Investigating the Impact of a Real-time, Multimodal Student Engagement Analytics Technology in Authentic Classrooms

Sinem Aslan[1], Nese Alyuz[1], Cagri Tanriover[1], Sinem E. Mete[2], Eda Okur[1], Sidney K. D'Mello[3], Asli Arslan Esme[1]

[1]Intel Corporation
Hillsboro, OR, USA
{sinem.aslan, nese.alyuz.civitci, cagri.tanriover, eda.okur, asli.arslan.esme}@intel.com

[2]Bahcesehir University
Istanbul, Turkey
sinememine.mete@stu.bahcesehir.edu.tr

[3]University of Colorado Boulder
Boulder, CO, USA
sidney.dmello@colorado.edu

ABSTRACT

We developed a real-time, multimodal Student Engagement Analytics Technology so that teachers can provide just-in-time personalized support to students who risk disengagement. To investigate the impact of the technology, we ran an exploratory semester-long study with a teacher in two classrooms. We used a multi-method approach consisting of a quasi-experimental design to evaluate the impact of the technology and a case study design to understand the environmental and social factors surrounding the classroom setting. The results show that the technology had a significant impact on the teacher's classroom practices (i.e., increased scaffolding to the students) and student engagement (i.e., less boredom). These results suggest that the technology has the potential to support teachers' role of being a *coach* in technology-mediated learning environments.

CCS CONCEPTS

• Applied computing -> Education -> Learning management systems • Human-centered computing -> Human computer interaction (HCI) -> Empirical studies in HCI

KEYWORDS: Learning Analytics, Affective Computing, Student Engagement, Real-time, Dashboards

ACM Reference format:
Sinem Aslan, Nese Alyuz, Cagri Tanriover, Sinem E. Mete, Eda Okur, Sidney K. D'Mello and Asli Arslan Esme. 2019. Investigating the Impact of a Real-Time, Multimodal Student Engagement Analytics Technology in Authentic Classrooms. In *2019 CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland, UK.* ACM, New York, NY, USA. Paper 304, 12 pages. https://doi.org/10.1145/3290605.3300534

CHI 2019, May 4-9, 2019, Glasgow, Scotland, UK

ACM ISBN 978-1-4503-5970-2/19/05...$15.00
https://doi.org/10.1145/3290605.3300534

1 INTRODUCTION

Students' increasing access to technology in classrooms brings forth new challenges for teachers. How can a teacher act as a *coach,* supporting student learning in technology-mediated learning environments? How can a teacher ensure that students are engaged while learning from technology? And how can a teacher identify the students in need of help and provide just-in-time personalized support to each of them?

Related research shows that providing real-time learning analytics to teachers facilitates their instrumental and emotional support to students in classrooms [1, 2], which in turn improves students' engagement, experience, and performance [3, 4]. Student engagement is an important factor for teachers to consider when personalizing learning experience [5, 6, 7, 8], as it is linked to major educational outcomes such as persistence, satisfaction, and academic achievement [6]. Towards this end, we developed a multimodal system – Student Engagement Analytics Technology (SEAT) – to detect students' engagement-related states in real-time and provide this information to teachers so they can implement just-in-time personalized interventions. We discuss the technology and an initial evaluation study in this paper.

2 BACKGROUND AND RELATED WORK

2.1 Modeling Engagement

Contemporary researchers adopt a multi-componential perspective for engagement and have operationalized person-oriented engagement in terms of affective, cognitive, and behavioral states [9]. The majority of the research investigating student engagement during learning is based on unimodal systems relying on either sensor-free data collected from student interactions and event logs or on unimodal physiological and behavioral data (see the review by [9]).

Multimodal approaches are quite rare though ostensibly quite beneficial since leveraging students' appearance (e.g., facial expressions) in addition to other contextual cues such as learning progress can be used to make better judgments about their engagement [10].

In contrast to the unimodal intelligent systems, there are some promising research studies taking a multimodal approach. One such example is the affect-sensitive AutoTutor, which tracks students' affective states using information derived from log files, facial expressions, and body posture to generate a variety of dialog moves including motivational feedback to complement pedagogical actions [11].

Using a similar multimodal approach, we developed machine learning models that use student appearance and interaction logs as two modalities to detect engagement in real-time [12, 13, 14]. For model development, we collected 500+ hours of multimodal student data from authentic classrooms and rigorously labeled the data using a Human-Expert Labeling Process (HELP) [15]. Our sensing technology is able to support various usage scenarios including providing engagement states to (1) students for improving self-awareness; (2) integrating this information in educational platforms (e.g., Intelligent Tutoring Systems, Learning Management Systems, etc.) for personalizing or evaluating their content; and (3) providing input to teachers for implementing personalized interventions. In this study, we focus on interventions with a human-touch by providing engagement analytics to a teacher to provide just-in-time support to students.

2.2 Empowering Teachers with Engagement Analytics

Our approach is motivated by the assumption that teachers, as the human experts in the feedback loop, can utilize engagement analytics to identify students in need of help and provide instrumental and emotional support (i.e., a human touch ingredient), thereby expanding the support that an intelligent system could provide to a student. To this end, there is some exemplary research using dashboards to provide either offline or real-time analytics to teachers. These studies investigated how much these dashboards were instrumental to instructors of distance courses [16], whether these dashboards could support teachers for facilitating collaborative learning in physical spaces [17] as well as digital platforms [18], and how such dashboards could influence teachers' classroom practice and student learning [19]. The results of these studies indicate that such dashboards could support teachers to understand individual students or groups of students who are in need of teacher support [16, 17]. Teachers can also better diagnose problems concerning participation of students [18] and implement analytics-driven classroom practices (e.g., lesson planning) [19].

Our work adds to the current literature in that we detect comprehensive *behavioral* and *emotional* components of student engagement using *a multimodal* approach, and provide these in *real-time* to teachers for them to implement *just-in-time* personalized interventions in *authentic classrooms*. More importantly, we investigate the impact of the technology on the teacher's classroom practices and students' engagement, experience, and performance in a *semester-long study*.

3 STUDENT ENGAGEMENT ANALYTICS TECHNOLOGY (SEAT)

3.1 Student Engagement Detection Approach

We trained multimodal classifiers on two components of overall engagement. One classifier addresses behavioral engagement (i.e., whether a student is on-task or off-task during learning) and the other assesses emotional engagement (i.e., whether a student is satisfied, bored, or confused during learning). These two classifiers provide an assessment of *observed engagement* through unobtrusive sensing – similar to a teacher observing student engagement in a classroom. We did not train a classifier for cognitive engagement because pertinent information is available in activity logs (see below).

As illustrated in Figure 1, we utilized two modalities for each engagement components. The Appearance modality is based on a video of individual students captured by their built-in device cameras, whereas the Context-Performance modality uses contextual and student-specific performance data provided by the educational platform. Key features from both of these modalities are extracted in 8-second windows to enable real-time detection of both behavioral and emotional engagement: Appearance features are based on facial landmarks defining movements of facial features and head movements, whereas for the Context-Performance, extracted features are related to time spent, number of trials, and hint counts. Classifiers were pre-trained separately for each modality and for each engagement dimension on a dataset from the 9th grade students studying math on the educational platform used in this study. For further details, refer to our previous work – see [12, 13] for the behavioral classifier and [14] for the emotional classifier.

We leveraged these multimodal models, in the Student Engagement Analytics Technology (SEAT) using a *Design Thinking Process* for its user experience design and implementation.

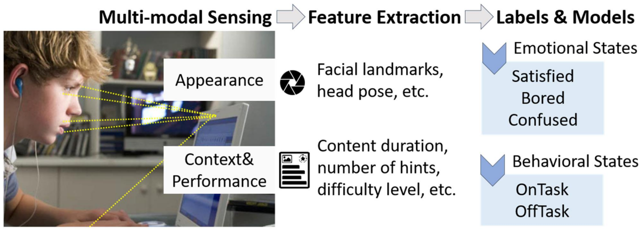


Figure 1. Student engagement detection approach

3.2 Design Thinking Process

To identify the features and functions specific to a real-time analytics dashboard, we utilized the *Design Thinking Process* consisting of the following stages: (1) empathize, (2) define, (3) ideate, (4) prototype, and (5) test [20]. Building on our preliminary literature reviews on features of analytics dashboards, we conducted a half-day workshop with six psychologists with educational experience. In this workshop, we first implemented the *braindump* method where each expert wrote their ideas about potential problems with engagement measurement in classes. We then had these experts *brainstorm* their findings as a team and ideate a number of possible technology solutions for providing engagement analytics to a teacher. Finally, we implemented the *sketchstorm* method to have each expert sketch what an engagement analytics dashboard could look like.

Leveraging findings from the literature reviews and the workshop, we created preliminary wireframes. Using these wireframes, a UI designer implemented the first graphical prototypes of the SEAT UI. We then used these prototypes to conduct a 2-hour workshop with four in-service math teachers. In this workshop, we implemented the *brainstorming* method with some guiding questions to gather data about the current prototypes and potential additional features and functions the teachers would like included.

This process enabled us to finalize the major features and functions of SEAT (e.g., engagement mapping, analytics formulations and representations). We implemented these in a working prototype to test in the field. Through several iterations during testing phase (with formative feedback from an educational researcher actively observing SEAT in classes), we created a final design prototype for evaluation on teachers and students.

3.3 The SEAT UI

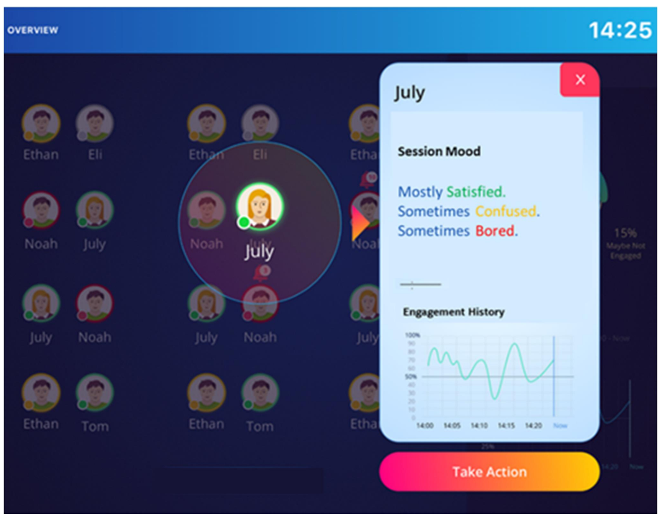
As illustrated in Figure 2, there are two major views in SEAT: (1) the overall class-view, and (2) the student-specific view. In the overall class view (see Figure 2a), a teacher can monitor real-time engagement of all the students in a class via color codes along with a “bell” notification icon on the top of each student’s avatar which indicates their contextual state (i.e., being on-platform vs. off-platform). The mapping used to display engagement states and the corresponding color codes are provided in Table 1, as adopted from [21]. However, in contrast to [21], based on the feedback gathered from the teachers, we changed the mapping of *on-task + bored* from *not engaged* to *maybe engaged* as the teachers indicated that as long as students are on-task, they should not be represented as not engaged.

In the overall class view (see Figure 2a), the teacher can also monitor the class’ overall engagement (e.g., 50% of the class is currently engaged) as well as an engagement history graph, displaying engagement changes of the class on a timeline (since the beginning of the class session). Based on the overall-class analytics, the teacher may click the *Class Intervention* button to log an intervention to the entire class. As shown in Figure 2 (b), when the teacher clicks on a student’s avatar, the teacher can view student-specific analytics with overall session affect (called mood in the figure) of that specific student and the student’s engagement history graph, demonstrating engagement changes of the student since the beginning of the session. Further, based on the student-specific analytics, the teacher may click on *Take Action* button to log an intervention to the student.

Table 1. Engagement Mapping and Color Codes

Contextual State	Behavioral State	Emotional State	Engagement Mapping	Color Code
On-Platform	On-Task	Satisfied	Engaged	Green
On-Platform	On-Task	Bored	Maybe Engaged	Yellow
On-Platform	On-Task	Confused	Maybe Engaged	Yellow
On-Platform	On-Task	Can’t Decide	Maybe Engaged	Yellow
On-Platform	Off-Task		Not Engaged	Red
Off-Platform			Not Engaged	Red (Bell)

(a)



(b)

Figure 2. UIs of the SEAT: (a) Overall class-view and (b) student-specific view

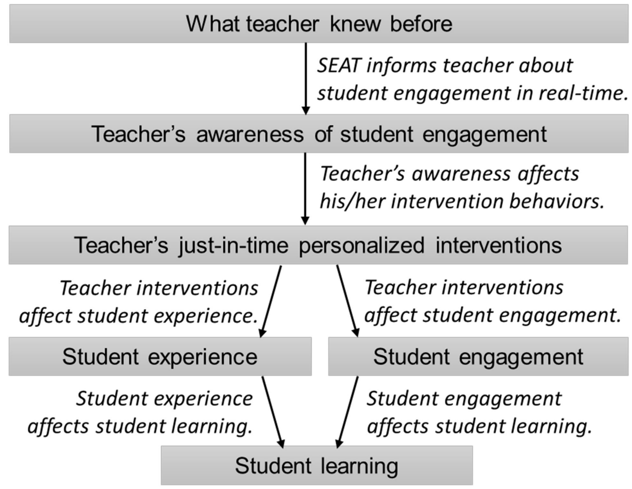


Figure 3. Impact model of the SEAT

4 METHODS

4.1 Research Context and Questions

In technology-mediated learning environments, teachers act as a coach to support students whenever needed while students learn on their own in a self-paced approach through digital content based on a preset curriculum. For such learning environments, we hypothesized an impact model (see Figure 3), inspired from the causal chain model introduced in [19], that captures teacher interventions based on real-time engagement analytics to understand whether and how such interventions have an impact on student experience, engagement, and performance.

To understand the impact of SEAT usage on teachers and students based on this model, we conducted a study in a high school in Turkey - the same country where the models were previously trained. For this study, we chose math as our target subject area. The administrators of the school, suggested an experienced math teacher (i.e., Melissa, pseudonym) to be involved in the study. At the time of the study, Melissa had two 9th grade classes and we randomly assigned one as the *control* class (19 Students: 11 Female, 8 Male) and the other as the *treatment* class (18 Students: 11 Female, 7 Male). Both Melissa and the school administration reported diverse academic achievement in math within each class (i.e., with low, moderate, and high achievers) and similar academic achievement in math across the two classes.

In this study, we aimed to address the following research questions:

1. What was the teacher's experience with the SEAT?
2. Did SEAT usage influence the teacher's classroom practices?
3. Did SEAT usage influence students' engagement, experience, and performance?

4.2 Multi-Method Research Design

We used a multi-method approach consisting of a quasi-experimental design with non-equivalent groups to evaluate the impact of SEAT combined with a single case study design to understand environmental and social factors surrounding the classroom setting. We started with three weeks of *Readiness* sessions for orientation of the teacher and the students. Next, we implemented three weeks of *Baseline* sessions in which the teacher did not use SEAT in any of her classes. The aim of these *Baseline* sessions was to understand any difference between these two classes before any treatment. After the *Baseline* sessions, the students in both of the classes completed a pre-test. In the next six weeks, we implemented *Comparison* sessions where we had the teacher use SEAT in the treatment class but not in the control class. At the end of these six weeks, the students in both of the classes completed a post-test. In the last four weeks, we implemented *Additional* sessions where the teacher used SEAT in both classes with the purpose of gathering additional qualitative data (See Figure 4 for an overview).

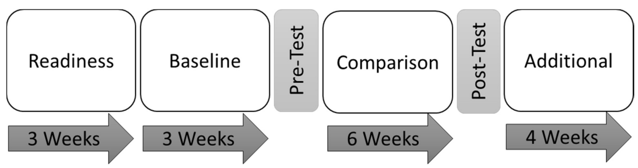


Figure 4. Summary of quasi-experimental research design phases

Because the first three weeks were only used to assess technological readiness, the data from these weeks was not included in the analysis. Moreover, the *Additional* sessions aimed to provide additional qualitative data and hence not included in the statistical analysis. We observed major end-of-semester effects on the students in both of the classes (i.e., attendance issues, distractions due to end of year exams, etc.) during these weeks.

4.3 Data Collection

Data collection for both classes took place on the same day of the week. Each class was around 80 minutes long with a 10-minute break in between two 40-minute sessions. In these classes, the students used their laptop computers to watch math-related instructional videos and solve questions on a publicly available educational content platform (see Figure 5 for the classroom setup). Our data collection applications ran in the background and collected the following information: (1) student appearance videos from built-in device cameras, (2) student screen capture videos, (3) student context and performance data from the educational content platform API, (4) URL logs from browser, (5) a class video from a fisheye camera, (6) class audio from a mic array, and (7) teacher audio from a lapel mic.

Figure 5. Sample classroom-setup views

An educational researcher and a technical support were present during all classes. The educational researcher collected the following additional data sources.

4.3.1. Semi-Structured Teacher Interviews. We conducted three sets of interviews with the teacher to understand her profile as a teacher, her perceived impact of SEAT, and her evaluation of the usability of SEAT.

4.3.2. Semi-Structured Student Interviews. To gather in-depth insights about perceived impact of SEAT from the students, we interviewed six students who volunteered – three from the control class and three from the treatment class.

4.3.3. Semi-Structured In-Class Observations. The educational researcher's semi-structured observations of the students and the teacher included pre-session notes (i.e., noting down any important events in the school, such as exam week) and in-session notes (i.e., observing the students and teacher for any important incidences).

4.3.4. Student Questionnaires. We administered three student questionnaires to identify any baseline differences between the two classes (before any treatment): (1) the Student Math Attitude Questionnaire [22, 23]; (2) the Student Intrinsic Motivation Questionnaire [24, 25]; and (3) the Student Flow Questionnaire [26, 27].

4.3.5. Achievement Tests for Pre-Post Assessment. The teacher selected and customized (e.g., changing numbers on problems) 10-multiple choice and open-ended questions from the educational content platform based on the topics planned to be covered in the weeks between the pre-test and post-test. The difficulty level of the questions were relatively high for the students as they were sampled from the topics that were not yet covered in the school curriculum.

4.4 Data Analysis

We utilized both qualitative and quantitative methods to analyze the collected data.

4.4.1. Observational and Interview Data. All observational notes were consolidated and organized for an in-depth analysis. The researchers transcribed all interviews verbatim. Content analysis was conducted to analyze the data in five major steps: (1) get familiarized with the data, (2) organize the data, (3) conduct initial coding, (4) validate codes (peer-review of the emerged codes and categories), and (5) interpret and report findings (for more details about the procedures implemented, see [28, 29, 30, 31]).

4.4.2. Questionnaire Data. For each questionnaire, we computed a total mean score for each student (negating the scores for reverse items). Additionally, we also computed subscale means (i.e., computing means for related sets of items) when applicable.

4.4.3. Achievement Tests Data. For each student, the difference post-score minus pre-score was computed and compared across classes.

4.4.4. Audio and Video Data. We used the Human-Expert Labeling Process (HELP) [15] to label (i.e., code) the video and audio data for student engagement and teacher interventions. Prior to labeling tasks, six human-experts with psychology/educational psychology background were recruited and trained based on the HELP methods.

The data used for labeling is provided on the study timeline in Figure 6. Due to data collection failures, there were missing weeks for the intervention labels. Moreover, due to the limited labeling resources, we had to select a subset of the data for engagement labeling and we mainly focused on emotional labels.

<u>Student engagement labeling</u>: The experts labeled student data for behavioral engagement (19 hours of data) and emotional engagement (200 hours of data). The dataset was divided into two parts and each part was labeled by three experts to enable reliability checking (i.e., multiple-expert labeling). The experts audio captured from student devices, individual student videos, and their desktop (screenshot) videos for labeling. Labels collected on the same data from each expert were divided into synchronous time segments for further comparison among each other (8-seconds of resolution as used by the engagement models). Majority voting was applied to obtain final behavioral and emotional labels, which then defined the overall engagement labels using the mapping provided in Table 1. Using these final labels, we obtained the durations of being in each state for each student and each week.

<u>Teacher interventions labeling</u>: We divided the 23 hours of data (classroom video with teacher audio) from the two classes into five parts and each expert labeled their assigned part (i.e., single-expert labeling). The experts were provided with a commercially available tool to label teacher interventions using classroom video from the fish-eye camera along with teacher/student audio from the teacher's lapel microphone.

The operational definitions of teacher interventions are provided in Table 2. We computed counts for each-intervention type for each student and for each week.

Table 2. Operational Definitions of Teacher Interventions

Type	Operational Definitions
Verbal Warning	The teacher intervenes with a verbal warning (e.g., "Stop talking with your friends!").
Positive Reinforcement	The teacher provides verbal positive feedback (e.g., "Good job, this week you are doing much better than the last week!").
Scaffolding	The teacher intervenes as she thinks that a student is confused or in need of help during learning (e.g., the teacher explains learning material/question to the student or provides hints to the student to solve a problem).
Close Monitoring	The teacher intervenes by standing by a student and observing him or her (e.g., the teacher attempts to look at the student's screen to see if s/he is on-task. The teacher stays physically close to the student for a while).
Other	If the teacher implements another intervention that does not belong to any of the categories above.

	Baseline				Comparison						
Week ID	4	5	6	Pre-test	7	8	9	10	11	12	Post-test
Intervention Labels		✓			✓	✓		✓	✓	✓	
Behavioral Labels		✓			✓						
Emotional Labels	✓	✓	✓		✓		✓	✓	✓	✓	

Figure 6. Data availability for labeling tasks

5 RESULTS

5.1 Meet the Teacher and the Students

The initial interview with Melissa revealed the following details about her as a teacher (See Figure 7).

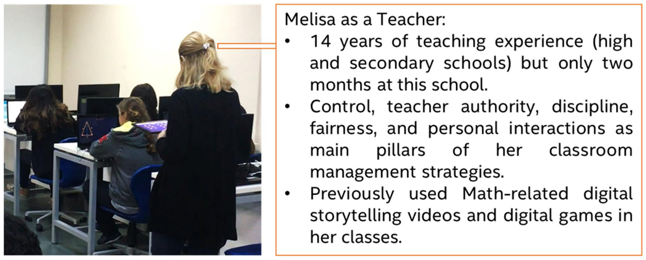


Figure 7. Teacher profile summary

We first analyzed any differences among the two classes with respect to students' attitudes, motivation, and flow. For this purpose, we utilized all three questionnaires and performed statistical analysis using Mann-Whitney U test. We used this test for all subsequent analyses since the data was not normally distributed. We adopt two-tailed tests with a significance criterion (alpha) of .05 for all analyses.

The results of Mann-Whitney U test indicated that there were no statistically significant differences between the two classes, in terms of the students' attitudes towards math, intrinsic motivation, and flow (i.e., all p-values are above 0.05). For motivation, we also checked subcategories focusing on interest, perceived competence, pressure, and perceived choice; and for flow, we checked subcategories of positive flow and worry. None of these subgroups provided any significant difference between the two classes. These findings are well aligned with what was previously reported by the school administrators and the teacher.

As described previously, to gather in-depth insights, we also interviewed six students who volunteered – three from the control class (pseudonyms: Maria, Ellie, Amber) and three from the treatment class (pseudonyms: Anne, Becca, and Judy). Note that based on the results from the Math Attitude Questionnaire, these six students had relatively positive attitude towards math in comparison to the class average (ranging 5%-23% above the average).

5.2 What was the teacher's experience with SEAT?

The in-class observations demonstrated that the teacher, Melissa, proactively used SEAT during her classes. Aligned with these observations, in her interview, Melissa stated that she commonly used all features and functions of SEAT during her classes. Specifically, she mentioned that she regularly checked the *Class Engagement Analytics* display (see Figure 2(a)) to get a sense of overall class engagement. If SEAT indicated a student was off-platform, she immediately implemented a relevant intervention. Similarly, she noted that she closely monitored color changes on the avatars for gauging the need for any other required interventions. She pointed out that she regularly checked the "Session Mood" analytics display of individual students (see Figure 2(b)). Based on these analytics, if she saw any concerning emotions, she implemented a relevant intervention for that student.

In her interview, Melissa also elaborated on her experience with SEAT. She stated that she found SEAT to be very instrumental for her as a teacher, especially when she compared her experience with and without it across the two classes. She noted that SEAT was helpful for actionable engagement analytics, easier real-time class-wide monitoring, and better detection of off-task, confused, or bored students. The detailed findings (based on the categories that emerged from her interview) along with sample quotes are outlined in Table 3.

Table 3. Findings and Sample Quotes from Melissa (Teacher Interview, 2017)

Findings and Sample Quotes
Actionable engagement analytics *"[Engagement analytics provided by SEAT] were spot on [with my in-class observations]… I greatly benefited [from SEAT]. I was able to witness how I could catch some emotional challenges of the students that I could not have anticipated [before]"*
Easier real-time class-wide monitoring *"[O]n the screen [of SEAT], one can monitor the entire class. [With SEAT], I was able to much more accurately detect the students who were confused, … [the students] who acted as if engaged but had a blank look or watching around, or the students who had an intention to open up an irrelevant video in the background …"*
Better detection of off-task, confused, and bored students *"One day, I was at the back of the classroom. I saw that a student was [off-platform] on SEAT – [probably] watching or listening something else. [When warned], he thought that I saw this [on my own] and told me that I had eyes like a hawk. But in fact, that never happened. I saw him [off-platform on SEAT]"* *"[With SEAT], I was able to much more accurately detect the students who were confused…When the students were stuck in understanding something … I went and asked them if there was something wrong. The students said 'I did not understand this' – may be they were too shy to ask. But I saw the students' emotional states on SEAT and had a chance to intervene accordingly"*. *"[With SEAT], I certainly better detected the students who were bored…When I took a closer look at these bored students, there were some students who had a higher tendency to transition to having a bell [go off-platform]"*.

Despite the fact that Melissa found SEAT very instrumental, she also emphasized that computers would not be aware of some important student context affecting their engagement (e.g., family issues, health conditions, disabilities, etc.). She indicated that she triangulated engagement analytics with student context before taking any actions. In her interview, she stated:

"We have to think of all variables [as teachers]. In other words, we need to combine our observations [with engagement analytics] ... The student may ... have a health issue with her eyes, or have an allergy... or she is sleepy, or having emotional challenges that day because of a family issue. At the end of the day, the subject of the task at hand is a human-being" (Melissa, Teacher Interview, 2017).

5.3 Did SEAT usage influence the teacher's classroom practices?

To understand whether SEAT had an impact on the teacher's classroom practices (i.e., interventions), we compared weekly counts of each intervention type across the classes. First, we compared two class averages over the *Baseline* week, where SEAT was not used in either of the classes. Mann-Whitney U test indicated that there was no significant difference between the two classes (p-values: Verbal Warning = 0.194, Scaffolding = 0.252, Close Monitoring= 0.933; no data was available for Positive Reinforcement due to insufficient occurrences).

Next, we compared two class averages over *Comparison* weeks (five weeks of data), where SEAT was used in the treatment class. To compute class averages, the number of interventions given to each student each week is counted, and then the mean over all students and weeks is calculated for each class. The average intervention counts per student per week are provided in Figure 8, separately for each intervention type.

Mann-Whitney U test indicated significant differences for scaffolding interventions ($median_{TREATMENT} = 0.55$, $median_{CONTROL} = 0.25$, $U = 160.5$, $p < 0.05$). This significant difference is also indicated in Figure 8, where the average scaffolding intervention count per student in a session is 2.9 times higher in the treatment class (mean of 0.83) than in the control class (mean of 0.29). The effect size (r = 0.37) calculated as a rank-biserial correlation coefficient suggested a medium effect [32]. This result suggests that SEAT could facilitate the teacher's instrumental and emotional support when the students needed help. These quantitative results are well supported qualitatively by both the teacher and the students in their interviews as noted in the following sample quotes:

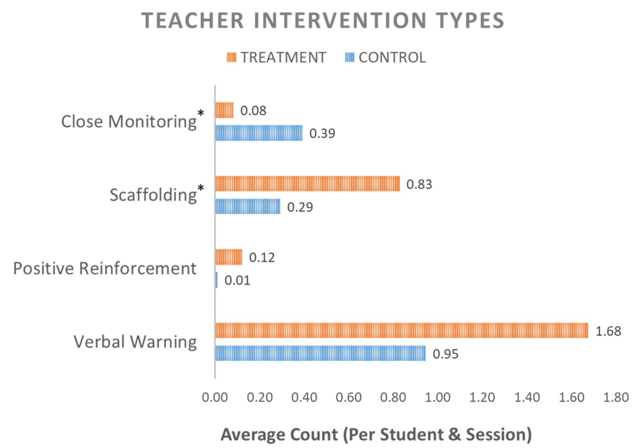


Figure 8. Intervention distributions for Comparison sessions (significant differences marked with *).

"[With SEAT], of course I was able to better detect students who needed help... [W]hile I was unable to discover these in the class where I did not use the SEAT, I detected the students' emotional challenges – confusion ... with the help of the SEAT" (Melissa, Teacher Interview, 2017).

[Before SEAT], sometimes the teacher could not see me as she was taking care of another student regardless of how much I was raising my hand when I did not understand a question since our class has 18 students. Thanks to this technology, the teacher right now understands what I'm doing, ... she can come to my desk immediately" (Anne, Student Interviews, 2017).

The results also showed a significant difference for close monitoring interventions with average counts being smaller in the treatment class (median = 0) than in the control class (median = 0.25) with $U = 244$, $p < 0.05$, $r = -0.43$. This indicates that the teacher was more likely to go around and get closer to the students to check their screens and status in the control class. However, in the treatment class, she seemed to rely more on SEAT's engagement analytics as suggested by the data and by the following sample quote from her interview:

"[SEAT] took the load off my shoulders in the classes... It significantly eased my burden. Obviously, [as teachers] we cannot handle all 20 students in the classroom at the same time...But on one screen [of SEAT], one can monitor the entire class" (Melissa, Teacher Interview, 2017).

In addition, Mann-Whitney U test conducted for the other intervention types showed that there were no significant differences for Positive Reinforcement (p-value = 0.061) or for Verbal Warning (p-value = 0.103) across classes.

5.4 Did SEAT usage influence students' engagement, experience, and performance?

5.4.1. Student Engagement. Since behavioral labels were only available for one *Baseline* week (see Figure 6 for data availability), we only focused on students' emotional states. For each student, we computed the average duration of each emotional state for the *Baseline* and *Comparison* weeks. Then, the change in emotional state durations were computed by subtracting each individual's *Baseline* average from his/her *Comparison* average. The class averages for emotional duration changes are shown in Figure 9. These results indicate a drop in satisfaction and confusion but an increase in boredom in both of the classes. A Mann-Whitney U test results indicated that the difference between the control and treatment classes was significant for boredom ($median_{CONTROL}$= 6.9, $median_{TREATMENT}$= 1.7, U= 282, $p < 0.01$, r = -0.65), but not confused or satisfied (see below for details on students' experience).

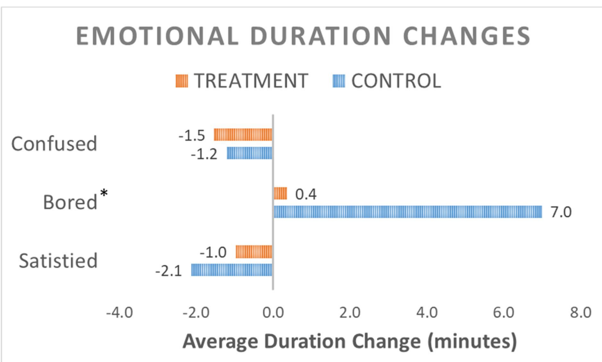


Figure 9. Class averages for emotional duration changes between Baseline and Comparison weeks (significant differences are marked with *).

5.4.2. Student Experience. All of the six students stated that SEAT's usage increased the teacher awareness of student engagement in the class. In other words, they all thought that by using SEAT, the teacher better identified which students were in need of help to provide support accordingly. In her interview, Judy explained this as follows:

"With this technology, our teacher is aware of our [engagement states]…. She can come and intervene to fix any issues. Normally, [the teacher] could not fix [it] because we are 18 students in the class. It was very difficult to pay attention to each of us. With this technology, she can pay attention to each of us individually" (Judy, Student Interviews, 2017).

Additionally, four of the six students (Anne, Becca, Judy, and Ellie) specifically pointed out the fact that SEAT helped the teacher to understand their emotions better in the class and intervene accordingly. Becca stated the following in her interview:

"[With SEAT], she can better understand when I get bored. For example, I was getting bored in some classes, she was coming to my desk immediately and asking me 'is there something wrong?, Is there a problem?'' (Becca, Student Interviews, 2017).

Similarly, three of the six students (Becca, Judy, and Maria) highlighted that SEAT usage increased the teacher's one on one interactions with the students. Becca stated the following:

"After starting using this technology [SEAT], my teacher started to come to my desk more often" (Becca, Student Interviews, 2017).

In their interviews, the students also elaborated on the impact of the teacher interventions as triggered by SEAT usage. Five of the six students (Anne, Becca, Judy, Maria, and Ellie) stated that teacher interventions resulted in increased engagement (as supported by the quantitative results in the previous section) where four of them (Anne, Becca, Judy, and Maria) indicated that it also increased their motivation. In her interview, Anne reported the following:

"After understanding that I am bored [through SEAT], getting the teacher to come to help me enabled me to pull myself together [and become engaged]" (Anne, Student Interviews, 2017).

Although the interview results showed positive student experience with SEAT, two of the students (Maria and Amber) expressed few concerns related to privacy. Maria reported discomfort from being monitored. Similarly, Amber indicated that she felt as if the teacher was always monitoring her but she added that it got better after a while. Contrary to these concerns, one student, Ellie, reported that she did not feel any discomfort of being monitored. She explained the reasons as follows:

"… I was not feeling uncomfortable at all. In fact, I wanted that. I was curious – wishing the teacher see this, do I have a weakness? In which topic? I was curious too. I was curious where I am good where I am not, therefore, I did not have any concerns" (Ellie, Student Interviews, 2017).

5.4.3. Student Achievement Tests. We first conducted a Mann-Whitney U test to compare pre-test scores of the control class and the treatment class (N=37 students). The results showed that there were no significant differences between two classes in terms of prior knowledge

($median_{PRE\_CONTROL}$= 32, $median_{PRE\_TREATMENT}$= 34.5, p-value = 0.564). Next a Mann-Whitney U test comparing the differences between post- and pre-test scores across the two classes showed no significant difference ($median_{DIFF\_CONTROL}$= -8, $median_{DIFF\_TREATMENT}$= 2, p-value = 0.09). That said, the difference was marginal in favor of the treatment class ($mean_{DIFF\_CONTROL}$= -4.8, $std_{DIFF\_CONTROL}$= 13.6; $mean_{DIFF\_TREATMENT}$= 3.4, $std_{DIFF\_TREATMENT}$= 11.6) and the lack of a significant difference might be attributable to our small sample size where power was likely insufficient to detect an effect. The histograms of score differences for the two classes are shown in Figure 10.

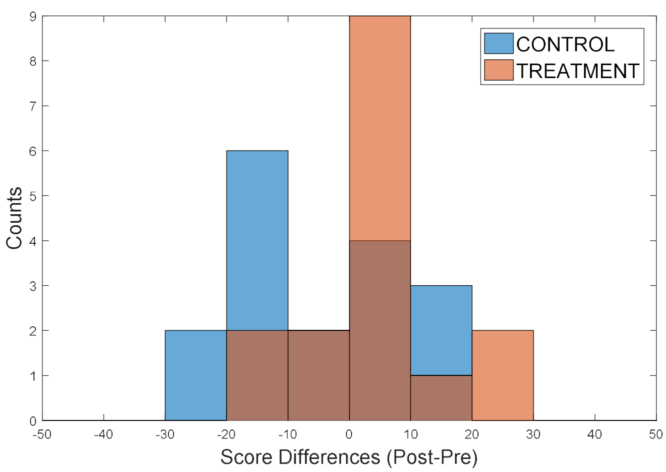


Figure 10. Distribution of achievement tests score differences for two classes are given.

Interestingly, in their interviews, both the teacher and the students reported positive impact of SEAT usage on learning:

"For example, when teaching the line graph [topic] in the class, from both of the classes, I got positive feedback such as: 'we learned this in class ... – we watched this [content]'. I think among all of these students – even if only one told this – it is a very positive feedback. In fact in the [treatment class], there were more students who pointed this out while in the [control class] some noticed this, too..." (Melissa, Teacher Interview, 2017).

"Use of this platform [SEAT] enabled me to be focused and increased my motivation. The teacher was able to see that I was disengaged through the technology, and then she would come up to me, and tell me what she saw and ask me to be more focused. Her feedback helped me to get re-engaged. I really believe that it also improved some aspects related to my performance" (Becca, Student Interviews, 2017).

However, as the teacher indicated further in her interview, there was an important confounding variable, instructional design of the educational content, which seemed to have a significant impact on the current results:

"For performance, I cannot say too much because the topics covered in the class were [very much] stand-alone and there were topics ... taught before besides those that were not. For performance, the things they learnt here would not be enough [without being part of the academic curriculum]. For instance, we selected "functions" as a topic that they have not been taught yet [in the academic curriculum]. We will better understand their comprehension of this topic next year – in other words, [the students will say], 'we covered this topic last year'. I think we planted a seed here, and it will grow next year (Melissa, Teacher Interview, 2017).

6 DISCUSSION

We utilized a multi-method approach consisting of qualitative and quantitative methods to address our research questions pertaining to SEAT. Our results suggest that the teacher implemented significantly more scaffolding in the treatment class than in the control class. Her interviews indicate that she used SEAT to identify key moments (e.g., being confused) to implement relevant scaffolds enabling students to achieve their tasks.

In the traditional sense, one expects a teacher to circulate through the classroom through close monitoring and ask questions to detect moments when students need help and provide scaffolds accordingly [33, 34]. In our case, the use of SEAT required significantly less close monitoring suggesting that the teacher seemed to trust and rely on the analytics in the treatment class.

The results from the student interviews also supported these changes in the teacher's classroom practices. The students indicated that the SEAT enabled the teacher to better identify which students were in need of help to facilitate timely support. The students also highlighted that SEAT usage increased the teacher's one on one interactions with them, which aligned with the in-class observations conducted by the educational researcher. This is a very important result as one of the criticisms of technology-mediated learning environments is that they could diminish the relationship between the students and the teacher. Our findings imply that SEAT could in fact orchestrate teacher time based on the students who need the most support, without compromising the overall student experience.

Our results also showed that SEAT had a positive impact on student engagement. Analyzing the student achievement tests, the improvement in the treatment class was limited and the difference between the two classes was only marginal ($p = .09$) albeit in the direction of the treatment class. Whereas the small sample size might have contributed to the marginal effect, there are also some concerns with the assessment items.

Thus, to further investigate reliability and validity of the achievement data used, we reviewed the instructional design and the test design in depth. We also analyzed the education platform logs to identify the amount of content covered by each student. As the teacher indicated in her interview, most of the topics covered in the learning sessions were new to students. The education platform logs showed that the students had to move to a different topic in each session according to the pre-designed lesson plan regardless of their progress. As a result, the students had limited time and opportunities to practice and reinforce their learning. Therefore, it is likely that student learning was not very deep and presumably involved the "remember" and "understand" levels according to Bloom's taxonomy of educational objectives [35]. In contrast, an analysis of the test design revealed that the test questions evaluated student learning on the "application" and higher levels in Bloom's taxonomy. This indicated a discrepancy between the instruction and assessment [36], which might explain the post-test results.

The results also suggest that instead of only relying on engagement analytics, the teacher considered multiple variables related to student context before taking any actions despite the fact that her perception of SEAT's accuracy was positive. We provided the teacher with a detailed description of how the technology worked. We believe this eliminated the AI-black box effect and gave the teacher enough transparency so that she felt empowered to use the technology to provide the best support to her students. We think it is very important from an ethical standpoint that educators are well-informed about the capabilities of underlying algorithms for learning analytics to guide their effective use of the technology.

Another important perspective about privacy came from the student interviews. We implemented a system that extracts features online and also runs inference online with no reliance on saving raw video data during normal operation. However, some of the students reported discomfort about being monitored. We believe there is a need for future research to study perceptions of privacy and how it is affected by technological design decisions.

The study was not without its limitations. First, we selected an exemplary math teacher with strong skills in pedagogical interventions. Therefore, the results reflect experience and classroom practice of a single teacher with certain characteristics. To increase the external validity of the results, aligned with the guidelines by [37], we obtained detailed descriptions of the teacher's profile to reflect how these results can translate to other teachers with different profiles. However, the next step would be to conduct a larger study with a diverse set of teachers across multiple classes with ideally random assignment to classes. Second, although learning sessions took place as a part of the actual course, we, along with the teacher, created a specific curriculum for this study based on the educational content platform. Because students were aware of the fact that they would not be officially graded, motivation could have been reduced. Therefore, future studies should be conducted in a way that content and assessment is a part of the school curriculum.

7 CONCLUSION

The results of our study indicate that a real-time student engagement technology can be instrumental to teachers to facilitate actionable engagement analytics, real-time class-wide monitoring, and better detection of off-task, confused, and bored students in real-time. Thus, without compromising student experience, SEAT could help allocate teacher time based on students who need most support and can influence teachers' classroom practices to support teachers' role of being a *coach* in technology-mediated learning environments.

REFERENCES

[1] Minović, M., & Milovanović, M. (2013, November). Real-time learning analytics in educational games. In *Proceedings of the First International Conference on Technological Ecosystem for Enhancing Multiculturality* (pp. 245-251). ACM.

[2] Ginda, M., Suri, N., Bueckle, A., & Börner, K. (2017). Empowering Instructors in Learning Management Systems: Interactive Heat Map Analytics Dashboard.

[3] Hattie, J., & Timperley, H. (2007). The power of feedback. Review of educational research, 77(1), 81-112.

[4] Federici, R. A., & Skaalvik, E. M. (2014). Students' Perceptions of Emotional and Instrumental Teacher Support: Relations with Motivational and Emotional Responses. *International Education Studies*, *7*(1), 21-36.

[5] D'Mello, S., Olney, A., Williams, C., & Hays, P. (2012). Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of Human-Computer Studies*, *70*(5), 377-398.

[6] Henrie, C. R., Halverson, L. R., & Graham, C. R. (2015). Measuring student engagement in technology-mediated learning: A review. *Computers & Education*, *90*, 36-53.

[7] Lee, M. J., & Ferwerda, B. (2017). Personalizing online educational tools. In Proceedings of the 2017 ACM Workshop on Theory-Informed User Modeling for Tailoring and Personalizing Interfaces (pp. 27-30). ACM.

[8] Vail, A. K., Wiggins, J. B., Grafsgaard, J. F., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2016). The Affective Impact of Tutor Questions: Predicting Frustration and Engagement. In *EDM* (pp. 247-254).

[9] D'Mello, S., Dieterle, E., & Duckworth, A. (2017). Advanced, analytic, automated (AAA) measurement of engagement during learning. *Educational psychologist*, *52*(2), 104-123.

[10] D'Mello, S. K., & Graesser, A. (2010). Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-adapted Interaction*, *20*(2), 147-187.

[11] Graesser, A. C., Cai, Z., Morgan, B., & Wang, L. (2017). Assessment with computer agents that engage in conversational dialogues and trialogues with learners. *Computers in Human Behavior*, *76*, 607-616.

[12] Okur, E., Alyuz, N., Aslan, S., Genc, U., Tanriover, C., & Arslan Esme, A. (2017, June). Behavioral engagement detection of students in the wild. In International Conference on Artificial Intelligence in Education (pp. 250-261). Springer, Cham

[13] Alyuz, N., Okur, E., Genc, U., Aslan, S., Tanriover, C., & Arslan Esme, A. (2017, November). An unobtrusive and multimodal approach for behavioral engagement detection of students. In Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education (pp. 26-32). ACM.

[14] Alyuz, N., Okur, E., Oktay, E., Genc, U., Aslan, S., Mete, S. E., Arnrich, B., & Arslan Esme, A. (2016, October). Semi-supervised model personalization for improved detection of learner's emotional engagement. In Proceedings of the 18th ACM International Conference on Multimodal Interaction (pp. 100-107). ACM.

[15] Aslan, S., Mete, S. E., Okur, E., Oktay, E., Alyuz, N., Genc, U. E., Stanhill, D. & Arslan Esme, A. (2017). Human expert labeling process (HELP): Towards a reliable higher-order user state labeling process and tool to assess student engagement. *Educational Technology*, 53-59.

[16] Mazza, R., Dimitrova, V.: CourseVis: a graphical student monitoring tool for supporting instructors in web-based distance courses. Int. J. Hum. Compu. Stud. **65**(2), 125–139 (2007).

[17] Maldonado, R. M., Kay, J., Yacef, K., & Schwendimann, B. (2012, June). An interactive teacher's dashboard for monitoring groups in a multi-tabletop learning environment. In *International Conference on Intelligent Tutoring Systems* (pp. 482-492). Springer, Berlin, Heidelberg.

[18] Van Leeuwen, A., Janssen, J., Erkens, G., & Brekelmans, M. (2014). Supporting teachers in guiding collaborating students: Effects of learning analytics in CSCL. *Computers & Education*, *79*, 28-39.

[19] Xhakaj F., Aleven V., McLaren B.M. (2017). Effects of a Teacher Dashboard for an Intelligent Tutoring System on Teacher Knowledge, Lesson Planning, Lessons and Student Learning. In: Lavoué É., Drachsler H., Verbert K., Broisin J., Pérez-Sanagustín M. (eds) Data Driven Approaches in Digital Education. EC-TEL 2017. Lecture Notes in Computer Science, vol 10474. Springer, Cham.

[20] Plattner, H. (2010). An Introduction to Design Thinking Process Guide. The Institute of Design at Stanford.

[21] Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., & Picard, R. (2009). Affect-aware tutors: recognising and responding to student affect. *International Journal of Learning Technology*, *4*(3-4), 129-164.

[22] Askar, P. (1986). Matematik dersine yonelik tutumu olcen likert-tipi bir olcegin gelistirilmesi. Egitim ve Bilim, 62, 31-36.

[23] Avci, E., Coşkuntuncel, O., & İnandi, Y. (2011). Ortaöğretim on ikinci sınıf öğrencilerinin matematik dersine karşı tutumları. Mersin Üniversitesi Eğitim Fakültesi Dergisi, 7(1).

[24] Deci, E. L. & Ryan, R. M. (1982), Intrinsic Motivation Inventory, Retrieved September 01, 2017 from, http://www.psych.rochester.edu/SDT/measures/word/IMIfull.doc

[25] Karagol, D. (2008). Promoting learner autonomy to increase the intrinsic motivation of the young language learners. Yayınlanmamis yuksek lisans tezi) Çukurova Universitesi Sosyal Bilimler Enstitusu, Adana.

[26] Rheinberg, F., Vollmeyer, R. & Engeser, S. (2003). Die Erfassung des Flow-Erlebens [The assessment of flow]. In J. Stiensmeier-Pelster & F. Rheinberg (Eds.), Diagnostik von Motivation und Selbstkonzept (Tests und Trends N.F. Bd. 2) (pp. 261-279). Göttingen: Hogrefe.

[27] Rheinberg, F. (2015). Die Flow-Kurzskala (FKS) übersetzt in verschiedene Sprachen The Flow-Short-Scale (FSS) translated into various languages. . 10.13140/RG.2.1.4417.2243.

[28] Creswell, J. W. (2009). Research design: Qualitative, quantitative, and mixed methods approaches. Thousand Oaks, CA: Sage Publications.

[29] Weber, R. P. (1990). *Basic Content Analysis*. Newbury Park, CA: Sage Publications.

[30] Krippendorff, K. (2004). Content analysis: An introduction to its methodology (2nd ed.). Thousand Oaks, CA: Sage.

[31] Aslan, S. (2012). Investigating" the coolest school in America": A study of a learner-centered school and educational technology in the information age (Doctoral dissertation, Indiana University).

[32] Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.), Hillsdale, NJ: Erlbaum.

[33] Saye, J.W. & Brush, T. (2002). Scaffolding critical reasoning about history and social issues in multimedia-supported learning environments. Educational Technology Research and Development 50(3): 77–96.

[34] Simons, K.D., Klein, J. & Brush, T. (2004). Instructional strategies utilized during the implementation of a hypermedia, problem-based learning environment: A case study. Journal of Interactive Learning Research 15: 213–233.

[35] Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. New York, NY: David McKay Company.

[36] Crowe, A., Dirks, C., & Wenderoth, M. P. (2008). Biology in bloom: implementing Bloom's taxonomy to enhance student learning in biology. *CBE—Life Sciences Education*, *7*(4), 368-381.

[37] Merriam, S. B. (1991). Case study research in education: A qualitative approach. San Francisco: Jossey-Bass.