Longitudinal Analysis of Instructor Actions in an Active, Blended, and Collaborative Classroom Environment

A. Zadoks[1], N. A. Stites[1], D. Evenhouse[1], N. Patel[1], R. Kandakatla[1],
E. Berger[1,2], J. F. Rhoads[2], J. DeBoer[1,2]
[1]School of Engineering Education
[2]School of Mechanical Engineering
Purdue University
West Lafayette, Indiana
freeform@purdue.edu, purdue.edu/freeform

***Abstract*—In recent years, there has been a call for Engineering Education to adopt innovative pedagogical techniques. One such innovation is Purdue University's *Freeform* environment for teaching undergraduate engineering courses. *Freeform* aims to enhance student learning with an Active, Blended, and Collaborative (ABC) learning environment. The majority of research into the *Freeform* environment thus far has focused on student outcomes and the tools *Freeform* uses rather than instructional practice [1–3]. However, with the growth of *Freeform* and other ABC environments, the lack of actionable feedback available to instructors about their role and practices is becoming increasingly apparent. Therefore, with the goal of building on the characterization of *Freeform* and ABC pedagogies in general, this paper presents a one-semester longitudinal analysis of two instructors and comparisons of their actions in this environment. The study concludes that these two instructors retain individuality in their pedagogical choices and adapt their instruction to the material presented while staying within the *Freeform* ABC ethos throughout the semester.**

***Index Terms*—active; blended; collaborative; learning environment; video coding; instructor**

I. Introduction

Many STEM educators embrace the modern challenge of incorporating new and engaging instructional strategies in their classrooms, including Active, Blended, and Collaborative (ABC) techniques. A substantial and growing body of literature shows that undergraduates learn more in environments that deliberately stimulate engagement [4, 5]. However, one major obstacle in understanding and implementing environments that use combinations of ABC techniques is the lack of empirical studies describing the typical structure of ABC and other events during a class session.

For the past few years, our team has developed and disseminated an innovative undergraduate engineering learning environment called *Freeform*, which combines Active, Blended, and Collaborative instructional elements in a coherent system encompassing student resources (e.g., a discussion forum and videos) and in-class pedagogies. Most prior research on *Freeform* has focused on the physical and virtual learning materials it provides and the experience of students interacting with them. However, some new faculty adopters of *Freeform* have communicated difficulty adjusting to and implementing the *Freeform* system in the classroom. This challenge is exacerbated by a lack of common vocabulary and detailed data to discuss in-class actions with new adopters. With the aim of improving this situation, our group video recorded each class session taught by two veteran instructors for an entire semester, and we analyzed these instructors' actions using a previously-developed coding scheme [6].

In our group's previous study, we evaluated a library of videos recorded during each class meeting across a typical semester for the same two veteran *Freeform* instructors who are included in this study [7]. The data set was analyzed using a coding structure inspired by The Classroom Observation Protocol for Undergraduate STEM (COPUS) and amended specifically for ABC learning environments [6, 8]. To extract quantitative information, the lectures were analyzed with continuous video coding software, which allowed for the granular identification of events with fine temporal resolution. Previous statistical analyses of this data by our group exclusively discussed average time values for various actions over a semester. These average values may smooth and hide many intricacies present in the broader, longitudinal data set [7]. In this study, we aim to advance our understanding of pedagogical choices in *Freeform* by analyzing in-class instructor actions throughout a semester and comparing instructor actions and decisions by topic. The specific research question addressed in this paper is: *to what extent does pedagogical choice vary based on instructor and topic within a Freeform environment?*

II. Background

Freeform began about a decade ago as an attempt to improve the mechanical engineering dynamics course offered at Purdue University using ABC pedagogical techniques in a single course [3].

Active learning environments engage students in classroom activities in order to improve student learning outcomes compared to outcomes from traditional environments [9]. In this instructor-focused study, Active techniques or Active instructional practices are instructor actions intended to achieve participatory responses from students and bring their full attention to the topic. Some examples include collaborative problem solving, group quizzes, and asking questions to engage or challenge students.

Blended resources combine technology and in-person instruction to further engage students in course material both in and out of class [10]. These include, specifically for *Freeform*, worked example videos, homework solution videos, dynamics visualization videos, and a collaborative blog and discussion forum. Blended structures offer many affordances to students in terms of flexibility and convenience of access to support resources, and students in well-constructed Blended environments may perform better compared to those in a more traditional (i.e., non-Blended) instructional format [10].

Collaborative techniques aim to bring students together to work on in-class activities and on homework or other out-of-class activities. The online discussion forum for *Freeform* aims to facilitate collaboration outside of the classroom, while Collaborative in-class work includes problem solving or group quizzes. The benefits of Collaborative learning are extensive, including increased motivation, learning, and persistence, when compared to more traditional, individualistic learning [11–13].

The *Freeform* environment with its foundation in ABC pedagogies has shown promising results since its inception. It has reduced the average percentage of students who receive a D, F, or who withdraw from the course from approximately 20% to approximately 10%, and this result has been consistent across instructors and semesters at Purdue [2, 3]. *Freeform* is being adopted at new and distinct institutions, and our research group continues work to understand instructors' pedagogical options to assist them when implementing this new learning environment.

III. Methods

Data were collected from video recordings of nearly all in-class meetings of the Spring 2016 semester for two veteran instructors. The instructors were teaching Basic Mechanics II, a sophomore-level, core engineering class in the *Freeform* environment. Every lecture recording was coded using StudioCode, a continuous video recording software, and a coding scheme developed specifically for *Freeform* and similar ABC pedagogical environments. The coding scheme and its development are discussed in depth in [6], and a summary of the coding scheme is available in Table I [7].

The coding scheme groups instructor actions into nine mutually-exclusive instructor actions and class events. Also, the amount of time used for Active, Blended, and/or Collaborative techniques is tracked. Finally, the intended participation of students (and the associated amount of time) is coded for no (None) students being engaged, Some students being engaged, or All students being engaged in Active learning. To ensure interrater reliability (IRR), statistical analysis was conducted at regular intervals during coding. This analysis included Cohen's kappa, which is a method for measuring coder agreement while controlling for chance overlap [14]. All IRR tests returned kappa values larger than 0.6 (graders consistent) except for the first [14]. After discussing differences of opinion, the graders consistently agreed with each other.

TABLE I: Coding Scheme

Event	Code	Description
None	NONE	None of the students in the class are being engaged by the instructor, and the instructor is not attempting to engage with students actively (e.g., talking / lecturing).
Some	SOME	The instructor is attempting to engage with some of the class (e.g., a question).
All	ALL	All of the students in the classroom are directly engaged in classwork (e.g., an assessment).
Active	ACT	The students are actively interacting with each other or the instructor, or there is an expectation that they are.
Blended	BLEND	Blended resources are being referenced or used in the classroom.
Collaborative	COLL	The students are working collaboratively with each other.
Passive	PASS	The students are passively being instructed.
Conceptual Talking / Lecturing	CTL	The instructor is talking or lecturing about conceptual material in monologue.
Problem Solving Talking / Lecturing	PSTL	The instructor is solving a problem verbally, in monologue.
Conceptual Real Time Writing	CRTW	The instructor is discussing conceptual topics using written notes on the board or projector.
Problem Solving Real Time Writing	PSRTW	The instructor is solving a problem on the board or projector.
Questions	Q	The instructor asks a question of the class to which an answer is clearly expected, or a student asks a question of the instructor.
Graded Assessment	GA	The students are taking an assessment for a grade.
Ungraded Assessment	UGA	The students are taking an assessment that they know will not be graded
Demonstration	DEMO	The instructor is using physical or computational instructional materials.
Other / Administrative	OA	The instructor is conducting activities other than the 8 codes above or is doing administrative actions such as returning exams

Each of these measures uses the time (in seconds) devoted to each of the nine mutually exclusive codes in the coding structure. This mutual exclusivity is important because broader Active, Blended, Collaborative, Passive, None, Some, and All codes would be interesting to analyze, but they are not necessarily mutually exclusive.

Additionally, while the analysis discusses comparisons in terms of chapters, which is technically true, it is much more interesting and informative to think of the chapters not as just divisions of the course but as topics of instruction. Therefore, a table relating chapter number and the topic of instruction involved has been included in Table II.

TABLE II: Chapter Topics

Chapter	Topic
1	Point Kinematics
2	Planar Kinematics of Rigid Bodies
3	Moving Reference Frame Kinematics
4	Particle Kinetics
5	Planar Kinetics of Rigid Bodies
6	Vibrations

We used chi-squared testing, along with calculations for effect size, in two studies:

Study 1 — Between Instructors

This study included a set of chi-squared tests for independence comparing the two instructors for a given chapter in the course (between instructors). Additionally, Cohen's effect size was calculated for each test that was statistically significant [15]. These tests answer the question "do the actions and pedagogical choices when teaching a given chapter differ significantly between instructors?", and the effect size answers the question "when the two instructors are significantly different, how different are they?" Both these constituent questions and their answers help us understand instructor choice within a *Freeform* environment and the extent to which two instructors teaching nominally the same content can deliver that content differently.

Study 2 — Within Instructor

The second study included chi-squared tests for independence between two course chapters for a given instructor (within instructor). Cohen's effect sizes were also calculated for statistically significant tests. These tests answer the question "does a given instructor teach these two chapters in a statistically significantly different way?", and the effect size continues to characterize how much the chapters differ if the chi-squared tests return a significant result.

To conduct valid chi-squared tests for independence, the counts in each cell of the contingency table must be greater than five, and all categories must have all non-zero counts [16]. However, in many of the categories selected for testing in this paper, zero counts were present. To conduct valid tests, the zero-count categories were removed from the data sets being compared before the test was conducted. This lowers the degrees of freedom of the test, but allows for a test to be conducted. In the reported results, the degrees of freedom are reported for each test, so it is evident which tests required this correction given that a full contingency table has 8 degrees of freedom in this study.

Cohen's effect size (w) statistic is used extensively in this analysis because it places statistically significant results on a more practically useful scale. Therefore, understanding Cohen's suggested ranges for effect values and their interpretations is important. Cohen suggests effect sizes to be broken into four ranges: [0.0–0.1) no or trivial effect, [0.1–0.3) "small" effect, [0.3–0.5) "moderate" effect, [0.5–1.0] large effect [15].

IV. Results

All tests in Study 1 (instructor v. instructor) tested statistically significant with $p < 0.001$. The reason all tests returned significant results is likely the large sample size (count of seconds), which directly relates to how likely a chi-squared test is to have a significant result [16]. These results imply that for any given chapter in the course, the two instructors taught with a statistically significant difference. Or, in the language of the test, the instructors were not independent with respect to chapter in the course. These results are available in Table III.

TABLE III: Study 1
Instructor v. Instructor

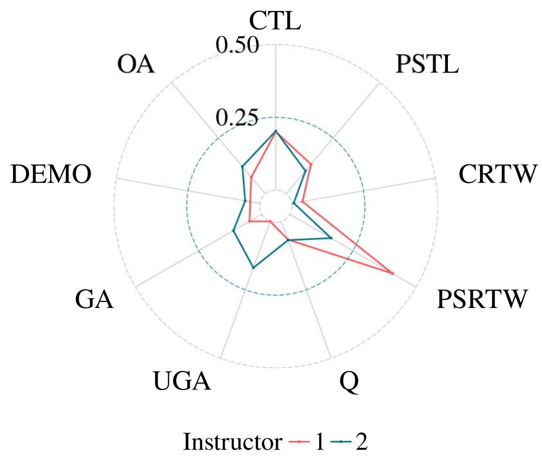
Chapter	χ^2	df	p	w
1	1495	6	<0.001	0.280
2	933	8	<0.001	0.197
3	1611	8	<0.001	0.240
4	3673	8	<0.001	0.299
5	1037	7	<0.001	0.164
6	4317	8	<0.001	0.319

Because all chi-squared tests reached the same conclusion about the instructors, effect sizes (w) were calculated for all tests. Effect sizes in this set of tests range from 0.164 to 0.319, which places the lower bound in Cohen's "small" effect size (0.1–0.3) and the upper bound in his "moderate" (0.3–0.5) effect size range. However, statistical significance and effect size are not necessarily intuitive indicators of the true magnitude of variance between the instructors in a practical sense. To better visualize the true difference in instruction, radar plots were constructed for the tests with the largest and smallest effect sizes in Study 1 (Figure 1). These plot the proportional distribution of time for the two data sets being compared in each test. For example, in Figure 2 (top), it is clear that the largest proportional difference in time between Instructor 2's Chapters 1 and 5 is the time spent in Other / Administrative activities. Figure 1 (bottom) shows how similarly Instructors 1 and 2 taught Chapter 5.

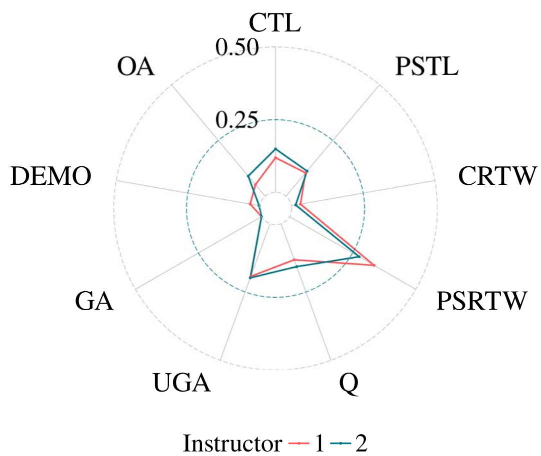
The tests that compared chapters within each instructor (Study 2) also all returned statistically significant outcomes with $p < 0.001$. These results imply that for a given instructor, any two chapters are taught statistically significantly differently from each other. These results are available in Table IV.

Effect sizes, shown in Table IV, were also calculated for each of the tests in Study 2, and radar plots were constructed (Figure 2).

The method of approaching these effect sizes utilized in this paper is an analysis of the minima and maxima from the two studies. This is intended as a qualitative representation of the trends in effect size given knowledge of the source lectures, chapter topics, and instructors.

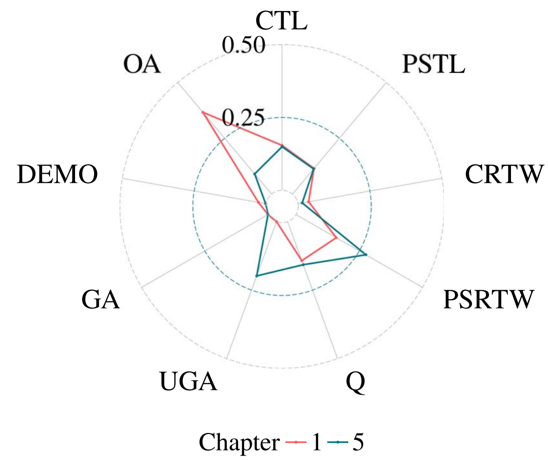


(a) Maximum (Chapter 6), w = 0.319

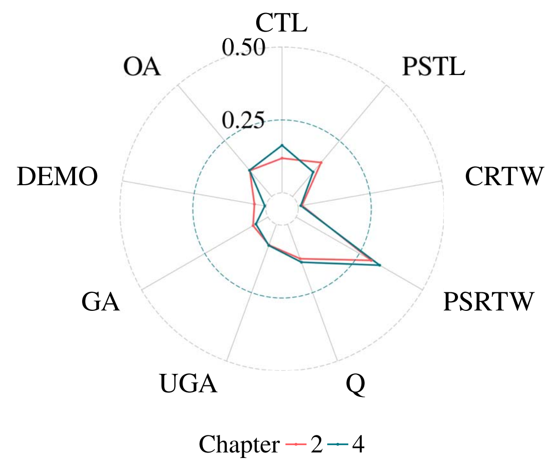


(b) Minimum (Chapter 5), w = 0.164

Fig. 1: Study 1 Effect Size



(a) Maximum (Instructor 2), w = 0.428



(b) Minimum (Instructor 2), w = 0.160

Fig. 2: Study 2 Effect Size

V. Discussion

To begin, it is helpful to understand the previous study we conducted using this data. In the first analysis of these data, our group conducted a series of chi-squared tests for independence comparing two veteran instructors' instructional habits averaged over an entire semester [7]. The analysis presented in [7] yielded some interesting results, concluding that overall, the two professors taught statistically differently ($p < 0.01$), but only with a small Cohen's effect size (0.189) [7]. The researchers further broke down the differences into the nine mutually exclusive coding categories and found that only the Conceptual Real Time Writing (RTW) and Problem Solving RTW categories were used with significant difference between the two instructors [7].

One yardstick missing in [7] is a reference relating statistical significance for chi-squared tests, Cohen's effect size, and the practical differences they describe. Because this paper extensively utilizes Cohen's effect size, it is important to contextualize the numbers presented in terms of practical pedagogical differences. Radar plots of the data that give the largest and smallest effect sizes for the two categories of tests conducted can be found in Figures 1 and 2. These plots help to visualize the difference associated with given effect sizes. In both Study 1 and Study 2, there is a large amount of overlap in the radar plots. In contrast, the maximum effect size plots show that for effect sizes in the moderate range, there is still reasonable similarity in instruction.

Although these plots give a good proportional representation of the differences between the two data sets being compared, they do not clarify the absolute time differences involved. Therefore, a table that presents the average differences in minutes for each of the four plots presented is provided in Tables V and VI. This table was calculated by first calculating the average number of minutes captured for each lecture over the span of the semester (46.6 minutes) and then multiplying

TABLE IV: Study 2
Chapter v. Chapter

Instructor 1				Instructor 2			
Chapters	χ^2	df	w	Chapters	χ^2	df	w
1,2	3907	8	0.398	1,2	2859	8	0.395
1,3	2039	7	0.284	1,3	2631	8	0.349
1,4	1178	7	0.189	1,4	3792	8	0.376
1,5	3650	7	0.337	1,5	4625	7	0.428
1,6	2239	8	0.271	1,6	4861	8	0.397
2,3	2036	8	0.274	2,3	653	8	0.162
2,4	3994	8	0.338	2,4	774	8	0.160
2,5	3731	8	0.331	2,5	2547	8	0.299
2,6	1250	8	0.197	2,6	2111	8	0.249
3,4	2267	7	0.252	3,4	931	8	0.167
3,5	5754	8	0.408	3,5	3364	8	0.325
3,6	2351	8	0.267	3,6	2963	8	0.281
4,5	5070	8	0.346	4,5	2069	8	0.236
4,6	2829	8	0.263	4,6	3860	8	0.301
5,6	3232	8	0.285	5,6	4530	8	0.332

this by the difference in proportional time spent on each of the nine mutually exclusive codes, averaged over the unit of interest. This table helps to roughly quantify the difference in the amount of time spent on each activity between two instructors or chapters. To show how to read the table, a few examples are included below.

As a note, the average number of minutes captured per lecture recording (46.6 minutes) is not equal to the length of each class period (50 minutes) because of miscommunication with the videographers. This mostly underrepresents assessments.

TABLE V: Study 1 (Instructor)
Average Minute Difference

	Chapter 6 (max)	**Chapter 5 (min)**
CTL	-0.79	-1.41
PSTL	1.82	-0.37
CRTW	1.44	0.78
PSRTW	9.50	2.76
Q	-0.29	-1.16
UGA	-4.90	-0.24
GA	-3.02	0.00
DEMO	-1.16	1.44
OA	-2.61	-1.80

TABLE VI: Study 2 (Chapter)
Average Minute Difference

	1 v. 5 (max) Instructor 2	**2 v. 4 (min) Instructor 2**
CTL	0.21	-2.07
PSTL	0.96	1.97
CRTW	1.04	0.18
PSRTW	-5.46	-1.55
Q	-0.69	-0.60
UGA	-9.23	-0.08
GA	0.00	0.51
DEMO	1.16	1.69
OA	12.87	-0.07

Example 1

The first entry in Study 1 (Chapter 6) shows that for Chapter 6, Instructor 1 spent an average of 9.50 more minutes on Problem Solving Real Time Writing (PSRTW) per class period than Instructor 2. Instructor 1 also spent on average 4.90 fewer minutes on Ungraded Assessment (UGA) per class than Instructor 2.

Example 2

The first entry in Study 2 (Chapter 1 v. 5) shows that in an average class period taught by Instructor 2, the instructor spent 12.87 more minutes in Chapter 1 on Other / Administrative (OA) tasks than Chapter 5. Also, the instructor spent 9.23 fewer minutes per class period on Ungraded Assessment (UGA) in Chapter 1 versus Chapter 5.

With this information, it is easier to contextualize the effect sizes calculated in the chi-squared tests in terms of true practical differences.

Study 1 concludes that instructors maintain their instructional autonomy when transitioning to an ABC or *Freeform* pedagogy. This Corroborates the dependence on instructor found in [7], every instructor comparison test conducted concluded a statistically different set of pedagogical actions for each instructor, for every chapter. However, this analysis has also uncovered further detail that was not available in [7].

The effect sizes calculated for each Study 1 test show a trend that the beginning and ending chapters are more different between instructors than the middle chapters. This shows that most of the divergence between instructors takes place at the beginning of the course, when the material is at least partially review and instructors require more time for administrative tasks, and the end of the course, when the material is not necessarily as related to the other material in the course (vibrations versus kinetics / kinematics). The middle chapters, particularly Chapters 2, 3, and 5, show much more similarity than Chapters 1 or 6, which means that for these chapters, the instructors are providing a more similar in-class experience. Chapter 4 does not fit this trend, and no clear explanation has yet been found for this. Future research should focus on the internal causes of Study 1 tests more generally, much like what was reported in [7].

Study 2 concludes that the material being taught has a considerable influence on how a veteran *Freeform* instructor teaches that material. This set of chi-squared tests was conducted comparing chapters for an individual instructor, testing every combination of chapters (chapter v. chapter). These tests are useful tools to find how much instructors adapt to the unique material of each chapter as well as to find how wide a pedagogical range a *Freeform* or ABC instructor has at their disposal. The answers are that these two *Freeform* instructors do have a wide instructional range, and they do seem to adapt their teaching decisions to the chapters they are teaching. The largest inter-chapter difference found had an effect size of 0.428, and the average effect size for these tests is 0.297.

The results of both Study 1 and Study 2 conclude that instruction varies in small to moderately large ways over the course of a semester, when broken down by chapter. The tests additionally show that within *Freeform* and potentially other similar ABC environments, instructors retain their own unique instructional style, generating at least a small inter-instructor difference, and at most a moderately large one.

There are some limitations to this analysis. Many lecture recordings were cut slightly short of the true 50-minute class period due to logistics. A typical missing segment was about 3.4 minutes. After some cursory investigation performed by watching the ends of a small sample of lecture recordings, it seems that the recordings usually cut off during quizzes. Therefore, the main categories affected by these cuts are Graded Assessment and Ungraded Assessment, each of which is probably underrepresented. The data presented here did not extrapolate into the missing time and ascribe these Assessments to this missing data, although the evidence suggests that doing so would be reasonable. This missing information generally skewed the data away from Collaborative time, which is common for Ungraded Assessment, and Passive time, which is typical of Graded Assessment.

VI. Future Work

By showing that there is indeed much more information that can be extracted by analyzing this data set longitudinally, this paper opens the doors to many different approaches and depths for future analyses. The most closely related direction forward is to analyze what aspects of instruction drive inter-instructor and inter-chapter differences the most. This could be conducted simply with more chi-squared tests for independence and Cohen's effect size statistics or, more rigorously, by using hierarchical linear modeling techniques.

Additionally, cluster analysis could be conducted to generate groups of the most similar types of class periods, perhaps illuminating some class period archetypes that could be further studied to characterize ABC environments.

More directions can be taken when the data set discussed here is combined with data from other sources and instructors. If a similar data set were to be collected at an adopter institution, comparisons between veteran / foundational instructors and adopting instructors could be made and quantified. More practically, student outcome data could be combined with this class characterization data to help conclude which instructional practices correlate with positive student outcomes.

VII. Conclusion

By leveraging the longitudinal aspect of a data set spanning a semester's worth of class meetings for an ABC environment, we have shown that instructors maintain pedagogical variety and individuality throughout the semester while using *Freeform*. Despite hesitation to conclude similarly when using semester-wide averages in an analysis conducted in [7], longitudinal analysis has helped to provide a more detailed view of the data.

These results have opened a large variety of options for future analyses and have shown the value of the longitudinal aspect of our data. Diving further into longitudinal research with smaller units of analysis and more robust tests and models shows promise to clarify the instructor's role in an ABC classroom. Additionally, combining the data discussed here with student outcome and other related data could enable further investigation into the root causes of improved DFW rates and student outcomes.

Acknowledgment

This Material is based upon work supported by the National Science Foundation under Grant No. 1525671. Any opinions, findings, and conclusions or recommendations expressed in this Material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] N. Stites *et al.*, "Analyzing an Abbreviated Dynamics Concept Inventory and Its Role as an Instrument for Assessing Emergent Learning Pedagogies," *123rd ASEE Annu. Conf. Expo.*, 2016.

[2] J. DeBoer *et al.*, "Work in Progress: Rigorously Assessing the Anecdotal Evidence of Increased Student Persistence in an Active, Blended, and Collaborative Mechanical Engineering Environment," *123rd ASEE Annu. Conf. Expo.*, 2016.

[3] J. F. Rhoads, E. Nauman, B. Holloway, and C. M. Krousgrill, "The Purdue Mechanics Freeform Classroom: A New Approach to Engineering Mechanics Education," *123rd ASEE Annu. Conf. Expo.*, 2016.

[4] M. Prince, "Does Active Learning Work? A Review of the Research," *J. Eng. Educ.*, vol. 93, no. 3, pp. 223231, 2004.

[5] J. K. Knight and W. B. Wood, "Teaching More by Lecturing Less," *Cell Biol. Educ.*, vol. 4, no. 4, pp. 298310, 2005.

[6] D. Evenhouse et al., "Development of a Video Coding Structure to Record Active, Blended, and Collaborative Pedagogical Practice," 7th Res. Eng. Educ. Symp., pp. 1-9, 2017.

[7] D. Evenhouse, C. Freitas, J. DeBoer, E. Berger, and C. M. Krousgrill, "What does an In-Class Meeting Entail? A Characterization and Assessment of Instructor Actions in an Active, Blended, and Collaborative Classroom," *124th ASEE Annu. Conf. Expo.*, 2017.

[8] C. J. Finelli et al., "A Classroom Observation Instrument to Assess Student Response to Active Learning," *2014 IEEE Front. Educ. Conf. Proc.*, pp. 14, 2014.

[9] S. Freeman et al., "Active Learning Increases Student Performance in Science, Engineering, and Mathematics," *Proc. Natl. Acad. Sci.*, vol. 111, no. 23, pp. 84108415, 2014.

[10] B. Means, Y. Toyama, R. Murphy, M. Bakia, and K. Jones, "Evaluation of Evidence-based Practices in Online Learning: A Meta-analysis and Review of Online Learning Studies," *U.S. Dep. Educ.*, p. 94, 2009.

[11] L. Springer, M. E. Stanne, and S. S. Donovan, "Effects of Small-group Learning on Undergraduates in Science, Mathematics, Engineering and Technology: A Meta-analysis," *Rev. Educ. Res.*, vol. 69, no. 1, pp. 21-51, 1999.

[12] P. Armbruster, M. Patel, E. Johnson, and M. Weiss, "Active Learning and Student-centered Pedagogy Improve Student Attitudes and Performance in Introductory Biology," *CBE Life Sci. Educ. 2*, vol. 8, pp. 203-213, 2009.

[13] R. W. Preszler, "Replacing Lecture with Peer-led Workshops Improves Student Learning," *CBE Life Sci. Educ. 2*, vol. 8, no. 3, pp. 182-192, 2009.

[14] M. L. McHugh, "Lessons in Biostatistics Interrater Reliability : the Kappa Statistic," *Biochem. Medica*, vol. 22, no. 3, pp. 276-282, 2012.

[15] J. Cohen, "A Power Primer," *Psychol. Bull.*, vol. 112, no. 1, pp. 155-159, 1992.

[16] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures, 5th ed.* Chapman and Hall / CRC, pp. 637-759, 2011.