Using Clickstream Data Mining Techniques to Understand and Support First-Generation College Students in an Online Chemistry Course

Fernando Rodriguez
School of Education
University of California, Irvine
fernanr1@uci.edu

Hye Rin Lee
School of Education
University of California, Irvine
hyerl@uci.edu

Teomara Rutherford
School of Education
University of Delaware
teomara@udel.edu

Christian Fischer
Hector Research Institute of
Education Sciences and Psychology
University of Tübingen
christian.fischer@uni-tuebingen.de

Eric Potma
Department of Chemistry
University of California, Irvine
epotma@uci.edu

Mark Warschauer
School of Education
University of California, Irvine
markw@uci.edu

ABSTRACT

Although online courses can provide students with a high-quality and flexible learning experience, one of the caveats is that they require high levels of self-regulation. This added hurdle may have negative consequences for first-generation college students. In order to better understand and support students' self-regulated learning, we examined a fully online Chemistry course with high enrollment (N = 312) and a high percentage of first-generation college students (65.70%). Using students' lecture video clickstream data, we created two indicators of self-regulated learning: lecture video completion and time management. Performing a k-means clustering on these indicators uncovered four distinct self-regulated learning patterns: (1) Early Planning, (2) Planning, (3) Procrastination, and (4) Low Engagement. Early Planning behaviors were especially important for course success—they consistently predicted higher final course grades, even after controlling for important demographic variables. Interestingly, first-generation college students classified as Early Planners achieved at similar levels as their non-first-generation peers, but first-generation students in the Low Engagement group had the lowest average grades among students. Overall, our results show that self-regulation may be an important skill for determining first-generation students' STEM achievement, and targeting these skills may serve as a useful way to support their specific learning needs.

CCS CONCEPTS

• **Applied computing → Learning management systems**; **E-learning**.

LAK21, April 12–16, 2021, Irvine, CA, USA
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8935-8/21/04.
https://doi.org/10.1145/3448139.3448169

KEYWORDS

Clickstream Data Mining, STEM, Online Learning, Self-Regulation, College Students, Underrepresented Students

ACM Reference Format:
Fernando Rodriguez, Hye Rin Lee, Teomara Rutherford, Christian Fischer, Eric Potma, and Mark Warschauer. 2021. Using Clickstream Data Mining Techniques to Understand and Support First-Generation College Students in an Online Chemistry Course. In *LAK21: 11th International Learning Analytics and Knowledge Conference (LAK21), April 12–16, 2021, Irvine, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3448139.3448169

1 INTRODUCTION

Online course offerings have been steadily increasing at two- and four-year colleges, as they help address issues with over-enrollment, improve students' degree completion, and reduce institutional costs [1, 46]. Now, in the middle of the COVID-19 pandemic, they have been the primary way students take college courses. From an instructional standpoint, research has found benefits to student learning and engagement from efforts to improve online content delivery (e.g., short and well-produced lecture videos)[13], personal interactions [21], and other course design features (e.g., splitting up assignments into manageable tasks)[6]. However, the main limitation of college-level online courses is that students have worse learning outcomes when compared to traditional face-to-face courses [45].

An added concern about online learning is that these performance gaps can be be worse for students who are underrepresented in higher education (African American, Latino(a), Native American, students from first-generation and/or low-income backgrounds) [19, 20, 45]. These gaps may be more pronounced when thinking about online STEM (Science, Technology, Engineering, Math) courses which have high learning demands, such as understanding complex diagrams, remembering numerous declarative and procedural information, understanding mathematical notation, and using equations to solve complex problems. There is a strong body of work that documents the various challenges underrepresented students face in college STEM courses [11, 17, 18, 22], but in the context of online learning, an added challenge involves the very nature how how learning takes place in this setting.

For asynchronous online courses, students must figure out when they plan to login to the course. They must watch course videos and study materials without any immediate instructor feedback. The lack of any physical presence of instructors and students also means that students must actively seek out instructors and peers if they have questions or need help. Lastly, students must also track assignments needing to be completed and stay on top of important deadlines [46]. Because students who are underrepresented in higher education are more likely to come from historically undeserved communities that lack strong higher education social capital (networks of people who understand the ins and outs of college) or financial resources [3, 38], they may be further burdened by having to adapt to these learning demands without adequate support or guidance.

One potential solution for addressing the additional burdens underrepresented students face in online STEM courses is to focus on understanding how students regulate their independent learning time. Self-Regulated Learning (SRL) is a theoretical framework for understanding the different strategies individuals utilize when engaged in self-directed learning behaviors [34]. These include internal and external motivations towards learning, goal setting, planning, monitoring, and reflection [37, 47]. SRL occurs when organizing the overall learning process, as well when and when engaging in actual learning [44]. Student SRL has long been related to learning outcomes, such as achievement and persistence, including in Chemistry courses [25, 39]. In the context of asynchronous online learning, planning and monitoring are particularly important indicators of learning and achievement [10].

SRL may be especially important for first-generation students. Naumann and colleagues [28] found that for first-generation college students, SRL was more predictive of college GPA than were standardized test scores. Although there is some evidence that first-generation students may have difficulty with SRL in college [41], first-generation students, like non-first-generation students, are not a monolithic group with respect to SRL [2]. Person-centered studies of SRL have found that learners exhibit different profiles with respect to different aspects of SRL dispositions and behaviors, and these profiles have differential relations with performance [8].

One promising method for understanding SRL in online learning is to observe students' behaviors within a course's Learning Management System (LMS) [5, 26, 37]. Some LMS services allow researchers to obtain student clickstream data, which are time-stamped logs that record every single click event in the course, such as the specific page or resource a student visited. A large online course of about 300 students can produce roughly 380,000 individual data points [5]. Because of the size and richness of the data, researchers have been able to take advantage of data mining techniques to develop sophisticated models of college student learning [16, 23]. These include the ability to capture the moment-by-moment fluctuations in click activity that occur throughout a course; using sophisticated data mining techniques, researchers can identify students who have sudden drops in course activity relative to the class as whole [31].

One benefit of using clickstream data to study SRL is that it can provide an objective window into understanding students' learning process [42]. Although the relationship between self-reported measures of SRL and academic outcomes is well-established [29], it is not always clear how these student self-reports correspond to learning behaviors over the length of a course. Some work shows that students tend to overestimate their SRL processes, whereby students who reported high levels SRL had lower learning outcomes than did students who underestimated their SRL skills [43]. Some recent work has also revealed that students' self-reported cramming and spacing behaviors, although correlated with grades, did not correspond to their LMS course activity as one would expect; students who stated they spaced their studying over the length of a course showed slightly higher levels of cramming behaviors than students who stated they consistently crammed [36].

There have already been successful efforts to measure SRL using clickstream data; these methods have provided valuable insights into students learning patterns [4, 24, 26, 30, 40]. For instance, informed by SRL theories, Park and colleagues [32] utilized data mining techniques in an online STEM course to detect the extent to which students utilized effective time management skills or procrastinated (e.g., waited until the last minute to complete assignments), and these behavioral patterns were strongly predictive of course success. In another study, Cicchinelli and colleagues used clustering methods to classify the extent to which students utilized different planning, monitoring, and regulating activities [14]. This clustering method successfully classified students as either continuously active, focused on testing and quizzing, procrastinating, or inactive. They found that students who were continuously active received higher quiz and exam scores than students in the other categories.

One of the trade-offs of using clickstream data to examine SRL is that data mining approaches can be somewhat difficult for instructors or other college stakeholders to implement. That is, although this work holds value in helping researchers advancing theories and methods, deriving practical insights that can benefit instructors is difficult, both in terms of clearly illustrating how students regulate their learning and in understanding how to best support students. Although sophisticated data mining models can help us better understand how underrepresented students learn in college STEM courses, in a hypothetical example, we can envision future scenarios in which work using data mining techniques may only go so far as to highlight performance discrepancies without careful consideration into understanding the nuances that exist within these groups.

To help address these issues, we focused on an large 10-week online Chemistry course with a high proportion of first-generation college students. The goal of our work was to use clickstream data to gain important insights about SRL and learning outcomes in online STEM courses. We also sought to take advantage of data mining approaches to uncover important patterns of learning that can help explain differences in student achievement. Additionally, we paid careful attention to examining the nuances within first-generation students' click behaviors, as doing so could provide provide a more complete picture of their learning process. Guided by the motivation to conduct work that could be readily understood by instructors and practitioners, we also focused our efforts on developing practical indicators of SRL, especially those that captured planning and monitoring behaviors. In the context of this course, this involved developing SRL measures from the prerecorded lecture video clicks and observing how many assigned lecture videos students visited (assignment completion), as well as the timing to which they visited the assignments (time management). The course we examined in this paper has been previously studied [24], but the data from the previous study used only a subset of students in the course, and the focus of the study did not examine first-generation students.

The research questions (RQs) for our study are as follows:

RQ1 Does clustering clickstream measures of SRL reveal meaningful learning patterns?
RQ2 Are the cluster groups associated with learning outcomes?
RQ3 Do first-generation and non-first-generation students have different learning outcomes in the course, and how do the cluster groups interact with first-generation status?
RQ4 Do the SRL cluster groups predict learning after accounting for demographics and prior achievement, and how does first-generation status change the prediction between the SRL cluster groups and learning outcomes?

2 STUDY SETTING

We studied an asynchronous, fully online Preparation for General Chemistry course, which was taught at large and highly-selective public research university in the United States. The course was intended for incoming freshmen that need to take the General Chemistry series for their major, but who had not met the prerequisites for entering the first course in the series. The goal of this course was to equip students with basic knowledge of Chemistry, mathematical skills, and problem solving techniques. No prerequisites were in place for enrolling in this course. This online course was taught prior to the COVID-19 pandemic.

The course consisted of 48 learning objectives. Each learning objective was covered in a lecture video, each of which consisted of a 5-10 minute lecture video, a summary of important concepts, interactive practice questions, and for-grade homework questions. The 48 lecture video segments were organized into four different themes, called modules, for a total of four modules. Module 1 covered basic mathematical techniques for solving Chemistry problems (nine lecture videos), Module 2 introduced the structure of atoms and molecules (13 lecture videos), Module 3 provided an introduction to reaction stoichiometry (14 lecture videos), and Module 4 covered several classes of chemical reactions in water (12 lecture videos). Students navigated the course by completing lecture video one, and then moving to lecture video two, and so on. The lecture videos were recorded in a green room studio. The edited videos featured the course instructor, who introduced a given topic and discussed several examples. The background behind the instructor was used to show visual representations, animations, worked-out examples, and summarizing statements. Students were given two weeks to complete each module. Due to the course schedule, students had three additional days to complete Module 1 and four additional days to complete Module 3. Although each module had a due date, students were allowed to complete them past the due date.

3 PARTICIPANTS

We studied 312 students enrolled in the course. The total enrollment was 319, but we excluded seven students who had incomplete institutional data from which we obtained their demographic information, indicators of prior achievement, and/or clickstream data. At the beginning of the course, the instructor provided students an information sheet describing the general aims of the study as well as the different data sources that we collected (background data, clickstream data, gradebook data). None of the students in the class declined to participate in the study. Students had an average age of 18.4 years (*SD* = 0.45) and 77% were women. A large percentage of students were first-generation college students (65.70%), which indicated that they were the first in their immediate family to attend college. Over half of the students also came from low-income households (55.44%). The students were ethnically diverse, with the largest groups consisting of Latino(a) (50.00%), Asian/Pacific Islander (31.08%), White (11.53%), and Black/African American (5.12%) students. Overall, 55.12% of students were considered URM (Underrepresented Minority), which were Latino(a) and Black/African American students in our sample.

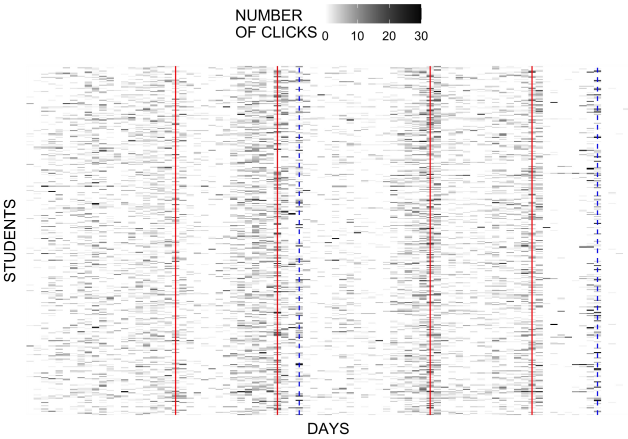


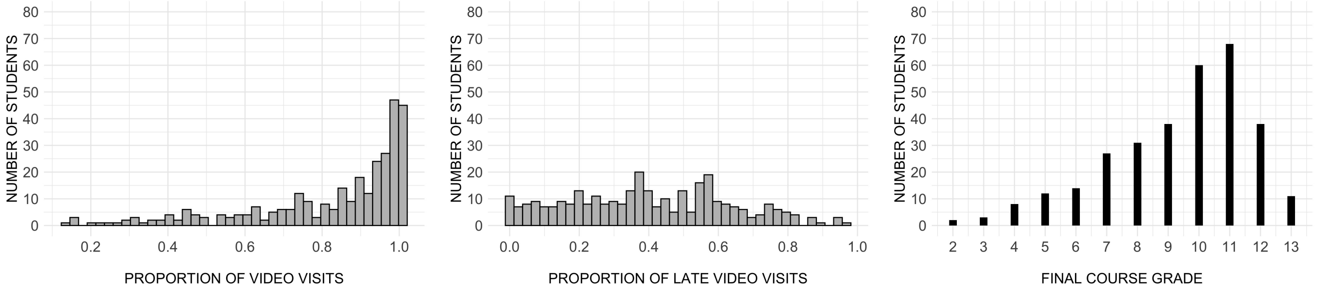
Figure 1: Student Lecture Video Clicks by Course Days (*N* = 312). Red lines represent module due dates (from left to right, Module 1 appears first and Module 4 appears last). Dashed blue lines represent exam days.

There was very high overlap between URM and first-generation status: URMs made up 83% of students who were first-generation college students. There was also notable overlap between low-income students and first-generation students (66.47%). Because of these overlaps, our study focused on first-generation college students as underrepresented learners, because they made up a large majority of the student body while also representing a large proportion of URM and low-income students. Thus, we use the terms *First-Generation* and *non-First-Generation* in our analyses and subsequent discussion.

4 DATA

4.1 Clickstream data

We obtained students' clickstream data from the course's LMS. When thinking about how to best utilize students' clickstream data, we considered *where* exactly SRL behaviors occurred and could



(a) Video Visits **(b) Late Video Visits** **(c) Final Course Grade**

Figure 2: Histograms of Self-Regulated Learning Measures and Final Course Grade (*N* = 312)

therefore be observed. In the context of this course, these were the pre-recorded lecture videos, as they contained most of the instructional content for this course. Additionally, students were required to watch the videos in order to complete corresponding assignments. For each of the four modules, we extracted the number of times per day a student clicked on the assigned lecture videos for a given module. As one can observe in Figure 1, students had an consistent distribution of lecture video clicks for Module 1, with a small uptick in activity as the deadline for completing the lecture videos approached. However, for Modules 2 through 4, we observed larger increases in students' lecture video visits in the days leading up to the due date. It is also important to point out that students had a large number of visits on the day past the module due date. Inspecting this data by time revealed these clicks occurred between one to three hours after the 11:59am deadline. The total number of lecture video clicks across all days averaged 77.50 (*SD* = 34.54, *Min* = 8, *Max* = 238).

4.2 Clickstream measures of SRL

Drawing on prior work examining SRL in clickstream data [14, 24, 26, 32], especially the work of Li and colleagues[24], we created two SRL indicators. The first examined studying on time, which we quantified as the extent to which students visited all of the assigned lecture videos before the Module deadlines. We label this indicator, *Proportion of Video Visits*. For every module, we calculated the proportion of the assigned lecture videos students visited by the module due date. For student i, the proportion of lecture video visits was calculated by $p' = x/n$, where x counts how many assigned lecture videos a student visited at least once and n is the total number of lecture videos that were assigned to a given module. Thus, a value of $p' = 1$ indicates a student visited all of the assigned lecture videos at least once, whereas a value of $p' = .5$ indicates a student visited only half of the assigned lecture videos.

Because we had four different modules, we first created a matrix $Y_{N \times M}$, where N is the number of students and M is the total number of modules. Thus, y_{im}, represents the proportion of video visit of student i on Module m, where $1 \leq i \leq N$ and $1 \leq m \leq M$. After creating this matrix, we calculated the weighted average of students' proportions across Modules 1-4. We used a weighted average to account for the fact that each module contained a different number of lecture videos. This weighted average served as our *Proportion of Video Visits* score (Figure 2a).

Our second indicator of SRL examined time management; we quantified this as the extent to which students delayed visiting the assigned lecture videos as the module due date neared. We label this indicator, *Proportion of Late Video Visits*. For student i, the proportion of late video visits was calculated by $p' = x/n$, where x is the number of lecture video clicks that occurred on the module due date and n is the total number of lecture video clicks for a given module from the time the module opened to the module due date. Note that the total number of lecture video clicks on the module date also included the clicks that occurred the day after. This was done because students were allowed to complete the modules past the due date, with many students completing the modules one to three hours past the 11:59pm deadline. A value of $p' = 0$ indicates a student did not visit any of the assigned lecture videos on the module due date, whereas a value of $p' = 1$ indicates that a student visited all of the lecture videos on the module due date. Similar to the Proportion of Video Visits, we calculated the weighted average of the proportion scores to create a single *Proportion of Late Video Visits* score (Figure 2b).

4.3 Student background data

We obtained student's background data from the university registrar. This included both demographic indicators and prior achievement. For the demographic indicators, this included student's gender (man, woman), first-generation status (non-first-generation college student, first-generation college student), low-income status (student did not reside in a low income household, student resided in a low income household), and URM (non-underrepresented minority, underrepresented minority). Student's Scholastic Aptitude Test (SAT) served as our indicator of prior achievement (*M* = 1613.33, *SD* = 132.94, *Min* = 1260, *Max* = 2040).

4.4 Learning outcome data

We used students' final course grade as our learning outcome (1 = F, 2 = D-, 3 = D, 4 = D+, 5 = C-, 6 = C, 7 = C+, 8 = B-, 9 = B, 10 = B+, 11 = A-, 12 = A, 13 = A+) (Figure 2c). Students in the course received an average final grade of 9.34, which corresponds to slightly above a B (*SD* = 2.32).

5 ANALYTICAL METHODS

For *RQ1*, we conducted a *k*-means clustering algorithm using our two clickstream measures of SRL, *Proportion of Video Visits* and *Proportion of Late Video Visits*. We chose a grouping of $K = 4$ with 24 initial random centroids. Both scores were scaled prior to running the *k*-means clustering algorithm. 76.2% of the total variance was explained by this clustering. We used two common methods to find the optimal number of K clusters: (1) The Elbow Method, and (2) The Bayesian Inference Criterion (BIC) for *k*-means [15]. The Elbow Method examined the percentage of variance explained based on the number of clusters, and we observed the point at which including an additional cluster K explained only a small amount of the variance. Our data showed this occurred between clusters $K = 3$ and $K = 4$. The BIC criterion generated 3 choice models, two of which $K = 4$ (VVI, 4 = -1428.62, VVE, 4 = -1433.36, VVI, 5 = -1439.32). We determined $K = 4$ to be the optimal number of clusters. Figure 3 shows the scatterplot between our two measures of SRL by the $K = 4$ cluster groups.

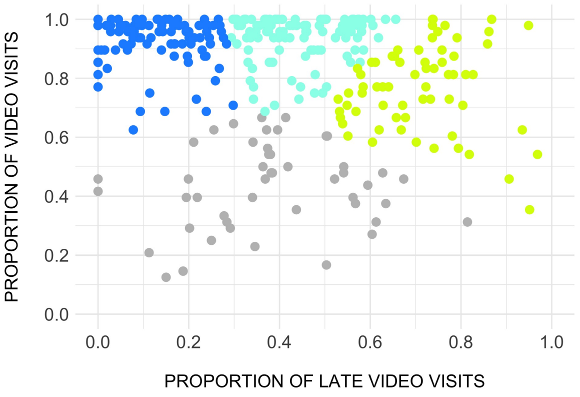


Figure 3: Scatterplot by Cluster $K = 4$ ($N = 312$)

To answer *RQ2* and *RQ3*, we conducted a 4 x 2 ANOVA with interaction terms to understand differences between the $K = 4$ cluster groupings and first-generation status (first-generation student vs. non-first-generation student) on final course grade. When checking the ANOVA assumptions, Levene's test revealed homogeneity of variance was violated. We therefore checked the ANOVA results as well as the results for the Tukey's HSD posthoc against the Welch's ANOVA test and Games-Howell for the post-hoc comparisons. Welch's ANOVA revealed the same results as the ANOVA test. This was also the case when we compared Tukey's HSD to the Games-Howell test. Because we did not find any discrepancies, we therefore report the ANOVA and Tukey's HSD results in the in the Results section.

For *RQ4*, we conducted a stepwise regression. This involved first predicting final course grade from the cluster grouping to understand whether the cluster groupings uniquely predicted final course grade before controlling for important covariates. We then added each of our covariates into a new model while also including covariates from the previous model. These covariates included demographic variables (Gender, First-Generation status, Low Income status, whether URM), SAT scores, and total video views as covariates to each subsequent model. Our models are as follows (the "..." indicates the model includes the coefficients from the previous model):

$$FinalGrade_i = B_0 + B_1(Cluster2)_i + \\ B_2(Cluster3)_i + B_3(Cluster4)_i + \epsilon_i \tag{1}$$

$$FinalGrade_i = \ldots + B_4(Gender)_i + \epsilon_i \tag{2}$$

$$FinalGrade_i = \ldots + B_5(FirstGeneration)_i + \epsilon_i \tag{3}$$

$$FinalGrade_i = \ldots + B_6(LowIncome)_i + \epsilon_i \tag{4}$$

$$FinalGrade_i = \ldots + B_7(URM)_i + \epsilon_i \tag{5}$$

$$FinalGrade_i = \ldots + B_8(SATTotalScore)_i + \epsilon_i \tag{6}$$

$$FinalGrade_i = \ldots + B_9(TotalVideoClicks)_i + \epsilon_i \tag{7}$$

In addition to building final predictive model (Model 7), we also wanted to understand how this model varied for First-Generation and non-First-Generation students, as this could reveal differences in how the cluster groupings predicted final course grade. To accomplish this, we ran two separate final regression models, stratified by First-Generation and non-First-Generation students.

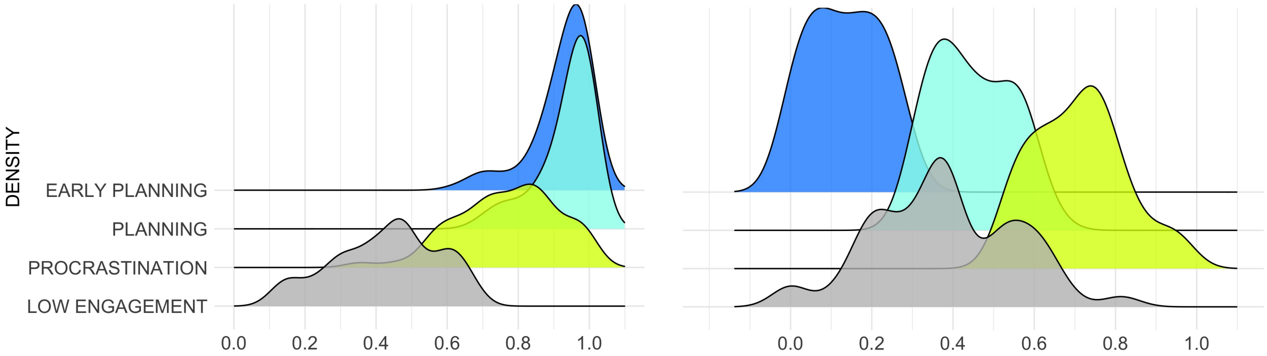
6 RESULTS

6.1 Identifying learning patterns from the SRL cluster groups (RQ1)

We first describe the distribution of our two clickstream measures of SRL by our $K = 4$ clusters. After examining the patterns of these distributions, we characterized the self-regulated behaviors as (1) *Early Planning*, (2) *Planning*, (3) *Procrastination*, and (4) *Low Engagement* (See Figure 4). Students in both the Early Planning and Planning groups tended to visit most of the assigned lecture videos as indicated by the density in the proportions of Lecture Video Visits. The difference between these groups was best captured when examining density in the proportions of their Late Video Visits. Students in the Early Planning group had a very low proportion of viewing the videos on the module due date. This proportion was higher for those in the Planning group. In contrast, students classified into the Procrastination group had lower proportion of Lecture Video Visits but a noticeably higher proportion of Late Video Visits, implying that many of these students waited until the due date to study for the course and did not get to viewing all of the videos. Students in the Low Engagement group had the lowest proportion for the Lecture Video Visits, whereas their proportion of Late Video Visits was more variable.

6.2 Comparing SRL cluster groups on student performance (RQ2)

We found a main effect of the clustering groups on final course grade, $F(3, 304) = 26.09$, $p < .001$, $\eta_p^2 = .18$ (See Table 1 and Figure 5). When comparing the Early Planners to the other three groups, post-hoc tests revealed that Early Planners had statistically significantly higher final course grades than students in the Low Engagement ($p_{\text{adjusted}} < .001$) and Procrastination groups. ($p_{\text{adjusted}} < .001$). There was no statistically significant difference between the Early Planning and Planning groups ($p_{\text{adjusted}} = .09$). The Planning group also



(a) Video Visits

(b) Late Video Visits

Figure 4: Density Plots of the SRL Clickstream Measures by Cluster Groups (N = 312)

had higher final course grades than students in the Procrastination ($p_{\text{adjusted}} < .01$) and Low Engagement groups ($p_{\text{adjusted}} < .001$). Finally, the Procrastination group had higher final course grades than the Low Engagement group ($p_{\text{adjusted}} < .001$).

Table 1: Final Course Grade Descriptive Statistics, Overall and by First-Generation Status

	Overall (N = 312)			
	n	*Percent*	*Mean*	*SD*
Early Planning	92	29.48	10.39	1.72
Planning	107	34.29	9.71	1.87
Procrastination	65	20.83	8.66	2.15
Low Engagement	48	15.38	7.41	2.95
	First-Generation (n = 205)			
Early Planning	55	26.82	10.12	1.86
Planning	68	33.17	9.52	1.79
Procrastination	48	23.41	8.47	2.26
Low Engagement	34	16.58	6.64	2.92
	non-First-Generation (n = 107)			
Early Planning	37	34.57	10.78	1.51
Planning	39	36.44	10.05	1.99
Procrastination	17	15.88	9.17	1.77
Low Engagement	14	13.08	9.28	2.12

6.3 Comparing first-generation status and cluster groups on student performance (RQ3)

There was a main effect of first-generation status on final course grade, $F(1, 304) = 13.52$, $p < .001$, $\eta_p^2 = .04$. First-Generation students received statistically significantly lower final course grades (M = 8.96, SD = 2.43) when compared to non-First-Generation students (M = 10.06, SD = 1.90). The interaction between the cluster groups and first-generation status on final course grade was also statistically significant, $F(3, 304) = 2.84$, $p < .05$, $\eta_p^2 = .03$. Post-hoc tests revealed that First-Generation students in the Low Engagement group received statistically significantly lower grades than non-First-Generation students in this same group ($p_{\text{adjusted}} < .001$). None of the other post-hoc comparisons between First-Generation students and non-First-Generation students in the same cluster group were statistically significant. When examining within-group differences for First-Generation students, post-hoc tests revealed students in the Low Engagement group had lower final course grades than students in the Procrastination, Planning, and Early Planning groups. Students in the Procrastination group also had lower final course grades than students in the Early Planning Group ($p_{\text{adjusted}} < .001$ for all post-hocs). Within non-First-Generation students, there were no statistically significant post-hoc comparisons between the SRL cluster groups and final course grade. (See Table 1 and Figure 6).

6.4 Predicting course performance and stratifying by first-generation status (RQ4)

As shown in Table 2, results for the stepwise regression revealed that, when examining the cluster groupings in isolation from other covariates (Model 1), they explained 19.3% of the variance and the model was statistically significant, $F(3, 308) = 24.65$, $p < .001$. Results from the stepwise regression showed R^2 increasing as we added each of our covariates, with the clustering groups remaining statistically significant predictors across all models. The final regression model (Model 7) explained 33.6% of the variance, and the model was a statistically significant predictor of final course grade, $F(9, 302) = 17.03$, $p < .001$.

Our cluster groups remained statistically significant predictors in the final model. We found that, when compared to Early Planning, the Low Engagement (B = -2.33, $p < .001$), Procrastination (B= -1.40, $p < .01$), and Planning (B = -0.70, $p < .05$) predicted lower final course grade. Other statistically significant predictors included SAT total score (B = 0.01, $p < 0.001$). We did not find that first-generation

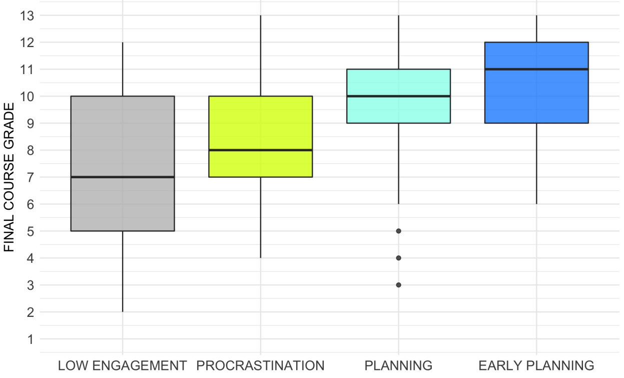


Figure 5: Final Course Grade by SRL Cluster Groups (N = 312)

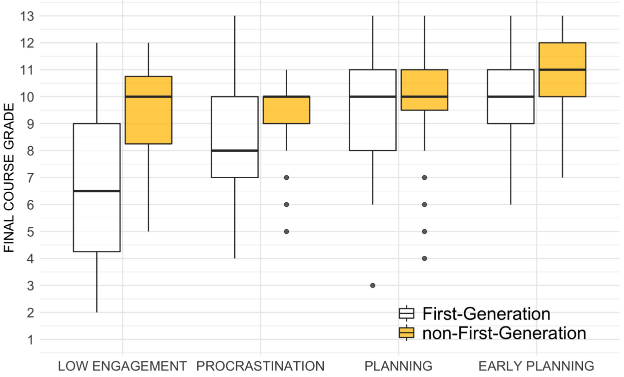


Figure 6: Final Course Grade by SRL Cluster Groups and First-Generation Status (N = 312)

status, low income status, URM status, or gender statistically significantly predicted final course grade when all of the covariates were taken into account in the final model. The final unstandardized predictive model was:

$$\widehat{FinalGrade} = 10.83 - 2.33(LowEngagement) - 1.40(Procrastination) - 0.70(Planning) - 0.37(Woman) - 0.42(FirstGeneration) + 0.17(LowIncome) - 0.22(URM) + 0.01(SATTotalScore) + 0.01(TotalVideoClicks)$$
$$N = 312,\ R^2 = 0.33$$

In order to understand how the models varied by first-generation status, we stratified our data by this indicator. We found that, for First-Generation students, Model 1 explained 24.1% of the variance and statistically significantly predicted final course grade, $F(3, 201) = 21.19$, $p < .001$. The final model (Model 7) explained 37.2% of the variance and statistically significantly predicted final course grade, $F(8, 196) = 14.54$, $p < .001$. For this final model, we found that, when compared to Early Planners, Low Engagement (B = -3.02, $p < .001$), Procrastination (B = -1.56, $p < .001$), and Planning (B = -0.75, $p < .01$) were negative predictors of final course grades. It is worth pointing out that, for First-Generation students, Low Engagement contributed to a notable negative prediction of 3.02 grade points when compared to Early Planners. This roughly translates to a full grade difference between both groups (e.g., C+ vs. B+). The only other statistically significant predictor was SAT Total Score. The final unstandardized predictive model for First-Generation students was:

$$\widehat{FinalGrade(FirstGen)} = 10.64 - 3.02(LowEngagement) - 1.56(Procrastination) - 0.75(Planning) - 0.65(Woman) + 0.22(LowIncome) - 0.29(URM) + 0.01(SATTotalScore) + 0.01(TotalVideoClicks)$$
$$n = 205,\ R^2 = 0.37$$

When examining non-First-Generation students, Model 1 explained 10.6% of the variance and statistically significantly predicted final course grade, $F(4, 102) = 3.04$, $p < .05$. When compared to Early Planning, Low Engagement negatively predicted final course grade (B = -1.49, $p < .01$). The final model (Model 7) was also a statistically significant predictor of final course grade, $F(8, 98) = 3.45$, $p < .001$, and explained 22.0% of the variance. However, the SRL cluster groups were not statistically significant predictors of final course grade in the final model. Only SAT total score was a statistically significant predictor of final course grade (B = 0.004, $p < .001$). The final undstandardized predictive model for non-First-Generation students was:

$$\widehat{FinalGrade(nonFirstGen)} = 9.71 - 0.79(LowEngagement) + 0.43(Procrastination) - 0.03(Planning) + 0.06(Woman) + 0.07(LowIncome) + 0.24(URM) + 0.01(SATTotalScore) + 0.01(TotalVideoClicks)$$
$$n = 107,\ R^2 = 0.22$$

7 DISCUSSION

As online college course offerings are ever increasing, there have been efforts to study how to best deliver content in an online setting. A prominent area of research has examined students' self-regulation skills in online courses [10, 12, 33]. Situated in the context of an online Chemistry course, the goal of this study was to uncover important relationships between SRL behaviors and learning outcomes. We were especially motivated to identify patterns of SRL behaviors that could inform how we understand and support First-Generation college students.

Our study focused on students' clickstream data obtained from the course's LMS, as these data could provide a window into how students managed their independent learning time. We turned to the SRL and clickstream literature to understand how to develop measures that reflected independent learning [24, 26, 32]. When constructing our measures of SRL, we were also mindful to ensure that these measures could be accessible to instructors and non-research audiences, both in terms of their real-life or face validity, and how we described and interpreted the results.

We constructed two key SRL measures from the clickstream data. The *Proportion of Video Visits* captured the extent to which students viewed all of the assigned lecture videos prior to the suggested course deadline and served as our indicator of assignment completion. The *Proportion of Late Video Visits* captured time management, especially from the perspective of completing or delaying the video assignments as the due dates approached. We employed a k-means

Table 2: Regression Results for Final Course Grade (*N* = 312)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Low Engagement	-2.974*** *-0.463* (0.218)	-3.008*** *-0.468* (0.370)	-2.909*** *-0.452* (0.365)	-2.909*** *-0.452* (0.366)	-2.895*** *-0.450* (0.364)	-2.724*** *-0.424* (0.345)	-2.338*** *-0.364* (0.399)
Procrastination	-1.729*** *-0.303* (0.339)	-1.751*** *-0.306* (0.337)	-1.628*** *-0.285* (0.333)	-1.628*** *-0.285* (0.334)	-1.667*** *-0.292* (0.333)	-1.607*** *-0.281* (0.314)	-1.401*** *-0.245* (0.331)
Planning	-0.671* *-0.137* (0.297)	-0.701* *-0.143* (0.296)	-0.665* *-0.136* (0.291)	-0.666* *-0.137* (0.292)	-0.625* *-0.128* (0.291)	-0.692* *-0.141* (0.275)	-0.707* *-0.144* (0.274)
Woman		-0.627* *-0.113* (0.281)	-0.551* *-0.099* (0.277)	-0.550* *-0.099* (0.277)	-0.520 *-0.094* (0.277)	-0.404 *-0.073* (0.262)	-0.375 *-0.067* (0.261)
First-Generation			-0.860*** *-0.176* (0.246)	-0.870** *-0.178* (0.280)	-0.676* *-0.138* (0.297)	-0.447 *-0.091* (0.283)	-0.429 *-0.087* (0.282)
Low Income				0.020 *0.004* (0.280)	0.058 *0.012* (0.265)	0.198 *0.042* (0.252)	0.170 *0.036* (0.251)
URM					-0.496 *-0.106* (0.257)	-0.238 *-0.051* (0.246)	-0.226 *-0.048* (0.245)
SAT Total Score						0.005*** *0.308* (0.001)	0.005*** *0.308* (0.001)
Total Video Clicks							0.007 *0.110* (0.004)
Constant	10.391*** (0.218)	10.896*** (0.313)	11.348*** (0.334)	11.344*** (0.339)	11.438*** (0.341)	10.962*** (0.332)	10.839*** (0.337)
R^2	0.193	0.206	0.236	0.237	0.246	0.328	0.336
Residual Std. Error	2.095 (df = 308)	2.082 (df = 307)	2.045 (df = 306)	2.048 (df = 305)	2.039 (df = 304)	1.927 (df = 303)	1.919 (df = 302)
F Statistic	24.650*** (df = 3, 308)	19.970*** (df = 4, 307)	19.000*** (df = 5, 306)	15.790*** (df = 6, 305)	14.180*** (df = 7, 304)	18.540*** (df = 8, 303)	17.030*** (df = 9, 302)

Note. The numeric/grade values for final course grade are as follows: 1 = F, 2 = D-, 3 = D, 4 = D+, 5 = C-, 6 = C, 7 = C+, 8 = B-, 9 = B, 10 = B+, 11 = A-, 12 = A, 13 = A+. The reference student is one who is in the Early Planning group, a man, non-First-Generation, not Low Income, and not an URM. For each predictor variable, the three reported coefficients are as follows (from top to bottom): Unstandardized Coefficient, Standardized Coefficient (italicized), Standard Error (in parentheses).

$^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$.

clustering algorithm on these measures to discover important patterns of SRL that would otherwise be overlooked.

The clustering algorithm successfully classified students into four meaningful groups: (1) *Early Planning*, (2) *Planning*, (3) *Procrastination*, and (4) *Low Engagement*. We also found meaningful and statistically significant relationships between the SRL cluster groups and final course grade. Students classified into the Early Planning group received the highest average grades in the course, followed by the Planning, Procrastination, and Low Engagement groups. This mirrors work by Cicchinelli and colleagues [14] who found similar relationships between their SRL cluster classifications and learning outcomes.

Our results also indicated that First-Generation students received lower course grades when compared to their non-First-Generation peers. This coincides with prior work that document the challenges of underrepresented students in navigating the learning demands of online courses [7, 10]. It additionally signaled the continued need to identify particular hurdles First-Generation students may encounter learning in online settings.

Examining First-Generation SRL patterns and their relation to course grades revealed both promising and concerning results. For the promising results, despite having lower course grades as a group, First-Generation students who utilized early planning behaviors fared very well in the course, earning comparable grades to non-First-Generation students. This shows that students who may be traditionally at-risk for failing STEM courses may benefit from utilizing SRL strategies that involve tracking completed assignments and doing so with the understanding that remaining assignments should not be put off or delayed.

For the concerning results, we found that First-Generation students in the Low Engagement group received the lowest grades in the course. These students also performed worse than non-First-Generation students in the Low Engagement group. In contrast, non-First-Generation earned similar grades regardless of their SRL behaviors. This finding aligns with what Greene and colleagues call the effort-outcome gap [17]. In their work, underrepresented student tended to report higher levels of engagement and effort in their coursework when compared to non-underrepresented groups, but had lower academic outcomes.

We additionally considered the role of other variables on final course grades using stepwise regression and found that the SRL cluster groups were statistically significant and consistent predictors of final course grades, even after controlling for demographic characteristics and prior achievement (SAT scores). Our stratified results revealed that SRL was a strong predictor for First-Generation students. However, SRL did not statistically significantly predict learning for non-First-Generation students.

Taken together, these results suggest that SRL skills are crucial for First-Generation students [7]. It is therefore important to ensure these students receive exposure and/or training in SRL or that courses are designed to scaffold the application of SRL. Indeed, work has found that interventions focused on teaching students self-regulation skills have the potential to reduce STEM achievement gaps [9, 35]. When thinking about how to best serve and support First-Generation students in STEM, especially in online course settings, our findings suggest that one should first keep in mind that students vary in their SRL behaviors. Therefore, online interventions or user-centered analytic dashboards that target students based on specific SRL behaviors may lead to greater learning outcomes [27].

This work had several limitations. Although these results provide an insightful and nuanced perspective about SRL and academic outcomes, there could be several unexplained reasons why First-Generation students who had low engagement fared so poorly. These include unfamiliarity with course concepts, internet access and technology issues, work and/or family obligations. Although our SRL measures help illustrate how these behaviors related to course success, low engagement or procrastination behaviors may not reflect a lack of SRL skills, but may simply be the artifact of the unique circumstances students faced that severely impacted their learning.

Our LMS data was also limited in that we were only able to observe whether students visited the videos, but not how students interacted with these videos (e.g., viewing coverage, number of times paused, number of replays). This additional information would have been useful in directly observing the process of how students learn when watching lectures. In terms of generalizing our results to other online college STEM courses, we want to caution that students in this study came from a large, selective, and research-intensive public university. This university, however, was unique in that it is also a minority-serving institution, where a large percentage of the student body were URM and First-Generation. Therefore, the climate of this institution may be somewhat different when compared to other selective universities (or universities with other student populations).

In summary, we have shown how theory-informed approaches to analyzing clickstream data can help instructors and other stakeholders better appreciate the different ways students regulate their independent learning time and how these behaviors relate to academic achievement. These results also highlight the variation in SRL patterns that exist among underrepresented students. Having low engagement was a risk indicator for First-Generation students, but we also found that early planning behaviors corresponded to high levels of achievement for this group. In other words, SRL really matters for First-Generation students. These results add to the growing body of work aimed at improving the outcomes of First-Generation college students. Designing targeted supports, whether through interventions or SRL-informed dashboards, can help support the students who need it most.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation grant, *Investigating Virtual Learning Environments* (Grant Number 1535300).

REFERENCES

[1] I Elaine Allen and Jeff Seaman. 2007. *Online nation: Five years of growth in online learning.* ERIC.

[2] Janeen Antonelli, Sara Jolly Jones, Andrea Backscheider Burridge, and Jacqueline Hawkins. 2020. Understanding the Self-Regulated Learning Characteristics of First-Generation College Students. *Journal of College Student Development* 61, 1 (2020), 67–83.

[3] Mara S Aruguete et al. 2017. Recognizing challenges and predicting success in first-generation university students. *Journal of STEM Education: Innovations and Research* 18, 2 (2017).

[4] Roger Azevedo, Daniel C Moos, Amy M Johnson, and Amber D Chauncey. 2010. Measuring cognitive and metacognitive regulatory processes during hypermedia learning: Issues and challenges. *Educational psychologist* 45, 4 (2010), 210–223.

[5] Rachel Baker, Di Xu, Jihyun Park, Renzhe Yu, Qiujie Li, Bianca Cung, Christian Fischer, Fernando Rodriguez, Mark Warschauer, and Padhraic Smyth. 2020. The benefits and caveats of using clickstream data to understand student self-regulatory behaviors: opening the black box of learning processes. *International Journal of Educational Technology in Higher Education* 17 (2020), 1–24.

[6] Sally J Baldwin and Yu-Hui Ching. 2019. An online course design checklist: development and users' perceptions. *Journal of Computing in Higher Education* 31, 1 (2019), 156–172.

[7] Lucy Barnard-Brak, Valerie Osland Paton, and William Y Lan. 2010. Profiles in self-regulated learning in the online learning environment. *International Review of Research in Open and Distributed Learning* 11, 1 (2010), 61–80.

[8] Lucy Barnard-Brak, Valerie Osland Paton, and William Y Lan. 2010. Self-regulation across time of first-generation online learners. *ALT-J* 18, 1 (2010), 61–70.

[9] Matthew L Bernacki, Lucie Vosicka, Jenifer C Utz, and Carryn Bellomo Warren. 2020. Effects of digital learning skill training on the academic performance of undergraduates in science and mathematics. *Journal of Educational Psychology* (2020).

[10] Jim Broadbent and Walter L Poon. 2015. Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *The Internet and Higher Education* 27 (2015), 1–13.

[11] Mitchell J Chang, Jessica Sharkness, Sylvia Hurtado, and Christopher B Newman. 2014. What matters in college for retaining aspiring scientists and engineers from underrepresented racial groups. *Journal of Research in Science Teaching* 51, 5 (2014), 555–580.

[12] Moon-Heum Cho, Yanghee Kim, and DongHo Choi. 2017. The effect of self-regulated learning on college students' perceptions of community of inquiry and affective outcomes in online learning. *The Internet and Higher Education* 34 (2017), 10–17.

[13] Ronny C Choe, Zorica Scuric, Ethan Eshkol, Sean Cruser, Ava Arndt, Robert Cox, Shannon P Toma, Casey Shapiro, Marc Levis-Fitzgerald, Greg Barnes, et al. 2019. Student Satisfaction and Learning Outcomes in Asynchronous Online Lecture Videos. *CBE—Life Sciences Education* 18, 4 (2019), ar55.

[14] Analía Cicchinelli, Eduardo Veas, Abelardo Pardo, Viktoria Pammer-Schindler, Angela Fessl, Carla Barreiros, and Stefanie Lindstädt. 2018. Finding traces of self-regulated learning in activity streams. In *Proceedings of the 8th international conference on learning analytics and knowledge.* 191–200.

[15] Ivo D Dinov. 2018. k-Means Clustering. In *Data Science and Predictive Analytics.* Springer, 443–473.

[16] Christian Fischer, Zachary A Pardos, Ryan Shaun Baker, Joseph Jay Williams, Padhraic Smyth, Renzhe Yu, Stefan Slater, Rachel Baker, and Mark Warschauer. 2020. Mining big data in education: Affordances and challenges. *Review of Research in Education* 44, 1 (2020), 130–160.

[17] Thomas G Greene, C Nathan Marti, and Kay McClenney. 2008. The effort—outcome gap: Differences for African American and Hispanic community college students in student engagement and academic achievement. *The Journal of Higher Education* 79, 5 (2008), 513–539.

[18] RB Harris, MR Mack, J Bryant, EJ Theobald, and S Freeman. 2020. Reducing achievement gaps in undergraduate general chemistry could lift underrepresented students into a "hyperpersistent zone". *Science Advances* 6, 24 (2020), eaaz5687.

[19] Shanna Jaggars. 2011. Online Learning: Does It Help Low-Income and Underprepared Students?(Assessment of Evidence Series). (2011).

[20] Shanna Jaggars and Thomas R Bailey. 2010. Effectiveness of fully online courses for college students: Response to a Department of Education meta-analysis. (2010).

[21] Shanna Smith Jaggars and Di Xu. 2016. How do online course design features influence student performance? *Computers & Education* 95 (2016), 270–284.

[22] Hannah Jordt, Sarah L Eddy, Riley Brazil, Ignatius Lau, Chelsea Mann, Sara E Brownell, Katherine King, and Scott Freeman. 2017. Values affirmation intervention reduces achievement gap between underrepresented minority and white students in introductory biology classes. *CBE—Life Sciences Education* 16, 3 (2017), ar41.

[23] KR Koedinger, S D'Mello, EA McLaughlin, ZA Pardos, and CP Rosé. 2015. Data mining and education. *Wiley interdisciplinary reviews. Cognitive science* 6, 4 (2015), 333.

[24] Qiujie Li, Rachel Baker, and Mark Warschauer. 2020. Using clickstream data to measure, understand, and support self-regulated learning in online courses. *The Internet and Higher Education* 45 (2020), 100727.

[25] Enrique J Lopez, Kiruthiga Nandagopal, Richard J Shavelson, Evan Szu, and John Penn. 2013. Self-regulated learning study strategies and academic performance in undergraduate organic chemistry: An investigation examining ethnically diverse students. *Journal of Research in Science Teaching* 50, 6 (2013), 660–676.

[26] Jorge Maldonado-Mahauad, Mar Pérez-Sanagustín, René F Kizilcec, Nicolás Morales, and Jorge Munoz-Gama. 2018. Mining theory-based patterns from Big data: Identifying self-regulated learning strategies in Massive Open Online Courses. *Computers in Human Behavior* 80 (2018), 179–196.

[27] Wannisa Matcha, Dragan Gašević, Abelardo Pardo, et al. 2019. A systematic review of empirical studies on learning analytics dashboards: A self-regulated learning perspective. *IEEE Transactions on Learning Technologies* 13, 2 (2019), 226–245.

[28] Wendy C Naumann, Deborah Bandalos, and Terry B Gutkin. 2003. Identifying variables that predict college success for first-generation college students. *Journal of College Admission* 181 (2003), 4.

[29] Ernesto Panadero. 2017. A review of self-regulated learning: Six models and four directions for research. *Frontiers in psychology* 8 (2017), 422.

[30] Abelardo Pardo, Feifei Han, and Robert A Ellis. 2016. Exploring the relation between self-regulation, online activities, and academic performance: A case study. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge.* 422–429.

[31] Jihyun Park, Kameryn Denaro, Fernando Rodriguez, Padhraic Smyth, and Mark Warschauer. 2017. Detecting changes in student behavior from clickstream data. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference.* 21–30.

[32] Jihyun Park, Renzhe Yu, Fernando Rodriguez, Rachel Baker, Padhraic Smyth, and Mark Warschauer. 2018. Understanding Student Procrastination via Mixture Models. *International Educational Data Mining Society* (2018).

[33] Mitchell Parkes, Sarah Stein, and Christine Reading. 2015. Student preparedness for university e-learning environments. *The Internet and Higher Education* 25 (2015), 1–10.

[34] Paul R Pintrich. 2000. Multiple goals, multiple pathways: The role of goal orientation in learning and achievement. *Journal of educational psychology* 92, 3 (2000), 544.

[35] Fernando Rodriguez, Mariela J Rivas, Lani H Matsumura, Mark Warschauer, and Brian K Sato. 2018. How do students study in STEM courses? Findings from a light-touch intervention and its relevance for underrepresented students. *PloS one* 13, 7 (2018), e0200767.

[36] Fernando Rodriguez, Renzhe Yu, Jihyun Park, Mariela Janet Rivas, Mark Warschauer, and Brian K Sato. 2019. Utilizing learning analytics to map students' self-reported study strategies to click behaviors in STEM courses. In *Proceedings of the 9th international conference on learning analytics & knowledge.* 456–460.

[37] Ido Roll and Philip H Winne. 2015. Understanding, evaluating, and supporting self-regulated learning using learning analytics. *Journal of Learning Analytics* 2, 1 (2015), 7–12.

[38] Elena Sandoval-Lucero, Johanna Maes, and Libby Klingsmith. 2014. African American and Latina (o) community college students' social capital and student success. *College Student Journal* 48, 3 (2014), 522–533.

[39] Şenol Şen. 2016. Modeling the structural relations among learning strategies, self-efficacy beliefs, and effort regulation. *Problems of Education in the 21st Century* 71 (2016), 62.

[40] Nora'ayu Ahmad Uzir, Dragan Gašević, Jelena Jovanović, Wannisa Matcha, Lisa-Angelique Lim, and Anthea Fudge. 2020. Analytics of time management and learning strategies for effective online learning in blended environments. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge.* 392–401.

[41] Peter E Williams and Chan M Hellman. 2004. Differences in self-regulation for online learning between first-and second-generation college students. *Research in Higher Education* 45, 1 (2004), 71–82.

[42] Philip H Winne. 2006. How software technologies can improve research on learning and bolster school reform. *Educational psychologist* 41, 1 (2006), 5–17.

[43] Philip H Winne and Dianne Jamieson-Noel. 2002. Exploring students' calibration of self reports about study tactics and achievement. *Contemporary Educational Psychology* 27, 4 (2002), 551–572.

[44] Christopher A Wolters, Paul R Pintrich, and Stuart A Karabenick. 2005. Assessing academic self-regulated learning. In *What do children need to flourish?* Springer, 251–270.

[45] Di Xu and Shanna S Jaggars. 2014. Performance gaps between online and face-to-face courses: Differences across types of students and academic subject areas. *The Journal of Higher Education* 85, 5 (2014), 633–659.

[46] Di Xu and Ying Xu. 2019. The Promises and Limits of Online Higher Education: Understanding How Distance Education Affects Access, Cost, and Quality. *American Enterprise Institute* (2019).

[47] Barry J Zimmerman. 1990. Self-regulating academic learning and achievement: The emergence of a social cognitive perspective. *Educational psychology review* 2, 2 (1990), 173–201.