Smart Learning Environments

RESEARCH Open Access

Incorporating teacher effect when modeling student engagement in smart STEM classrooms: a cluster analysis

Kelly Shreeve[1], Anthony Perry[2], Michael Cassidy[3], Kathryn Jessen Eller[4], Beth Price[5], Brandy Jackson[5], Leo Celi[6,7,8], Ismini Lourentzou[9] and Luk Hendrik[1*]

*Correspondence:
luk@aecern.com

[1] Data Science and Learning Group, Aecern LLC, Falmouth, MA, USA
[2] College of Education and Human Development, University of North Dakota, Grand Forks, ND, USA
[3] TERC, Cambridge, MA, USA
[4] The Concord Consortium, Concord, MA, USA
[5] Instructional Design and Program Delivery Group, Aecern LLC, Falmouth, MA, USA
[6] Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, MA, USA
[7] Division of Pulmonary, Critical Care, Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA
[8] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[9] School of Information Sciences, University of IL Urbana – Champaign, Champaign, IL, USA

Abstract

Student engagement during learning serves as a critical predictor of academic success and plays a pivotal role in nurturing interest and readiness for future careers. As digital platforms become increasingly important to learning, it is essential that we understand how the interactions that students have with them reflects their engagement with learning. Previous research has often modeled engagement in a fully online context, where students pursue lessons independently and outside the influence of the classroom, paced and structured by digital systems. However, in STEM (Science, Technology, Engineering, and Math) subjects—and many others—learning more frequently happens in a physical classroom setting, under the guidance of a teacher, and involves interactions with other students and tangible objects. Here digital materials are used to scaffold and support learning but are not typically the focus of where learning happens. To study how student interactions with digital materials in these settings might allow us to measure, evaluate and help teachers enhance engagement, we have developed and deployed a smart digital learning platform that guides instruction and captures real-time multimodal student learning events in the physical STEM classroom. Previously we have shown that a subset of student interactions measured with this platform can be used to model student learning and generate human-like insights into engagement. Here we report on the significant influence that teachers have on student interactions with our smart platform in the STEM classroom, and the impact that this has on evaluating their engagement with learning. In an analysis of 108 high school students that used the platform to complete a 19-lesson data science curriculum in 5 different classrooms, we found significant differences between teachers both in the measured time students spent on the lesson and the percentage of the lesson they completed. In this setting, taking teacher influence into account improves the outcomes of our machine learning clustering models that group students based on their level of engagement. These findings inform how we develop smart classroom technology and machine learning applications that are globally informed but locally relevant, and support teachers to enhance student engagement and learning outcomes in dynamic and highly variable STEM classroom learning environments.

Keywords: Digital learning platform, STEM learning environments, Student engagement, Cluster analysis, AI-driven teaching assistants

Introduction

The landscape of K-12 education, particularly in Science, Technology, Engineering, and Mathematics (STEM), has been drastically altered in recent years. Globally, the COVID-19 pandemic accelerated digital transformation in schools, while simultaneously exacerbating educational challenges, including widespread learning loss (NAEP, 2022), a shortage of qualified STEM educators (Darner & Boesdorfer, 2022), and increased disparities in access to high-quality instructional resources (Raugust & Berkman, 2022). In response, school systems made significant investments in educational technologies, creating opportunities to rethink and enhance the delivery of STEM education, especially through technology-supported, hands-on learning that enhances classroom content with digital materials (Raugust & Berkman, 2022).

Alongside this surge in technological integration, there has been increased interest in predicting and adapting to real-time student engagement in STEM instruction. Student engagement has long been recognized as a key predictor of academic success and long-term achievement (Fredericks et al., 2004; Sinatra et al., 2015; Lei et al., 2018) and has been shown to be a multidimensional construct encompassing behavioral, cognitive, emotional and agentic components (Reeve, 2013; Reschly & Christenson, 2022; Sinatra et al., 2015). These different expressions of engagement reflect different aspects of a student's relationships to learning and are manifested in different ways. Behavioral engagement reflects positive conduct and involvement in academic tasks, such as time-on-task or task completion. Cognitive engagement represents the investment in learning through effort to understand, use of flexible problem-solving techniques, and choice of challenging tasks. It can be manifested by the length or complexity of responses and the number and types of learning supports accessed. Emotional engagement explores the student's investment in the academic subject area and can be measured by self-reported interest and preparedness for lessons. Finally, agentic engagement describes whether a student seeks to enrich, personalize or request related learning experiences. It can manifest itself through collaboration and choice of tasks.

Particularly in STEM education, engaging students through relevant, inquiry-driven learning has been shown to foster deeper interest and better outcomes (Means et al., 2021). However, the dynamic, collaborative, and tech-intensive nature of modern STEM classrooms presents challenges in measuring and supporting student engagement, especially in ways that are practical and actionable for educators. Capturing these dimensions in real time, in active classroom environments, remains difficult. Traditional methods such as teacher observation are often inconsistent and unsustainable, especially for novice educators (Harris & Sass, 2011; King Rice, 2010).

At the same time, emerging technologies, including AI, machine learning (ML), and big data analytics, have introduced promising alternatives. Emotional engagement has been measured by applying computer vision to analyze facial expressions, mouse movement, and heart rate on students completing online activities (Monkaresi et al., 2017; Pantic & Rothkrantz, 2015). More recently, others used a combination of facial recognition, mouse movements, and clicks to model student emotional engagement while completing an online course (Altuwairqi et al., 2021). These video methods show promise, with some models achieving up to 95% accuracy in emotional recognition (Salloum et al., 2025). While well suited for adult online learners who consent to facial expression and mouse movement tracking to improve their engagement over time, in a K-12 setting, it is unlikely that they can be implemented at scale, as they bring ethical concerns and a high cost of implementation for resource-stretched school districts (Cetintas et al., 2010).

Others have explored the promise of log data, which are detailed information about students' interactions with digital learning systems that are captured during learning (Cocea & Weibelzahl, 2009). Critically, log data do not include direct visual or biometric student data and therefore are non-invasive of students' privacy and don't require any additional equipment to deploy in a K-12 context. Multiple methods of analysis of log data show promise in tracking student engagement, with research focused on detecting student disengagement with intelligent tutoring systems (Cetintas et al., 2010) and e-Learning systems (Cocea & Weibelzahl, 2009). We and others have found evidence that clustering analysis can be completed on log data, effectively grouping students into patterns of engagement (Liu, 2022; Shreeve et al., 2025). As log systems track how students interact with learning material, they represent readily available and non-invasive information on the behavioral aspect of engagement.

A limitation of research into log data techniques is that most has been predominantly conducted in fully online learning environments, where students interact with digital materials independently, in isolation from classroom influences. In such settings, models only consider individual learner characteristics, like prior knowledge or attention span, and don't account for external, systemic factors that can influence behavior.

Not all digitally-supported learning occurs online and in isolation. Smart classrooms—technology-enhanced physical spaces where teachers and peers are co-present—represent a growing paradigm that blends the benefits of digital tools with the instructional and social value of in-person learning (Saini & Goel, 2019). In these environments, student engagement is not shaped solely by personal characteristics or system design, but also by teacher behaviors, classroom culture, and peer interactions (Love et al., 2023). While this dynamic is well known, there has been little accounting for the impact of the classroom dynamics in machine learning model development predicting engagement. Existing engagement models in smart classrooms typically overlook teacher effects, specifically how instructional pacing, task emphasis, and pedagogical flexibility shape student interactions with digital tools. Not accounting for differences in these in-person contexts limits the relevance of current engagement models in more complex, real-world learning environments. As such, there is a need to develop models of engagement for smart classrooms, where teacher influence is both pervasive and often unaccounted for in existing models.

To support and model student engagement in smart classrooms, we have developed Scoutlier, a smart digital learning platform that scaffolds instruction and captures real-time multimodal student interaction data during classroom-based STEM learning (Shreeve et al., 2025). Over the preceding 3 years, we have used Scoutlier to deliver an experimental data science curriculum in a range of classroom settings, ultimately collecting a research grade, de-identified dataset of interaction data from 108 high school students across five classrooms (Larson et al., submitted; Jessen Eller, 2025; Jessen Eller et al., 2024). In this study, we explicitly focus on behavioral engagement, specifically task completion and time-on-task, as measurable, non-intrusive indicators of active participation. These aspects of engagement are particularly suited to K-12 classrooms, allowing for practical, ethically sound analysis of student interactions.

Analyzing this case study of high school students interacting with Scoutlier, we sought to understand how the live classroom affects students' engagement with online material, with a focus on traceable behavioral engagement, and to use those insights to build better machine learning models of engagement. Building on our prior work showing that student interactions with Scoutlier can serve as proxies for engagement during classroom STEM instruction (Shreeve et al., 2025), we now demonstrate that the teacher has a significant impact on how students interact with digital learning platforms in smart classrooms. Importantly, taking this "teacher effect" into consideration improves machine learning clustering models grouping students by engagement and results in more meaningful cluster interpretations. These findings underscore the differences in modeling student engagement and behavior in smart classrooms and demonstrate the importance of models that reflect the full complexity of in-person learning environments. The results suggest that the effect of the teacher on student engagement is a critical dimension that must be incorporated into the design of intelligent classroom technologies, learning analytics, and machine learning models.

The paper is structured as follows: Methods are detailed first, describing data collection, analysis, and clustering approaches. Results present quantitative and qualitative findings highlighting teacher effects on engagement. The Discussion interprets these findings, providing theoretical and practical insights. Limitations and future research suggestions follow, concluding with implications for practice.

Methods

Research design

This study follows a mixed-methods design integrating quantitative data collected from the Scoutlier platform and qualitative interviews conducted with teachers at the end of each semester. Quantitative data were collected from 108 high school students across five schools and three semesters from January 2023 to June 2024 (Shreeve et al., 2025). Classes participated in a 19-lesson Data Science, AI and You (DSAIY) in Healthcare curriculum delivered with support from Scoutlier (Larson et al., submitted; Jessen Eller, 2025; Jessen Eller et al., 2024). Teachers had full access to the course and led their students through the online curriculum over the course of a single school semester. Students progressed through the lessons with an individual learning profile and completed course material on desktops, laptops, or chromebooks with the supplementary aid and instruction of their teacher.

Semi–structured interviews were conducted with each of the teachers after the completion of the program. Interviews were approximately 45 min in length and focused on the teacher experience facilitating the data science course and interacting with the Scoutlier platform. Interviews were transcribed and coded for analysis. For analyses, the project team developed a preliminary codebook of predetermined codes, then two researchers coded each interview (Saladana, 2015). Quantitative and qualitative analysis were combined using parallel mixed model, where qualitative and quantitative data are collected within the same time frame and analyzed independently (Tashakkori & Teddlie, 1998).

Scoutlier curriculum

All students participated in DSAIY in Healthcare curriculum in Scoutlier, a web-based application for designing and delivering STEM lessons. Built around principles of accessible, data-informed instruction (Sana MIT, n.d.), Universal Design for Learning (Hall et al., 2004; Meyer et al., 2014), and the "positive deviance" concept (Gupta et al., 2022; Ruggeri & Folke, 2022), it supports creation of interactive STEM activities, delivery of lessons within or independent of learning management systems, and multi-modal student data collection (e.g., text, data, photo, audio, etc.) (Shreeve et al., 2025). The application was created to increase engagement and improve learning outcomes in STEM classrooms, with a teacher present to instruct and facilitate learning as students complete each lesson (Larson et al., submitted; Jessen Eller, 2025; Jessen Eller et al., 2024). Scoutlier records engagement metrics like time spent on task using timestamps and algorithms to calculate differences between begin and end of task and metrics like completion and student interaction by tracking when a student completes tasks or interacts with other students.

DSAIY lessons were divided into 4 units that introduce students to bias and machine learning in healthcare (Unit 1), review core K-12 data science concepts, like data collection, organization, visualization and analysis (Unit 2), introduce machine learning tools like Python and Jupyter notebooks so students can build their own algorithms (Unit 3), and synthesize and apply their learning via a summative project (Unit 4) and hands-on workforce experience datathon. Lessons were broken down into tasks and steps that asked students to participate in an activity and then respond through a paragraph, checkbox, screenshot, audio, or video responses on the online platform. Students were asked to do tasks such as code in Google Colab then write a paragraph about their learning on Scoutlier. All of the response types supported multimodal student learning and provided the teacher with multiple ways to assess understanding.

Participants

Five teachers were purposely selected from participants recruited to our NSF-funded project, ensuring representation of diverse instructional settings, student demographics, and pedagogical practices to maximize variability and explore broad engagement dynamics. Dean, Darren, and Lena taught two semesters each, while Heather and Paige taught a single semester. For purposes of analysis, students were combined within teachers who taught more than one semester. This resulted in a final sample size of 108 students spread across five teachers. All student participants were in high school and ranged from 15 to 19 years old. Schools represented a range of populations of students from Rhode Island. Two schools were in urban communities, two in suburban communities, and two in the suburban/urban ring. Five schools were public and one private. Students came from both extensive computer science backgrounds and limited coding experience, and schools had varying teaching philosophies from traditional public schools to exploratory learning and assessment policies.

Dataset description

Scoutlier tracked 20 different types of student interactions while they progressed through the course, including task completion, time on task, response length, response content, and peer interactions. Lessons consisted of 10–20 unique tasks, and Scoutlier recorded multiple engagement metrics for each task in real-time. Scoutlier also recorded lesson-level summary data of total time spent on lesson and the total percentage of tasks completed for each lesson. This resulted in a dataset of 108 rows and over 1500 columns of engagement metrics, approximately 75 engagement metrics per lesson. Deidentified log data were analyzed according to the Institutional Review Board (IRB) approved protocol.

Quantitative variables

Two main variables were extracted as dependent variables for use in this analysis. The first was percentage of the lesson completed, which represents the percentage of the online Scoutlier tasks a student completed within each lesson. The second was time spent on the lesson, calculated as the difference in timestamp between a student opening their Scoutlier assignment and closing the assignment for the day. We selected these specific metrics due to their straightforward interpretation, practical feasibility in real classroom settings, and their minimally invasive nature. They are understandable and quantifiable metrics of behavioral engagement that summarize total engagement for each student and are readily comparable across teachers. While the other metrics captured by Scoutlier representing emotional or cognitive engagement also show promise for comparison, they are outside the scope of the current analysis. The pair were extracted for each of the 19 lessons, resulting in 38 dependent variables, two for each lesson. The main independent variable for each lesson is the teacher teaching the class.

Statistical methods

The purpose of this analysis is to determine whether significant differences exist between teachers in time spent on the lessons and the percentage of the lesson completed, then carry these findings into machine learning model development. To determine whether significant differences existed, we employed the The Kruskal–Wallis test, a non-parametric statistical test developed to compare the distributions of three or more independent samples. It tests the null hypothesis that the medians are equal across the groups and an alternative hypothesis that at least one of the medians is different. As a non-parametric test, the Kruskal–Wallis test does not have normality and sample size assumptions. Thus, it is appropriate to compare the differences of groups when the data is non-normal and/or the sample sizes are small (Ostertagova et al., 2014). We used a cutoff of $p < 0.05$ for statistical significance. As the purpose was to determine whether a difference existed overall, less focus was given to which teachers differed from each other, and post-hoc tests were not conducted.

Results from the Kruskal–Wallis test informed our machine learning cluster analysis of student engagement based on time spent on the lesson and the percentage of the lesson completed. Cluster analysis is a method of unsupervised learning used to create meaningful groups of observations within a dataset. The purpose is to group unlabeled data points into clusters of similar characteristics where observations in a cluster are more similar to each other than to other observations (Jain et al., 1999). Specifically, this analysis employs k-means clustering (Lloyd, 1957; MacQueen, 1967), which creates clusters that minimize the sum of the squared differences (error) between the cluster centers and data points (Jain et al., 1999). It partitions the data into k distinct clusters that best group the data. To identify the optimal number of clusters for our analysis, scree plots illustrating within-cluster sum of squares (WCSS) were generated for all 19 lessons. The optimal solution (k = 4) was chosen as it represented a leveling off of the reduction of cluster sum of squares in the WCSS curve consistently across all lessons. Representative WCSS curves for Lesson 6 and Lesson 15 both depict an elbow point at four clusters, demonstrating the appropriateness of a four-cluster solution, indicating this as the optimal choice balancing explanatory power and simplicity (see Supplementary Figs. S1 and S2).

Cluster centroids were then fed through a Large Language Model (LLM), GPT-4, to translate the cluster centers into interpretable engagement profiles (Achiam et al., 2023; Shreeve et al., 2025). We provided GPT-4 with a dataset of cluster centroids (mean standardized completion and time-on-task values) and prompted the LLM with a description of the dataset and to interpret the cluster characteristics with descriptive labels characterizing engagement patterns and specific teaching interventions for each cluster. Generated descriptions underwent expert validation by qualitative researchers and curriculum specialists to ensure interpretive accuracy and alignment with educational theory. All statistical and machine learning analysis was conducted with Python 3.10.

Data preprocessing

Missing data

Due to the large and longitudinal nature of the dataset, data missingness was a concern. Missing data largely resulted from students not attending every lesson. A missing data analysis was conducted and found that over 10% of students were missing from class on some days. In these cases, single imputation is not recommended, and multiple imputation is recommended only if missing cases are below 5% (Dettori et al., 2018). Therefore, missing data were addressed through pairwise deletion of students from lessons they did not attend. Students who missed a lesson were deleted from only that lesson and kept in the analysis for lessons they were present.

Data standardization

Clustering requires an additional preprocessing step of data standardization. This process adjusts scores by the mean and standard deviation of the variable resulting in data that is on the same scale across variables and scores that represent standard deviations from the mean (James et al., 2021). Once scores are standardized, scores close to zero represent participants who are close to the mean and scores far away from zero represent participants who are far away from the mean. We standardized scores using the z-score formula:

$$z = \frac{x - \overline{x}}{s}$$

x: Student's score, $\bar{x}$: Mean, s: Standard deviation.

To assess the impact of the teacher on the clustering models, two clustering models were fit with z-scores calculated in two ways: (1) standardization across the full dataset, and (2) standardization within each classroom. In the first model, scores were standardized over the full dataset, calculated by subtracting the mean of all 108 students and dividing by the standard deviation of all students. This resulted in students who were far away from the mean of all students, across all classrooms, receiving higher standardized scores, and those who were close to the mean of all students receiving standardized scores close to zero.

In the second model, the data preparation was informed by differences in engagement across classrooms. Instead of standardizing across the full dataset, students' scores were standardized *within each classroom*. These standardized scores were calculated by subtracting the mean of the respective teacher's class and dividing by the standard deviation for that class. With this second normalization technique, students with high standardized scores were now relatively highly engaged *in their class* and those with low standardized scores were relatively less engaged *in their class*. The purpose of this paper is to justify that this second type of standardization within the classroom is more useful in detecting patterns in student engagement in smart classrooms.

While z-score standardization is sensitive to outliers, an outlier analysis was performed prior to clustering. Outliers were defined as datapoints $+/-3$ standard deviations away from the mean. One outlier was found in Lesson 6 and removed from the dataset. No transformations beyond z-score normalization were applied. Standardization was applied to the data only before clustering and was not applied prior to the Kruskal–Wallis test.

Results

Sample size by lesson

Given the longitudinal nature of the study, varying numbers of students attended each lesson. Table 1 displays the number of students who completed each lesson for each of the 19 lessons in the course, by the teacher who taught them. Students missing from a lesson are represented by a lower sample size for that lesson. Students per teacher ranged from 9 to 39. Lena had the largest number of students combined across semesters ($n=40$), followed by Dean ($n=28$), Paige ($n=20$), Darren ($n=11$), and Heather ($n=9$). The total sample size for each lesson ranged from 63 to 101.

Classroom differences in lesson completion

To determine whether the teacher impacts behavioral engagement in the classroom, we first examined the differences in lesson completion across all DSAIY lessons between the five teachers. Figure 1 displays the differences in the average percentage of the lesson completed between teachers across all 19 lessons in the course. Each line represents the average percentage of the lesson students completed within that teacher. The figure depicts large differences in the lines, representing large differences in the percentage of lessons students completed depending on the teacher. There are substantial differences in all lessons, though these differences are especially large in Lesson 10 through Lesson 15. Examining Lesson 11, Dean's and Lena's students completed only 51.0% and 55.0%

Table 1 Number of students completing each lesson by teacher

Lesson	Students completing lesson (n)					Total n
	Darren	Dean	Heather	Paige	Lena	
Lesson 1	10	25	8	19	39	101
Lesson 2	11	28	NA	20	35	94
Lesson 3	10	27	9	19	36	101
Lesson 4	10	28	9	20	36	103
Lesson 5	11	28	9	20	33	101
Lesson 6	11	28	9	20	33	101
Lesson 7	10	26	9	20	36	101
Lesson 8	11	25	9	20	37	102
Lesson 9	11	24	9	20	29	93
Lesson 10	11	24	9	19	16	79
Lesson 11	10	25	9	19	38	101
Lesson 12	10	10	9	8	32	69
Lesson 13	10	25	9	20	35	99
Lesson 14	11	25	9	20	34	99
Lesson 15	11	25	9	20	30	95
Lesson 16	10	18	9	20	32	89
Lesson 17	11	NA	9	18	12	50
Lesson 18	10	22	9	20	18	79
Lesson 19	10	11	8	18	16	63

NA represents a lesson that a teacher skipped or otherwise did not have data

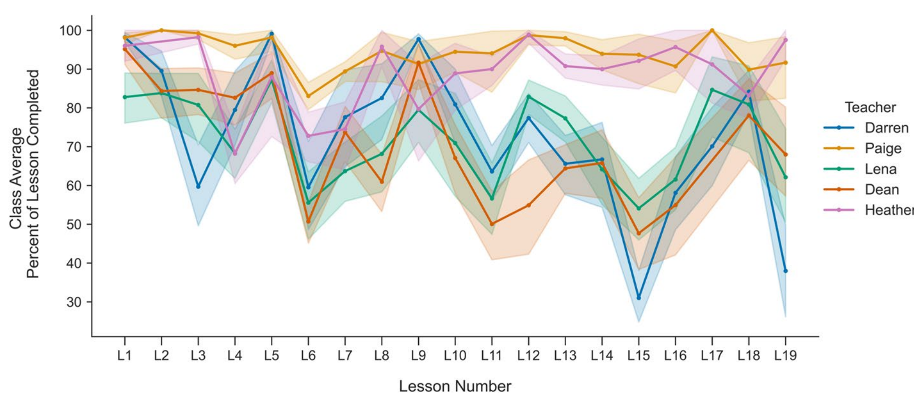

Fig. 1 Differences in average percentage of each lesson completed between teachers. *Note* Error bars represent the 95% CI for the mean

of the lesson on average, while Paige's students completed much more, averaging 95.0%. When examining trends across the lessons, Paige's and Heather's students consistently completed most of the lesson, while the other three classrooms completed much less of the online material.

These differences are further confirmed through Kruskal–Wallis Tests comparing differences in the median percentage of the lesson completed between the five teachers for each lesson (Table 2). Significant differences were found in Lesson 2–Lesson 18 (see Table 2 for p values), representing a significant difference in the median percentage of the lesson completed for at least one of the teachers in each lesson. Lesson 1, within

Table 2 Kruskal–Wallis test comparing median percentage of each lesson completed between teachers

Lesson	Class median percentage of lesson complete					H	df	*p* value
	Darren	Dean	Heather	Paige	Lena			
Lesson 1	100.00	100.00	100.00	100.00	100.00	8.73	4	.068
Lesson 2	90.00	90.00	NA	100.00	90.00	20.78	3	<.001
Lesson 3	53.50	92.00	100.00	100.00	96.00	23.90	4	<.001
Lesson 4	88.00	90.00	77.00	100.00	77.00	30.39	4	<.001
Lesson 5	100.00	100.00	100.00	100.00	100.00	14.92	4	.005
Lesson 6	55.00	50.00	75.00	80.00	65.00	27.77	4	<.001
Lesson 7	86.50	84.50	78.00	85.00	75.50	22.58	4	<.001
Lesson 8	90.00	72.00	100.00	100.00	86.00	39.44	4	<.001
Lesson 9	100.00	100.00	95.00	100.00	95.00	9.89	4	.042
Lesson 10	90.00	77.50	95.00	100.00	87.50	26.04	4	<.001
Lesson 11	70.00	51.00	92.00	100.00	55.00	39.32	4	<.001
Lesson 12	76.50	45.00	100.00	100.00	90.00	35.86	4	<.001
Lesson 13	66.00	73.00	93.00	100.00	86.00	44.69	4	<.001
Lesson 14	72.00	77.00	94.00	100.00	77.00	32.29	4	<.001
Lesson 15	27.00	45.00	95.00	100.00	63.00	47.48	4	<.001
Lesson 16	57.00	61.00	100.00	100.00	71.00	35.90	4	<.001
Lesson 17	77.00	NA	100.00	100.00	88.00	26.21	3	<.001
Lesson 18	84.00	92.00	84.00	100.00	92.00	9.86	4	.043
Lesson 19	41.50	66.00	100.00	100.00	83.00	29.79	4	<.001

NA represents a lesson that a teacher skipped or otherwise did not have data. They were removed from the comparison for that lesson, reflected in the df for that test

which a median of 100% of the lesson was completed across all classrooms, was the only lesson that did not show a significant difference in percentage of the lesson completed. Paige's students completed 80–100% of the tasks for each lesson, while Dean's students consistently completed only 45–85% of the tasks. These results signify that, for all but the first lesson, the teacher significantly impacted the percentage of the lesson students in that classroom completed.

Differences in lesson completion evidences higher or lower behavioral engagement depending on the teacher. Importantly, differences reflect students' behaviors and do not necessarily reflect emotional or cognitive aspects of engagement. These differences in student behavior are likely influenced by intentional pedagogical adaptations by teachers to match instructional content, pacing, and emphasis to their classroom contexts and student needs.

Classroom differences in time spent on the lesson

We then compared a second construct of behavioral engagement across the five teachers, time spent on the lesson. Figure 2 displays the differences in the time spent on the lesson between the five teachers across all 19 lessons. Again, large differences were displayed between the lines, indicating large differences in the time spent on each lesson depending on the teacher (Fig. 2). Overlapping with the percentage of lessons completed (Fig. 1), there are especially large differences in Lesson 12–Lesson 15. Lesson 15 shows one of the largest differences; Heather's students spent 312 min on this lesson on

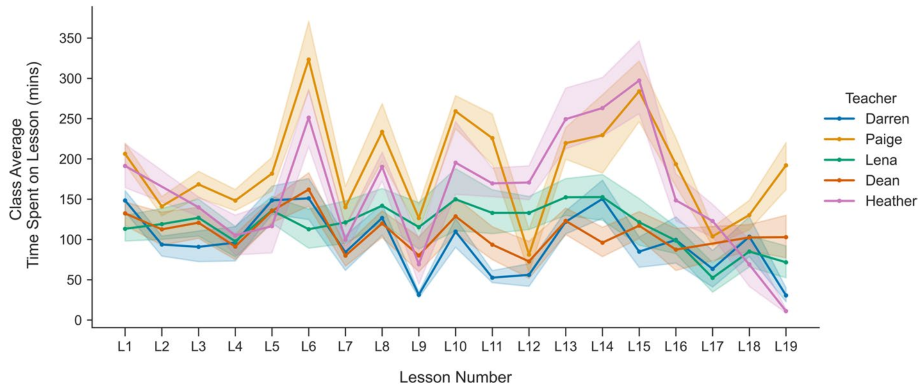

Fig. 2 Differences in average time spent on each lesson between teachers. *Note* Error bars represent the 95% CI for the mean

Table 3 Kruskal–Wallis test comparing median time spent on the lesson between teachers

Lesson	Median time spent on the lesson (mins)					H	df	*p* value
	Darren	Dean	Heather	Paige	Lena			
Lesson 1	154.21	120.79	196.68	207.83	73.66	39.65	4	<.001
Lesson 2	95.00	111.03	NA	131.20	113.23	8.58	3	.036
Lesson 3	73.95	128.66	136.80	164.50	103.27	12.45	4	.014
Lesson 4	115.58	96.93	109.15	150.47	90.66	12.57	4	.014
Lesson 5	149.07	145.44	146.90	190.46	137.67	10.29	4	.036
Lesson 6	160.13	175.04	250.28	312.42	89.43	51.74	4	<.001
Lesson 7	89.50	55.91	90.42	157.63	95.87	9.01	4	.061
Lesson 8	152.03	117.30	184.85	221.47	129.79	27.00	4	<.001
Lesson 9	31.00	48.75	82.30	125.01	65.38	24.76	4	<.001
Lesson 10	126.35	112.71	199.47	257.38	99.40	26.43	4	<.001
Lesson 11	49.31	61.43	154.25	221.08	116.58	27.93	4	<.001
Lesson 12	62.06	60.78	167.08	63.20	135.16	23.61	4	<.001
Lesson 13	123.79	127.35	229.03	212.62	131.77	34.04	4	<.001
Lesson 14	141.23	89.73	249.92	163.89	123.18	29.73	4	<.001
Lesson 15	82.05	118.88	268.35	312.48	81.15	43.61	4	<.001
Lesson 16	115.70	94.82	129.78	193.27	91.35	25.79	4	<.001
Lesson 17	37.38	NA	115.50	94.42	25.36	22.68	3	<.001
Lesson 18	101.05	105.35	37.80	125.84	100.28	10.20	4	.037
Lesson 19	28.49	73.75	11.09	203.89	60.53	34.17	4	<.001

NA represents a lesson that a teacher skipped or otherwise did not have data. They were removed from the comparison for that lesson, reflected in the df for that test

average, while Darren's students spent an average of only 82 min. As observed for the percentage of lessons completed, Heather and Paige's students typically spent more time on each lesson than the other three teachers.

Kruskal–Wallis Tests comparing the median time spent on the lesson between the teachers for each lesson confirmed the differences in time spent based on the teacher. Significant differences were found in Lessons 1–6 and Lessons 8–19 (see Table 3 for *p* valuess), representing a significant difference in the median time spent on the lesson for at least one teacher. Lesson 7 was the only lesson that did not show a significant difference. The results indicate that students spent significantly more or less time on the lesson material depending on the teacher teaching the course.

These results again reflect a difference in students' behaviors depending on the teacher. More time spent does not necessarily equate to higher quality engagement, but it does evidence a difference in students' interactions with the material based on the teaching strategies of their teacher.

Qualitative evidence of teacher impact

Qualitative evidence obtained through semi-structured interviews suggested that the observed differences between teachers was due to the teacher's instructional strategy. We examined this more closely for Heather and Darren, who are on opposite ends of the distribution of time spent and percentage complete. Heather's class completed more of the lessons and spent more time on Scoutlier, while Darren's class completed less of the lessons and spent less time on the platform. In qualitative interviews, Heather describes feeling that students should complete all parts of every lesson. For example, when her students video recorded their answers, "most of the time [the students] didn't want to record themselves…when it came to the final project, I told them that they had to have a video recording with themselves on video". Heather requires her students to follow the lesson as written, even for lessons and tasks the students would possibly skip if on their own. Her teaching philosophy is reflected in her students completing almost 100% of every lesson.

Darren, on the other end of the spectrum, utilized more flexibility in the lesson. He also brought up video recordings, but stated that, "a lot of students didn't want to record a video…and we made the decision to not record videos at a certain point, as a result of that". Being open to skipping certain steps in the lesson resulted in his students completing less of the lesson and spending less time on the lesson overall. Students who would have been required to complete a video if in Heather's class would be allowed to skip the video if in Darren's class. These insights support the notion that differences in student interactions with digital learning platforms across classrooms represent not only individual student variations but are significantly influenced by the teaching strategy of their teacher. While qualitative examples focus on two teachers for illustrative purposes, similar variations in pedagogical strategy were consistently observed across all participating teachers, underscoring broader instructional trends.

Cluster analysis of student engagement

Findings on the differences in engagement between classrooms were then carried into machine learning model development. A k-means cluster analysis with two variables was developed to identify groups of students based on the time they spent engaging with the lesson and the percentage of the lesson they completed for each lesson. In this context, clusters identify students who are highly engaged, defined as high time spent and large percentage complete, from those who are less engaged, low time spent and low percentage complete. Cluster results are shown for two methods of data pre-preparation: (1) standardizing scores across all students in the data, and (2) standardizing scores instead within each classroom. We focus here on Lesson 6 and Lesson 15 as representative lessons, though similar patterns were found in all 19 lessons.

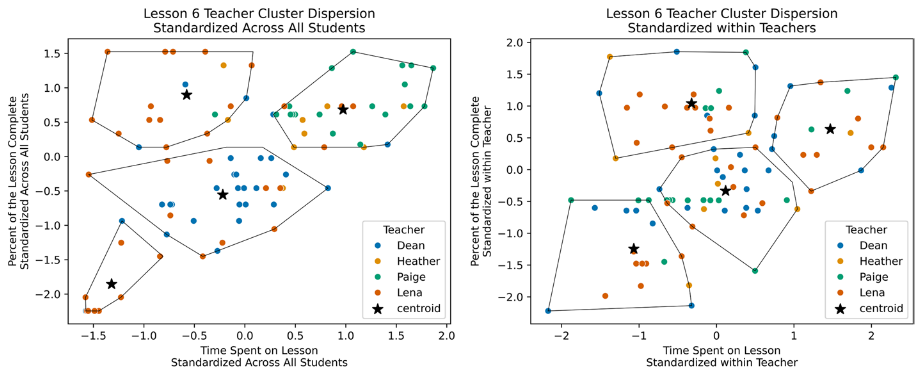

Fig. 3 Lesson 6 cluster dispersion for all standardized versus teacher standardized methods *Note* Lines represent cluster boundaries

Table 4 Lesson 6 cluster dispersion by teacher comparing all student standardized data versus teacher standardized data

Teacher	Standardization method	Students per cluster n (%)				Total
		Cluster 0	Cluster 1	Cluster 2	Cluster 3	
Darren	All students	2 (18.2)	7 (64.6)	2 (18.1)	0 (0.0)	11 (100.0)
	Within teacher	1 (9.1)	4 (36.4)	3 (27.3)	3 (27.3)	11 (100.0)
Dean	All students	3 (10.7)	21 (75.0)	2 (7.1)	2 (7.1)	28 (100.0)
	Within teacher	5 (17.9)	12 (42.9)	6 (21.4)	5 (17.9)	28 (100.0)
Heather	All students	2 (7.4)	1 (11.1)	6 (66.7)	0 (0.0)	9 (100.0)
	Within teacher	1 (11.1)	4 (44.4)	1 (11.1)	3 (33.3)	9 (100.0)
Paige	All students	1 (5.3)	0 (0.0)	18 (94.7)	0 (0.0)	19 (100.0)
	Within teacher	3 (15.0)	9 (45.0)	4 (5.0)	4 (5.0)	20 (100.0)
Lena	All students	13 (30.2)	9 (26.5)	4 (11.8)	8 (23.5)	34 (100.0)
	Within teacher	8 (23.5)	8 (23.5)	7 (20.6)	11 (32.4)	34 (100.0)

Lesson 6

The first cluster analysis of Lesson 6 used a standard pre-processing method of standardizing scores over the full sample to determine whether clusters would be able to separate students within the same class or pick up only on differences between teachers. The graph in Fig. 3 (Left) displays the dispersion of teachers' students across clusters when data are standardized across the full sample. The black bounds represent the bounds of each cluster, the stars represent cluster centers, and the colors indicate the teacher of the student.

In this standard pre-processing, students of the same teacher are clumped together in one cluster or another. The pattern is especially apparent in Dean (blue) and Paige (green). The model has grouped 64.6% of Dean's students in cluster 1 and has not assigned any of his students to cluster 3 (Table 4). Almost all of Paige's (green) students are in cluster 2 (94.7%), and no students have been assigned to cluster 1 or cluster 3. These clusters favor one teacher or another and result in a clustering model grouping students based on apparent large differences in engagement predominantly driven by the instructional approaches of the teachers, rather than identifying student differences within each classroom.

The second cluster analysis for Lesson 6 instead standardized students within each teacher to determine if eliminating the between-teacher differences would allow the model to better separate students within each class. Figure 3 (right) shows marked improvement in the dispersion of each teacher's students across the four clusters. Dean's (blue) students are now dispersed across all four clusters with 17.9%, 42.9%, 21.4% and 17.9% of students in each cluster, respectively (Table 4). The pattern holds for Paige (green), with students now distributed across all four clusters with 15.0%, 45.0%, 5.0%, and 5.0% in each cluster, respectively. This normalization method removed the large differences between teachers and allowed the model to pick up student differences within each classroom. These clusters now represent students who are high and low engagers in their class and is evidence of a model that can distinguish students who have high and low engagement relative to the other students in their class.

To translate the teacher-standardized clusters into interpretable teaching recommendations, the cluster representations (centroids) were fed into an LLM (GPT-4) to create interpretable labels and cluster characteristics of each cluster. The LLM characterized the four clusters as "Struggling or Disengaged Learners", represented by below-average time spent and lesson completion, "Slow but Persistent Learners", students who spend more time on the lesson but do not complete the most, "Efficient and Engaged Learners", represented by moderate completion and moderate time investment, and "Distracted or Unmotivated Learners", represented by spending a moderate amount of time but not completing much of the lesson (Table 5). Because the scores are standardized within each teacher's classroom, these represent students who fit the cluster characteristics, not across the full dataset, but compared to the other students in their class. While cluster interpretation always has a subjective element, the names and characteristics output by GPT-4 were validated by domain experts.

Lesson 15

To confirm the results, the cluster analysis of percentage of the lesson completed and time spent on the lesson was then repeated on Lesson 15. As before, the first model standardized scores across the full dataset. With this standardization method, teachers' students are again grouped around one cluster or another (Fig. 4, Left). Heather's (orange) students were only assigned to two clusters, cluster 0 (44.4%) and cluster 2

Table 5 LLM cluster interpretations for Lesson 6

Cluster	Cluster name	Cluster description
Cluster 0	Struggling or disengaged learners	Below-average time spent and lesson completion in their classroom, indicating possible difficulty with the material or lack of engagement
Cluster 1	Slow but persistent learners	These students spend significantly more time on the lesson but do not have the highest completion rates, suggesting difficulty in processing content
Cluster 2	Efficient and engaged learners	These students complete a large portion of the lesson with a moderate time investment, suggesting they grasp concepts quickly
Cluster 3	Distracted or unmotivated learners	These students spend a moderate amount of time on the lesson but do not complete much of it, which may indicate distractions or lack of motivation

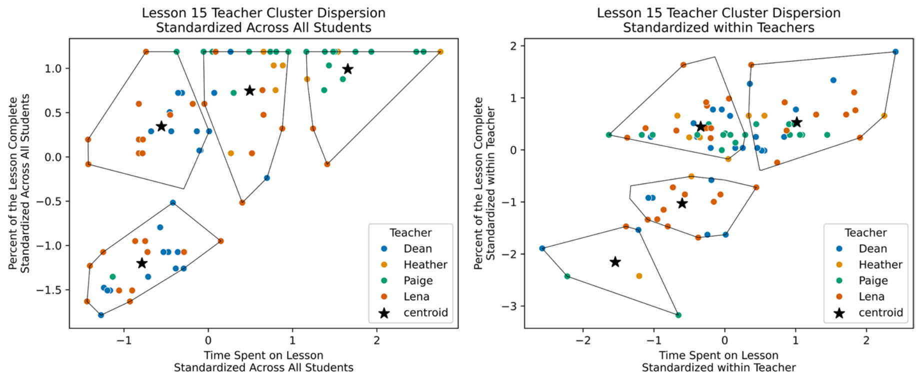

Fig. 4 Lesson 15 cluster dispersion for all standardize versus teacher standardized methods. *Note* Lines represent cluster boundaries

Table 6 Lesson 15 cluster dispersion by teacher comparing all student normalized data versus teacher normalized data

Teacher	Standardization method	Students per cluster n (%)				Total
		Cluster 0	Cluster 1	Cluster 2	Cluster 3	
Darren	All students	0 (0.0)	0 (0.0)	9 (81.8)	2 (18.2)	11 (100.0)
	Within teacher	2 (18.2)	1 (9.1)	5 (45.5)	3 (27.3)	11 (100.0)
Dean	All students	0 (0.0)	3 (12.0)	13 (52.0)	9 (36.0)	25 (100.0)
	Within teacher	9 (36.0)	5 (20.0)	2 (8.0)	9 (36.0)	25 (100.0)
Heather	All students	4 (44.4)	5 (55.6)	0 (0.0)	0 (0.0)	9 (100.0)
	Within teacher	2 (20.0)	3 (30.0)	3 (30.0)	1 (10.0)	9 (100.0)
Paige	All students	11 (55.0)	7 (35.0)	1 (5.0)	1 (5.0)	20 (100.0)
	Within teacher	6 (30.0)	5 (25.0)	7 (35.0)	2 (10.0)	20 (100.0)
Lena	All students	2 (6.7)	7 (23.3)	11 (36.7)	10 (33.3)	30 (100.0)
	Within teacher	9 (30.0)	10 (33.3)	1 (3.3)	10 (33.3)	52 (100.0)

(55.6%), and no students were assigned to cluster 1 or cluster 3. Similarly, Dean's (blue) students were only assigned to cluster 1 (12.0%), cluster 2 (52.0%), and cluster 3 (36.0%), while no students were assigned to cluster 0. This model again picked up on the differences between teachers and assigned cluster membership based on relative classroom engagement rather than differences between students in the same classroom.

To determine if we could improve the cluster results, the model was fit a second time on scores standardized within each teachers' class. The second model again showed large improvement, with students within each classroom more evenly distributed across the four clusters. In Fig. 4 (right) each teacher's students are shown dispersed across the four clusters. Dean's (blue) students are now spread across all four clusters, with 36.0%, 20.0%, 8.0%, and 36.0% of students in each cluster, respectively (Table 6). Heather's (orange) students are also spread across all four clusters, with 20.0%, 30.0%, 30.0%, and 10.0% of students in the four clusters, respectively.

Cluster representations (centroids) from Lesson 15 were then fed into GPT-4 to create interpretable cluster meanings that could be integrated into an online platform. The model labeled the four clusters as, "Steady Strivers", represented by spending a relatively long time on the lesson and completing a large portion of the lesson, "Quick and Curious", who spend less time on the lesson but still complete a moderate portion, "Wandering Learners", who spend an average amount of time by complete less of it, and "Lost and Overwhelmed", distinguished as spending little time on the lesson and completing little (Table 7). As these cluster descriptions are based on the teacher-standardized values, they represent students who are excelling or falling behind in each class and are useful for teachers to identify who is engaged or disengaged in their classroom. Cluster names and interpretations were again validated by domain experts with teaching expertise.

Discussion

Across nearly every lesson in our 19-unit data science curriculum, students' behavioral engagement, operationalized as time on task and percentage of the lesson completed, varied significantly by teacher. Based on a deeper dive into two teachers who varied significantly in how they delivered instruction through a digital learning platform, these differences were substantial and supported by post-instruction interviews. These results demonstrate that, in the smart classroom, students' behavioral engagement is significantly influenced by their teacher. Students in classrooms whose teachers follow structured use of the digital platform completed significantly more of the material compared to students whose teachers used the platform only as a loose guide to their instruction. As a result, when clustering analysis was standardized across all classrooms, these differences in apparent student engagement between teachers largely swamped out any meaningful differences between student engagement within that teacher's classroom. This type of analysis provides very little insight that is useful and actionable at the individual classroom level.

In contrast, when clustering analysis removed differences between teachers by standardizing scores within each teacher's classroom, engagement patterns were revealed among students within the same instructional context. This distinction is critical as it provides information on learning and possible interventions that are actionable by the teacher and targeted to specific groups of students in their classroom. Indeed, we were able to show that the output of "teacher centered" engagement modeling can be passed through GPT-4 to translate cluster centers into human-readable engagement profiles and recommendations. Cluster interpretations provided by GPT-4 were reviewed by domain experts (curriculum developers and qualitative researchers), who validated their educational relevance. Using GPT-4 over a

Table 7 LLM cluster interpretations for Lesson 15

Cluster	Cluster name	Cluster description
Cluster 0	Steady strivers	These students spend significantly more time on the lesson and complete a large portion of it. They are persistent learners who work diligently
Cluster 1	Quick and curious	These students spend less time than average but still complete a moderate portion of the lesson. They may grasp concepts quickly
Cluster 2	Wandering learners	These students spend a typical amount of time on the lesson but complete less of it, possibly due to distractions or motivation issues
Cluster 3	Lost and overwhelmed	These students spend very little time on the lesson and complete almost none of it, suggesting disengagement or struggle

human-in-the-loop interpretation allows these cluster interpretations to be built into software applications that can automatically cluster and interpret student behavior, lightening the load on teachers. These interpretations are presented as exploratory and illustrative, rather than definitive, and provide a foundation for future refinement.

Our findings provide evidence that, in the STEM classroom, the teacher exerts an important, measurable effect on how students interact with digital learning platforms and that this significantly impacts how we should model and interpret individual student engagement with learning. This study adds to a growing body of work that seeks to understand student engagement during learning as a multi-dimensional, dynamic, and contextually embedded process through their interactions with digital learning platforms (Dewan et al., 2019). Our focus is on STEM classrooms—and similar physical learning environments—that blend these digital learning platforms with a wide range of learning materials and technologies, as well as in-person and interpersonal instruction (Shreeve et al., 2025). These are settings where students can build awareness, agency, and access to emerging technology fields that will influence their future lives and careers. In STEM learning environments, student interactions with digital learning platforms and engagement with learning is influenced by factors that are not present in more static, digital-only settings that have traditionally been the focus of this research (Liu, 2022; Cetintas et al., 2010; Cocea & Weibelzahl, 2009).

This teacher effect highlights a critical challenge in applying machine learning and clustering techniques to classroom-based education data: the inherent complexity of the physical classroom. Smart classrooms are not simply digital environments—they are rich, interactive ecosystems shaped by teacher expectations, instructional strategies, classroom norms, and peer dynamics (Love et al., 2023). Understanding the factors that influence patterns in our data can lead to substantial improvements in our data science models and the technology developed for smart classrooms. This technique allows teachers to see which students have high and low engagement in their classrooms and better tailor interventions for individual students.

Our findings underscore the need to reframe how engagement is modeled in smart classrooms. Traditional approaches often center on individual learner characteristics—such as prior knowledge or attention span—and have not accounted for the context of in-person learning. However, our study suggests that such models must evolve to include systemic and contextual influences, especially in smart classrooms where teaching style and classroom culture play a pivotal role in shaping student engagement with digital tools.

Though this is a single case study, this pattern likely extends throughout technologies that are used cross-classrooms. By introducing teacher-aware modeling techniques—such as within-classroom standardization—we not only improve the technical performance of engagement clustering but also take a meaningful step toward more context-sensitive and learning analytics. This shift from student-centric to classroom-situated engagement models allows us to identify and support learners within the instructional environments where their behavior actually unfolds. Practically, teachers should intentionally align digital platform usage with instructional goals and classroom needs. Educational technologists should ensure platforms are sufficiently adaptable to diverse teaching methods. Developers are encouraged to incorporate design flexibility to effectively support varied instructional approaches and classroom contexts.

Limitations and future research

While this study offers valuable insights into how teacher influence shapes student engagement in smart STEM classrooms, several limitations should be considered when interpreting the findings. First, the study was conducted with a relatively small and regionally concentrated sample—five teachers across six schools in one geographic area. Although the sample included a range of school types and student demographics, its scope limits generalizability. Our findings are exploratory and primarily generalizable within similar high school STEM classroom contexts using comparable digital platforms. Generalization across differing educational contexts, subjects, age groups, or technologies requires additional research and replication. Future research should replicate and expand this work across broader geographic, cultural, and institutional contexts to determine the extent to which teacher effects on engagement are consistent across diverse learning environments.

Second, our engagement analysis focused exclusively on behavioral indicators—specifically, time on task and percentage of the lesson completed—derived from digital log data. While these metrics offer useful snapshots of student activity, they do not fully represent the multidimensional nature of engagement, which also includes cognitive, emotional, and agentic dimensions (Reeve, 2013; Reschly & Christenson, 2022; Sinatra et al., 2015). Future studies should integrate richer, multimodal data sources—such as student reflections, classroom video, real-time affective sensing, or patterns of peer interaction—to create a more comprehensive model of engagement.

Finally, while our clustering models demonstrated the value of within-classroom standardization, they relied on only two behavioral features. There is considerable opportunity to extend this modeling approach by incorporating additional engagement signals and exploring alternative machine learning techniques capable of adapting to the evolving context of classroom instruction. Doing so would support the development of predictive models that are not only technically robust, but also context-aware and responsive to the dynamic nature of teaching and learning.

Conclusion

As STEM classrooms increasingly integrate digital platforms and smart technologies, there is a growing need to rethink how student engagement is modeled in these complex, blended learning environments. Traditional approaches, often developed in online or individual learning contexts, tend to overlook the layered realities of in-person instruction, where engagement is shaped not only by individual traits or system design, but also by the instructional strategies and classroom culture established by the teacher.

This study offers a critical perspective on that dynamic. Drawing on both quantitative metrics and qualitative teacher interviews, we found that students' behavioral engagement with digital materials varied significantly based on the instructional context provided by their teacher. These differences were consistent across lessons and aligned closely with each teacher's pedagogical approach, highlighting the profound influence of teacher behavior on how students engage with digital tools in the classroom.

Our findings demonstrate that accounting for these classroom-level dynamics—specifically through within-teacher standardization—enables machine learning models to uncover more meaningful and actionable patterns of student engagement. Rather than comparing students to a generic benchmark, this approach provides insight into how individual learners are engaging relative to their peers in the same instructional environment. This shift enables more targeted support for students and more relevant feedback for educators. Our results reinforce theoretical assertions that behavioral engagement measures, like task completion and time-on-task, are particularly sensitive to instructional contexts. Conversely, cognitive or emotional dimensions may differ in their responsiveness to teacher practices, an area warranting further exploration.

The study offers both a methodological innovation and a conceptual reframing. Methodologically, it introduces a more context-aware and equitable approach to clustering engagement data. Conceptually, it challenges the field to move beyond static or universal models of engagement and toward more nuanced, classroom-situated frameworks that reflect the lived dynamics of teaching and learning in smart STEM classrooms. Ultimately, this work contributes to the development of more accurate, equitable, and responsive systems that support both students and educators in modern STEM classrooms.

As digital technologies continue to reshape the educational landscape, engagement models must evolve in parallel. Effective systems will need to do more than track activity—they must interpret behavior in context, recognize the influence of instruction, and deliver insights that support real-time decision-making. By centering teacher influence in the modeling of engagement, this work contributes to the design of smarter, more responsive educational tools that empower teachers and help all students thrive. Future research should examine correlations between identified teacher-adjusted engagement patterns and academic outcomes, including achievement and STEM persistence. Investigations into how professional development influences teachers' utilization of engagement analytics, and identifying which behavioral indicators best predict long-term student success, would significantly enhance practical applicability.

Abbreviations

DSAIY	Data Science, AI, and You
LLM	Large Language Model
STEM	Science, Technology, Engineering, and Math
GPT	Generative Pre-Trained Transformer

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40561-025-00405-1.

Supplementary material 1.

Supplementary material 2.

Acknowledgements

We would like to thank the district administrators and teachers that participated in our research program, as well as the students who took on the might task fo learning about Data Science, Machine Learning and AI. We would also like to thank Dr. Anushree Bopardikar (TERC, Cambridge, MA) for her valuable assistance collecting and analyzing classroom implementation data.

Author contributions

KS performed the statistical analysis and data modeling for this paper and was a major contributor in writing the manuscript. AP provided guidance on the interpretation of student engagement in STEM classrooms. MC led the qualitative analysis of teacher approaches to deliver the DSAIY curriculum. KJE coordinated work under the NSF grant that supported this study. BP and BJ designed the DSAIY curriculum and led training to prepare teachers to implement it in their classrooms. LH led the analysis of student engagement through interactions with Scoutlier and contributed to writing the manuscript. All authors read and approved the final manuscript.

Funding
This study is supported by the NSF under Grant No. 2148451. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

Availability of data and materials
Because students, parents, and teachers consented that no raw data would be made public, we cannot provide data. Under our IRB policy, the raw data is for our project team. Scoutlier is a free-to-use digital learning platform and can be accessed at scoutlier.com. The DSAIY curriculum can be accessed on Scoutlier upon request.

Declarations

Competing interests
The authors declare that they have competing interests.

Received: 9 April 2025 Accepted: 21 July 2025
Published online: 10 September 2025

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I, Aleman, F.L., Almeida, D., & Zoph, B. (2023). Gpt-4 technical report. *arXiv preprint*. https://doi.org/10.48550/arXiv.2303.08774.

Altuwairqi, K., Jarraya, S. K., Allinjawi, A., & Hammami, M. (2021). Student behavior analysis to measure engagement levels in online learning environments. *Signal, Image, and Video Processing, 15*, 1387–1395. https://doi.org/10.1007/s11760-021-01869-7

Cetintas, S., Lou, S., Ping Xin, Y. P., & Hord, C. (2010). Automatic detection of off-task behaviors in intelligent tutoring systems with machine learning techniques. *IEEE Transactions on Learning Technologies, 3*(3), 228–236. https://doi.org/10.1109/TLT.2009.44

Cocea, M., & Weibelzahl, S. (2009). Log file analysis for disengagement detection in e-Learning environments. *User Model User-Adapted Interaction., 19*, 341–385. https://doi.org/10.1007/s11257-009-9065-5

Darner, R., &, Boesdorfer, S. (2022). Online science teacher preparation and nontraditional recruitment: New strategies for addressing the STEM teacher shortage? *Innovations in Science Teacher Education*, *7*(2). Retrieved from https://innovations.theaste.org/online-science-teacher-preparation-and-nontraditional-recruitment-new-strategies-for-addressing-the-stem-teacher-shortage/

Dettori, J. R., Norvell, D. C., & Chapman, J. R. (2018). The sin of missing data: Is all forgiven by way of imputation? *EBSJ Special Section: Science-in-Spine, 8*(8), 892–894. https://doi.org/10.1177/2192568218811922

Dewan, M. A. A., Murshed, M., & Lin, F. (2019). Engagement detection in online learning: A review. *Smart Learning Environments*. https://doi.org/10.1186/s40561-018-0080-z

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*(1), 59–109. https://doi.org/10.3102/00346543074001059

Gupta, C., Jogdand, D. S., & Kumar, M. (2022). Reviewing the impact of social media on the mental health of adolescents and young adults. *Cureus*. https://doi.org/10.7759/cureus.30143

Hall, N. C., Hladkyj, S., Perry, R. P., & Ruthig, J. C. (2004). The role of attributional retraining and elaborative learning in college students' academic development. *The Journal of Social Psychology, 144*(6), 591–612. https://doi.org/10.3200/SOCP.144.6.591-612

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics, 95*(7–8), 798–812.

Hudson, R. A., Mojica, G. F., Lee, H. S., & Casey, S. (2024). Data moves as a focusing lens for learning to teach with CODAP. *Computers in the Schools*. https://doi.org/10.1080/07380569.2024.2411705

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*. https://doi.org/10.1145/331499.331504

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning. *Springer*. https://doi.org/10.1007/978-1-0716-1418-1

Jessen Eller, K. (2025). Data science, AI, and you in healthcare: Teaching students about medical data bias. Concord Consortium. Retrieved March 28, 2025, from https://concord.org/newsletter/2025-winter/data-science-ai-and-you-in-healthcare-teaching-students-about-medical-data-bias/

Jessen Eller, K., Cassidy, M., Bopardikar, A., Price, E., Jackson, B., Henrik, L., Gichoya Wawira, J., & Celi L. A. (2024). *Hive learning ecosystem: Introducing high school students to rapidly evolving cross-disciplinary skills and careers in data science, AI, and healthcare* [Conference presentation]. National Science Teaching Association (NSTA), NSTA National Conference on Science Education, New Orleans, LA, United States.

King Rice, J. (2010). The impact of teacher experience examining the evidence and policy implications. In *ERIC* (pp. 1–8). National Center for Analysis of Longitudinal Data in Education Research, Urban Institute. https://files.eric.ed.gov/fulltext/ED511988.pdf

Larson, A., Price, E., Jackson, B., Hendrik L., Cassidy M., Celi, L.A., & Jessen Eller, K. (2025). Increasing data literacy in high school students through data science, AI and healthcare. *The Science Teacher*, Submitted.

Lei, H., Cui, Y., & Zhou, W. (2018). Relationships between student engagement and academic achievement: A meta-analysis. *Social Behavior and Personality, 46*(3), 517–528. https://doi.org/10.2224/sbp.7054

Liu, R. (2022). Data analysis of educational evaluation using k-means clustering method. *Computational Intelligence and Neuroscience*. https://doi.org/10.1155/2022/3762431

Lloyd, S. P. (1957). Least squares quantization in PCM. *IEEE Transactions on Information Theory., 28*(2), 129–137. https://doi.org/10.1109/TIT.1982.1056489

Love, H. B., Dickmann, E. M., & Fisher, E. R. (2023). What is an "ArchinatorTM?" A paradigm shift in teaching, facilitation, and learning: The impact of different types of coursework expectations on classroom network structures. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0288136

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297). University of California Press.

Means, B. M., & Stephens, A. (Eds.). (2021). *Cultivating interest and competencies in computing: Authentic experiences and design factors*. National Academies Press. https://doi.org/10.17226/25912

Meyer, A., Rose, D. H., & Gordon, D. (2014). *Universal design for learning: Theory and practice*. CAST Professional Publishing.

Monkaresi, H., Bosch, N., Calvo, R. A., & D'Mello, S. K. (2017). Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Transactions on Affective Computing, 8*(1), 15–28. https://doi.org/10.1109/TAFFC.2016.2515084

NAEP Long-Term Trend Assessment Results: Reading and Mathematics. The Nation's Report Card. (2022). https://www.nationsreportcard.gov/highlights/ltt/2022/

Ostertagova, E., Ostertag, O., & Jozef, K. (2014). Methodology and application of the Kruskal–Wallis test. *Applied Mechanics and Materials, 611*, 115–120. https://doi.org/10.4028/www.scientific.net/AMM.611.115

Pantic, M., & Rothkrantz, L. J. M. (2015). Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Transactions on Affective Computing, 6*(1), 15–28. https://doi.org/10.1109/T-AFFC.2014.2352715

Raugust, K., & Berkman, R. (2022). *K-12 science market outlook*. Simba Information.

Reeve, J. (2013). How students create motivationally supportive learning environments for themselves: The concept of agentic engagement. *Journal of Educational Psychology, 105*(3), 579–595.

Reschly, A. L., & Christenson, S. L. (Eds.). (2022). *Handbook of research on student engagement*. Springer. https://doi.org/10.1007/978-3-031-07853-8

Ruggeri, K., & Folke, T. (2022). Unstandard deviation: The untapped value of positive deviance for reducing inequalities. *Perspectives on Psychological Science, 17*(3), 711–731. https://doi.org/10.1177/17456916211017865

Salloum, S. A., Alomari, K. M., Alfaisal, A. M., Aljanada, R. A., & Basiouni, A. (2025). Emotion recognition for enhanced learning: Using AI to detect students' emotions and adjust teaching methods. *Smart Learning Environments*. https://doi.org/10.1186/s40561-025-00374-5

Saini, M. K., & Goel, N. (2019). How smart are smart classrooms? A review of smart classroom technologies. *ACM Computing Surveys, 52*(6), 1–28. https://doi.org/10.1145/3365757

Shreeve, K., Jessen Eller, K., Cassidy, M., Price, B., Jackson, B., Perry, A. M., Celi, L., Lourentzou, I., & Hendrik, L. (2025). Development of a machine-learning-driven digital teaching assistant that utilises student engagement data to improve access to and success in K-12 STEM education. *Learning Letters*. https://doi.org/10.59453/ll.v4.31

Sinatra, G. M., Heddy, B. C., & Lombardi, D. (2015). The challenges of defining and measuring student engagement in science. *American Psychological Association, 50*(1), 1–13. https://doi.org/10.1080/00461520.2014.1002924

Saldana, J. (2015). *The coding manual for qualitative researchers*. Sage Publications.

Sana MIT Website. (n.d.). https://sana.mit.edu/

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches*. Sage Publications Inc.

Publisher's Note