Toward Measuring Student Engagement: A Data-driven Approach

Andrew Koster, Tiago Primo, Állysson Oliveira, Fernando Koch

SAMSUNG Research Institute Brazil {andrew.k, tiago.t, allysson.o, fernando.koch}@samsung.com

Abstract. This paper presents an empirical analysis of an automated measure for in-classroom student engagement. The paper presents (1) a novel learning metric to measure student engagement and (2) a curated data set, collected with a novel tool for logging usage data from a tablet-based digital teaching platform. We show that in this data, the student engagement metric has a weak positive correlation with students' self-reported engagement. We further show that teaching style and content type have a large influence on engagement, and how the automated measurement of engagement can help an educator in tailoring the lesson plan and content to engage students.

1 Introduction

The use of technology in the class is widely regarded as the way to provide cutting edge education while lowering the costs. Nevertheless, it is not a panacea, and educators and students alike want to see "clear educational or social value in using technologies, and [are] resistant to attempts to integrate technology for technology's sake" — Waycott [8]. One of the problems we have identified with contemporary technology-enhanced learning solutions are their tendency to circumvent the educator. Educators indicate that initiatives such as 'One Laptop Per Child', flipped classrooms, intelligent tutoring systems or MOOCs miss the mark, because the educators themselves are left out of the loop. This is problematic, because *educators matter* [7]. Educators are in charge of their own pedagogical plans and methodology, and thus the keystone to how technology is used in classroom education. For technology to fulfill a clear educational purpose, it must *empower the educator*, in order to empower the student.

This work makes a start at providing tools within a Digital Teaching Platform (DTP) to bring valuable information to the educator about the students in the class, such as their engagement, learning styles or emotional wellbeing. We envision that this information can be further used to provide timely alerts and recommendations about changes in student behaviour, or potentially interesting material and didactic techniques for specific classes. We show how we are curating a data set with detailed logs of how students and educators use a DTP, and how this data can be used to measure student engagement.

The idea of analysing detailed usage logs to infer useful student metrics is not new. The PSLC's DataShop [4] maintains a repository of log data from Intelligent Tutoring Systems, which contains datasets with detailed usage logs of students, the order they go through the material, the tutor responses, and performance metrics. Similarly, MOOCs can be considered Big Data, and while public datasets, such as EdX logs [3], contain mostly aggregate statistics, the courses generate terabytes of data from thousands of students. Similar programs are under way in universities for traditional education [1]. Nevertheless, most current data collection, and hence repositories, focus on the way individual students learn (using an intelligent tutor or in a MOOC), and fail to capture social interactions within a classroom, and the vital role of the educator, his or her material selection and didactic method [9]. In contrast, the DTP we used to collect the data enables the use of tablets in a traditional "chalk and talk" methodology, causing the least possible disruption in the way educators are already used to running their class while offering advantages such as automated roll calls, exercise correction and of course, the use of learning analytics to inform the educator about important metrics.

In the next section we briefly discuss the DTP we used, before describing the data set and the metric of engagement we infer from it.

2 Tablet-based Digital Teaching Platform

We use a prototype Digital Teaching Platform (DTP), with the main distinction that it is designed to be a single platform for the composition of classroom material using a repository of learning objects, a lightweight distribution server that can push this material to tablets over a wireless LAN in the classroom, and an application on the tablet to display this material and log the users' actions [5].

In Table 1 we list a description of all the events that are logged. For all events, we record the user id, the timestamp, the event type and the event-specific values in a JSON object. The tablets' timestamps are synchronized to within an error of 15 seconds.

One of the experimental functionalities of the DTP is to allow the teacher to view "student engagement" during the class. This can be visualised as a graph, giving a moving average over time, or numerically in a corner of the teacher's screen, giving the engagement at that moment. Student engagement is computed using the `appPause`, `appResume`, `openResource`, `closeResource`, `pageNavigation` and `viewPhotoInGallery` events. Using these events we can compute what learning object a user has opened (if any at all), and an individual is considered engaged if he or she is studying the same learning object as the teacher. Class engagement is a percentage: the number of students who are engaged at that time. Because this is noisy, we use a smoothing filter: we average the class engagement over the last 5 seconds.

This metric of engagement is intended to measure what Fredricks et al. define as *behavioural engagement*; one of the three principal components that they identify for school engagement [2]. Behavioural engagement attempts to describe

Event type	Description
annotation	User annotates a Learning Object
appPause	User paused (closed) the Player app
appResume	User resumed (opened) the Player app
assessDynaObjInter	User interacts with a dynamic object embedded in a questionnaire
assessmentFinish	User closes a questionnaire
assessmentStart	User opens a questionnaire
assessQuestionAnswer	User answers a question in a questionnaire
assessQuestionSelect	User selects a question in a questionnaire
blockAssignment	Educator blocks students from answering a questionnaire
classFinished	Educator closes the classroom material
classStarted	Educator opens the classroom material
closeResource	User closes a learning object
dynamicObjectInteraction	User interacts with a dynamic learning object
enterClass	Student opens the classroom material
enterFullscreen	User views a video, image or gallery learning object in fullscreen
evaluateResource	User evaluates a learning object (“like”, or register a question)
eyeTracking	User is looking at, or away from, the tablet
highlightText	User highlights text with the stylus
leaveClass	Student closes the classroom material
leaveFullscreen	User stops viewing a learning object in fullscreen mode
openResource	User opens a learning object
pageNavigation	User navigates between pages (chapters)
pauseMedia	User pauses playback of a video or audio object
playMedia	User starts/resumes playback of a video or audio object
scrollDetector	User scrolls in the content
sharedContent	User sends a learning object and associated message to another user
unblockAssignment	Educator allows students to answer a questionnaire
userTouch	User touches the screen somewhere with no other action associated
viewPhotoInGallery	User browses between images in a gallery object
viewSharedContent	User opens a learning object that was shared by another user

Table 1. Possible event types and their description

how a student's behaviour with regards to the educational tasks can be characterized as engagement. Time on task and homework completion are mentioned as metrics, but most methods use self-reporting surveys or observational studies to measure engagement. In the next section we outline the experiments we performed to assess whether our metric corresponds to this type of engagement.

3 Data Collection and Results

We conducted in-classroom field trials of the DTP described in the previous section, collecting data from 8 different classes at 3 different schools in Brazil; 5 classes at 2 schools in Manaus, and the remaining 3 classes at a school in Campinas. The classes administered at the schools in Manaus were part of the regular curriculum in mathematics, with on average 32 students per class. One of the classes in Campinas was also on mathematics, and the other was on physics, but all were extra-curricular classes with on average 12 students per class. In total, we collected approximately 223 student-hours worth of data, from approximately 13 hours of classes with in total approximately 200 different students, in classrooms with 3 different teachers and circumstances. This resulted in a set of 74,700 logged events of the types as in Table 2. This data set is available here: `http://will.be.made.public.in.camera.ready` From analysing the

Event type	**Total**	**Event type**	**Total**
annotation	8	enterFullscreen	2850
appPause	1513	evaluateResource	29
appResume	1876	eyeTracking	21610
assessDynaObjInter	1554	highlightText	1563
assessmentFinish	496	leaveClass	151
assessmentStart	684	leaveFullscreen	2609
assessQuestionAnswer	2740	openResource	1963
assessQuestionSelect	4466	pageNavigation	338
blockAssignment	3	pauseMedia	9
classFinished	10	playMedia	14
classStarted	12	scrollDetector	782
closeResource	1513	unblockAssignment	4
dynamicObjectInteraction	11008	userTouch	4625
enterClass	689	viewPhotoInGallery	11581

Table 2. Event types that were logged during the trials of the DTP

raw events, we can already correlate some of the events with the teaching style. If we divide the events up by school, we see they correspond proportionally to the number of hours taught there. There were 8 hours of lessons at School A in Manaus, corresponding to 60% of the total. Nevertheless, because the class size was larger, this corresponds to approximately 160 student-hours of data, or approximately 70% of total student-hours logged. Despite this, these lessons

only generated a total 43,521 events, or approximately 60% of the total events, indicating that overall activity level of the these students was lower than average. However, the number of appPause and appResume events in these classes are proportionally far higher, approximately 90% of all appPause and appResume events. We discovered that this is due to the teacher, and subsequently the students, taking notes by using the tablet's built in note-taking functionality: this pauses the content player app, and allows note taking. The notes are saved to disk, and the app is resumed. For each teacher and class, we can find such specific patterns in the data, corresponding to *how* they use the app and navigate through the content. The events are also highly dependent on the content, of course: the teacher in Campinas relied far more heavily on the use of virtual laboratories, implemented in the player as dynamic objects[1]. Therefore over 80% of the dynamic object interaction events were logged in these 3 class hours.

3.1 Experiment 1: Educator vs. Material

Given that both material and educator have a significant influence on the type of signals produced in a class, it is interesting to know which has a greater effect. All the classes at both School A and B in Manaus used the same material, barring some minor personal touches, whereas the classes in Campinas had the same teacher, but one was with different material. The hypothesis we test is that the *the educator has a greater influence than the material on the type of events generated in a class.*

To test this, we compute normalized vectors of the events per classroom (as percentages of the total number of events per classroom), and cluster them into 3 groups. If the educator has a greater influence, we would expect the Euclidean distance between School A and School B in Manaus to be greater than the distance between the classes with different material in Campinas, and thus to find one cluster per school. If, however, material has a greater influence, we would expect the reverse: School A and B would be clustered together, whereas the classes in Campinas would be split along the line of different material.

Using k-Means to cluster our data results consistently in it being clustered along *material* lines. We evaluate the statistical significance of this clustering by considering each type of event as a random sample from a Gaussian distribution, allowing us to generate synthetic data of other possible event configurations from classes. We can then perform a Monte Carlo analysis of our clusters and approximate the p-value of significance of our clustering. We use a parametric bootstrap, consisting of 1000 Monte Carlo realizations of event sets, where the population mean and standard deviation are set equal to the sample mean and standard error of our real data set, respectively. As a test statistic, we use the silhouette of the clustering, and we can then calculate the Monte Carlo estimate p_m of the probability that a sample gives a cluster silhouette as good or better than the silhouette of the clustering of the real data. The clustering according

[1] A dynamic object is an HTML5 app that is contained within the digital classroom material

to material is significant with $p_m = 0.002$, or in other words, only 7 of our 1000 random data sets resulted in an equal or better clustering than clustering the data according to material. An added benefit of this method is that we can also test whether the clustering according to teacher would have been significant. That is not the case ($p_m = 0.852$) (see Figure 1).

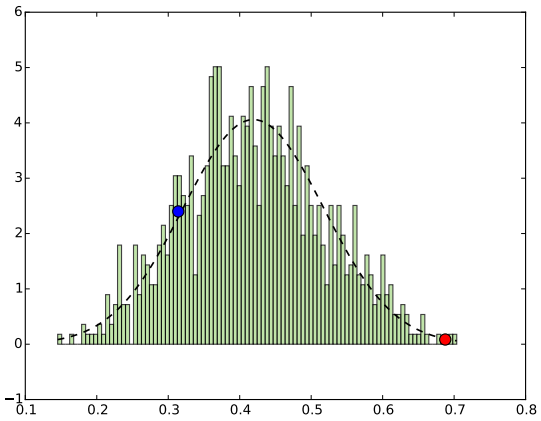

Fig. 1. Histogram with results of Monte Carlo estimate of the distribution of silhouettes, and a Gaussian pdf plotted in it. The red dot represents the position of the outcome of k-Means clustering on the real data. The blue dot the value of the silhouette for a clustering based on the professor.

While the parametric bootstrapping gives us confidence in these results, it is nevertheless based on a small data set. Further, and more diverse, experiments are needed in different locations to verify that this holds for a larger array of materials and teaching styles. Nor does the number and type of events generated within the DTP by itself say anything about the quality of the student's learning. For that, we need to look at other metrics, which we can derive from these events. The point of this experiment is mostly to highlight that, in our data set so far, the influence of the teacher, and his or her teaching style, is of secondary importance of the type of learning objects used; this is useful, because the latter is easily analysed, whereas the former is extremely hard to capture automatically.

Regardless of what the underlying cause is for the variance between classes, the variance is noise to learning analytics, which we desire to represent a meaningful result regardless of who is teaching and what is being taught.

3.2 Experiment 2: measuring student engagement

Within the DTP, the educator has the ability to monitor the class' engagement. This engagement is computed as a moving average of individual students' engagement at any given time as described in Section 2. We are interested in

evaluating whether this engagement value actually represents students' engagement. We can evaluate a post-hoc mean of individual engagement using the same principles as the online metric, by computing the average engagement per student over the full class time. Table 3 contains the average engagement per class in Manaus. The metric is computed for each individual as the percentage of time that the class was in session that the student was viewing the same learning object as the educator. We only use the data from these four classes, because

	School A	School B1	School B2	School B3
Automated	34.3	49.1	39.3	41.7
Survey	4.14	4.27	4.36	4.31

Table 3. Average overall engagement per class for the two datasets from Manaus. The automated value in percentages, and the survey as an average on a Likert scale from 1-5

these are the classes where we had time for a brief survey after the class in which students could self-report their engagement. This survey was a shortened version of the User Engagement Scale [6] with seven questions on different aspects of engagement. Because the classes were of different sizes and different variance, we test for whether the data from all classes can reasonably be considered as a single data set using the non-parametric Kruskal-Wallis test, and we reject the null hypothesis that there are significant differences between the data sets ($p >> 0.05$). We do the same for our self-reported survey, and here as well we can reject the null hypothesis ($p >> 0.05$).

Because we asked seven questions in our survey, we further analyse whether these all refer to a single concept using the Crohnbach α statistic. The α statistic for our data is 0.76, and a generally accepted value for data describing a single concept is $\alpha > 0.7$, so our survey can be seen to describe a single concept. We can thus compute the mean for each student over the survey questions and consider this a metric for self-reported engagement. The average per class is reported in Table 3.

The main question we want to resolve, however, is whether the automated engagement metric accurately represents student engagement. To answer this, we use Pearson's correlation statistic between the automated metric and self-reported engagement. The Pearson r-statistic shows a weak, but nevertheless significant (at the $p = 0.05$ level), positive correlation: 0.227. This is further illustrated in Figure 2.

We thus conclude that the automated measurement can be valuable in the classroom: it is correlated with the self-reported engagement, but further studies are necessary. First, it is a weak correlation, and the search continues for additional metrics that can be combined with this engagement measure. Second, the self-report survey had very positive answers, possibly due to a number of known positive biases, such as acquiescence bias and social desirability bias, but we also suspect that the novelty factor from using tablets in the classroom could lead

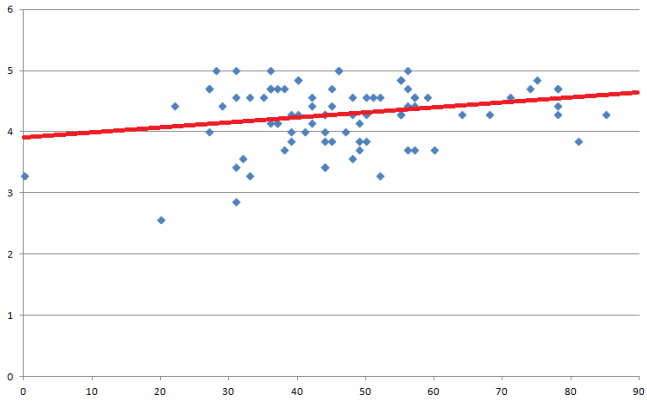

Fig. 2. Scatterplot between engagement as measured by self-reporting surveys (y-axis) and automated measurement (x-axis), with linear trend line

to increased engagement. We aim to follow this up with a longitudinal study in which we can more accurately determine the true value for student engagement (through a combination of self-reporting and observations).

3.3 Experiment 3: juxtaposing engagement and content type over time

In the first experiment we determined that differing material was the primary motivator for changes in data sets. In the second experiment we assumed that our engagement measure is invariant to such "noise". In this experiment we test the effect of content type on our automated engagement measure. An initial test shows that content type definitely influences overall engagement: when we compute the metric for the remaining classes, the Kruskal-Wallis analysis of variance test shows we should not discard the null hypothesis ($p << 0.01$): there is a significant difference within our data sets. However, a subgroup analysis within the Campinas classes (with two different content types) shows no significant variation: it is only when we compare classes from our Campinas and our Manaus tests with each other that there is a significant difference. It thus might be simply due to the different environment. In Campinas the classes were extracurricular in a smaller group than the regular curriculum classes in Manaus. Small group-size has been shown to lead to greater engagement on its own. However, it is also possible that the higher engagement level from these classes is due to the increased use of interactive (dynamic) materials. To analyse this, and to further help educators understand their students' engagement level, we have created an infographic that plots average class engagement against the content type (in rough categories) that the educator is explaining, or expecting the students to be studying. This infographic has been plotted in Figure 3. In these graphs we can clearly see some effects of the content type on engagement. First, engagement as

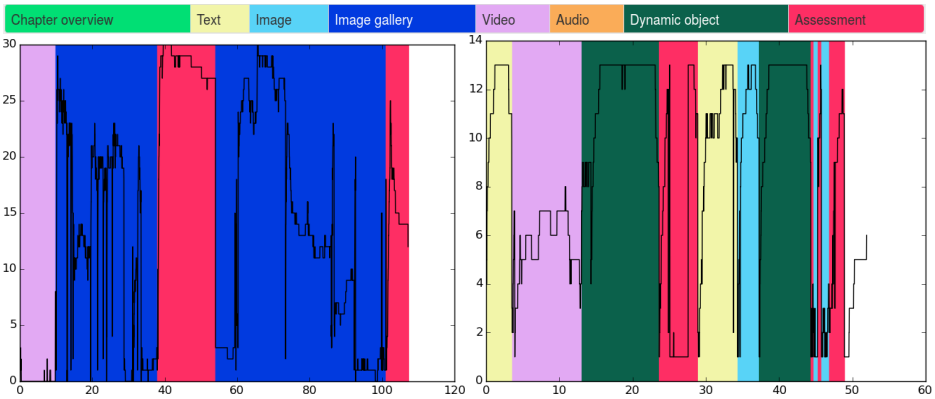

Fig. 3. Class engagement juxtaposed with content type for a class in Manaus (on the left) and Campinas (on the right). Above is the color legend for the background colors, representing what material the educator had opened at each time

measured by our metric is lowest, particularly in the Manaus class, but also in the Campinas class, when watching a video. This is easily explained by observing that the video content is disabled on students' tablets in the classroom, and they are expected to watch it on the television. Thus, what material they have opened in their tablet is in these instances a bad measurement of engagement. Using eye tracking data could be used in these instances: we expect an engaged student to be looking away from the tablet during a video. Second, engagement is maximal when working with the dynamic objects, in Campinas. This may very well explain the discrepancy in overall engagement between the two types of classes better than the actual material: both of the Campinas lesson plans relied heavily on the use of dynamic content. It further seems to validate this use of technology: there is no real need for a tablet in the classroom if its only purpose is as an ebook reader. The real value of technology-enhanced education is in novel applications such as dynamic content and interactive questionnaires. During questionnaires we can also see that engagement was higher than average.

In conclusion, our engagement measure is very much dependent on the material if averaged over the entire class. However, we expect that within a class, individual engagement is mostly independent of the material: a disengaged student may spend less time exploring a dynamic object than an engaged student. Alternatively a more fine-grained approach might be required that monitors activity within such objects, which the current version of the DTP is not capable of. In our follow-up studies we will also monitor individual student engagement across multiple different materials and classes. Furthermore we will test whether we should report individual deviations from the mean: we hypothesise that while the mean engagement is dependent on the content type, individual deviations from the mean are significant, and the educator may want to be notified and take actions if an individual's engagement spikes downward during the class.

4 Conclusion

In this work we explored the potential of a novel data collection methods for in-classroom learning analytics, by describing the data set collected so far, and how it was used in an evaluation of the real-time computation of engagement. We have only just scratched the surface of this data collection and analysis. Creating detailed logs of user actions in digital educational content is not novel, but insofar as we know this is the first time it has been applied in a classroom setting, in which we can analyse relations between the actions of different students, and with the professor. Our data-driven analysis supports pedagogical work on the influence of the teacher in the classroom and that any learning analytics ignoring this effect are likely to be missing a critical influence on the student's performance, learning style and indeed engagement with the class.

We also introduce a novel learning metric to measure engagement, and present an empirical analysis of this metric. A weak positive correlation between the metric and a self-report survey indicates that it can be used to measure engagement, but for a more reliable measurement should be refined, or combined with other metrics. Interesting uses of this metric could be in warning the teacher that class engagement is dropping, that a specific student is abnormally disengaged or as a tool in lesson planning: we show how engagement fluctuates with the content type that the teacher is using, and the teacher can use this in refining classroom material.

References

1. Arnold, K.E.: Signals: Applying academic analytics. EDUCAUSE Quarterly **33**(1) (2010) `http://www.educause.edu/library/EQM10110`.
2. Fredricks, J.A., Blumenfeld, P.C., Paris, A.H.: School engagement: Potential of the concept, state of the evidence. Review of educational research **74**(1) (2004) 59–109
3. Ho, A.D., Reich, J., Nesterko, S.O., Seaton, D.T., Mullaney, T., Waldo, J., Chuang, I.: HarvardX and MITx: The first year of open online courses, fall 2012 – summer 2013. HarvardX and MITx Working Paper No. 1 (2014)
4. Koedinger, K.R., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J.: A data repository for the EDM community: The PSLC datashop. In Romero, C., Ventura, S., Pechenizkiy, M., Baker, R., eds.: Handbook of Educational Data Mining. CRC Press (2010) 43–55
5. Koster, A., Primo, T., Koch, F., Oliveira, A., Chung, H.: Towards an educator-centred digital teaching platform: The ground conditions for a data-driven approach. In: Proceedings of the 15th IEEE Conference on Advanced Learning Technologies (ICALT), Hualien, Taiwan, IEEE (2015) 74–75
6. OBrien, H., Cairns, P.: An empirical evaluation of the User Engagement Scale (UES) in online news environments. Information Processing & Management **51**(4) (2015) 413–427
7. Slater, H., Davies, N., Burgess, S.: Do teachers matter? Measuring the variation in teacher effectiveness in england. Oxford Bulletin of Economics and Statistics **74**(5) (2012) 629–645

8. Waycott, J., Bennett, S., Kennedy, G., Dalgarno, B., Gray, K.: Digital divides? Student and staff perceptions of information and communication technologies. Computers & Education **54**(4) (2010) 1202–1211
9. Wenglinsky, H.: The link between teacher classroom practices and student academic performance. Education Policy Analysis Archives **10**(12) (2002) 1–30