Student Engagement with Pre-recorded Lecture Videos in a Flipped-class Format

David M. Pabst
Dept. of Mechanical Engineering
University of Pittsburgh
Pittsburgh, PA, USA
dmp114@pitt.edu

Vincent J. Taverna
Dept. of Mechanical Engineering
University of Pittsburgh
Pittsburgh, PA, USA
vjt7@pitt.edu

Samantha E. Wismer
Dept. of Mechanical Engineering
University of Pittsburgh
Pittsburgh, PA, USA
sew105@pitt.edu

Lee A. Dosse
Dept. of Mechanical Engineering
University of Pittsburgh
Pittsburgh, PA, USA
lad132@pitt.edu

David Schmidt
Dept. of Mechanical Engineering
University of Pittsburgh
Pittsburgh, PA, USA
des53@pitt.edu

Matthew M. Barry
Dept. of Mechanical Engineering
University of Pittsburgh
Pittsburgh, PA, USA
mmb49@pitt.edu

***Abstract*—This research-based work-in-progress (WIP) addresses the use of flipped class formats (FCFs) in a collegiate environment, particularly student engagement with pre-recorded lecture videos (PLVs). FCFs have gained popularity in the past decade, for the student-centered instruction method provides pupils with additional opportunities to engage with the course material, their peers, and/or instructor, typically more than they would have received through a traditional didactic lecture. Although shown to increase student engagement, satisfaction, and performance, FCFs, in particular on-line lecture videos, have not been scrutinized in terms of student interaction with PLVs. To this end, it is hypothesized that by providing a means of interaction during a lecture video, students will be more engaged with the digital content and subsequently will perform better than their peers who did not engage with online media.**

During the fall 2022 term, 183 students enrolled in a sophomore-level Statics and Mechanics of Materials course were taught using a FCF. The students viewed a total of 65 PLVs encompassing the content for 23 lectures on the university's approved video platform, Panopto, and integrated in a Learning Management System, Canvas. Oftentimes there were multiple videos for one lecture, as to minimize the length of the video. The PLVs had a total run-time of 271 minutes and 21 seconds; the average video length was 4 minutes and 10 seconds, and the mode of the lecture video length was 3 minutes and 26 seconds. Twenty-one pre-recorded lectures had an accompanying set of questions administered through Top Hat that tested the students' conceptual understanding of the material presented during the videos. These questions took three common forms—multiple-choice, true or false and matching—and totaled 162. There were an average of 7.5 questions per lecture, and the mode of questions per lecture was 8. Hints for incorrect answers were provided to students.

Student engagement with the lecture videos was analyzed through quantitative Panopto and Top Hat data. Student engagement metrics such as normalized video viewership and the answering of Top Hat questions (answering habits, correctness to attempts ratio) were determined. Correlations and trends between video viewership, interactions with Top Hat, and course performance were elucidated using statistical methods, namely H-tests and analyses of variance. Students who answered Top Hat video questions after watching lecture videos performed half a letter grade better than those who answered the questions prior to viewing the videos. The top and middle terciles of normalized video viewership cohorts had a corresponding 6% and 3.6% increase in performance over the lowest tercile group; the more a student watched lecture videos, the better they performed. Students who had high correctness scores in comparison to their attempts on Top Hat video questions (namely the top and middle tercile cohorts) performed 10.6% and 5.2% better than the lowest tercile group, respectively. Lastly, students who answered all Top Hat video questions on average earned a 5% higher score than students who skipped questions.

***Index Terms*—Flipped class format, online lecture videos, engagement with digital media**

I. Introduction

The use of FCFs has begun to be an ever increasing popular pedagogical trend in collegiate engineering programs. In a FCF, students are introduced to the lecture topics prior to coming to class. This is typically accomplished through the use of PLVs. Students can view these videos at their convenience before class, and even multiple times before and after class. The FCF allows the instructor to provide additional instruction and guidance by spending class time answering questions, leading discussions, conducting in-class group activities, and providing additional examples. This format also has the benefit of providing the instructor and/or teaching assistants with additional interaction with the students, especially those who may be struggling [1]. This classroom format has been shown to be beneficial for student performance in the classroom [2], and some studies suggest that students even prefer it to a traditional lecture, as will be detailed in the following.

In 2013, a study [3] was conducted on a software engineering course of second year students (n = 120). For one specific topic the students would utilize a FCF. The students were instructed to watch specific videos from YouTube related to this topic. They were given learning objectives and were encouraged to re-watch the videos as many times as necessary. The students were then given a worksheet to work on individually that contained questions pertaining to the topic being discussed. Following the worksheet the students were asked to answer two questions about the section. In order to encourage students to partake in this activity, participation points were awarded to each student. The two questions being asked were, "Do you have any other questions to clarify regarding this topic?" and "Do you like this way of learning the new topic?" In response to the latter question, 85.8% of the students said they enjoyed this lecture format and found it interesting, 11.7% found it acceptable but preferred the traditional format of lecture, and 2.5% said they disliked the FCF. A study [4] done in a Controls course in 2013 involving a FCF had similar results. The majority of the students said they would like to see more flipped courses at their university. Another study [5] also revealed that students would like to see more flipped courses for their STEM classes.

Flipped classrooms appear to have a positive effect on student performance as well. A study [6] conducted at the University of California compared the test results of a traditional lecture and a

FCF engineering course. Students in the FCF course outperformed the traditional cohort on all three exams by more than 10%. Interestingly, the largest difference in scores was between students that ranked in the bottom two-thirds of the class. The bottom two-thirds of the FCF cohort scored approximately 20% higher than the traditional cohort equivalent. A prior study [7] involving an introductory Computer Design course showed similar results. Their FCF cohort also outperformed the traditional cohort in both midterm and final scores.

The majority of these courses utilized PLVs to be watched prior to attending lecture, as these would replace the traditional lecture component of class time. But how can we be sure that the students are watching these videos prior to class, or just at all? A study [8] was conducted on a FCF Computer Science course that analyzed the viewing habits of the students through multiple semesters. The three metrics they used were videos viewed, content coverage, and punctuality expressed in hours. Videos viewed is the percentage of the videos the students accessed (out of 25 videos). The content coverage is the proportion of each video actually watched by the student. Finally, punctuality is the amount of time in hours prior to or after class the video was watched. Positive values are hours prior to class while negative values correspond to hours after the class. During the first semester of the study, the average videos viewed from the students was about 49%, content coverage was about 37%, and the punctuality was -88 hours. For the next semester the instructor added a self-reflection question to the end of each video in an attempt to gain more student engagement. This was, at first, not mandatory. Halfway through the semester the instructor decided to enforce a new policy that the students had to have at least one post in response to each video. This increased the videos viewed to about 94%, content coverage to 80% and punctuality to 0.6 hours. For the third semester of the study, the responses were mandatory for the entire year. The videos viewed was 98%, content coverage was 85% and the punctuality in hours was about 35.

To see what impact this all had on student performance, the researchers compared these metrics to the students' final course grade using Spearman correlation coefficients. For the first semester of the study, all three metrics had statistically significant positive correlations with the final course grade. This is deemed reasonable, as the majority of students were not participating in the course via video viewership, those that were would outperform them. For the other two semesters of the study, the only metric that had any statistically significant positive correlation was punctuality. Though this provides some insight on students' interactions with the lecture videos, it does not thoroughly quantify how they were interacting with the PLVs.

With an overwhelming body of research that shows a FCF increases student performance, and a simultaneous underwhelming amount of research attempting to provide correlations or trends as to how high-performing students are capitalizing on this lecture format, and how it benefits low-performing students, this study aims to investigate the relation between students' engagement with PLVs and video questions, and their course performance. Currently, there exists no such trend. Thus, by quantifying students' video viewership habits, and their engagement with the videos through the use of online video questions, we hope to elucidate trends as to what has the greatest affect on student performance.

II. Methodology

Students in a flipped Statics and Mechanics of Materials I course were expected to watch PLVs and answer corresponding video questions prior to attending lectures. There were a total of 65 PLVs that covered the 23 lectures for the semester. A total of 162 video questions were administered with the videos to assess the students' conceptual understanding of the course material, namely their ability to recall basic requisite information covered in the videos (i.e., the first level of Bloom's taxonomy [9]). A detailed overview of the course administration is detailed by Pabst et al. [10].

The video questions were accessible through Top Hat. These questions were graded on participation, not correctness, to encourage students to give an earnest attempt without the worry of losing points. Students had five attempts to answer each question. Often times, the number of attempts a student had per question exceeded the number of available answers to that question. Top Hat tracked the correctness score and the number of attempts a student made to answer each question, and also time-stamped their responses.

To see if the students' final grades were affected by their interaction with the PLVs and video questions, four variables were created and observed. For each variable, students were categorized into different groups so they could be analyzed. The first variable is denoted "Top Hat Video." This variable quantifies when the students answered the Top Hat video questions with regards to watching the PLVs. This variable encompasses three different groups: "before," "during," and "after." "Before" means the student answered the majority of their Top Hat questions prior to watching the PLVs. "During" refers to the student answering the majority of the questions while watching the PLVs, and "after" means the majority of the questions were answered following the completion of the PLVs.

The next variable was called "Normalized Views." This variable calculated the student's total viewership of an individual lecture video divided by the length of the video. Anything over a value of one may be a result of the video being watched more than once. That is to say they may have watched two - thirds of the video twice, while never actually watching the video to completion. Each normalized video view was then totaled and divided by the total number of lecture videos (65) to give an average for each student. To calculate the "Correctness to Attempts Ratio," the third variable of interest, the number of video questions correctly answered was divided by the number of attempts needed to achieve the correct answer. The results are bounded between zero and one. A value of one corresponds to all the video questions being answered correctly on the first attempt. As the ratio decreases towards zero, this indicates either more attempts were used to achieve the correct answer, or possibly the correct answer was never selected for some questions. A value of zero indicates a student never correctly answered one of the 162 video questions. For these two variables, "Normalized Views" and "Correctness to Attempts Ratio," the students were broken up into three equal groups based on their respective percentile. For the "Normalized Views" grouping, the top third watched more than 67% of all the PLVs. The middle third watched between $22\%-67\%$, and the bottom third watched less than 22% of all the PLVs. In terms of "Correctness to Attempts Ratio," the top third had a ratio that was greater than 71%. The middle third had a ratio between 58% and 71%, while the bottom third was below 58%.

Though the video questions were solely graded on participation, one question that arises is if those who went out of their way to answer all the questions performed better in the course. The final variable observed was "Did Not Answer" (DNA) as constructed in an attempt to answer this question. This variable was binned into two categories. One group was made up of students who answered all 162 questions. That is to say that their DNA percentage was 0%. The other group was made up of students who did not answer every question. In other words, their DNA percentage was greater than 0%.

The results for differences between the groups and final course

grades were analyzed with one-way between subjects analysis of variance (ANOVA), one for each variable. Before utilizing the one-way ANOVA, all assumptions were checked to ensure accuracy of results. These assumptions include normality, homogeneity of variance, and significant outliers, all of which were examined using IBM SPSS Statistics. The Shapiro-Wilk and Levene's test were used to check normality and homogeneity of variance, respectfully. If there was a statistically significant difference found at a level of 0.05 between the groups, then a post-hoc comparison was conducted to determine the differences. Tukey's Honestly Significant Difference (HSD) test was used for this post-hoc pairwise comparison.

III. Results and Discussion

In the following, the four aforementioned variables were analyzed, and results are presented in order of appearance. Prior to running a one-way ANOVA between "Top Hat Video" groups (i.e., "before," "during" and "after"), Shaprio-Wilk was used to check normality. Unfortunately, the three groups had varying sample sizes and the "after" group departed significantly from normality (p = 0.0140). Therefore, a Kruskal-Wallis H-test was used to determine if there was any statistically significant differences between the groups. Kruskal-Wallis H-test is a non-parametric test that does not assume normality [11]. It is noted that the sample sizes for "Top Hat Video," "Correctness to Attempts Ratio," and "Did Not Answer" (n = 117) are less than "Normalized Views" (n = 183). This is because 66 students un-enrolled from the Top Hat course after completing the course and prior to data collection, and the data was unavailable during analysis. The data has been recovered and will be used in future analyses.

The results showed that there was a statistically significant difference in final grades between the groups, $H(2) = 8.083$, $p = 0.018$. These test results are summarized in Tab. I. Since these results were found statistically significant, Dunn's test was used to determine exactly which groups were different by way of pairwise comparisons [12]. Bonferroni correction was used to made an adjustment to the p-values. Because multiple hypotheses are being tested, the probability of observing a significant difference increases. In other words, a type I error (i.e., a false positive) is more likely to occur, and the Bonferroni correction accounts for this [13]. A statistically significant difference was found between the before and after group with an adjusted p-value of 0.023.

As seen in Fig. 1, the students who answered the video questions "after" watching the PLVs performed better in the course than the students who did so "before" and "during," and on average half a letter grade better. It is interesting to note the "during" group, although having a lower maximum and higher minimum than the other two groups, had a comparable mean to the "before" group, as well as a very narrow distribution of final grades. Although this distribution was narrower, the standard deviation is comparable to the other groups due to the small sample size. The minimum of the "during" cohort is above the first quartile of the "before" group, and nearly two letter grades higher than the minimum of the same group. Additionally, the minimum of the "during" group is approximately one letter grade higher than the minimum of the "after" cohort.

To see if the students' "Normalized Views" had an impact on their final course grade a one-way ANOVA was conducted between the three groups. The middle and top tercile groups both departed significantly from normality with p-values of 0.042 and 0.016, respectively. However, because the sample sizes across the three groups were identical (n = 61) and are all greater than 40 [14], [15], normality was deemed acceptable. Levene's Test for Equality of Variances was used to check the homogeneity of variance; all

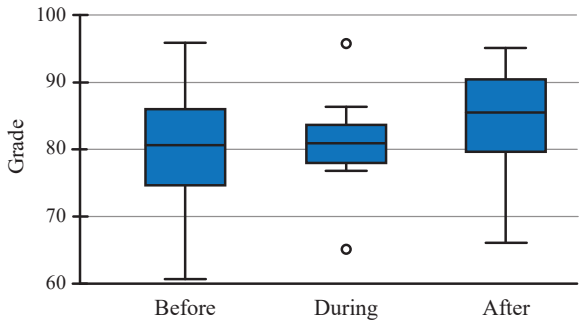

Fig. 1. "Top Hat Video" question answering habits ("before," "during" and "after") in relation to lecture video viewership versus grade.

groups passed. Some outliers were present in both the middle and top third groups, as seen in Fig. 2, but none were determined to be significant; the outliers were included in the analysis. Therefore, it was deemed acceptable to run the one-way ANOVA. The results showed that there was a statistically significant difference between the groups ($F(2, 182) = 8.728$, $p < 0.001$). These results are displayed in Tab. I. Tukey's HSD showed that the differences were between the middle third and bottom third (p = 0.040) groups, as well as the top third and bottom third ($p < 0.001$) groups.

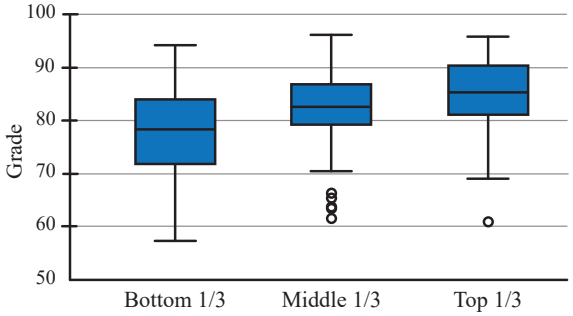

Fig. 2. "Normalized Views" versus grade for bottom (< 22% total video viewership), middle (22%–67% total video viewership), and top (> 67% total video viewership) terciles.

As seen in Fig. 2, students in the top and middle terciles of "Normalized Views" outperformed those in the bottom. The middle third performed about 3.6% better than the bottom third and the top third performed about 6% better than the bottom third (about half a letter grade). In other words it would appear that students who watched a higher portion of the videos, possibly re-watching videos, performed better in the class than those who watched a smaller portion of the videos. Additionally, the minimum score of the middle and top terciles performed as well as the first quartile of the bottom third. The median grade of the top third is approximately equal to that of the third quartile of the bottom third.

Two unique variables were created and observed to see if the students' interactions with the video questions had any impact on their final grade. The first variable is the "Correctness to Attempts Ratio." The assumption of normality was met as the groups passed the Shapiro-Wilk test. No significant outliers were detected. However, the groups violated the assumption of homogeneity of variance which was tested using Levene's. Because of this, Welch's ANOVA was used to calculate the results [16]. It was determined that the groups did differ significantly from one another, $F(2, 72.86) = 28.646$, $p <$

TABLE I
DESCRIPTIVE STATISTICS AND COMPARISON OF MEANS OF DATA SET

		Group Statistics			Levene's Test for Equality of Variances	Kruskal-Wallis H-Test		One-Way ANOVA		Welch's ANOVA		Shapiro-Wilk Test of Normality
Variable	Group	n	Mean	σ	Significance	H	Asymptotic Sig.	F	Significance	F	Significance	Significance
Top Hat Video	Before	43	80.34	8.29	-	2.235	0.017	-	-	-	-	0.667
	During	11	80.97	7.46								0.356
	After	63	84.69	7.44								0.014
Normalized Views	Top 1/3	61	84.56	7.23	0.325	-	-	8.728	< 0.001	-	-	0.016
	Middle 1/3	61	82.08	8.24								0.042
	Bottom 1/3	61	78.53	8.53								0.563
Attempt to Correctness Ratio	Top 1/3	39	88.00	5.00	0.024	-	-	-	-	28.646	< 0.001	0.329
	Middle 1/3	39	82.83	7.67								0.369
	Bottom 1/3	39	77.39	7.29								0.846
Did Not Answer	= 0%	71	84.99	7.37	0.867	-	-	16.149	< 0.001	-	-	0.017
	$> 0\%$	46	79.27	7.72								0.808

0.001. Seen in Tab. I is the summary of the results discussed. Due to homogeneity of variances being violated, a Games-Howell post-hoc test was used for the pairwise comparisons [17]. This revealed difference between the top third and middle third of the course (p = 0.002) and the top third and bottom third of the course ($p < 0.001$).

Students who had higher ratios outperformed their peers with lower ratios. The difference between the top and middle third of the "Correctness to Attempts Ratio" was half a letter grade (5.2%). The difference between the top and bottom third of the course was equivalent to a whole letter grade (10.6%). One potential reason for this is the students with the lower ratio scores may have been guessing randomly instead of taking earnest additional attempts. However, the questions were made up of various forms to include: multiple choice, true or false, and matching. This notion could be the case for questions that were multiple choice and matching questions, but would not be true for true or false questions. It is important to remember that the students had a maximum of five attempts per question. Students who maxed out their attempts while guessing an answer would have a low "Correctness to Attempts Ratio," and those who chose to skip certain questions would receive a zero for that respective question.

Figure 3 depicts the variation of student performance for the three terciles. It is worth noting a few things. The minimum of the top third is approximately the same value as the median of the bottom third, and the minimum of the middle third is approximately the same value as the first quartile of the bottom third. The median of the top third is approximately equal to the third quartile of the middle third, and the median of the middle third is approximately the same value as the third quartile of the bottom third. The maximum of the bottom third is approximately equal to the third quartile of the top third. The difference of minimums between each tercile is nearly a letter grade. Conversely, the maximums between each tercile are approximately the same. It is evident that high "Correctness to Attempt Ratios" significantly impacted the students' overall performance.

The second variable relating to student interactions with video questions was DNA. The two groups associated with this variable failed normality according to the Shapiro-Wilk test. Due to the sample sizes being greater than 40 in both groups, normality was again assumed. Both groups also had no significant outliers and passed Levene's Test of homogeneity of variance, therefore a one-way ANOVA was conducted. This showed that there was a statistically significant difference between the groups, $F(1, 115) = 16.149$, $p < 0.001$. These results are displayed in Tab. I. Comparing means, it appears that the students who attempted all the questions (i.e., the 0% cohort) performed about half a letter grade higher than those that did not (i.e., the > 0% cohort). This data does not consider whether the questions were answered correctly, but merely answered. It appears that just attempting all the questions would lead to better performance in the course. This trend has been seen with interactive, on-line textbooks; the act of simply interacting with the content leads to increased performance [18].

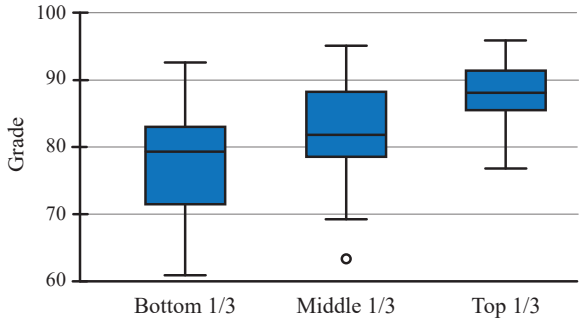

Fig. 3. Top Hat video "Correctness to Attempt Ratio" versus grade for bottom (< 58%), middle (58% –71%), and top (> 71%) terciles.

Figure 4 provides insight into the the variation of student performance when considering answering habits. Although the maximums of the two groups are nearly the same, the median of those who answered all questions (i.e., = 0%) approximately equals the third quartile of those who did not (i.e., > 0%). The median of those who did not answer all questions is about equal to the first quartile of those who answered all questions. The difference between minimums is nearly a letter grade with the group answering all questions having the higher grade.

IV. CONCLUSION

By investigating how students interacted with both PLVs and video questions, it was hoped insight into their affects on student performance would be elucidated. Variables created from quantitative data, provided by both Top Hat and Panopto allowed for the assessment, to some degree, on how students utilized the aforementioned videos and video questions. This data was used to see what their impacts were on the students' performance in the course. This was done through the use of various statistical methods, which include, but are not limited to, analyses of variance, H-tests, and non-parametric tests. Firstly, there was a difference in performance between students who decided

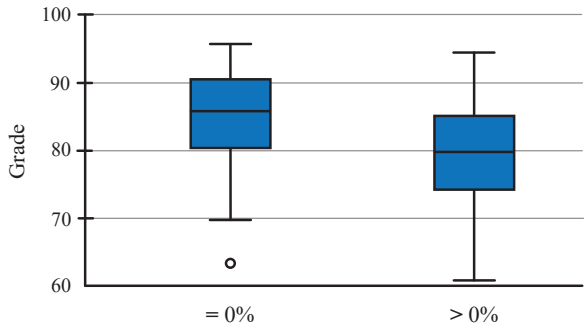

Fig. 4. Answering habits, characterized by the variable "Did Not Answer," versus grade for groups who answered all Top Hat questions (= 0%), and those who did not answer all Top Hat questions (> 0%).

to watch the PLVs before answering the Top Hat video questions, and the students who answered the Top Hat video questions prior to watching the lecture videos. Students who answered the Top Hat lecture video questions after watching the lecture videos earned on average half a letter grade higher than students who attempted to answer the lecture video questions before watching the lecture videos.

Secondly, students who watched a higher portion of the PLVs, which includes re-watching, performed better in the course than those that watched a smaller portion. The top tercile (> 67%) of "Normalized Views" earned grades that were approximately 6% higher, and the middle tercile (22%–67%) performed on average 3.6% better, than the lowest tercile (< 22%). Thirdly, it was also found that students who had higher "Correctness to Attempts Ratios" in regards to the Top Hat video questions performed better. The most significant difference was between the top (> 71%) and bottom tercile (< 58%) cohorts. The top tercile performed 10.6% better than the lowest tercile. The middle tercile (58%–71%) earned grades 5.2% higher than the bottom tercile. Lastly, students who went out of the way to at least answer all of the Top Hat video questions, regardless of attempts, performed better than those who did not answer all of the questions, earning 5% more.

This WIP provides insight into student engagement with PLVs in a FCF, and provides a basis for future explorations. In the future, the researchers would like to make some adjustments to the current study. Though it was known when the videos were watched in reference to when the video questions were answered, it would be prudent to adjust the definition of the "during" section. It would be helpful to apply a small buffer to the time period defining "during" as there may be students who are answering the video questions immediately after watching the video. As the study was run, there may have been a potential inflation of the "after" group of that variable, and an under-representation of the "during" group. In addition to investigating the duration of the buffer (e.g., a five to ten minute grace period), the study will be re-run to increase the sample size. In doing such, some groups may no longer fail normality or homogeneity of variance. This current sample size prevented the running a two-way ANOVA on any of the variables, which could help identify which variable or variables together have the largest impact on student performance. The researchers would also like to use student grade point average as a covariate.

Further investigation may provide insight into the students habits when watching the videos. Namely, are the students watching the video all the way through on their first watch, or do they pause the video once or multiple times. Knowing if they rewind certain parts to re-watch or skip through certain parts, in relation to their number of views, as well as answering the video questions, may be insightful. Although "Normalized Views" is indicative of multiple video views, it would be helpful to know how often the students are accessing the videos. Do they watch a video multiple times, and if so which ones? Are they re-watching the videos solely before the due date (i.e., prior to class, when they are due) or prior to assessments? With this knowledge in mind, it may help guide the instruction of FCFs as to what task or combinations of tasks a student must undertake to maximize their performance.

References

[1] J. Bergmann and A. Sams, "Flip your classroom: Reach every student in every class every day," Eugene, Or.: International Society for Technology in Education, 2012.

[2] R. M. Clark, M. Besterfield-Sacre, D. Budny, Daniel, W. W. Clark, B. A. Norman, R. S. Parker, J. F. Patzer and W. S. Slaughter, "Flipping engineering courses: A school wide initiative," *American Society for Engineering Education: Advances in Engineering Education*, vol. 5, no. 3, pp. 1–39, 2016.

[3] P. Kiat and Y. Kwong, "The flipped classroom experience," in 2014 IEEE 27th Conference on Software Engineering Education and Training, pp. 39–43, 2014.

[4] D. Jean-Pierre and M. Egerstedt, "Flipping the controls classroom around a MOOC," in 2014 American Control Conference, pp. 2557–2562, 2014.

[5] G. B. Johnson, "Student perceptions of the Flipped Classroom," T, University of British Columbia, 2013.

[6] Q.-V. Dang and D. D. Gajski, "Bringing In-Class Online - A Hybrid Solution," in Fourth Interdisciplinary Engineering Design Education Conference, pp. 12–17, 2014.

[7] A. Amresh, A. R. Carberry, and J. Femiani, "Evaluating the effectiveness of flipped classrooms for teaching CS1," in 2013 IEEE Frontiers in Education Conference (FIE), pp. 733–735, 2013.

[8] S. L. Dazo, N. R. Stepanek, R. Fulkerson, and B. Dorn, "An Empirical Analysis of Video Viewing Behaviors in Flipped CS1 Courses," *ACM Inroads*, Vol. 7, no. 4, pp. 99–105, 2016.

[9] D. Krathwohl and L. Anderson, "Merlin C. Wittrock and the revision of Bloom's taxonomy," *Educational psychologist*, vol. 45, no. 1, pp. 64–65, 2010.

[10] D. M. Pabst, L. A. Dosse, S. E. Wismer and M. M. Barry, "A Comparative Study on Student Performance using Traditional and Interactive Textbooks,"; in American Society of Engineering Education (ASEE) 2023 Annual Conference and Exposition, 2023.

[11] Laerd Statistics, "Kruskal-Wallis H Test in SPSS Statistics | Procedure, output and interpretation of the output using a relevant example.," Laerd.com, 2014. [Online]. Available: https://statistics.laerd.com/spss-tutorials/kruskal-wallis-h-test-using-spss-statistics.php. [Accessed: May 21, 2023].

[12] O. J. Dunn, "Multiple Comparisons Among Means," *Journal of the American Statistical Association*, Vol. 56, no. 293, pp. 52–64, 1961.

[13] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 57, no. 1, pp. 289–300, 1995.

[14] A. Elliott and W. Woodward, *Statistical analysis quick reference guidebook: With SPSS examples*, Sage, 2007.

[15] J. Pallant, *SPSS survival manual: A step by step guide to data analysis using IBM SPSS*, McGraw-hill education (UK), 2020.

[16] H. Liu, "Comparing Welch ANOVA, a Kruskal-Wallis test, and traditional ANOVA in case of heterogeneity of variance," Virginia Commonwealth University, 2015.

[17] "One-Way ANOVA Post Hoc Tests," www.ibm.com. [Online]. Available: https://www.ibm.com/docs/en/spss-statistics/SaaS?topic=anova-one-way-post-hoc-tests. [Accessed: May 21, 2023].

[18] S. E. Wismer, L. A. Dosse and M. M. Barry, "Evaluation of an Interactive Top Hat Text for Engaged Learning," in 2021 IEEE Frontiers in Education Conference (FIE), pp. 1–5, 2021.