From Clicks to Insights: Exploring SRL Behaviors Longitudinally using Institutional Data

Heeryung Choi, Chris Steadman, Caitlin Mills, Panayiota Kendeou

University of Minnesota

heeryung.c@gmail.com, stead032@umn.edu, cmills@umn.edu, kend0040@umn.edu

Abstract

The collection and use of student data in higher education institutions have increased in recent years with the widespread adoption of learning management systems (LMSs). These large-scale data have the potential to provide scholarly and practical insights about student learning and how their learning evolves over time. We used a learning analytics approach on an institutional dataset to cluster students (N=2293) based on behavioral characteristics, such as engagement and self-regulated learning (SRL), and observed how these clustering patterns changed over time and influenced academic performance and student retention. Our results highlight the importance of long-term SRL behaviors for academic success and demonstrate that SRL behaviors can act as early predictors of at-risk students. (113 words)

Objectives

It is a common practice for higher education institutions to collect and store data generated by students' interactions with learning management systems (LMS) over multiple years. These large-scale data have the potential to provide scholarly and practical insights through rich longitudinal information from various contexts (Brooks et al. 2023). However, drawing insights from these datasets is challenging. Specifically, hypothesis testing as a research approach can be easily interrupted by potential noises within a massive dataset. Even if researchers identify noise, efforts to control them often result in complex models that are difficult to analyze and interpret. In this paper, we address these challenges by applying a data mining approach to an institution-wide student dataset with the goal of developing a high-level understanding of individual students' behavior across multiple courses and institutional contexts. We specifically focused on relationships between student engagement, self-regulated learning (SRL) behaviors, academic performance, and student retention, which provides important insights into risk detection involving self-regulated behaviors.

Theoretical Framework

Learning is a dynamic and complex process, with learners adapting to numerous internal (e.g., prior knowledge) and external (e.g., task difficulty) contexts. Educational research has placed great importance on establishing controlled study contexts to reduce the impacts of confounding variables. While such practices are essential for research rigor, it also complicates application of its findings to real-world educational contexts.

Applying a learning analytics approach to institution-level data could ease this complication by offering more general and applicable insights from diverse contexts and student populations (Bernacki, 2025). For example, previous studies have shown mixed results about the

relationship between SRL and learning outcomes such as retention and academic performance (Hertel et al., 2024; Leite et al., 2022; Martin & Craigwell, 2022). These studies were often conducted in a small number of courses, which often share course designs or subject contexts. Data mining in institution-level data could contribute to addressing the mixed findings by aggregating behavioral patterns across diverse academic contexts, course designs, and student populations, thereby reducing the influence of course-specific confounding variables that may have obscured the relationship between SRL and academic outcomes in previous studies. Additionally, the use of a large, naturally occurring dataset that includes a longitudinal aspect allows us to find patterns or trends that might otherwise be missed in a smaller scale study.

Present Study

In this study, we conducted exploratory data analysis on a large-scale longitudinal institutional dataset, seeking behavioral indicators of SRL, including active engagement with course LMSs, as well as learning outcomes including academic performance and student retention. Specifically, we posed two primary research questions.

RQ1. What clusters can be identified based on engagement and SRL behavior features, and what can the clusters tell us about patterns of the student behaviors?

RQ2. How do longitudinal trajectories of the clusters relate to student retention and academic performance?

Methods

Dataset and Feature Extraction

We focused on behavioral and learning outcome data from full-time undergraduate students of a 4-year university in North America. First, we randomly sampled about 10%[1] of the

[1] We used a hash function for random selection, which often selects less data than threshold (10%) due to modulo bias effects. Thus, we used a slightly higher threshold (13%) to pull enough data.

students who were enrolled in the Fall 2022 semester (N=2293) and then extracted data for these students for the subsequent four semesters. The four datasets included weekly behavioral features and semester learning outcomes. For efficient and scalable feature selection, we followed two parallel processes inspired by Deininger et al. (2025). One was a theoretically guided process based on constructs relevant to engagement and SRL, while the other was data-driven and based on preliminary correlations between possible features and target variables. We iterated these two processes until we finalized features and target variables to examine. Table 1 presents the final eight selected features and four target variables.

Principal Component Analysis

To better understand students' engagement and SRL behavioral patterns (RQ1), K-means clustering was conducted on each semester dataset[2]. To address high dimensionality and multicollinearity among features, Principal Component Analysis (PCA) was applied prior to clustering (Jolliffe & Cadima, 2016; Abdi & Williams, 2010). All features were z-scored to standardize scales before conducting PCA. The number of principal components (PC) per semester was determined by examining the scree plot of explained variance and applying the elbow criterion, a standard approach (Cattell, 1966).

K-means Clustering to Group Students per Semester (RQ1)

K-means clustering was conducted on the PCs. The optimal number of clusters per semester was determined by three validation metrics: the Silhouette coefficient (Rousseeuw, 1987), the Davies-Bouldin index (Davies & Bouldin, 1979), and the Calinski-Harabasz criterion (Calinski & Harabasz, 1974). Once the optimal number of clusters was determined, the K-means

[2] Note that all data analyses to answer research questions were performed in Python (v3.135) using the SciPy library (v1.15.0) and scikit-learn (v1.6).

clustering was performed with five different random seeds per semester to ensure the stability of cluster labels for each student.

Non-parametric Kruskal-Wallis tests were employed to investigate how coursework planning, academic achievement, and enrollment status were respectively related with behavioral clusters. Follow-up pairwise comparisons were performed using Dunn's post hoc procedure with Bonferroni adjustment, when statistically significant main effects were found.

Hierarchical Clustering to Group Students' Longitudinal Trajectory (RQ2)

To detect patterns within students' longitudinal trajectories of SRL, engagement, retention, and academic performance, the sequence of cluster labels assigned each semester was concatenated to create a trajectory profile per student. There were missing cluster assignments since some students' data did not appear in all four semesters. The absence of data could indicate one of three cases: students stopped enrolling before earning their degree (dropout), students who graduated from their program with a degree (graduated), and students who transferred to other institutions (transferred). Although distinguishing between these cases is important, our dataset only included binary enrollment status data (enrolled or not) for each semester. Thus, we differentiated students based on the record of degree conferral dates; if a student had a conferral date before their semester-long absence, they were labelled as "graduated." Students who did not have such records were labelled as "dropout", which included both dropout and transferred cases. We then assigned -1 to fill missing clustering assignments of dropout students, and -2 to students who graduated.

Once missing data was filled, agglomerative hierarchical clustering (Hamming distance and average linkage method) was conducted on the trajectory profiles. The optimal number of trajectory clusters was determined by visually inspecting its dendrogram. Kruskal-Wallis tests

(non-parametric group comparisons) and Dunn's post hoc test with Bonferroni correction were then employed to examine the relationship between trajectory clusters and SRL behaviors, as well as academic predictors (i.e., academic performance and student retention).

Results

Our **first research question** focused on exploratory analysis of students' behavioral cluster assignment per semester. To answer this question, we conducted K-means clustering on PCA results. For all four semesters, three PCs consistently emerged as the optimal and interpretable solution, with high cumulative explained variances ranging from 87.4% to 96.7% (Greenacre et al., 2022). In addition, the four semesters exhibited similar loading distribution patterns, supporting the robustness of the PCA results (Table 2). We summarized the characteristics of each PC by focusing on the features with loadings larger than |0.3| (Li et al., 2022; Padilha et al., 2024). PC1 represented students' interaction with application files (e.g., .docx, .pdf), PC2 indicated students' overall LMS interactions both at day and night, and PC3 demonstrated LMS interactions at night time.

For K-means clustering, a three-cluster solution was consistently found to be optimal across all four semesters, with similar cluster sizes and characteristics (Table 3) indicating stable clustering. We then labelled each cluster based on their characteristics. **Regulars**, the largest clusters, displayed small negative PC1 and PC2 scores, indicating generally low interactions with application files (PC1) and low student activity levels and LMS interaction at night (PC2). **Explorers**, the medium size clusters, showed positive PC1 scores and especially high PC2 scores, indicating high student activity level. Their LMS interaction at night is also active, and they frequently view application files. Finally, **Observers**, the smallest clusters, exhibited the highest positive PC1 values and negative PC2 and PC3 values. This pattern indicates their

generally low student activity level with engagement concentrated mainly in interactions with application files. A Kruskal-Wallis H-test and post hoc Dunn's test showed that student behavioral clusters had statistically significant differences between all pairwise comparisons in PC1 ($H=1263.74$, $p<.001$) and PC2 ($H=1533.27$, $p<.001$) scores. Regarding relationships with target variables (Table 4), Kruskal-Wallis H-test and post hoc Dunn's test pairwise comparisons indicated that Explorers, throughout all four semesters, consistently showed significantly higher semester GPAs, higher coursework planning, and also a lower missing assignment ratio compared to Regulars. Additionally, the Observers tended to have the lowest GPAs and lowest coursework planning.

Our **second research question** focused on exploring the longitudinal trajectory patterns of behavioral clusters over four semesters (Figure 1). The student population (n=2293) consisted of 1481 (64.58%) students who stayed enrolled, 420 (18.31%) students who graduated, and 392 (17.09%) students who stopped enrolled. Agglomerative hierarchical clustering with a four-cluster solution was conducted on the sequences of four behavioral clusters.

The findings showed four trajectory clusters with distinguishing characteristics: **Dropout** (n=246, 10.72%, 14 (13.46%) unique sequences), **Mostly Explorer** (n=619, 26.99%, 36 (34.61%) unique sequences) **Graduated** (n=345, 15.04%, 13 (12.50%) unique sequences), **Mostly Regular** (n=1083, 47.23%, 41 (39.42%) unique sequences). Table 5 shows the most common trajectories for each cluster.

Regarding the relationships between each trajectory cluster and target variables, Kruskal-Wallis tests showed that there were significant differences in semester GPAs ($H=88.74$ $p<.001$) and coursework planning ($H=109.32$, $p<.001$) (Table 6). According to Dunn's tests on

GPAs, the **Mostly Regular** group and **Dropout** group did not have significant differences in semester GPAs ($p=.44$), which was the only non-significant pairwise post-hoc comparison.

Regarding the relationship with target variables of SRL behaviors, the **Dropout** group showed a significantly lower average frequency of coursework planning compared to all other groups ($p<.001$). Additionally, the **Mostly Explorer** group showed a significantly lower level of missing assignment ratio, another metric of SRL, when compared to both the **Dropout** and **Mostly Regular** groups. However, there was no significant difference between the **Mostly Regular** and **Dropout** groups ($p=.21$).

Scientific or scholarly significance of the study or work

This study leveraged a large longitudinal dataset to show that students' SRL behaviors represent clearer indicators of engagement level and enrollment status than semester GPAs. The most engaged students (**Explorers**) consistently did more coursework planning within the LMS when compared to students with lower engagement (**Regulars**). While the **Mostly Regular** group did not show significant differences in semester GPAs compared to the **Dropout** group, the **Mostly Regular** group showed a higher engagement with coursework planning over time.

The finding of no significant differences between **Mostly Regular** and **Dropout** groups suggests a detection approach based primarily on semester GPA would not be the most effective for detecting students at-risk of dropping out. Instead, focusing on their SRL behaviors may allow for earlier detection and timely intervention. The findings also demonstrated a consistent positive relationship between SRL behaviors (coursework planning, missing assignment ratio) and academic performance (semester GPA) across all semesters (**Explorer** > **Regular** > **Observer**), highlighting the importance of long-term SRL behaviors, and therefore SRL interventions for academic success (Xu et al., 2022).

It is important to note that this study was limited to Canvas LMS data, which may have constrained our understanding of student behaviors and limited generalizability to other platforms. Future research should incorporate additional data sources (e.g., offline behaviors) and conduct multi-institutional validation studies. Despite this limitation, by focusing on a holistic student-level analysis and leveraging cross-contextual data (e.g., various courses, semesters, and class standing), we have increased the applicability of our findings and provided more broad actionable insights. These findings lay the foundation for similar future work that can be broadly applied to many higher education contexts because we used data that is commonly collected at an institutional level. Thus, this study serves as a key step forward toward more context-specific, granular analysis for researchers and practitioners. (1944 words)

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(4), 433–459. https://doi.org/10.1002/wics.101

Bernacki, M. L. (2025). Leveraging learning theory and analytics to produce grounded, innovative, data-driven, equitable improvements to teaching and learning. *Journal of Educational Psychology, 117*(1), 1–11. https://doi.org/10.1037/edu0000933

Borrella, I., Caballero-Caballero, S., & Ponce-Cueto, E. (2019). Predict and intervene: Addressing the dropout problem in a MOOC-based program. *In Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale* (pp. 1-9).

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. C*ommunications in Statistics, 3*(1), 1–27. https://doi.org/10.1080/03610927408827101

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research*, *1*(2), 245-276.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Deininger, H., Pieronczyk, I., Parrisius, C., Plumley, R. D., Meurers, D., Kasneci, G., Nagengast, B., Trautwein, U., Greene, J. A., & Bernacki, M. L. (2025). Using theory-informed learning analytics to understand how homework behavior predicts achievement. *Journal of Educational Psychology, 117*(1), 12–37.

Greenacre, M., Groenen, P. J., Hastie, T., d'Enza, A. I., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers, 2*(1), 100.

Hertel, S., Reschke, K., & Karlen, Y. (2024). Are profiles of self-regulated learning and intelligence mindsets related to students' self-regulated learning and achievement?. *Learning and Instruction, 90,* 101850.

Jayaprakash, S. M., Moody, E. W., Lauría, E. J., Regan, J. R., & Baron, J. D. (2014). Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics, 1*(1), 6-47.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Leite, W. L., Kuang, H., Jing, Z., Xing, W., Cavanaugh, C., & Huggins-Manley, A. C. (2022). The relationship between self-regulated student use of a virtual learning environment for algebra and student achievement: An examination of the role of teacher orchestration. *Computers & Education, 191,* 104615.

Li, X., Lin, X., Zhang, F., & Tian, Y. (2022). What matters in online education: exploring the impacts of instructional interactions on learning outcomes. *Frontiers in Psychology, 12,* 792464.

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education, 54*(2), 588-599.

Martin, H., Craigwell, R., & Ramjarrie, K. (2022). Grit, motivational belief, self-regulated learning (SRL), and academic achievement of civil engineering students. *European Journal of Engineering Education, 47*(4), 535-557.

Padilha, J. M., Costa, P., Sousa, P., & Ferreira, A. (2024). Clinical virtual simulation: predictors of user acceptance in nursing education. *BMC medical education, 24*(1), 299.

Peterson, A. D., Ghosh, A. P., & Maitra, R. (2018). Merging K-means with hierarchical clustering for identifying general-shaped groups. *Stat (International Statistical Institute), 7*(1), e172. https://doi.org/10.1002/sta4.172

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20,* 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Rust, M. M., & Motz, B. A. (2023). Incorporating an LMS learning analytic into proactive advising: Validity and use in a randomized experiment. *EdArXiv Preprints*.

Xu, Z., Zhao, Y., Zhang, B., Liew, J., & Kogut, A. (2022). A meta-analysis of the efficacy of self-regulated learning interventions on academic achievement in online and blended environments in K-12 and higher education. *Behaviour & Information Technology*, *42*(16), 2911–2931.

Table 1

Descriptions of Study Features and Target Variables

Feature	***Description***
Mean student activity level	Mean of weekly student activity score (Rust & Motz, 2023).
Mean student activity level (Octile)	Mean percentile of z-scored weekly student activity score compared to other students in the course.
Mean app file viewed	Mean of weekly number of the unique 'application' file (e.g., docx, pdf) viewed. File types were pre-categorized by the data platform before our analysis. Application file type does not include image, video, and audio files.
SD app file viewed	Standard deviation of weekly number of the unique 'application' file (e.g., docx, pdf) viewed.
Mean app file views	Mean of weekly number of the unique 'application' file that was viewed by individual students.
SD app file views	Standard deviation of weekly number of the unique 'application' file that was viewed by individual students.
Mean duration of night activity	Mean weekly total duration (in seconds) of students' LMS interaction sessions between midnight and 5:00 AM over the course of the semester. The definition of interaction sessions follows the definition of '30 minute session' as described by the data platform.
Mean counts of night activity	Mean weekly total counts of students' LMS interaction sessions between midnight and 5:00 AM over the course of the semester. The definition of interaction sessions is the same as the one above.
Target Variable	***Description***
Semester GPA	(Academic achievement) Individual students' GPA per semester
Coursework planning	(SRL) Mean weekly frequency of students checked LMS deadline pages and moved to linked course pages of corresponding tasks (e.g., assignments, quiz,...)
Missing assignment ratio	(SRL) Mean weekly count of assignment submitted / Count of assignment due in the week

Enrollment status	(Retention) Categorical variable of enrollment status showing if students stayed enrolled, graduated, or dropped out in the semester. Note that unlike three other target variables, enrollment status was used to identify students who graduated and stopped enrollment.

Note. All features were extracted on a weekly basis. For each course, weekly feature values were summarized by calculating either the mean (Mean features) or the standard deviation (SD features). This resulted in a single feature value per unique student per course. For example, the student activity score was first averaged weekly within each course, and then these weekly averages were further averaged across the course duration to produce one overall activity score per student per course.

Table 2

Average PC Loadings for Fall 2022 to Spring 2024

PC	Mean Activity	Mean Activity (Octile)	Mean App Files viewed	SD App Files viewed	Mean File view	SD File view	Mean Night Duration	Mean Night Count
PC1	.62	.62	1.84	1.85	1.86	1.86	.84	.84
PC2	1.99	1.98	-.75	-.70	-.73	-.70	1.66	1.78
PC3	-1.89	-1.91	-.22	-.07	-.17	-.30	2.17	1.98

Note. Throughout the four semesters, the distribution of PC loadings remained consistent.

Table 3

Average Principal Component Scores per Behavioral Cluster of each Semester

Cluster	Mean PC1 Score	Mean PC2 Score	Mean PC3 Score	Cluster Size (n, %)
Fall 2022				
Regular	-.87	-.75	-.03	1307 (56.99%)
Explorer	.60	1.39	.07	888 (38.72%)
Observer	6.40	-2.85	-.27	98 (4.27%)
Total Enrolled Students				2293
Spring 2023				
Regular	-.82	-.86	.06	1273 (58.18%)
Explorer	.38	1.46	-.05	826 (37.35%)
Observer	6.79	-1.92	-.28	89 (4.06%)
Total Enrolled Students				2188
Fall 2023				
Regular	-.72	-.84	-.02	999 (57.44%)
Explorer	.51	1.32	.01	701 (40.31%)
Observer	9.25	-2.28	-.36	39 (2.24%)
Total Enrolled Students				1739
Spring 2024				
Regular	-.67	-.96	.05	803 (54.22%)
Explorer	.38	1.32	-.04	652 (44.02%)
Observer	12.83	-2.89	-.50	26 (1.75%)
Total Enrolled Students				1481

Table 4

Descriptive Statistics and Dunn's Test Results of the Target Variables by Behavioral Clusters

Cluster	Mean (SD)	Median	Pattern
Coursework Planning (z-scored)			
Regular	-.28 (.65)	-.46	Observer = Regular < Explorer
Explorer	-.04 (.92)	-.32	
Observer	-.26 (.90)	-.60	
Semester GPA			
Regular	2.45 (.90)	2.50	Observer < Regular < Explorer
Explorer	2.75 (.82)	2.84	
Observer	2.10 (.85)	2.09	
Missing Assignment Ratio			
Regular	.39 (.33)	.30	Explorer < Observer = Regular
Explorer	.24 (.26)	.15	
Observer	.28 (.32)	.15	

Note 1. The statistics here are calculated with all behavioral clusters of four semesters (3 clusters X 4 semesters = 12 clusters). All four semesters showed highly similar patterns of clusters and their relationships with target variables.

Note 2. Bonferroni-adjusted $\alpha = .0167$ for 3 Dunn's test pairwise comparisons between trajectory clusters. If Bonferroni-adjusted p-value < .0167, it was considered significant.

Table 5

The Top Three Most Frequent Sequences per Trajectory Group

Sequence Rank	Mostly Regular	Mostly Explorer	Graduated	Dropout
1st Sequence	R-R-R-R (n=418)	E-E-E-E (n=207)	R-R-G (n=169)	R-R-S (n=68)
2nd Sequence	R-R-R-E (n=72)	R-E-E-E (n=60)	E-E-G (n=64)	R-S (n=64)
3rd Sequence	R-R-E-R (n=62)	E-E-E-R (n=54)	R-E-G (n=33)	E-S (n=37)

Note. R = Regular cluster, E = Explorer cluster, G = Graduated, S = Stopped Enrolling

Table 6

Descriptive Statistics and Dunn's Test Results of Target Variables per Trajectory Clusters

Trajectory Cluster	Mean (SD)	Median	Pattern
Coursework Planning (z-scored)			
Dropout (DR)	-.54 (.55)	-.79	DR < MR = GR = ME
Mostly Explorer (ME)	-.09 (.77)	-.30	
Graduated (GR)	-.22 (.63)	-.39	
Mostly Regular (MR)	-.21 (.60)	-.32	
GPAs			
Dropout (DR)	2.34 (.88)	2.35	DR = MR < ME < GR
Mostly Explorer (ME)	2.65 (.61)	2.71	
Graduated (GR)	2.83 (.61)	2.88	
Mostly Regular (MR)	2.48 (.68)	2.55	
Missing Assignment Ratio			
Dropout (DR)	.38 (.34)	.25	DR < MR = GR < ME
Mostly Explorer (ME)	.25 (.20)	.22	
Graduated (GR)	.31 (.32)	.19	
Mostly Regular (MR)	.37 (.24)	.35	

Note 1. SD: Standard deviation

Note 2. Bonferroni-adjusted $\alpha = .0083$ for 6 Dunn's test pairwise comparisons between trajectory clusters. If Bonferroni-adjusted p-value < .0083, it was considered significant.

Figure 1

Trajectory of Behavioral Clusters over Four Semesters