Exploring the Differences in Students' Behavioral Engagement With Quizzes and Its Impact on their Performance in a Flipped CS1 Course

Ashish Aggarwal
ashishjuit@ufl.edu
Department of Engineering Education
University of Florida
Gainesville, Florida, USA

Akshay Ashok
akshayashok@ufl.edu
Department of Comp. & Info. Science and Engineering
University of Florida
Gainesville, Florida, USA

ABSTRACT

There is an increase in the adoption of flipped classroom pedagogy for introductory programming (CS1) courses. In a flipped course, students watch the content videos and complete an accountability quiz before the class, then do active learning activities during the class. The role of students' behavioral engagement and its impact on learning outcomes is widely studied in education, but little is known about its effect in flipped CS1 courses. This paper analyzes factors related to students' behavioral engagement with quizzes, such as how much time they spend on quizzes, when they choose to submit the quizzes, and how consistently they space their weekly quiz submissions over a fifteen-week semester. Firstly, group differences based on GPA, gender and prior programming experience (PPE) are explored to understand how behavioral engagement varies among different student populations. Secondly, we analyze the association of behavioral engagement with students' performance using exam averages. We find that behavioral metrics do not vary based on GPA, PPE, and gender. Further, we find that while the time taken on quizzes and weekly consistency is not correlated with students' performance, students who submit the quizzes earlier tend to have statistically higher exam averages than those who complete them near the deadlines. These results align with earlier findings and will help instructors understand students' behavioral approaches to flipped CS1 courses, which can help them tailor their instructions accordingly.

CCS CONCEPTS

• **Social and professional topics → CS1**; **Model curricula**; *Computational thinking*.

KEYWORDS

student behavior, submission times, self-regulation, time management, CS1, flipped classroom, active learning

Koli 2022, November 17–20, 2022, Koli, Finland

ACM ISBN 978-1-4503-9616-5/22/11…$15.00
https://doi.org/10.1145/3564721.3564740

ACM Reference Format:
Ashish Aggarwal and Akshay Ashok. 2022. Exploring the Differences in Students' Behavioral Engagement With Quizzes and Its Impact on their Performance in a Flipped CS1 Course. In *Koli Calling '22: 22nd Koli Calling International Conference on Computing Education Research (Koli 2022), November 17–20, 2022, Koli, Finland*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3564721.3564740

1 INTRODUCTION

An essential goal for CS educators and researchers is to foster better learning outcomes among students. Students' learning outcomes are a function of many latent constructs like motivation, ability, instruction, level of practice, etc. These latent constructs are further influenced by multiple factors like a student's course load, life events, time management, study habits, etc. Thus, studying causal or even correlational factors associated with learning outcomes is a study of complex individual characteristics and their interaction with each other.

Various studies have explored the effects of multiple factors on students' performance, especially in the context of undergraduate university courses [4, 33, 40]. One such area of exploration is identifying factors that predict academic success [31]. A review conducted by Hellas et al. at the ITiCSE-2018 Working Group, classified the features used to predict student performance into five categories: demographic (e.g., age, gender), personality (e.g., self-efficacy, self-regulation), academic (e.g., high-school performance, course performance), behavioral (e.g., log data) and institutional (e.g., high-school quality, teaching approach) [15]. They found that the *"majority of the articles used academic data for prediction (e.g., predicting course performance based on high-school performance)"* and that while the use of log data as a metric for student behavior is increasing, it *"is still relatively rare"*.

With the rising enrollments in introductory programming courses and relatively higher attrition rates [2, 21, 37], it has become important for CS educators to review current pedagogical and instructional practices to improve learning outcomes [5, 26, 29]. Many categories described above, like demographics, are fixed and cannot be altered to influence outcomes. Other factors like self-efficacy are difficult to address in a semester-long setting for an individual course. Instructors can work on altering the instructional strategy and tailor it to students' needs. Still, the context of large CS classes having students with varying levels of ability and prior experience challenges the extent to which every student's learning needs can be addressed [20, 38]. One such instructional change that has been in progress in recent years has been the adoption of flipped classroom pedagogy [3, 12, 19]. In a flipped course, students watch content videos and complete an accountability quiz before the class. This accountability quiz reinforces the importance for watching the videos before the class. They then engage in active learning activities during the class, where the instructor and other course staff can assist them. This flipped class pedagogy has resulted in higher learning outcomes across many domains like Physics [9], Chemistry [13] and Computer Science [5, 12, 16]. Various studies have explored factors that affect students' performance in flipped courses [22], and factors that contribute to creating successful flipped classes [13, 16].

A factor that has been traditionally less studied but has increasingly received more attention in CS Education literature is understanding student behavior and its influence on performance [24, 34, 35, 39]. Through the growing use of various tools and learning management systems (LMSs), accessing behavioral data that shows not just what students are doing but also but also when they are learning has become more accessible [18, 20, 27, 35]. While there have been studies comparing the learning outcomes of a flipped and a traditional course [17], little is known about the effect of students' behavioral engagement on their learning and performance. Given that in a flipped course, students watch the videos and complete the accountability quiz at their own pace, analyzing students' behavioral engagement with the content can provide further insights into how the course structure and policies can be tailored for increasing learning outcomes [8, 12, 13].

The term *"student behavior"* or just *"behavior"* is a construct that can entail many independent and dependent observations. A recent systematic review on students' study behavior in computing education by Loras et al. provides a comprehensive categorization of student behavior into broadly four categories, i.e., process, strategies, habits, and tactics [25]. They define these categories as follows:

- **Process**: Cognitive engagement with study activities, that is, students' internal approaches to studying, learning and learning styles.
- **Strategies**: Level of cognitive control over student activities, that is, metacognition, self-regulation, motivation and affective constructs.
- **Habits**: Consistency and actualization of study activities without considering the interaction with the environment, like time spent on activities (time engagement), frequency of study sessions etc.
- **Tactics**: Individual learning tools used by students while studying like note taking, self testing and video viewing.

So, while students' learning behavior can be classified and studied at many levels [7, 24, 25], we are interested in exploring secular behavior trends associated with students' natural study habits and time engagement. Increasingly, instructional approaches like online or hybrid learning, flipped classroom approach, or mastery-based learning, all allow self-paced exploration of content that students can engage with at their convenience [12, 29]. Such approaches not only provide students flexibility but also expect them to plan their learning effectively. Thus, as students are responsible for their time management, it would be valuable to explore when and how consistently students engage with the content and whether that influences their performance.

When we refer to content, we are specifically interested in understanding when students complete the accountability quiz due before the class. We believe that the behavioral metrics related to this quiz can provide insights into understanding students' organic behavior as the videos and quizzes are available long before they are due. Also, completing the accountability quizzes is a proxy for students' video viewing time, as the quizzes are based on the content videos. In this paper, we are interested in exploring: When do students do their weekly accountability quizzes in a flipped CS1 course? How much time do they spend on these quizzes? How consistent are their quiz submission times across multiple weeks? And do these factors affect their performance? We conduct our analysis based on three factors, students' gender, GPA, and prior-programming experience (PPE), to understand if there are group differences across any of these factors.

2 BACKGROUND AND RELATED WORK

Student behavior has been of interest to educational researchers and psychologists. Multiple studies on time management, self-regulation, consistency in behavior, and procrastination have advanced our understanding of the relationship between behavior and learning [1, 30, 36]. In a recent large-scale study spanning across multiple semesters and courses, it was found that procrastination behavior varies across different sociodemographic groups and that it has a negative association with performance across groups [32]. In CS Education, multiple studies have explored the role of students' behavior on their performance [24, 34, 39] and have found mixed results, with some studies suggesting that it is significant [5, 16], and some finding it insignificant [13]. Prior programming experience has been an influential factor in students' performance, especially in introductory CS courses [38].

Regarding study habits, Soohyun et al. analyzed differences in behavioral approaches between higher and lower-performing students. They found that while procrastination does not impact higher performing students, it affects lower performing students as they are unable to complete the assignment and get the required assistance [24]. In another subsequent study, Soohyun et al. found that high-performers more frequently practice some study habits (like self-regulation) than low-performers. Further, when accounting for students' prior programming experience, they conclude that *"although prior experience translates to better class performance, it is not associated with the same study habits as lower- and higher-performers, suggesting that prior experience and study habits are separately associated with better student performance"* [23].

Willman et al. studied the effect of study habits on students' performance in CS1 and found that *"students who receive the highest grade start and finish their work early, do not work on weekends, and do not work at night...."* [39]. Edwards et al., and Shaffer et al., also found that students who begin programming projects earlier perform better than those who begin closer to the due date [10, 34]. Given that it is mostly accepted that students who start earlier on assignments perform better [35], little is known about the effect a student's quiz submission time could have on their performance in a flipped CS1 class.

In the context of the flipped classroom, Dazo et al. studied the impact of video viewing behaviors in a flipped CS1 class [8]. They found, firstly, that viewing videos is not guaranteed and should not be assumed. Secondly, they also found students who watch all the videos earlier and completely perform better in the course [8], thus implying the importance of not only watching the videos but also watching them early. Moore et al. also conducted a similar analysis to study video watching behavior in a flipped CS1 course and found that students with prior programming watched fewer videos and that *"while video watching is an important learning component for some of the students in a flipped CS1 course, other students can chose not to watch the videos and still learn via the other components of the course such as reading, active learning, problem solving, and working on programming assignments"* [28].

Further research indicates students in flipped classrooms are more engaged with their learning and become timelier with their preparation, leading to an increased performance [5, 11]. The flipped classroom model is designed to increase consistent engagement with course content, which minimizes the amount of *"cramming"* style studying [13]. Participating within a flipped classroom has also been linked to exam performance [13], but the timing of engagement and quality of engagement is largely unknown. Gross et al. indicated that accurately answering homework questions in the two weeks before the exams can lead to higher exam performance [13], which suggests that time is an influential factor in flipped classrooms.

Therefore, grounded in this above prior work, we believe that exploring the patterns and influence of just the quiz submission times as a proxy for student behavior may provide some guidance on overall and group based difference.

3 COURSE CONTEXT

The data used in this study comes from a two credit introductory programming course for engineering students that is taught in MATLAB. The course was taught using the flipped classroom model in the spring semester of 2021 (after the start of the COVID-19 pandemic) at an R1 research university in the US. This course had a total of 3 sections, all of which offered an in-person class subsection (capped at 10 students for social distancing purposes) and an online class subsection. Each of the 3 main sections contained around 40 students, and most students opted to attend the class virtually even if they were enrolled in the in-person section.

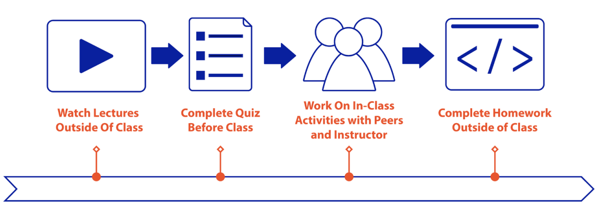

Figure 1: Weekly plan of the flipped CS1 course

Figure 1 presents the weekly timeline for the course. Since the course style used the flipped classroom model, students were expected to watch module lectures before class, as well as complete the module quiz before class. Its important to note that this quiz was mostly for accountability purposes and students were given two attempts with each attempt capped at twenty minutes. The quizzes

What will be the output of the following program:

```
clc; clear;

for ii = 3:3:7
    jj = ii;
    while jj < 5
       jj = jj + 1;
       continue;
       fprintf('%g', jj);
    end
    jj = jj + 2;
    break;
end

fprintf('%g', jj)
```

Figure 2: Sample quiz question from Module-6 (loops)

mostly had five to seven questions, mostly output-prediction questions (Figure 2) that were either multiple choice or free-response questions. The quizzes accounted for 10% of the overall grade and there were eleven quizzes over a semester of fifteen weeks. The class was a combined 2-period block of 50 minutes each, totaling 100 minutes with a 15 minute break in between. Students spent the first 30 minutes reviewing the topic and peer instruction activities. They then worked on solving three smaller-programming problems (Figure 3) individually in the remaining time. They were encouraged to discuss their potential approaches and questions with their peers. They could also ask for help from the undergraduate teaching assistant and the professor. After the class, students were expected to work on the module homework assignments outside class, which was due at the end of the school week.

Activity B

Given an image of coins on a table, count the number of pennies! Note that you are provided three images to test with, but your code should work for any similar image. (Assume that the images were taken with the same camera, at the same distance from the table) **Note: You many not use any automatic commands for finding circles in the image!**

Figure 1: Two sample images of coins on a table.

Program Inputs

- **Enter image name:**
  - *The user will always enter a valid filename*

Program Outputs

- **Number of pennies: XXX**
  - *Replace XXX with the number of pennies in the image*

Test Case 1:

```
Enter image name: Penny_Alpha.png
Number of pennies: 1
```

Figure 3: In-class programming problem from Module-11 (images)

The key programming concepts covered in the modules included input/output, if statements, while loops, for loops, vectors, string, images, and thresholding. The course was divided into 14 modules with eight programming-related homework assignments and two exams (a midterm and a final). The last homework assignment was a cumulative final project.

4 DEMOGRAPHICS

There were 121 students across the three sections of the introductory programming course out of which data of 3 students was discarded due to missing information. Out of 118 students, a total of 60 students (51%) had no prior programming experience while 58 students (49%) had some form of prior programming experience (PPE). The sections consisted primarily of mechanical and/or aerospace engineering students (61%), civil and/or environmental engineering students (14%), biomedical engineering students (9%), and other students (16%) as shown in Figure 4. There were 79 (67%) male students and 39 (33%) female students. Based on students' academic year, 72 students (61%) were in the first year of their undergraduate program, 29 students (25%) were in their second year, 14 students (12%) were in their third year, 2 students were in the fourth year, and 1 student was in the first year of their graduate program (Figure 5).

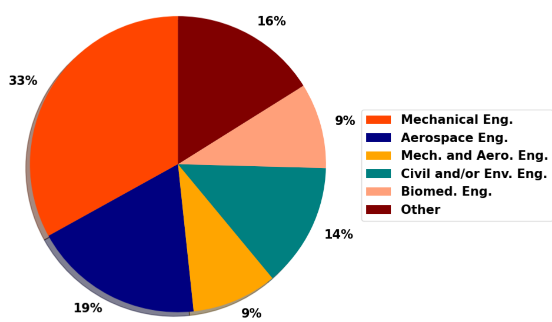

Figure 4: Students' majors within the course

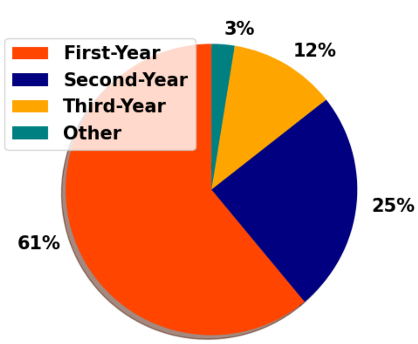

Figure 5: Students' academic year within the course

Students were also categorized based on prior Grade Point Average (GPA). At the institution at which this study was conducted, on the GPA scale, an 'A' grade translates to a 4.00, a 'B' grade translates to a 3.00, a 'C' grade translates to a 2.00, a 'D' grade translates to a 1.00, and an 'E' or 'F' grade translates to a 0.00. In most courses at the institution, a grade of 'C' or above is considered passing. GPA is calculated as the average of a student's course grades, weighted based on the number of credits for each course. We set the GPA ranges in Table 1 according to minimum and maximum GPA of students in the sample and traditional perceptions of low, average, and high GPA.

Table 1: Organization of GPA Groups

Group	GPA Range	Count	Percent of Sample
0	Not available	9	7.6%
1	(1.1, 2.8]	6	5.1%
2	(2.8, 3.5]	29	24.6%
3	(3.5, 4.0]	74	62.7%

5 MOTIVATION AND RESEARCH QUESTIONS

At an R1 research university in the US where this study was conducted, the flipped classroom model helped lower the student-to-teacher ratio. Consequently, it facilitated an individualized support mechanism where the instructor and teaching assistants could better help students while they were working on programming problems. Though research supports the notion that such active learning is more important in STEM courses [5, 11], it is still unknown exactly what other influential factors could enhance or detract from a student's learning experience.

In a flipped environment, there is increased responsibility on the student to complete the assigned lectures and learn the material before coming to class every week. With this in mind, it is important to know what impact this increased responsibility has on a student's learning experience [12, 13, 19]. Managing their own time, some students will regularly submit the required weekly quiz more than a day earlier than some of their peers. With the submission time of quizzes varying between students, it is important to be able to identify if this impacts a student's overall performance in the course. A starting point for understanding student behavior is their approach toward self-scheduling the commitments required for a flipped course. Although every student has a different schedule impacting when assignments are completed (alongside other factors), this paper strives to understand more about how students' behavioral approach vary based on groups related to GPA, gender and PPE, and how it impacts their performance.

Knowing more about students' learning behaviors and which types of behaviors contribute to higher performance can help instructors curb the higher failure rate in computer science courses [2, 4, 37]. With an understanding of factors influential to a student's overall performance, instructors can better alter instructional design to encourage high-performing behaviors in their students. With this in mind, we established two guiding inquiries for this paper:

- Are there differences in quiz submission behavior or performance between groups based on GPA, gender, and prior programming experience?
- Do proactive and consistent quiz submission times affect students' performance in a flipped classroom CS1 course?

5.1 Variables

Based on the guiding inquiries, we derive three variables to account for behavioral engagement and one variable to account for students' performance in order to further investigate the topic.

- **Time Spent**: This independent variable serves as a metric for how long a student usually takes to complete their weekly quiz.
- **Proactive Behavior Index (PBI)**: This independent variable serves as a metric for how early a student usually submits their weekly quiz with regard to the deadline. This essentially translates to a measure of procrastination with regard to weekly quizzes.
- **Inconsistent Behavior Index (IBI)**: This independent variable serves as a metric for how variant a student is in terms of submitting their quiz at similar times every week. This essentially translates to a measure of self-regulation (or lack thereof).
- **Exam Average**: This dependent variable serves as an indicator of a student's performance in the course.

5.2 Research Questions

Incorporating the three behavioral variables and one performance variable, we arrived at two formal research questions:

1. **Are there differences in Time Spent, Proactive Behavior Index, Inconsistent Behavior Index, and Exam Average between groups based on GPA, prior programming experience, and gender?**
2. **Is there a relationship between Exam Average and Time Spent, Proactive Behavior Index, and Inconsistent Behavior Index?**

6 METHODOLOGY

Self-reported data on students' gender, prior programming experience, GPA, and major was collected via an introductory survey during the first week of the course. Data on students' quizzes was retrieved via the learning management system for the class. The Institutional Review Board's (IRB) approval was taken before analyzing the data. Out of the 121 students in the course, three were removed from the data set due to missing information, resulting in a sample size of 118 students. Data of all the 118 students was available for the entire semester and no student dropped the course in between.

There were a total of 11 quizzes and two exams (a midterm and a final) administered throughout the course. Since students had two attempts for every quiz, we considered only the first submission time and time spent on the first attempt for this analysis. This is because the behavioral metric associated with the first attempt is more indicative of students' behavioral engagement and they may or may not use the second attempt.

For each quiz, start time was subtracted from the corresponding finish time to produce a result in terms of minutes. This value will be referred to as the "time taken" for each submission.

Then, for each quiz, submission time was subtracted from the corresponding due date to produce a result in terms of hours. This value will be referred to as "time remaining" for each submission. It is important to note here that this time remaining has strong potential to characterize a student's behavioral approach toward the course. This value helps in comparative analysis of students based on how much time earlier they submitted the assigned quiz before it was due. While it may be possible that a student completes a particular assignment well ahead of the deadline and does not submit it until closer to when it is due, a collection of multiple such time remaining values can still indicate a dominant behavioral approach, especially when analyzed over a semester's coursework. Thus, these time remaining values form the basis of our characterization of a student's approach toward the course, which we believe to be valuable for its potential impact on students' performance in a flipped classroom environment.

- **Time Spent** was calculated for each student by taking the median of all "time taken" values for that student.
- **Proactive Behavior Index** was calculated for each student by taking the median of all "time remaining" values for that student.
- **Inconsistent Behavior Index** was calculated for each student by taking the inter-quartile range of all "time remaining" values for that student.
- **Exam Average** was calculated for each student by taking the mean of the two exam scores for that student and multiplying by 100 to obtain a percentage.

6.1 (RQ 1) Differences in Variables Between Groups Based on GPA, PPE, and Gender

We are first interested in finding differences in Time Spent, Proactive Behavior Index, Inconsistent Behavior Index, and Exam Average between student groups based on GPA, prior programming experience, and gender. After performing Chi-Square tests for goodness of fit for normal distribution, we determined that Time Spent, PBI, IBI, and Exam Average all yield significantly non-normal distributions (as is visible in Figures 6 and 7). As such, we opted for nonparametric tests for differences. Note that for all tests, we used a p value threshold of 0.05.

To find differences in the four variables between GPA groups, we used a multiple comparison method. Comparisons for all variables failed Levene's test for homogeneity of variance. As such, we used the Kruskal-Wallis H test, a nonparametric equivalent to the one-way ANOVA. Note that Kruskal-Wallis does not necessarily compare means or medians. Rather, each data point is transformed to a rank among all data in the test (smaller values ranking lower and larger values ranking higher), and mean rank for each group is compared. We performed one Kruskal-Wallis H test between GPA groups for each of the four variables. The null hypothesis for each test was that mean rank for the given variable was the same between the GPA groups. The alternate hypothesis for each test was that the given variable differed between the GPA groups. After we performed the Kruskal-Wallis test, we applied Dunn's test for

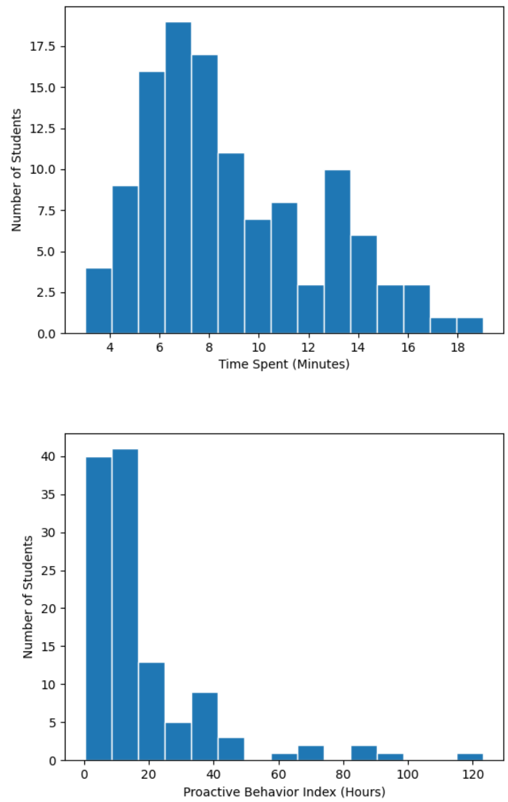

Figure 6: Distributions of Time Spent and Proactive Behavior Index (PBI) in the sample

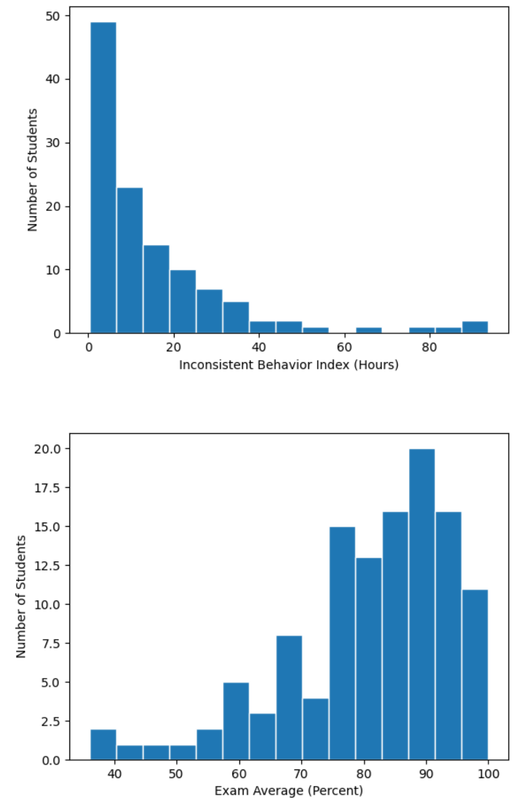

Figure 7: Distributions of Inconsistent Behavior Index (IBI) and Exam Average in the sample

post-hoc analysis of pairwise differences between the groups. To reduce the risk of Type I error, we performed a Bonferroni correction on the significance values for each pairwise comparison. Note that after the Bonferroni correction, the p value threshold for each pairwise difference is 0.008333 (distinct from the p value threshold of 0.05 used for the overall Kruskal-Wallis test). The null hypothesis for Dunn's test is that there is not a significant difference in the given variable between any two specific groups in the comparison. The alternate hypothesis is that there is a difference in the given variable between at least one pair of groups in the comparison.

To find differences in the four variables between students with and without PPE, we used the Mann-Whitney U test, a nonparametric equivalent to the two-sample t test. Note that the Mann-Whitney U test does not necessarily compare means or medians. Like Kruskal-Wallis, each data point is transformed to a rank among all data in the test, and mean rank for each group is compared. We performed one Mann-Whitney U test between students with and without PPE for each of the four variables. The null hypothesis for each test was that mean rank for the given variable was the same between students with and without PPE. The alternate hypothesis for each test was that the given variable differed between students with and without PPE.

To find differences in the four variables between male and female students, we again used the Mann-Whitney U test. We performed one Mann-Whitney U test between male and female students for each of the four variables. The null hypothesis for each test was that mean rank for the given variable was the same between male and female students. The alternate hypothesis for each test was that the given variable differed between male and female students.

6.2 (RQ 2) Relationship Between Performance and Behavioral Variables

As previously stated, Exam Average was used as an indicator for students' performance in the course. Therefore, to gauge how students' performance in the course related to behavioral variables, we constructed an ordinary least-squares (OLS) multiple regression model for inference. In this regression model, Exam Average was the response variable and Time Spent, Proactive Behavior Index, and Inconsistent Behavior Index were the regressors. Note that for all tests, unless otherwise specified, we used a p value threshold of 0.05.

The original regression model did not meet all assumptions of multiple linear regression. The first assumption failed was normality of residuals. The second assumption failed was homoscedasticity (homogeneity of variance) of residuals, specifically for PBI and IBI.

This means that for the two variables, as their value increased, the variance of their residuals tended to decrease. The third assumption failed was linearity, again for PBI and IBI. As such, we transformed PBI and IBI to try to resolve the three failed assumptions. After we tested several transformations, checking regression assumptions for each, we decided to use a natural logarithm transformation for PBI and IBI to resolve the failed assumptions.

With transformed PBI and IBI, we again checked the assumptions for OLS multiple regression, as follows. Independence of observations and homoscedasticity of residuals were verified by analyzing the residual plots. Linearity was verified by analyzing scatterplots of Exam Average versus each regressor. Normality of residuals was verified with a Chi-Square test for goodness of fit for normal distribution (p = 0.06447). The null hypothesis for this test was that the residuals follow a normal distribution, while the alternate hypothesis was that the residuals have a non-normal distribution. We calculated variance inflation factor (VIF) for all three independent variables and found all values to be non-significant (VIF < 2.5), meaning that there was no significant multicollinearity between the regressors. Altogether, we determined the model to be appropriate for analysis.

After fitting the regression model, we performed a Bonferroni correction on the significance values for each predictor. Note that after Bonferroni correction, the p value threshold for each individual predictor is 0.01667 (distinct from the p value threshold of 0.05 used for the overall regression). For the overall model, the hypotheses are as follows. The null hypothesis is that there is no correlation between Exam Average and any of the regressors. The alternate hypothesis is that there exists a significant correlation between Exam Average and at least one of the regressors. For the relationship between Exam Average and each individual regressor, the hypotheses are as follows. The null hypothesis is that there is no correlation between Exam Average and the predictor variable (the coefficient of the variable in the regression equation is zero). The alternate hypothesis is that there is a significant correlation between Exam Average and the given variable.

7 FINDINGS

7.1 Differences in Time Spent Between Groups Based on GPA, PPE, and Gender

For all students in the sample, the statistics for Time Spent were as follows. Mean Time Spent was 8.91 minutes. Median Time Spent was 8.024 minutes. The minimum Time Spent was 2.99 minutes, and the maximum was 19.05 minutes. Standard deviation of Time Spent was 3.48 minutes.

The mean, median, and standard deviation of Time Spent for each GPA group (Group 0, Group 1, Group 2, Group 3 in Table: 1) are as follows: [μ = 9.27, Mdn = 8.07, σ = 3.79] minutes for Group 0, [μ = 9.35, Mdn = 8.66, σ = 3.53] minutes for Group 1, [μ = 8.48, Mdn = 7.85, σ = 3.20] minutes for Group 2, and [μ = 9.00, Mdn = 8.11, σ = 3.59] minutes for Group 3. *The Kruskal-Wallis H test for di erences in Time Spent between GPA groups is not statistically signi cant [H(3) = 0.60, p = 0.9276].* As such, we fail to reject the null hypothesis that Time Spent is uniform across GPA groups.

The mean, median, and standard deviation of Time Spent for students with and without PPE are as follows: [μ = 8.47, Mdn = 7.53, σ = 3.48] minutes for students with PPE and [μ = 9.34, Mdn = 8.76, σ = 3.45] minutes for students without PPE. *The Mann-Whitney U test for di erence in Time Spent between students with and without PPE is not statistically signi cant [U = 1451.5, p = 0.2389].* As such, we fail to reject the null hypothesis that Time Spent is the same for students with and without PPE.

The mean, median, and standard deviation of Time Spent for male and female students are as follows: [μ = 8.61, Mdn = 7.72, σ = 3.48] minutes for male students and [μ = 9.51, Mdn = 8.85, σ = 3.44] minutes for female students. *The Mann-Whitney U test for di erence in Time Spent between male and female students is not statistically signi cant [U = 1290.5, p = 0.2869].* As such, we fail to reject the null hypothesis that Time Spent is the same between male and female students.

Altogether, there are no significant differences in Time Spent between any groups of students based on GPA, PPE, or gender. Thus, we affirm that past academic performance, prior programming experience, and gender do not determine how much time a student usually takes to complete a quiz.

7.2 Differences in Proactive Behavior Index Between Groups Based on GPA, PPE, and Gender

For all students in the sample, the statistics for Proactive Behavior Index were as follows. Mean PBI was 17.63 hours. Median PBI was 11.45 hours. The minimum PBI was 0.19 hours, and the maximum was 123.33 hours. Standard deviation of PBI was 20.78 hours.

The mean, median, and standard deviation of Proactive Behavior Index for each GPA group are as follows: [μ = 14.64, Mdn = 13.03, σ = 10.70] hours for Group 0, [μ = 9.13, Mdn = 2.35, σ = 17.03] hours for Group 1, [μ = 14.21, Mdn = 7.35, σ = 24.45] hours for Group 2, and [μ = 20.03, Mdn = 12.51, σ = 20.32] hours for Group 3. *The Kruskal-Wallis test for di erences in PBI between GPA groups is statistically signi cant [H(3) = 11.09, p = 0.01126].* As such, we reject the null hypothesis and affirm that PBI differs between GPA groups. *However, a er Bonferroni correction was performed, Dunn's post-hoc test showed no signi cant pairwise di erences between groups (p ≥ 0.05).* Therefore, we fail to reject the null hypothesis that there are no pairwise differences between groups in the comparison. Given the lack of a statistically significant result for pairwise differences, the most significant difference in PBI was between Group 2 and Group 3 (p = 0.01025).

The mean, median, and standard deviation of Proactive Behavior Index for students with and without PPE are as follows: [μ = 20.49, Mdn = 12.31, σ = 24.56] hours for students with PPE and [μ = 14.88, Mdn = 10.61, σ = 16.06] hours for students without PPE. *The Mann-Whitney test for di erence in PBI between students with and without PPE is not statistically signi cant [U = 1523, p = 0.4033].* As such, we fail to reject the null hypothesis that PBI is the same between students with and without PPE.

The mean, median, and standard deviation of Proactive Behavior Index for male and female students are as follows: [μ = 20.13, Mdn = 13.05, σ = 23.42] hours for male students and [μ = 12.57, Mdn = 9.86, σ = 12.84] hours for female students. *The Mann-Whitney* *test for difference in PBI between male and female students is not statistically significant [U = 1194, p = 0.1119].* As such, we fail to reject the null hypothesis that PBI is the same between male and female students.

Overall, the only differences in Proactive Behavior Index are between GPA groups. However, even this is contentious because the source of the difference cannot be confirmed. This means that a student may usually submit their quizzes farther from or closer to the deadline based on past academic performance. On the other hand, it is also likely that how early a student tends to submit their quizzes is not based on any of the factors provided here.

7.3 Differences in Inconsistent Behavior Index Between Groups Based on GPA, PPE, and Gender

For all students in the sample, the statistics for Inconsistent Behavior Index were as follows. Mean IBI was 15.41 hours. Median IBI was 8.70 hours. The minimum IBI was 0.27 hours, and the maximum was 93.74 hours. Standard deviation of IBI was 19.40 hours.

The mean, median, and standard deviation of Inconsistent Behavior Index for each GPA group are as follows: [μ = 9.62, Mdn = 7.54, σ = 6.45] hours for Group 0, [μ = 14.40, Mdn = 11.81, σ = 15.23] hours for Group 1, [μ = 15.98, Mdn = 11.08, σ = 16.39] hours for Group 2, and [μ = 15.98, Mdn = 8.67, σ = 20.36] hours for Group 3. *Kruskal-Wallis for differences in IBI is not statistically significant [H = 0.26, p = 0.9667].* As such, we fail to reject the null hypothesis that IBI is uniform across GPA groups.

The mean, median, and standard deviation of Inconsistent Behavior Index for students with and without PPE are as follows: [μ = 13.15, Mdn = 7.08, σ = 16.71] hours for students with PPE and [μ = 17.61, Mdn = 12.46, σ = 19.79] hours for students without PPE. *Mann-Whitney for difference in IBI between students with and without PPE is not statistically significant [U = 1461, p = 0.2583].* As such, we fail to reject the null hypothesis that IBI is the same between students with and without PPE.

The mean, median, and standard deviation of IBI for male and female students are as follows: [μ = 14.98, Mdn = 7.90, and σ = 18.78] hours for male students and [μ = 16.28, Mdn = 12.45, σ = 17.83] hours for female students. *Mann-Whitney for difference in IBI between male and female students is not statistically significant [U = 1446, p = 0.6894].* As such, we fail to reject the null hypothesis that IBI is the same between male and female students.

Altogether, there are no differences in Inconsistent Behavior Index between any groups of students based on GPA, PPE, or gender. Thus, we affirm that past academic performance, prior programming experience, and gender do not determine how variant a student is with respect to when they submit their quizzes.

7.4 Differences in Exam Average Between Groups Based on GPA, PPE, and Gender

For all students in the sample, the statistics for Exam Average were as follows. Mean Exam Average was 80.99%. Median Exam Average was 84.25%. The minimum Exam Average was 36.00%, and the maximum was 100.00%. Standard deviation of Exam Average was 13.56%.

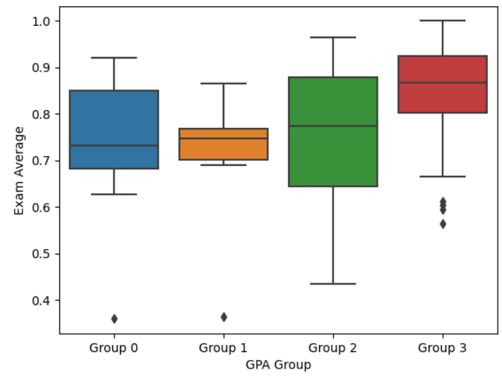

Figure 8: Exam Average based on GPA groups

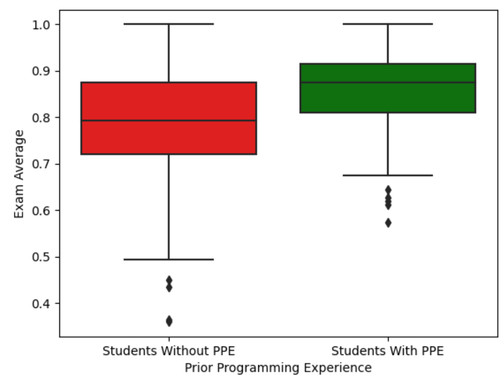

Figure 9: Exam Average based on PPE

The mean, median, and standard deviation of Exam Average for each GPA group are as follows: [μ = 72.30, Mdn = 73.20, σ = 16.84]% for Group 0, [μ = 69.75, Mdn = 74.65, σ = 17.27]% for Group 1, [μ = 75.78, Mdn = 77.50, σ = 15.35]% for Group 2, and [μ = 85.01, Mdn = 86.75, σ = 10.29]% for Group 3 (Figure 8). *We find that between GPA groups, Kruskal-Wallis for differences in Exam Average is statistically significant [H(3) = 16.30, p = 0.0009829].* As such, we reject the null hypothesis and affirm that Exam Average differs between GPA groups. *We also find that Dunn's test for difference in Exam Average between Group 2 and Group 3 is statistically significant (p = 0.003871).* Therefore, we reject the null hypothesis and affirm that Exam Average differs between students in Group 2 and students in Group 3.

The mean, median, and standard deviation of Exam Average for students with and without PPE are as follows: [μ = 84.86, Mdn = 87.50, 10.53]% for students with PPE and [μ = 77.26, Mdn = 79.25, σ = 15.12]% for students without PPE (Figure 9). *Mann-Whitney for difference in Exam Average between students with and without PPE is statistically significant [U = 1182.5, p = 0.008836].* As such, we reject the null hypothesis and affirm that Exam Average differs between students with and without PPE.

The mean, median, and standard deviation of Exam Average for male and female students are as follows: [μ = 81.26, Mdn = 85.00, σ = 14.42]% for male students and [μ = 80.46, Mdn = 82.00, σ = 11.76]% for female students. *Mann-Whitney for difference in Exam Average between male and female students is not statistically significant [U = 1395.5, p = 0.5656].* As such, we fail to reject the null hypothesis that Exam Average is the same for male and female students.

Overall, Exam Average differs between students in different GPA groups and between students with and without PPE. This suggests that a student's exam scores may vary based on their prior academic performance and prior programming experience.

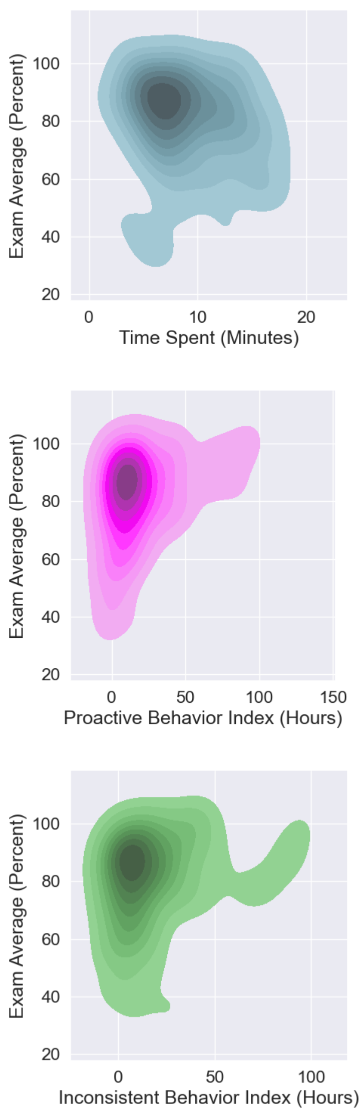

Figure 10: Two-dimensional distribution plots of Exam Average vs. Time Spent, PBI, and IBI

7.5 Regression of Exam Average Versus Time Spent, Proactive Behavior Index, and Inconsistent Behavior Index

Ordinary least-squares multiple linear regression was used to test if Time Spent, Proactive Behavior Index, and Inconsistent Behavior Index significantly predict Exam Average. The fitted regression model is:

$$Exam\ Average = 0.7752 + 0.03481 * PBI - 0.004611 * IBI - 0.003713 * Time\ Spent$$

We find that the overall regression is statistically significant [R^2 = 0.1119, F(3, 114) = 4.79, p = 0.003531]. As such, we reject the null hypothesis and posit that there exists a significant correlation between Exam Average and at least one of the regressors. After we performed a Bonferroni correction, we analyzed individual interactions. *There is a statistically significant relationship between Exam Average and Proactive Behavior Index (β = 0.03481, p = 0.01227).* Therefore, we reject the null hypothesis and affirm that PBI at least partially predicts Exam Average.

There is not a statistically significant relationship between Exam Average and Time Spent (β = -0.003713, p = 0.2928). As such, we fail to reject the null hypothesis that the coefficient for Time Spent in the regression equation is zero.

Similarly, there is no statistically significant relationship between Exam Average and Inconsistent Behavior Index (β = -0.004611, p = 0.7345). Thus, we fail to reject the null hypothesis that the coefficient for IBI in the regression equation is zero.

The overall regression model suggests some interplay between Exam Average and Time Spent, Proactive Behavior Index, and Inconsistent Behavior Index. This interplay accounts for 11.19% of the variance in Exam Average. However, with further tests it is indicated that the only significant predictor of Exam Average is Proactive Behavior Index, with a positive coefficient in the regression equation. This means that as the usual time remaining between quiz submission and the corresponding deadline increases, a student's performance in the course tends to increase. We also find that the amount of time a student usually spends completing a quiz and the variance in a student's quiz submission times do not affect their performance in the course.

8 DISCUSSION

From these findings we observe that broadly, there are no statistically significant differences in Time Spent, Proactive Behavior Index, and Inconsistent Behavior Index based on GPA, prior programming experience, and gender. This implies that students' decision to do a quiz at a particular time is independent of their GPA, PPE, and gender. Their behavioral engagement with these quizzes might be a function of some other unknown factors. The fact that this aligns with the findings of some of the prior studies [23] confirms that finding correlational factors with students' decision on when to do a particular activity needs further investigation and research. The findings also indicate that there are statistically significant differences in Exam Averages based on PPE and GPA which is again a confirmation of the prior findings [38].

Additionally, the results of multiple regression indicate that quiz submission times (calculated as PBI) affect the exam average, implying that earlier submission times correlate with higher exam averages. This also aligns with earlier findings [14, 34] that earlier submission times affects students' performance.

In formal course settings, instructors can optimize students' learning by influencing their experience and behavior based on instruction and course policies. We believe that given how much impact students' behavior has on students' performance, CS educators should focus more on coming up with interventions that impact student behavior in all respects. At the end, students' behavior (both cognitive and habitual) is what has the potential to impact their learning [6]. We can't directly impact or change a student's demographic factors, socio-economic status, self-efficacy, or innate potential for better outcomes. But what we can influence is their behavioral approaches (time based engagement being one of them) by providing adequate scaffolding, effective tools, practice with effective feedback and course policies that help them succeed. However, one thing which requires further investigation and research is factors that affect student behavior. The current findings suggest that GPA, PPE and gender do not affect their quiz submission time in flipped CS1 class and it requires further analysis based on a combination of these and other factors.

9 THREATS TO VALIDITY AND LIMITATIONS

The current analysis explores the group-based differences in submission times of quizzes that are due prior to the class every week. Students' quiz submission times can be a function of varied work load, academic motivation, individual priorities, and other factors that are different from the ones used to understand the group-based differences. Additionally, students may opt to watch the weekly video lecture earlier and submit the quiz later, or even submit the quiz without watching the videos as they have two attempts available. This brings up additional complexities in understanding what influences their decision to submit a quiz at a particular time. Moreover, the data of 118 students were analyzed for the current study and when these are divided into multiple groups, the number of students may not be sufficient especially when there are more than two groups. Also, the current data is collected from a CS1 course that is tailored towards engineering students. That may or may not generalize with other other CS1 courses, especially with CS students. These all are limitations of the current analysis.

10 CONCLUSION AND FUTURE WORK

As the use of flipped class pedagogy increases in CS courses, and as more students with mixed abilities and varied prior programming experience enroll in these courses, it is valuable to understand how behavioral factors related to study habits affect students' performance. An instructor's ability to better understand students' approaches to flipped CS courses lays a foundation for designing more informed and effective instruction. The analysis presented in the paper makes the following three contributions: firstly, it finds no significant differences in students' behavioral engagement based on gender, GPA or prior programming experience; secondly, it confirms differences in students' exam averages based on PPE and GPA; thirdly, it finds that quiz submission time is significantly correlated with students performance, indicating that students who submitted quizzes earlier had higher exam averages. These findings suggest that while behavior does impact performance, further research is required in order to understand the potential correlational and causal factors that influence students' behavior.

For future work, a combination of gender, PPE and GPA can be used to model and predict behavior as they might not have individual differences but may have interaction effect to understand variations in student behavior. This can be accompanied by a qualitative analysis that explores how students decide when to watch videos and complete quizzes. Another possibility is to study other types on behaviors described by Loras et al. [25], their group differences and their impact on learning to identify most influential behavioral factors that affect learning outcomes.

REFERENCES

[1] Dan Ariely and Klaus Wertenbroch. 2002. Procrastination, deadlines, and performance: Self-control by precommitment. *Psychological science* 13, 3 (2002), 219–224.

[2] Jens Bennedsen and Michael E. Caspersen. 2007. Failure Rates in Introductory Programming. *SIGCSE Bull.* 39, 2 (June 2007), 32–36. https://doi.org/10.1145/1272848.1272879

[3] Jacob Bishop and Matthew A Verleger. 2013. The Flipped Classroom: A Survey of the Research. In *2013 ASEE Annual Conference & Exposition*. ASEE Conferences, Atlanta, Georgia, 1–18. https://peer.asee.org/22585

[4] Juris Borzovs, Laila Niedrite, and Darja Solodovnikova. 2015. Factor Affecting Attrition among First Year Computer Science Students: the Case of University of Latvia, In Proceedings of the International Scientific and Practical Conference. *Environment. Technology. Resources* 3. https://doi.org/10.17770/etr2015vol3.174

[5] Ricardo Caceffo, Guilherme Gama, and Rodolfo Azevedo. 2018. Exploring Active Learning Approaches to Computer Science Classes. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education* (Baltimore, Maryland, USA) *(SIGCSE '18)*. Association for Computing Machinery, New York, NY, USA, 922–927. https://doi.org/10.1145/3159450.3159585

[6] Adam S. Carter, Christopher D. Hundhausen, and Olusola Adesope. 2015. The Normalized Programming State Model: Predicting Student Performance in Computing Courses Based on Programming Behavior. In *Proceedings of the Eleventh Annual International Conference on International Computing Education Research* (Omaha, Nebraska, USA) *(ICER '15)*. Association for Computing Machinery, New York, NY, USA, 141–150. https://doi.org/10.1145/2787622.2787710

[7] Kevin Casey and David Azcona. 2017. Utilizing student activity patterns to predict performance. *International Journal of Educational Technology in Higher Education* 14, 1 (2017), 1–15.

[8] Suzanne L. Dazo, Nicholas R. Stepanek, Robert Fulkerson, and Brian Dorn. 2016. An Empirical Analysis of Video Viewing Behaviors in Flipped CS1 Courses. In *Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education* (Arequipa, Peru) *(ITiCSE '16)*. Association for Computing Machinery, New York, NY, USA, 106–111. https://doi.org/10.1145/2899415.2899468

[9] Louis Deslauriers, Logan S. McCarty, Kelly Miller, Kristina Callaghan, and Greg Kestin. 2019. Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences* 116, 39 (2019), 19251–19257. https://doi.org/10.1073/pnas.1821936116 arXiv:https://www.pnas.org/content/116/39/19251.full.pdf

[10] Stephen H. Edwards, Jason Snyder, Manuel A. Pérez-Quiñones, Anthony Allevato, Dongkwan Kim, and Betsy Tretola. 2009. Comparing Effective and Ineffective Behaviors of Student Programmers. In *Proceedings of the Fifth International Workshop on Computing Education Research Workshop* (Berkeley, CA, USA) *(ICER '09)*. Association for Computing Machinery, New York, NY, USA, 3–14. https://doi.org/10.1145/1584322.1584325

[11] Scott Freeman, Sarah L. Eddy, Miles McDonough, Michelle K. Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth. 2014. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences* 111, 23 (2014), 8410–8415. https://doi.org/10.1073/pnas.1319030111 arXiv:https://www.pnas.org/content/111/23/8410.full.pdf

[12] Michail N. Giannakos, John Krogstie, and Nikos Chrisochoides. 2014. Reviewing the Flipped Classroom Research: Reflections for Computer Science Education. In *Proceedings of the Computer Science Education Research Conference* (Berlin, Germany) *(CSERC '14)*. Association for Computing Machinery, New York, NY, USA, 23–29. https://doi.org/10.1145/2691352.2691354

[13] David Gross, Evava Pietri, Gordon Anderson, Karin Moyano Camihort, and Mark Graham. 2015. Increased Preclass Preparation Underlies Student Outcome

Improvement in the Flipped Classroom. *CBE life sciences education* 14 (12 2015). https://doi.org/10.1187/cbe.15-02-0040

[14] Leslie Harvey III and Ashish Aggarwal. 2021. Exploring the Effect of Quiz and Homework Submission Times on Students' Performance in an Introductory Programming Course in a Flipped Classroom Environment. In *2021 ASEE Virtual Annual Conference Content Access*. ASEE Conferences, Virtual Conference. https://peer.asee.org/37149

[15] Arto Hellas, Petri Ihantola, Andrew Petersen, Vangel V Ajanovski, Mirela Gutica, Timo Hynninen, Antti Knutas, Juho Leinonen, Chris Messom, and Soohyun Nam Liao. 2018. Predicting academic performance: a systematic literature review. In *Proceedings companion of the 23rd annual ACM conference on innovation and technology in computer science education*. Association for Computing Machinery, New York, NY, USA, 175–199.

[16] Antti Herala, Erno Vanhala, Antti Knutas, and Jouni Ikonen. 2015. Teaching Programming with Flipped Classroom Method: A Study from Two Programming Courses. In *Proceedings of the 15th Koli Calling Conference on Computing Education Research* (Koli, Finland) *(Koli Calling '15)*. Association for Computing Machinery, New York, NY, USA, 165–166. https://doi.org/10.1145/2828959.2828983

[17] Diane Horton, Michelle Craig, Jennifer Campbell, Paul Gries, and Daniel Zingaro. 2014. Comparing Outcomes in Inverted and Traditional CS1. In *Proceedings of the 2014 Conference on Innovation & Technology in Computer Science Education* (Uppsala, Sweden) *(ITiCSE '14)*. Association for Computing Machinery, New York, NY, USA, 261–266. https://doi.org/10.1145/2591708.2591752

[18] Petri Ihantola, Arto Vihavainen, Alireza Ahadi, Matthew Butler, Jürgen Börstler, Stephen H. Edwards, Essi Isohanni, Ari Korhonen, Andrew Petersen, Kelly Rivers, Miguel Ángel Rubio, Judy Sheard, Bronius Skupas, Jaime Spacco, Claudia Szabo, and Daniel Toll. 2015. Educational Data Mining and Learning Analytics in Programming: Literature Review and Case Studies. In *Proceedings of the 2015 ITiCSE on Working Group Reports* (Vilnius, Lithuania) *(ITICSE-WGR '15)*. Association for Computing Machinery, New York, NY, USA, 41–63. https://doi.org/10.1145/2858796.2858798

[19] Aliye Karabulut-Ilgu, Nadia Jaramillo Cherrez, and Charles T. Jahren. 2018. A systematic review of research on the flipped learning method in engineering education. *British Journal of Educational Technology* 49, 3 (2018), 398–411. https://doi.org/10.1111/bjet.12548

[20] Hassan Khosravi and Kendra M.L. Cooper. 2017. Using Learning Analytics to Investigate Patterns of Performance and Engagement in Large Classes. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education* (Seattle, Washington, USA) *(SIGCSE '17)*. Association for Computing Machinery, New York, NY, USA, 309–314. https://doi.org/10.1145/3017680.3017711

[21] Päivi Kinnunen and Lauri Malmi. 2006. Why Students Drop out CS1 Course?. In *Proceedings of the Second International Workshop on Computing Education Research* (Canterbury, United Kingdom) *(ICER '06)*. Association for Computing Machinery, New York, NY, USA, 97–108. https://doi.org/10.1145/1151588.1151604

[22] Celine Latulipe, Audrey Rorrer, and Bruce Long. 2018. Longitudinal Data on Flipped Class Effects on Performance in CS1 and Retention after CS1. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education* (Baltimore, Maryland, USA) *(SIGCSE '18)*. Association for Computing Machinery, New York, NY, USA, 411–416. https://doi.org/10.1145/3159450.3159518

[23] Soohyun Nam Liao, Kartik Shah, William G. Griswold, and Leo Porter. 2021. A Quantitative Analysis of Study Habits Among Lower- and Higher-Performing Students in CS1. In *Proceedings of the 26th ACM Conference on Innovation and Technology in Computer Science Education V. 1* (Virtual Event, Germany) *(ITiCSE '21)*. Association for Computing Machinery, New York, NY, USA, 366–372. https://doi.org/10.1145/3430665.3456350

[24] Soohyun Nam Liao, Sander Valstar, Kevin Thai, Christine Alvarado, Daniel Zingaro, William G. Griswold, and Leo Porter. 2019. Behaviors of Higher and Lower Performing Students in CS1. In *Proceedings of the 2019 ACM Conference on Innovation and Technology in Computer Science Education* (Aberdeen, Scotland Uk) *(ITiCSE '19)*. Association for Computing Machinery, New York, NY, USA, 196–202. https://doi.org/10.1145/3304221.3319740

[25] Madeleine Lorås, Guttorm Sindre, Hallvard Trætteberg, and Trond Aalberg. 2021. Study Behavior in Computing Education—A Systematic Literature Review. *ACM Trans. Comput. Educ.* 22, 1, Article 9 (oct 2021), 40 pages. https://doi.org/10.1145/3469129

[26] Charles Mcdowell, Linda Werner, Heather Bullock, and Julian Fernald. 2006. Pair programming improves student retention, confidence, and program quality. *Commun. ACM* 49 (08 2006), 90–95. https://doi.org/10.1145/1145293

[27] Keir Mierle, Kevin Laven, Sam Roweis, and Greg Wilson. 2005. Mining Student CVS Repositories for Performance Indicators. *SIGSOFT Softw. Eng. Notes* 30, 4 (May 2005), 1–5. https://doi.org/10.1145/1082983.1083150

[28] Colin Moore, Lina Battestilli, and Ignacio X. Domínguez. 2021. Finding Video-Watching Behavior Patterns in a Flipped CS1 Course. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education* (Virtual Event, USA) *(SIGCSE '21)*. Association for Computing Machinery, New York, NY, USA, 768–774. https://doi.org/10.1145/3408877.3432359

[29] Cindy Norris. 2016. An Examination of Layers of Quizzing in Two Computer Systems Courses. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education* (Memphis, Tennessee, USA) *(SIGCSE '16)*. Association for Computing Machinery, New York, NY, USA, 48–53. https://doi.org/10.1145/2839509.2844608

[30] Ted O'Donoghue and Matthew Rabin. 1999. Doing it now or later. *American economic review* 89, 1 (1999), 103–124.

[31] Keith Quille and Susan Bergin. 2019. CS1: how will they do? How can we help? A decade of research and practice. *Computer Science Education* 29, 2-3 (2019), 254–282.

[32] Sunil Sabnis, Renzhe Yu, and René F. Kizilcec. 2022. Large-Scale Student Data Reveal Sociodemographic Gaps in Procrastination Behavior. In *Proceedings of the Ninth ACM Conference on Learning @ Scale* (New York City, NY, USA) *(L@S '22)*. Association for Computing Machinery, New York, NY, USA, 133–141. https://doi.org/10.1145/3491140.3528285

[33] Merilin Säde, Reelika Suviste, Piret Luik, Eno Tõnisson, and Marina Lepp. 2019. Factors That Influence Students' Motivation and Perception of Studying Computer Science. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education* (Minneapolis, MN, USA) *(SIGCSE '19)*. Association for Computing Machinery, New York, NY, USA, 873–878. https://doi.org/10.1145/3287324.3287395

[34] Clifford A. Shaffer and Stephen H. Edwards. 2011. Scheduling and Student Performance. In *Proceedings of the 16th Annual Joint Conference on Innovation and Technology in Computer Science Education* (Darmstadt, Germany) *(ITiCSE '11)*. Association for Computing Machinery, New York, NY, USA, 331. https://doi.org/10.1145/1999747.1999842

[35] Jaime Spacco, Paul Denny, Brad Richards, David Babcock, David Hovemeyer, James Moscola, and Robert Duvall. 2015. Analyzing Student Work Patterns Using Programming Exercise Data. In *Proceedings of the 46th ACM Technical Symposium on Computer Science Education* (Kansas City, Missouri, USA) *(SIGCSE '15)*. Association for Computing Machinery, New York, NY, USA, 18–23. https://doi.org/10.1145/2676723.2677297

[36] Yaacov Trope and Ayelet Fishbach. 2000. Counteractive self-control in overcoming temptation. *Journal of personality and social psychology* 79, 4 (2000), 493.

[37] Christopher Watson and Frederick W.B. Li. 2014. Failure Rates in Introductory Programming Revisited. In *Proceedings of the 2014 Conference on Innovation & Technology in Computer Science Education* (Uppsala, Sweden) *(ITiCSE '14)*. Association for Computing Machinery, New York, NY, USA, 39–44. https://doi.org/10.1145/2591708.2591749

[38] Chris Wilcox and Albert Lionelle. 2018. Quantifying the Benefits of Prior Programming Experience in an Introductory Computer Science Course. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education* (Baltimore, Maryland, USA) *(SIGCSE '18)*. Association for Computing Machinery, New York, NY, USA, 80–85. https://doi.org/10.1145/3159450.3159480

[39] Salla Willman, Rolf Lindén, Erkki Kaila, Teemu Rajala, Mikko-Jussi Laakso, and Tapio Salakoski. 2015. On study habits on an introductory course on programming. *Computer Science Education* 25, 3 (2015), 276–291. https://doi.org/10.1080/08993408.2015.1073829

[40] Brenda Cantwell Wilson and Sharon Shrock. 2001. Contributing to Success in an Introductory Computer Science Course: A Study of Twelve Factors. *SIGCSE Bull.* 33, 1 (Feb. 2001), 184–188. https://doi.org/10.1145/366413.364581