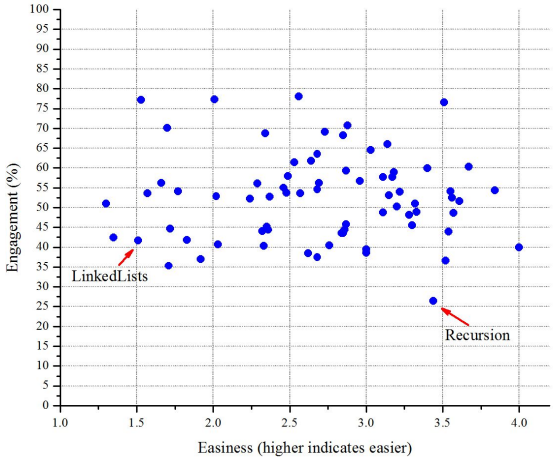
(a) engagement across activities (b) histogram of engagement across activities

To further understand how engagement is affected by the design of activities, we analyzed the relationship between engagement and correctness. Figure 4 presents engagement as a function of correctness and easiness across all individual activities. As Figure 4(a) shows, a higher engagement leads to a higher correctness, but the relationship is not linear. Two outliers are of particular interest (marked as red circles), which provide insight into the design of specific activities. One of the highlighted outliers is a recursion activity that had a low engagement of 26.45% but a high correctness of 87.71%. A correlation with the activity's easiness shown in Figure 4(b) reveals that the particular activity was perceived to be one of the easiest ones. Another highlighted outlier is an activity on linked lists. This activity had low engagement and low correctness. By cross-examining Figure 4(b), and also interpreting the notes made by the preceptors, we attribute the failure of this activity to its difficulty (easiness=1.51 in Figure 4 (b)). Alternatively, a low correctness could be due to the allocation of insufficient time, but in those cases one would expect a higher engagement. Figure 4(b) shows the engagement as a function of the activity easiness. One does not see a clear correlation between the two metrics. Some activities that were perceived to be hard achieved high engagement, whereas for others the engagement was low. Similarly, some of the easier activities produced low engagement. In general, we observe that activities of medium easiness (between 2.5 and 3) were those that produced the highest engagement.

Fig. 4. Engagement as a function of correctness and easiness for individual activities. Outliers are marked as red circles.



(a) engagement vs. correctness



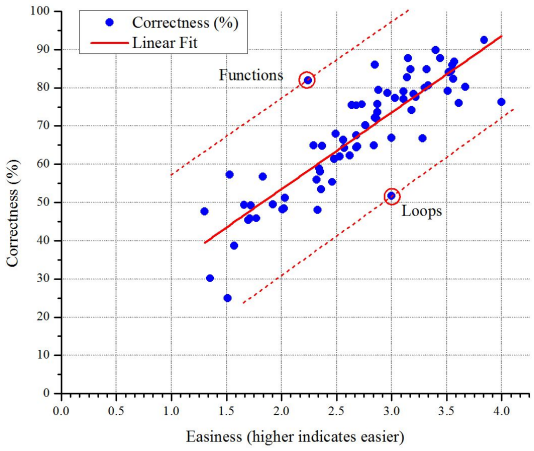
(b) engagement vs. easiness

What Drives Correctness

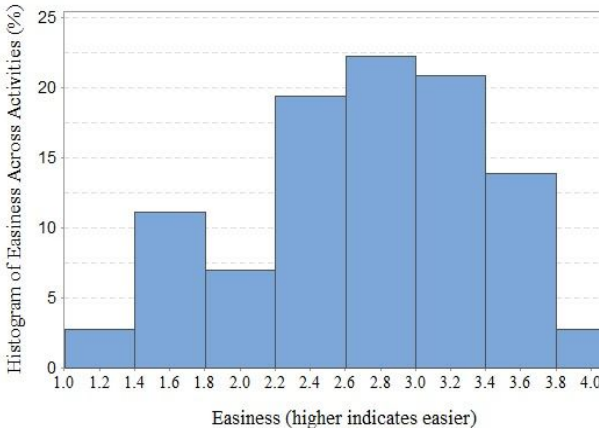
We further analyzed the data collected using the FEAL instrument to identify what type of activities are likely to be correctly completed by the students. Figure 5(a) shows the activity correctness as a function of the activity easiness for each of the recorded activities. A linear fitting is also shown. For the majority of activities, we observe a positive correlation (the slope of the linear fitting is 20.03 with a standard error of 1.36) between easiness and correctness. This correlation is significantly stronger than between engagement and correctness (see Figure 4(a)). Two outliers are of particular interest. An activity on functions shows a significantly higher

correctness relative to its easiness. Upon further examination, this activity was the last activity for that class day, and students already completed several similar activities on functions.

Fig. 5. Easiness of individual activities during the semester. red solid line is a linear fitting for all activities.



(a) correctness vs. easiness



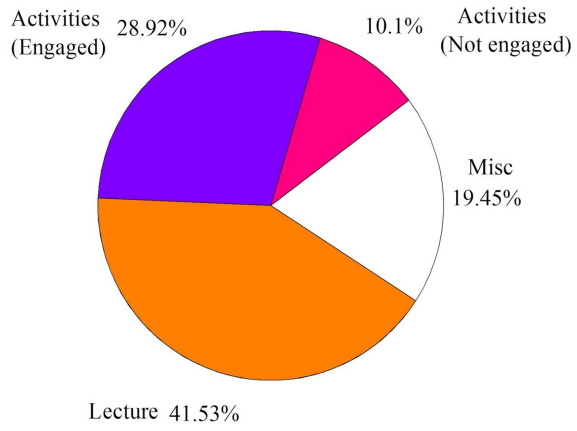
(b) histogram of easiness across activities

Our analysis can be used by an instructor to determine if the time allocated to each activity achieved the desired goals. For example, an instructor may determine that the time for an activity was not beneficial, as students had already sufficiently mastered the concepts. Notably, such a conclusion must take into account the instructor's intention for the activity, but nevertheless, the data collected with FEAL can be used to provide evidence to guide an instructor's decision. Another notable outlier is an activity on loops that had a correctness lower than expected, given the easiness of the activity. This activity was the first activity introducing loops, which could account for low correctness. The low correctness score may indicate that the activity was too big of a leap in knowledge at the time that it was conducted. Figure 5(b) shows the normalized histogram of activity easiness. The very easy and very difficult activities represent less than 5% of the activities. Overall, the class appears to have a good distribution of activity difficulty, with most activities being scored higher than 2.

Categorizing Learning Pedagogies

The FEAL instrument can also be used to provide a rough estimate of how time is spent in class, similar to COPUS but with a coarser granularity. Figure 6 shows the aggregated breakdown of class time during which students were engaged in activities, not engaged during activities, listening to lecture, or performing other miscellaneous tasks. Overall, activities in which students were engaged corresponded to 28.9% of the total class time, with only 10% of class time corresponding to non-engaged students. Notably, lecture time occupied a significant portion of the overall class time, which indicates that the instructors employed more of a mixed approach, combining lecture and active-learning techniques. The miscellaneous category primarily consists of time allocated to answering questions about homework, project, or other course requirements.

Fig. 6. Pie chart for time distribution.



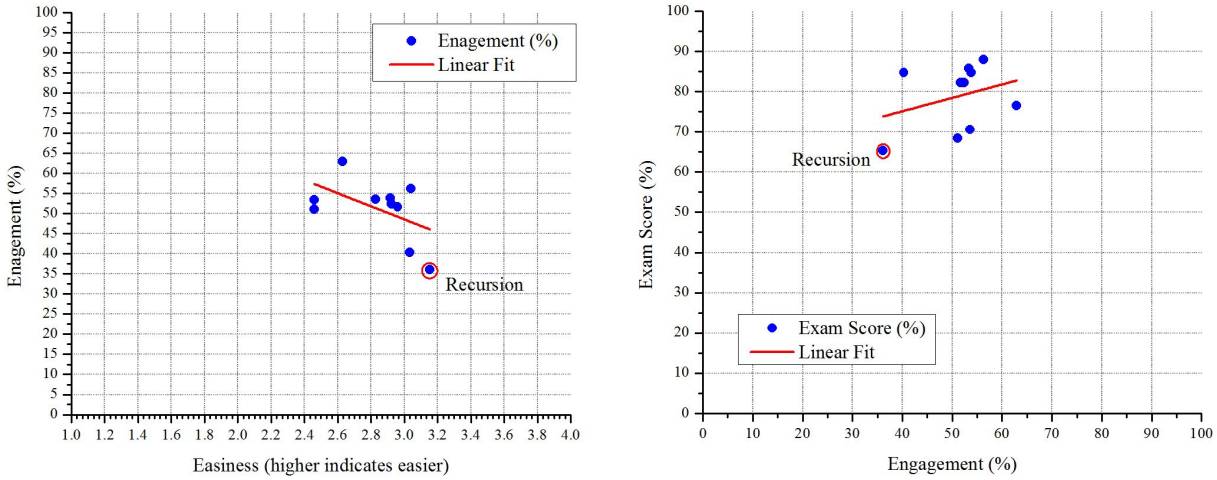
Improving Class Design Through Concept-driven Assessment

A primary goal for FEAL is to allow instructors to correlate the impact of in-class activities to student performance and concept understanding. This correlation can serve as a key metric of the activity quality. To evaluate this correlation, we analyzed the relationship between exam scores, total lecture time, total activity time, engagement, easiness, and correctness. This analysis was performed per concept (i.e., data from different activities was aggregated per concept of Table 3) and the concepts were restricted to those that were covered by at least one exam question. If a concept was covered by multiple exam questions, the exam score on a specific concept was averaged across all relevant exam questions. Similarly, engagement, easiness, and correctness were aggregated (averaged) over all activities for each concept. The total lecture time and total activity time were computed as the sum of lecture time and sum of activity time from all activities on this concept, respectively.

Figure 7(a) shows the engagement as a function of the activity easiness aggregated per concept, along with a linear fitting line. We observe that easier activities tend to have lower student engagement (slope of linear fitting is -16.29 with standard error of 9.45), whereas such relationship was not observed from the analysis of individual activities (Figure 4(b)). This indicates that when an activity on a given concept is fairly easy (score of 3 or above), students tend to disengage fairly fast. Figure 7(b) shows the average exam score as a function of engagement based on concept. As it is evident from the graph, higher engagement leads to higher exam scores (the slope of linear fitting is 0.33 with standard error of 0.35). The concept of recursion is again highlighted as one of particular importance, as it appears to be an outlier. From Figure 7(a) and Figure 7(b), we observe that students were particularly disengaged during the time that recursion was covered in class (engagement of 36.12%). This led to the lowest score compared to any concept in the exam (65.34%). From Figure 7(a), we infer that the recursion activities had the highest easiness of 3.15. Students did well in the class activities, but they did not perform well on the exam. The reason for this is twofold. First, the exam questions on recursion were far more difficult than the in-class activities. Second, the total time devoted to recursion activities was quite low, so the concept was not well covered. The FEAL instrument

enables an instructor to quickly identify these outliers, and revise activities related to a particular concept to adequately prepare students for the exams.

Fig. 7. Engagement as a function of easiness on the concepts of Table 3 and influence of engagement on the exam scores. Solid red line indicate a linear fit of the data.

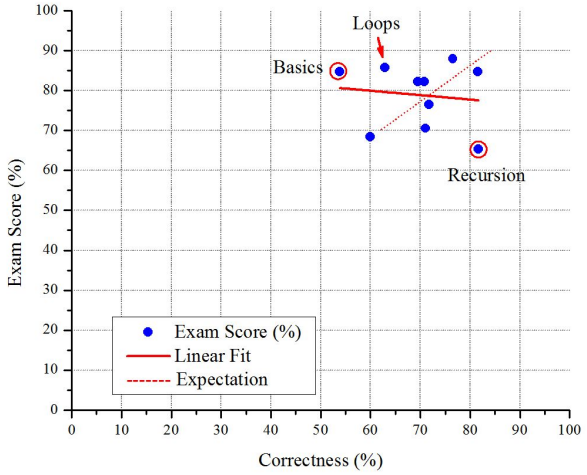


(a) engagement as a function of easiness (b) exam score as a function of engagement

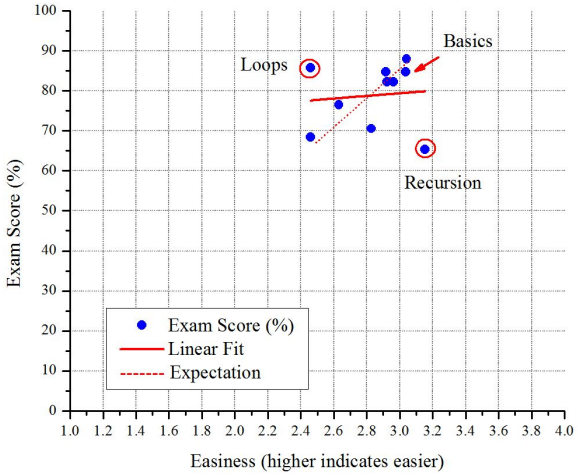
Figure 8(a) shows the average exam score as a function of correctness for each concept, with a linear fitting. We expected that higher correctness during the in-class activities would lead to better performance in the exam. However, the fitted line is far different from this expectation (linear fitting slope is -0.11 with standard error of 0.31), mainly affected by two outliers. For the basics concept students achieved low correctness in the activities, but had high exam scores. This is mainly because the basics concepts are used in almost all other activities, so students get continuous practice. While this finding is not new, it reinforces the common conception that activities can be designed to both focus on new concepts while reinforcing previous ones lead to better understanding and success. This suggests that instructors should design activities that simultaneously cover multiple concepts. The other outlier is again recursion, which is the same outlier as in Figure 7. The low exam score on recursion despite the high correctness is the one that dampens the positive correlation between correctness and student performance in the exams. This is a indication of an in-class activity that could be better designed and a high disparity between concept coverage and exam expectations.

Figure 8(b) shows the average exam score as a function of easiness, with a linear fitting. We expected that easier concepts, as categorized by the easiness of their respective activities would lead to higher exam scores. Whereas this is true for most concepts, two outliers (recursion and loops) make the fitted line deviate from this expectation (linear fitting slope is 3.27 with standard error of 11.6). The loops concept has low easiness for in-class activities, but students achieved a high score in the exam. This is because the loops concept had a very long lecture time (82.5 min) and activity time (89 min) in the class, which led us to further analyze the relationship between the allocation of class time and both exam scores.

Fig. 8. Average exam score as a function of correctness and easiness based on concepts. Solid red line is a linear fit of the data. Dashed red line is the expectation trend, outliers are marked as red circles.

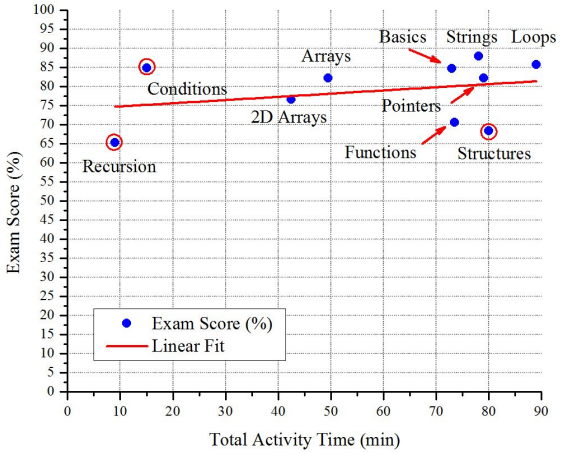


(a) exam score as a function of correctness

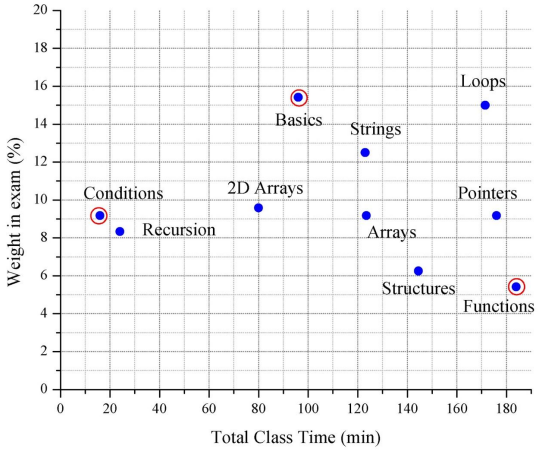


(b) exam score as a function of easiness

Fig. 9. The influence of total activity time on exam and concept weight distribution in the exam as a function of total class time, based on concepts. Solid red line is a linear fit of the data, outliers are marked as red circles.



(a) exam score as a function of total activity time



(b) concept weight distribution in the exam

Figure 9(a) shows the average exam score as a function of the total activity time, with a linear fitting. The fitting shows that longer total activity time lead to higher exam scores (linear fitting slope is 0.083 with standard error of 0.09), but the correlation is not very strong. For the conditions concept, students achieved high exam scores but the total activity time was very low. We believe this is because conditions is an easier concept to master and is introduced very early in the semester. On the other hand, a large amount of time was devoted to the structures concept but students did not perform as expected on the exam. As such, spending more total activity time on structures concept is not likely to improve exam scores. Instead, the activities on structures may need to be redesigned to better teach that concept. Figure 9(b) shows the concept weight

distribution in the exam and the relative amount of time devoted in class. Note that some concepts are absent because they were not directly tested in the exam. The total class time includes both the total activity time and total lecture time on a specific concept. From Figure 9(b), we observe that the basics concept had the highest weight in the exam. However, the points allocated to exam questions on conditions and functions does not align well with the time allocated for class time. Conditions, which has the least class time, accounted for 9.17% of the exam grade, whereas functions accounted only for 5.42%, despite devoting the highest amount of class time. This analysis empowers instructors to design fair exams based on their in-class time allocation or adjust the in-class activities to reflect the exam expectations.

Preceptor Survey

To measure the overhead of the FEAL form administration and its impact on the preceptors' ability to remain engaged, we conducted an anonymous survey. The survey asked preceptors to estimate the time required to complete the form for each activity by selecting from one of the following options: less than 1 minute, 1-2 minutes, and longer than 2 minutes. The survey also asked preceptors to rate on a Likert scale whether they agreed or disagreed with the following statement: "Completing the form did not interfere with my ability to follow and understand the flow of the class". The survey was conducted across two semesters. We received 7 responses out of 8 preceptors that were part of the class during the Fall 2016 semester (for which the FEAL data was collected and analyzed in this paper) and 16 responses out of 21 preceptors for the current Spring 2017 semester. We note that only 3 preceptors were common to both semesters.

For the Fall 2016 semester, 28.6% of the preceptors required less than 1 minute to complete the FEAL form, 28.6% required 1-2 minutes, and 42.9% required more than 2 minutes. In the Spring 2017 semester, half the preceptors required less than 1 minute and half required 1-2 minutes.

In the Fall 2016 semester, 28.6% of preceptors somewhat disagreed with the statement that completing the form did not interfere following the class flow. 42.9% neither agreed nor disagree, and 28.6% somewhat agreed. In the Spring 2017 semester, 81.3% strongly agreed or somewhat agreed, and only 18.8% somewhat disagreed. The difference between the two semester lies in the class size and responsibilities areas of each preceptor. For Fall 2016, every preceptor was responsible for three tables of up to six students each. In Spring 2017, every preceptor is responsible for six tables of up to three students. As averaging of FEAL scores occurs per table, the operation of completing the relevant form fields takes more time than it did prior.

Limitations

The FEAL instrument for assessing and observing collaborative learning activities does not require additional observers, which saves labor and reduces complexity, but this instrument does introduce some overhead for preceptors.

The subjectivity of the preceptors affects the inter-rater reliability. Several efforts have been made to reduce this subjectivity, including training, averaging evaluation, and interspersed

preceptor surveys. However, additional analysis is needed to evaluate the effectiveness of these efforts and further quantify both the subjectivity and inter-rater reliability.

Another potential limitations of FEAL is that preceptors, who have already mastered the material, tend to rate the easiness of activities higher than the actual easiness for new learners. As such, the raw easiness value by itself may not provide sufficient insight, but can be useful only when combined with other factors. Conversely, preceptors and instructors with more experience may adjust their ratings to account for their knowledge of topics with which students struggle. Thus, there is a need for further analysis to understand how subject expertise and teaching experience affect the ratings when using FEAL.

The FEAL instrument does not capture other class components that contribute to learning such as weekly labs, preparatory reading assignments, and homework assignments, which also affect the student performance.

Conclusions and Future Work

This paper presented a measurement instrument for observing and assessing in-class collaborative learning activities for large classes, by recording lecture time, activity time, easiness, engagement, and correctness. Through analyzing the relationship between these factors and their impact on students' performance in exams, we showed that the FEAL instrument can be employed to evaluate the quality of the in-class activities. The instrument was used to observe a single semester, but the full potential of the proposed instrument is in analyzing changes made in response to the data, which requires further observation and analysis across semesters. Additionally, our initial analysis only evaluates the observable data for the activities and does not yet consider the instructor's intent. As such, future work includes incorporating data for the instructor's expected easiness, expected engagement, and expected correctness. Additionally, we hope to incorporate Bloom's taxonomy for each activity to code the intention of the activity (e.g., knowledge recall, analysis, application).

Acknowledgement

This research was supported by the University of Arizona AAU Undergraduate STEM Education Project funded by the Helmsley Trust.

References

[1] Freeman, S., S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, M. Wenderoth. *Active learning increases student performance in science, engineering, and mathematics.* Proceedings of the National Academy of Sciences of the United States of America. PNAS 2014 111 (23), 8410-8415, 2014.

[2] Prince M. *Does active learning work? A review of the research.* Journal of Engineering Education, 93:223–231, 2004.

[3] Knight J.K., Wood, W.B. *Teaching more by lecturing less.* Cell Biology Education, 4(4), 298-310, 2005.

[4] Michael J. *Where's the evidence that active learning works?* Advances in Physiology Education, 30(4), 159-67, 2006.
[5] McConnell, J.. *Active learning and its use in computer science.* In Proceedings of the 1st ACM Conference on Integrating technology into computer science education (ITiCSE '96). ACM, pp. 52-54, 1996.
[6] Simon, B., J. Parris, and J. Spacco. *How we teach impacts student learning: peer instruction vs. lecture in CS0.* Proceeding of the 44th ACM technical symposium on Computer science education, 2013.
[7] Porter, L., D. Bouvier, Q. Cutts, S. Grissom, C. Lee, R. McCartney, D. Zingaro, B. Simon. *A Multi-institutional Study of Peer Instruction in Introductory Computing.* Proceedings of the 47th ACM Technical Symposium on Computing Science Education (SIGCSE '16), 2016.
[8] Lyman, R.. *Think-pair-share: An expanding teaching technique.* Maa-Cie Cooperative News 1.1, pp. 1-2, 1987.
[9] Azlina, N. A., Nik, A. *CETLs: supporting collaborative activities among students and teachers through the use of think-pair-share techniques.* International Journal of Computer Science Issues, 7(5), 18-29, 2010.
[10] Mazur, E. *Peer instruction.* 2013.
[11] EDUCAUSE, *7 Thing You Should Know about Collaborative Learning Spaces*, https://net.educause.edu/ir/library/pdf/ELI7092.pdf, Accessed 2015.
[12] Graetz, K. A., & Goliber, M. J. *Designing collaborative learning places: Psychological foundations and new frontiers.* New Directions for Teaching and Learning, 92, 13-22, 2002.
[13] West. E., C. Paul D. Webb, W. Porter. *Variation of instructor-student interactions in an introductory interactive physics course*. Physical Review Physics Education Research, 9, 010109, 2013.
[14] Smith, M., F. Jones, S. Gilbert, C. Wieman. *The Classroom Observation Protocol for Undergraduate STEM (COPUS): a new instrument to characterize university STEM classroom* practices. CBE-Life Sciences Education, 12(4), 618-627, 2013.
[15] Lysecky, R., F. Vahid. *Programming in C*. zyBooks, 2016.