Adaptive Review for Mobile MOOC Learning via Multimodal Physiological Signal Sensing - A Longitudinal Study

Phuong Pham [*]
Microsoft
Redmond, WA, USA
phuong.pham@microsoft.com

Jingtao Wang [*]
Google Cloud AI
Beijing, China
jingtaow@google.com

ABSTRACT

Despite the great potential, Massive Open Online Courses (MOOCs) face major challenges such as low retention rate, limited feedback, and lack of personalization. In this paper, we report the results of a longitudinal study on AttentiveReview[2], a multimodal intelligent tutoring system optimized for MOOC learning on *unmodified mobile devices*. AttentiveReview[2] continuously monitors learners' *physiological signals*, *facial expressions*, and *touch interactions* during learning and recommends personalized review materials by predicting each learner's perceived difficulty on each learning topic. In a 3-week study involving 28 learners, we found that AttentiveReview[2] on average improved learning gains by 21.8% in weekly tests. Follow-up analysis shows that multimodal signals collected from the learning process can also benefit instructors by providing rich and fine-grained insights on the learning progress. Taking advantage of such signals also improves prediction accuracies in emotion and test scores when compared with clickstream analysis.

CCS CONCEPTS

• **Human-centered computing→Ubiquitous and mobile computing→Ubiquitous and mobile computing systems and tools**

KEYWORDS

MOOC; Heart Rate; Facial Expressions; Intelligent Tutoring System; Physiological Signal; Affective Computing; Multimodal; Mobile Interface.

ACM Reference format:

Phuong Pham and Jingtao Wang. 2018. Adaptive Review for Mobile MOOC Learning via Multimodal Physiological Signal Sensing - A Longitudinal Study. In *Proceedings of 20th ACM International Conference on Multimodal Interaction (ICMI'18), October 16-20, 2018, Boulder, CO, USA.* ACM, NY, NY, USA, 10 pages.
DOI: https://doi.org/10.1145/3242969.3243002

ICMI '18, October 16–20, 2018, Boulder, CO, USA

ACM ISBN 978-1-4503-5692-3/18/10…$15.00 https://doi.org/10.1145/3242969.3243002

1 INTRODUCTION

The rise of Massive Open Online Courses (MOOCs) presents both opportunities and challenges to knowledge dissemination at scale. By December 2017, MOOCs have attracted more than 81 million registered learners [12]. When taking MOOCs, learners can control their learning process and have access to high quality learning material at low cost [20]. In a recent survey involving 52 thousand MOOC learners, researchers also found that MOOCs were particularly beneficial to economically and academically disadvantaged populations [6]. However, despite the promising growth, today's MOOCs also suffer from challenges such as low retention rate (e.g. around 4.0% in Coursera [6], and 7.7% in edX [7]), low engagement with the learning materials (52.0% in-video dropout rate [21]), and more importantly, lack of personalization [27]. As a result, today's MOOCs are still an inferior choice when compared with one-on-one tutoring or even traditional classroom teaching.

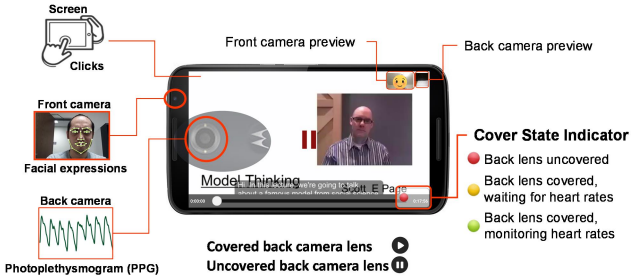

Figure 1: The primary interface of AttentiveReview[2].

Paradoxically, advantages in today's MOOCs are often the fundamental causes of the challenges in MOOCs. First, pre-recorded lecture videos are easy to distribute to tens of thousands of learners. Meanwhile, the passive, one-size-fits-all videos also reduce learners' engagements and isolate instructors from important cues in traditional classrooms, such as facial expressions or raised hands to assess teaching effectiveness. Although clickstream analysis [21], quizzes, and post-lecture/course surveys [6][20] can be used to analyze the learning process, such post-hoc techniques are usually coarse-grained and highly delayed [42]; Second, the scalability and ubiquity of MOOCs (e.g. 7,902 participants per course in [7]) also

[*] This research was in-part conducted when PP and JW were at the University of Pittsburgh. JW is currently with Google AI China Center.

restricted the choice of technologies that can be adopted. For example, it will be hard for most of the affect-based Intelligent Tutoring Systems (ITS) that require additional hardware [1][11][36] to be deployed in MOOC environments due to challenges in equipment costs and portability.

In response to these challenges, we propose AttentiveReview^2 (Figure 1), a multimodal intelligent mobile learning system optimized for MOOC learning on unmodified smartphones. AttentiveReview^2 is a multimodal adaptive learning technology built on top of the sensing infrastructure enabled by AttentiveLearner^2 [33]. AttentiveReview^2 uses on-lens finger gestures for both tangible video control and implicit Photoplethysmography (PPG) sensing. A learner covers and holds the back camera lens to play a lecture video, uncovering the lens will pause the video. Furthermore, AttentiveReview^2 leverages the front camera for real-time Facial Expression Analysis (FEA). AttentiveReview^2 then infers the learner's perceived difficulty towards each topic in a lesson from both collected PPG signals and facial expressions, and recommends an appropriate reviewing topic that would benefit her learning outcome. The idea of supporting personalized review to improve mobile MOOC learning without additional hardware requirements has been explored by AttentiveReview [31] in the past. However, there are two major improvements when compared with AttentiveReview. First, AttentiveReview only relies on PPG signals to predict learners' cognitive states while our system demonstrates the feasibility and efficacy of a multimodal intelligent tutoring system for adaptive review on unmodified smartphones; Second, there was only one 24-minute learning session on an introductory topic in the study of AttentiveReview. In comparison, we explore the effectiveness of our intervention in a more realistic setting by conducting a 3-week longitudinal study on a more challenging (math heavy) learning topic. As detailed in the follow-up sections, we found that instead of recommending the most difficult/confusing topic for review as in AttentiveReview, adaptive reviewing of more complex learning topics should take into account both the *absolute difficulty* of learning topics and the learner's zone of proximal development (ZPD).

This paper offers three significant contributions:

- The design, prototyping, and evaluation of a multimodal adaptive intervention technology optimized for enabling personalized MOOC learning on unmodified smartphones.
- A 28-participant longitudinal study to investigate the feasibility, efficacy, and challenges of affect-aware interventions in informal learning environments.
- A direct, quantitative comparison of three modalities, i.e. PPG signals, facial expressions, and touch landing points as feedback channels to measure learning outcome in the context of mobile MOOC learning.

2 RELATED WORK

2.1 Learning Activities in MOOCs

Clickstream analysis [15][21][38] and user-generated content (UGC) analysis [45] are the two most popular techniques for researchers to understand learning activities in MOOCs. For example, by analyzing mouse click logs in 6.9 million video watching sessions on edX, Kim et al. [21] discovered that the logarithmic value of video length can predict the in-video dropout rate. Informed by quantitative log analysis of learning activities in MOOCs, Guo et al. [15] proposed a set of video production recommendations to create more engaging contents for MOOCs. Van der Sluis et al. [38] revealed the negative impact of contents' difficulty level on video watching time, arguing that tutorial videos should be personalized for each learner to reduce in-video drop-outs. By combining learners' in-video activities with their posts in course-specific discussion forums, Yang and colleagues [45] found that students who discuss more frequently are less likely to drop a course. Although clickstream analysis and UGC analysis can reveal key insights from existing activity logs, they work better for capturing the *aggregated trends* from thousands of learners, rather than enabling *personalized interventions* for individual learners.

Researchers have explored various interaction techniques [8][22][23] to facilitate MOOC learning. For example, Kovacs [22] designed a quiz-driven video navigation interface for MOOCs and found such interface can help learners to seek for answers in MOOC videos. Coetzee et al. [8] proposed the use of a real-time chatroom to facilitate discussions during MOOC learning. Krause and colleagues [23] integrated social gamification mechanisms with MOOCs and found a 25% increase in video watching time and a 23% increase in test scores.

Researchers have also explored the idea of personalized learning in MOOCs [3][28][35]. Brinton et al. [3] proposed a personalized schedule via learners' browsing history and found this technique led to a 70% increase in the number of lessons viewed. Miranda and colleagues [28] explored adaptive assessment questions based on learners' performance on previous questions in MOOCs. Raghuveer et al. [35] proposed a technique to customize learners' paths of learning based on corresponding learning objectives. In summary, most personalization techniques for today's MOOCs either rely on clickstream analysis, which can be sparse within a single lecture video or require learners' active participation (e.g. taking quizzes, reporting learning objectives). In comparison, we explore the implicit collection of learners' physiological signals as well as facial expressions in MOOC learning and provide adaptive learning experiences by inferring learners' cognitive and affective states in learning from signals beyond clickstream analysis.

2.2 Affective Computing in Education

Affective computing [34] aims to design, implement, and evaluate computing techniques for recognizing, interpreting, and responding to human affects. Since learners' cognitive and affective states have a direct impact on learning [40], affective

computing is important in both understanding the learning process and designing intelligent tutoring systems in education. During the past decade, researchers have investigated various modalities and physiological signals [1][31][41] to identify learners' affective states in teaching and learning.

Szafir and Mutlu [36] used learners' electroencephalogram (EEG) signals to predict attention in MOOCs. Afergan et al. [1] explored the dynamic adjustment of task difficulty in a path planning task by analyzing participants' brain activities from functional Near Infrared Spectroscopy (fNIRS). D'Mello and colleagues [11] showed the feasibility of inferring students' mind wander moments from eye gaze moment patterns. Grafsgaard et al. [14] classified learners' engagement and frustration events via facial expressions. Pham and Wang [32] explored the detection of learners' Mind Wandering (MW) events from photoplethysmography (PPG) signals implicitly captured from unmodified smartphones. Xiao and Wang [41] further improved the reliability of AttentiveLearner by predicting extreme personal events and aggregated learning events.

Various intervention technologies have been proposed based on the inferred cognitive/affective states. The learning content could be adjusted adaptively based on learners' perceived difficulty [1]. Reorienting pop-up messages were a widely used intervention technology and can be triggered when the system found learners mind wandered [11] or disengaged [41]. Adaptive review is another effective intervention technology [13]. An adaptive review algorithm is usually composed of two parts: 1) Choosing the reviewing content based on learners' attention [36], perceived difficulty [31], or the number of mind wandering events [11]; 2) Determining the reviewing schedule. According to Dunlosky [13], the spaced rereading approach was more effective than massed rereading (review immediately) in the reading comprehension context.

Previous work [9][29] showed the benefits in prediction accuracies by combining multiple channels of signals into a multimodal system. D'Mello and Graesser [9] achieved a 0.2 increase in Kappa for predicting learners' emotions by combining facial expressions, posture data, and dialog cues. Monkaresi et al. [29] achieved higher accuracy in detecting engagement by ensembling models of heart rate and models of facial expressions. Unfortunately, most of the existing multimodal research require additional sensors for collecting signals from learners. The cost, availability, and portability of such equipment have prevented the wide adoption of such systems in MOOCs.

2.3 Facilitating Mobile MOOC Learning

Researchers have explored the idea of designing affect-aware interfaces to support MOOC learning on mobile devices in the past [31][32][33][41][42]. In the AttentiveLearner project, researchers were able to infer learners' mind wandering events [32], boredom, and confusion [42] via implicit PPG sensing on unmodified smartphones. Built upon AttentiveLearner, AttentiveReview [31] demonstrated the effectiveness of adaptive review by predicting learner's perceived difficulty levels of corresponding learning materials. C2F2 [41] explored pop-up reminders during learning sessions to re-engage learners when a disengagement is detected before an important learning topic. AttentiveLearner^2 [33] supplemented AttentiveLearner with real-time facial expression analysis (FEA) via the front camera on an unmodified smartphone and achieved a 6.4% improvement in Accuracy averaging across 6 emotions. Nevertheless, AttentiveLearner^2 was only evaluated via offline benchmarking and it was unclear whether such a multimodal system can lead to direct learning gain in a real learning environment.

By comparison, our new system supplements AttentiveReview [31] with real-time facial expression analysis from the front camera to improve the robustness and accuracy of affect prediction in learning. Further, we deploy AttentiveReview^2 in a three-week longitudinal study to verify its usability and efficacy in a more realistic MOOC learning task and conduct a direct, quantitative comparison of the multiple streams of signals collected. To the best of our knowledge, AttentiveReview^2 is the first multimodal learning system that enables personalized MOOC learning experiences via a combination of implicit PPG sensing and real-time facial expression analysis on today's unmodified mobile devices.

3 DESIGN OF ATTENTIVEREVIEW^2

AttentiveReview^2 has three main components: 1) the tangible video control channel, 2) a triple stream signal sensing module, and 3) algorithms for supporting adaptive review.

3.1 Tangible Video Playback Control

AttentiveReview^2 uses on-lens finger gestures for video control, i.e. the video is played when a learner covers and holds the back camera lens with her fingertip while uncovering the lens will pause the video (Figure 1). We used the Static LensGesture algorithm [44] for lens covering detection. Existing research [42] found this control mechanism easy to learn and responsive to use.

3.2 Triple Stream Signal Sensing

AttentiveReview^2 collects three complementary streams of signals, i.e. PPG signals, facial expressions, and on-screen touch interactions, implicitly from learners during MOOC learning. As a by-product of the tangible video control, AttentiveReview^2 can extract a learner's waveforms of heart beats (i.e. PPG signals) from the back camera. The underlying mechanism is: in each cardiac cycle, the arrival and withdrawal of fresh blood change a learner's skin transparency, including her fingertip covering the back camera lens. AttentiveReview^2 uses the LivePulse algorithm [17] to compute NN intervals from raw PPG waveforms collected. By detecting the peaks and valleys of these skin transparency changes, LivePulse can extract the normal to normal (NN) intervals in each heartbeat.

At the same time, AttentiveReview^2 utilizes the front camera to capture a learner's facial expressions while watching lecture videos. As a result, AttentiveReview^2 enables the automatic collection of both PPG signals and facial expressions implicitly

during mobile MOOC learning. We used Affdex SDK [26] for real-time facial expression analysis through the front camera.
Last but not least, AttentiveReview[2] collects on-screen touch interaction events as its third channel of feedback signals.

3.3 Adaptive Review Algorithm

Our adaptive algorithm is based on two educational principles: 1) reviewing relevant and appropriate tutorial materials will improve learning [13] and 2) learners' performance depends on the appropriate difficulty level of learning materials [1][31]. AttentiveReview[2] recommends a learner to review a learning topic based on her perceived difficult levels. We will discuss how to extract features from both PPG signals and facial expressions, and train the adaptive review algorithm in detail. Our algorithm suggests the learner review either the most difficult topic or the easiest topic depending on the reviewing strategies in use.

3.3.1 Feature Extraction.

After discarding the first and the last 10s of each learning topic, AttentiveReview[2] extracts Heart Rate Variability (HRV) features using a global window and non-overlapping sliding windows (Figure 2). For each window type (global or sliding), 8 dimensions of HRV features are extracted from the normal to normal (NN) intervals: 1) AVNN (average NN); 2) SDNN (standard deviations of NN); 3) pNN60 (percentage of adjacent NN with a difference > 60 ms); 4) rMSSD (root mean square of successive differences); 5) SDANN (standard deviation of averages of NN within an m-second segment); 6) SDNNIDX (mean of the standard deviations of NN within an m-second segment); 7) SDNNIDX / rMSSD; 8) MAD (median absolute deviation). Only top 8 HRV features are chosen using univariate ANOVA.
Different from existing research [9][14] that use the existence and frequency of static facial expressions as features. We propose a new set of Action Unit Variability (AUV) features to capture the temporal dynamics of facial expression within a given amount of time. The notations of AUVs are in part inspired by HRV features reported above.

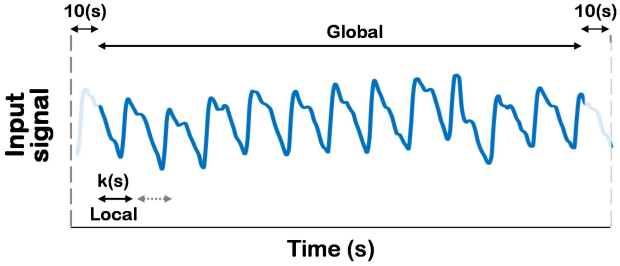

Figure 2. Feature extraction from the PPG signal of a topic (facial features were extracted using the same method).

AttentiveReview[2] extracts top 8 dimensions of AUV features from facial expressions within a global window and non-overlapping sliding windows. For each window type, 8 dimensions of AUV are extracted: 1) AVAU (average action unit value); 2) SDAU (temporal standard deviations of action unit value); 3) MAXAU (the maximum value of action unit value); 4) rMSSD; 5) SDAAU (standard deviation of the averages of action unit value within an m-second segment); 6) SDAUIDX (mean of the standard deviations of action unit within an m-second segment); 7) SDAUIDX / rMSSD; 8) MAD.

3.3.2 Perceived Difficulty Ranking.

AttentiveReview[2] uses a ranking SVM model with a linear kernel [31] to determine a learners' perceived difficulty of each topic in a lesson. The ranking SVM model uses the top 8 HRV features and top 8 AUV features selected by the highest F-ratios from a univariate ANOVA test. We train the ranking SVM using data from Pham and Wang [33]. The training dataset contains PPG signals and facial expressions of 26 users collected while they were watching 6-minute tutorial videos on a Nexus 6 smartphone. The users reported their perceived difficulty of each video after watching the video. We optimized the tradeoff margins for hyper-parameter tuning.

4 USER STUDY

4.1 Experimental Design

Our longitudinal study lasted 3 weeks and there were two lessons per week. We chose three weeks to evaluate AttentiveReview[2] for two reasons: First, we wanted to investigate the *course level* performance of adaptive review in MOOC learning and it is possible to finish a small MOOC course within three weeks. Second, as reported by Gütl et al. [16], most of the dropouts in MOOCs occur within the first 3 weeks. We wanted to take a closer look at how students learn in a technology instrumented MOOC environment.

4.1.1 Learning Material.

We chose a well-received but math-intensive course – "*Model Thinking*" by Professor Scott E. Page from the University of Michigan in this study. The original course has been offered in Coursera since 2012. We split the course into six lessons, with three topics in each lesson (3 × 6 = 18 topics in total). We offered two lessons per week. Each topic was modified to fit within 6 minutes (i.e. 18 minutes per lesson). In addition to the lecture videos, there were 6 multiple choice questions for each topic (3 for pretest and 3 for weekly test).

4.1.2 Reviewing Methods.

Pham and Wang [31] showed the effectiveness of reviewing the most difficult topic over reviewing the easiest topic in a single-lesson user study. However, in a multi-session learning setting, we hypothesize that reviewing the easiest topic would also be beneficial in certain situations. For example, if a learner could not understand any topics in a lesson, starting to review the easiest topic would be more effective than starting to review the most difficult topic. Therefore, we evaluated two reviewing conditions: reviewing the most difficult topic (Hard-Review) and reviewing the least difficult topic (Easy-Review). Since both Hard-Review and Easy-Review were significantly better or equivalent to a No-Review baseline [31], we removed the No-Review condition in this study to make the scale of the longitudinal study manageable.

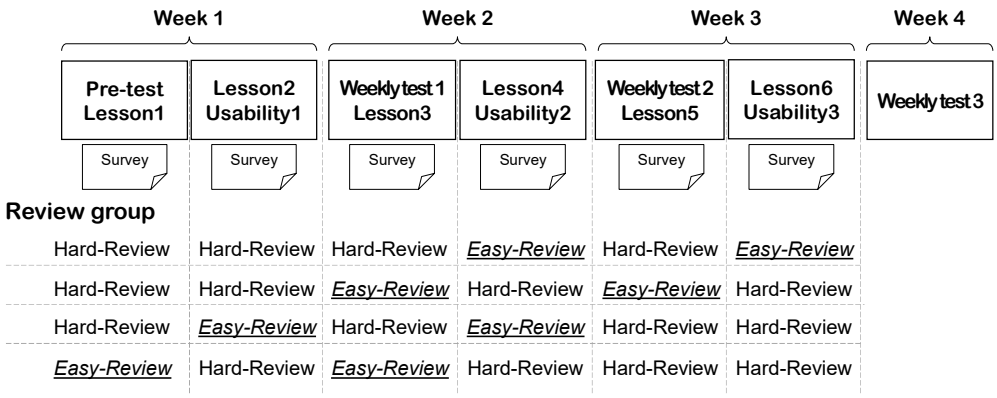

Figure 3. Procedure of the user study.

Pham and Wang [31] found the Easy-Review condition was significantly worse than a full review condition. Therefore, the performance and motivation of a participant would be negatively affected if she was exposed to Easy-Review condition multipletimes in this longitudinal study. To reduce the effects of this confounding factor, we used a within-subject design to alternate the reviewing condition for each participant instead of assigning a single reviewing condition to a particular group as in a between-subjects design. We made two additional modifications of the within-subject design to further relax the negative effects (if any) from the Easy-Review condition. First, we reduced the number of Easy-Review compared to Hard-Review by using a single-subject design which assigns 2 Easy-Reviews and 4 Hard-Reviews to each participant. Second, we interleaved a Hard-Review between 2 Easy-Reviews. The locations of Easy-Review were distributed across 6 lessons. As a result, we had 4 groups of participants: HHHEHE, HHEHEH, HEHEHH, and EHEHHH; given H stands for Hard-Review and E means Easy-Review (Figure 3).

4.2 Procedure

Figure 3 showed the procedure of this study. Each participant visited our lab 6 times to take MOOC courses. The participant also visited our lab one more time for the final exam. To simulate self-paced learning in MOOCs, we let participants select the schedule by themselves.

Before starting a new lesson, participants review a topic (suggested by AttentiveReview[2]) of the previous lesson. This spaced reviewing approach has shown to be more effective than instant reviewing in reading comprehension [39]. Participants took a weekly test after taking 2 lessons in a week. The weekly test was conducted before the start of the next lesson, except for the last weekly test. We also collected weekly usability from participants. The usability survey includes 10 questions of the System Usability Scale (SUS) [4]. The usability survey was collected at the end of each week allowing participants to have more time to experience the system, especially in the first week. A pretest was conducted before lesson 1 of the study.

After each lesson, we collected participants' self-reports about their emotions, e.g. curiosity, boredom, and confusion, towards each topic in the lesson using 7-point Likert scale questions.

4.3 Participant and Apparatus

28 subjects from a local university (12 females), of whom average age was 26.3 (σ = 3.9), participated in the user study. 20 participants have taken at least one MOOC in the past. 18 participants have had the experience of watching tutorial videos on their smartphones.

The user study was conducted on a Nexus 6 smartphone (Android 7.0) with a 5.96 inch, 2560 x 1440 pixel display and a 2.7 GHz quad-core processor. The phone was equipped with a 13-megapixel back camera and a 2-megapixel front camera.

5 RESULTS

5.1 Subjective Feedback

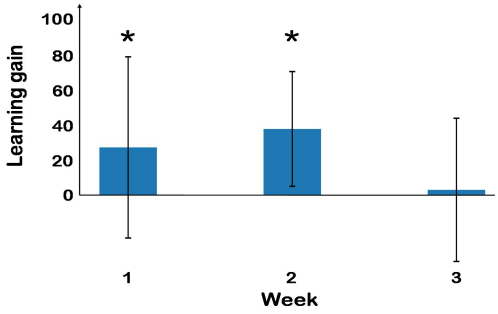

Figure 4. Learning gains of 3 weeks. * indicates a significant better performance than the pretest.

The average SUS over 3 weeks of this study was 80.5 (σ = 11.8), in which week 1 was 79.2 (σ = 10.6), week 2 was 80.5 (σ = 12.4), and week 3 was 81.6 (σ = 12.4). Note that previous research found the average SUS from 500 products was 68.0 [36] and an 80-ish SUS indicates a good product [2]. Even though there was a small increase of SUS after every week, the difference was not

statistically significant. The result suggests that AttentiveReview[2] was easy to learn and enjoyable to use.

Besides the SUS, we also collected subjective feedback from participants. In general, participants like AttentiveReview[2] because of its responsiveness and personalized recommendations. Some reoccurring positive feedback include: "*I like the auto pause feature [on-lens finger gesture]*", "*A lot of functions [a]b[o]ut video play very smoothly*", "*very easy to use and learn, very responsive*", and "*personalized review recommendations*".

On the other hand, there was also negative feedback to both AttentiveReview[2] and the learning material, such as ("*A video is long*" or "*can only review 1 sub session among 3. There is a possibility that I didn't learn well for 2 or 3 of them*") and the front camera widget ("*Face detection was not stable which consistently made me distracted*").

5.2 Learning Outcome

5.2.1 Learning Gain

We use the normalized learning gain, i.e. (weekly test - pretest) / (1 - pretest), as the performance metric. On average, participants in this study gained positive learning outcomes, i.e. mean learning gain of weekly test = 21.8% (σ = 0.4) when using AttentiveReview[2]. As shown in Figure 4, the average learning gains of week 1 was 26.1% (σ = 0.5), week 2 was 36.3% (σ = 0.3), and week 3 was 2.9% (σ = 0.4). Using one-sampled t-tests, we found the test scores of week 1 and week 2 were significantly better than pretest ($p < 0.01$). While the test score of week 3 was comparable to pretest ($t(27)=0.54$, $p = 0.29$).

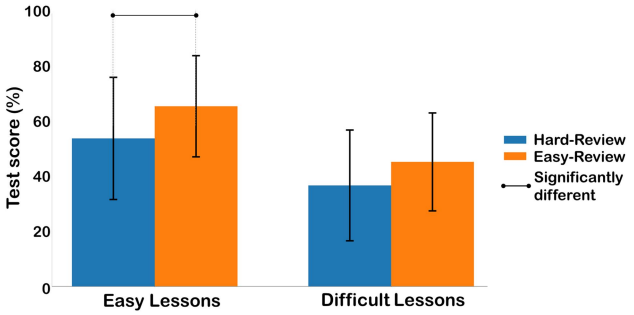

Figure 5. Average test scores of Hard-Review and Easy-Review in 2 groups: easy lessons and hard lessons.

5.2.2 Review Effectiveness.

In general, using AttentiveReview[2] led to a positive learning gain for our participants. However, we hypothesize that different recommending strategies (Hard-Review vs. Easy-Review) have different effectiveness when applied to learning materials with different difficulty levels (difficult topics vs. easy topics). To evaluate the reviewing effectiveness in different difficulty levels, we group the lessons based on the average pretest scores, into two groups: easy lessons (4, 2, 1) and difficult lessons (6, 3, 5).

Figure 5 showed the average weekly test scores of Hard-Review and Easy-Review in easy lessons and difficult lessons. The average score of Hard-Review on easy lessons was 53.8% (σ = 0.2) and on difficult lessons was 36.7% (σ = 0.2). While the mean score of Easy-Review on easy lessons was 65.5% (σ = 0.2) and on difficult lessons was 45.2% (σ = 0.2). Applying Hard-Review on easy lessons was significantly worse than applying Easy-Review ($t(64)=-2.38$, $p < 0.05$). However, the performance of the Hard-Review was comparable with the Easy-Review in difficult lessons as there were no significant differences between them ($t(64)=-1.88$, $p = 0.06$). This result suggested that, Hard-Review was not as effective as Easy-Review when reviewing easy lessons.

An explanation for this observation comes from Vygotsky's ZPD (zone of proximal development) theory [5]. The ZPD theory argues that a learner can only benefit from scaffolding if the lesson is still within her ZPD (not completely mastered, or too difficult). We hypothesize that all topics in the difficult lessons in this study were out of the participants' ZPD hence participants cannot benefit from any adaptive reviewing methods. On the other hand, the easy lessons may contain topics that were either within the ZPD or beyond (too difficult). Consequently, Easy-Review was more effective than Hard-Review in the easy lessons setting of our study as participants could review topics within the ZPD. This hypothesis also explains the difference in findings between this study and Pham and Wang [31] where Hard-Review was found more effective than Easy-Review. We hypothesize that the simple introduction of law topics in [31] would lie within participants' ZPD. Therefore, the authors found reviewing the most difficult topic (Hard-Review) was more effective than reviewing the easiest topic (Easy-Review). However, this hypothesis needs to be validated in follow-up studies.

5.3 Signal Analysis

In-depth signal analysis shows that the multimodal signals from AttentiveReview[2] can provide fine-grained feedback and benefit MOOC instructors. Moreover, these signals outperform the traditional clickstream analysis in predicting learners' emotion ratings and learning outcomes.

5.3.1 Facial Expression.

Figure 6 showed types of emotions expressed by each participant across six lessons. Each row is a lesson and each column is a participant. A 3x3 square indicated which emotion type a participant expressed in a lesson. An empty cell in the square means the participant did not express that emotion anytime during the lesson. Emotion types are (from left to right, top to bottom): anger, fear, sadness, surprise, joy, disgust, contempt, engagement, and attention. The output range of these 9 emotions in Affdex is [0, 100] which implies the detection confidence. We discarded all outputs less than 50 to avoid noisy predictions. Three emotions (contempt, engagement, and attention) were expressed by all participants in all lessons. Besides contempt, another negative emotion (disgust) was also expressed by many participants. In fact, AttentiveLearner[2] [33] found that both contempt and disgust expressions are helpful to detect learners' confusion, which frequently appears [9] and has a positive impact in learning [24]. From Figure 6, an instructor can quickly identify which lesson received the most negative emotions or whether a participant is getting bored when taking more lessons from the course.

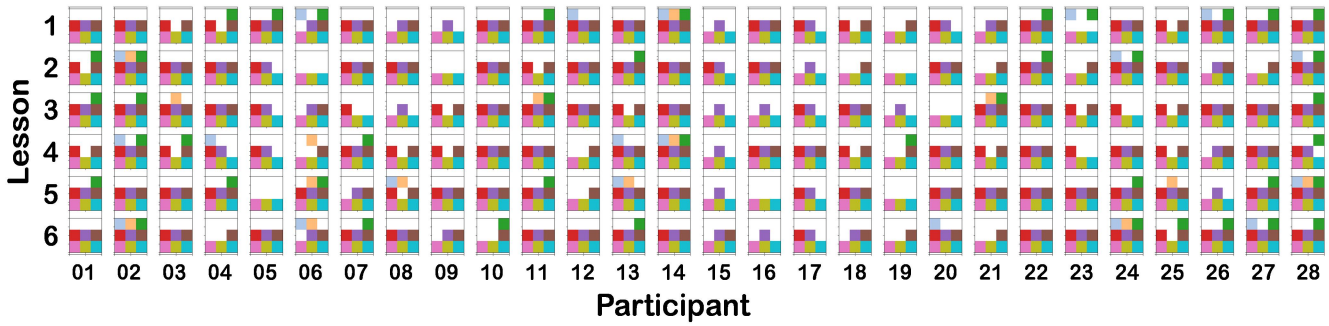

Figure 6. Facial emotions expressed by participants. A 3x3 square indicates which emotion types expressed by a participant: anger, fear, sadness, surprise, joy, disgust, contempt, engagement, and attention (left to right, top to bottom).

In addition to feedback from individual participants, the aggregated values of each emotion type can be valuable to instructors. Figure 7 shows the percentage of participants expressing engagement in every 30s of each lesson. We observed there was a sudden drop in participants expressing engagement (only within the first 30s) at the beginning of all lessons. The high engagement expression drop at the beginning of each lesson can be explained as - all participants adjusted their postures to make sure the facial recognition works before learning. On the other hand, each lesson has different temporal locations where most participants stay engaged, e.g. lesson 1: the 9.5th minute with 18 participants or lesson 3: the 11th minute with 20 participants. These high peaks in Figure 7 could serve as examples of good instruction for later deployments. By contrast, not many participants were engaged throughout lesson 4, which could raise an immediate alert to the course instructors implicitly.

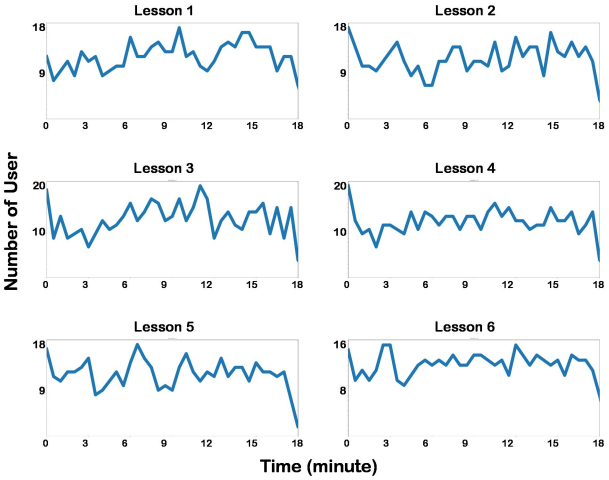

Figure 7. Number of participants showed Engagement expressions in every 10s across six lessons.

One major limitation of facial expression analysis in our study was the missing data. Compared to PPG signal, FEA experienced significantly more missing data. We defined a time t as a missing moment of a modality (facial data or PPG) when t lasts longer than 2s and AttentiveReview[2] did not receive any data from the modality during t. This 2-second threshold is quite conservative considering that the framerate of the back and the front cameras is around 30 frames per second. Using pairwise t-tests, we found the average missing data of for facial expression analysis (223.60s, σ=281.14) was significantly longer than that of PPG signal (5.69s, σ=6.33) with $t(54)=4.03$, $p < 0.01$.

5.3.2 Touching Data.

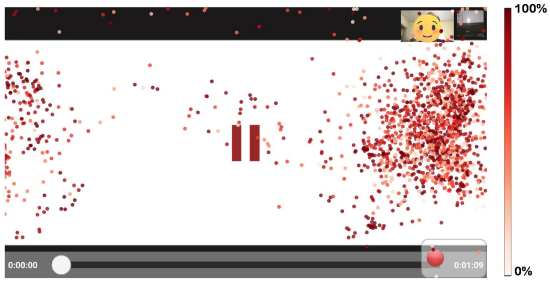

Figure 8. On-screen click locations and timestamps of all participants in all lessons.

Since AttentiveReview[2] uses on-lens finger gestures for video play back, clicking on the touch screen is not necessary during MOOC learning. Unexpectedly, participants still clicked on the touch screen frequently throughout the study (on average, there were 10.25 clicks per participant per lesson). Figure 8 showed the locations and timestamps of finger touches in this study. The darker a click is plotted, the later the click was done in a lesson (normalized by the lesson's length). Most of the clicks were not directly on the interface's widgets but located on the left-hand side and the right hand sides of the lecture videos.

Figure 9 showed the temporal touching distribution of 28 participants. The figure was sorted by the total number of touching moment of each participant. More clicks were done in the later lessons, e.g. lesson 5 and lesson 6, than at the first lessons. We also saw more clicks at the end of a lesson than at the beginning. From our observations during the study and follow up interviews, participants clicked the touch screen to check the tutorial's remaining time from the pop-up progress bar. By selecting the extreme groups of clicking participants, i.e. top 25.0% (Figure 9, top row) and bottom 25.0% (Figure 9, bottom row), we found a correlation between their curiosity ratings and

number of clicks using Spearman correlation (rho = 0.17, p < 0.1). The result suggested that when a participant clicked a lot while watching a tutorial video, she would lose curiosity about the lesson and only wait for the end of the lesson.

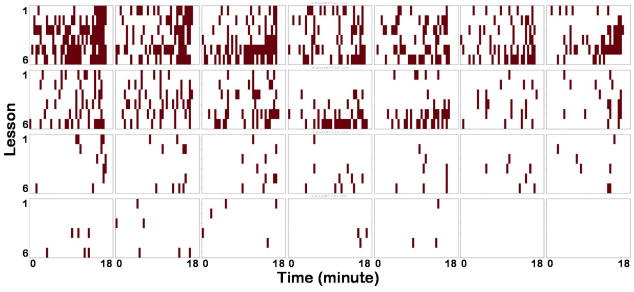

Figure 9. Touching data from 28 participants across 6 lessons. In each subplot, the horizontal axis is the lesson length and vertical axis is lesson number. Subplots were sorted by the number of click moments.

We compared the performance of screen touches (traditional clickstream analysis in MOOCs) and other fine-grained modalities (PPG signals and facial expressions) in predicting Curiosity and results in weekly tests. As in [33], we used 16 HRV features (8 global features and 8 local features). With facial features, we selected top 16 AUV features [33] in each tutorial video. For clickstream, we extracted 9 features (for both global and local sliding windows): total touches, mean of touching moment, standard deviation of touching moment, max of touching moment, min of touching moment, mean of latency between adjacent touches, standard deviation of latency between adjacent touches, max of latency between adjacent touches, and min of latency between adjacent touches. We selected the top 16 clickstream features to balance total of features between different modalities. A feature fusion model was created by concatenating 16 HRV features, 16 AUV features, and 16 clickstream features. All feature selections were done using univariate regression analysis. All features were fed into a linear regression model to predict curiosity ratings and weekly test scores.

Table 1 showed the mean square errors (MSEs) of unimodal and multimodal models. Overall, the fine-grained channels from AttentiveReview^2 outperformed the traditional clickstream channel. PPG signals had the best performance in predicting Curiosity when it was marginally better than facial features ($t(27)=-1.75$, $p < 0.1$) and clickstream features ($t(27)=-1.98$, $p < 0.1$). We did not find any significant difference between the performance of screen touch and facial data ($t(27)=1.06$, $p = 0.29$). Similar to this result, Pham and Wang [33] also found PPG signals gave a better performance than facial feature when predicting Curiosity in mobile MOOC learning. While PPG signals also gave the best performance in predicting weekly test score, we did not find any significant differences between these modalities. We found the feature fusion model were comparable with the best unimodal models in all tasks. When predicting Curiosity, the fusion model was comparable with PPG signals ($t(27)=-0.09$, $p = 0.93$). There was no significant difference between the fusion model and the PPG-based model in predicting weekly test score ($t(27)=1.22$, $p = 0.23$).

Table 1. Mean Square Error (MSE) of unimodal and feature fusion models in Curiosity rating and Weekly Test score. The reported results are MSE (standard deviation).

Modality	Curiosity	Weekly Test
Clickstream	1.96 (±1.72)	5.41% (±0.03)
PPG	1.87 (±1.63)	5.21% (±0.03)
Facial	1.93 (±1.69)	5.24% (±0.03)
Fusion	1.86 (±1.62)	5.41% (±0.03)

6 FUTURE WORK

There are at least three major directions in the near future. First, although we conducted a longitudinal study for three weeks, the study was still completed in a lab setting. The findings in this project may be biased to the environment in our lab. We plan to deploy AttentiveReview^2 in the wild to observe how to learners use AttentiveReview^2 in everyday environments.

Second, Verkoeijen et al. [39] found the reviewing performance depends on the distance between the learning time and the reviewing time, e.g. reviewing after 3.5 weeks did not give advantage compared to reviewing after 4 days. We also identified another factor affecting the reviewing performance, i.e. the relationship between a topic's difficulty and the learner's current ZPD. We plan to study the effectiveness of reviewing strategies which take both reviewing time and learners' ZPD into account.

Last but not least, we are working on a new design for the facial widget. The facial widget intents to be an awareness channel revealing whether AttentiveReview^2 can capture learners' facial expressions reliably. However, many participants reported the widget to be distracting because it revealed too many details from the learning environment. We plan to design an icon style facial widget that can facilitate facial data collection without disclosing private information around learners.

7 CONCLUSIONS

We presented AttentiveReview^2, a multimodal intelligent tutoring system running on unmodified smartphones. AttentiveReview^2 collects rich learning signals from three modalities: PPG signals, facial expressions, and clickstream. Through a 3-week longitudinal study with 28 participants, we found that AttentiveReview^2 on average improved learning gains by 21.8% in weekly tests. Follow-up analysis showed that multimodal signals collected from the learning process can also benefit instructors by providing rich and fine-grained insights on the learning progress. In summary, AttentiveReview^2 showed the feasibility and potential of a multimodal affect-aware intelligent tutoring system for MOOC learning on today's smartphones without additional hardware modifications.

REFERENCES

[1] Afergan, D., Peck, E.M., Solovey, E.T., Jenkins, A., Hincks, S.W., Brown, E.T., Chang, R. and Jacob, R.J., 2014, April. Dynamic difficulty using brain metrics of workload. In Proceedings of the 32nd annual ACM conference on Human factors in computing systems (pp. 3797-3806). ACM.

[2] Bangor, A., Kortum, P. and Miller, J., 2009. Determining what individual SUS scores mean: Adding an adjective rating scale. Journal of usability studies, 4(3), pp.114-123.

[3] Brinton, C.G., Rill, R., Ha, S., Chiang, M., Smith, R. and Ju, W., 2015. Individualization for education at scale: MIIC design and preliminary evaluation. IEEE Transactions on Learning Technologies, 8(1), pp.136-148.

[4] Brooke, J., 1996. SUS-A quick and dirty usability scale. Usability evaluation in industry, 189(194), pp.4-7.

[5] Chaiklin, S., 2003. The zone of proximal development in Vygotsky's analysis of learning and instruction. Vygotsky's educational theory in cultural context, 1, pp.39-64.

[6] Chen, C., Alcorn, B., Christensen, G., Eriksson, N., Koller, D. and Emanuel, E., 2015. Who's benefiting from MOOCs, and Why. Harvard Business Review, 22.

[7] Chuang, I. and Ho, A.D., 2016. HarvardX and MITx: Four Years of Open Online Courses--Fall 2012-Summer 2016.

[8] Coetzee, D., Fox, A., Hearst, M.A. and Hartmann, B., 2014, March. Chatrooms in MOOCs: all talk and no action. In Proceedings of the first ACM conference on Learning@ scale conference (pp. 127-136). ACM.

[9] D'Mello, S.K. and 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), pp.147-187.

[10] D'Mello, S. and Graesser, A., 2012. Dynamics of affective states during complex learning. Learning and Instruction, 22(2), pp.145-157.

[11] D'Mello, S., Kopp, K., Bixler, R.E. and Bosch, N., 2016, May. Attending to attention: Detecting and combating mind wandering during computerized reading. In Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (pp. 1661-1669). ACM.

[12] Dhawal Shah. 2018. By The Numbers: MOOCS in 2017. Retrieved April 2018 from https://www.class-central.com/report/mooc-stats-2017

[13] Dunlosky, J., Rawson, K.A., Marsh, E.J., Nathan, M.J. and Willingham, D.T., 2013. Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. Psychological Science in the Public Interest, 14(1), pp.4-58.

[14] Grafsgaard, J., Wiggins, J.B., Boyer, K.E., Wiebe, E.N. and Lester, J., 2013, July. Automatically recognizing facial expression: Predicting engagement and frustration. In Educational Data Mining 2013.

[15] Guo, P.J., Kim, J. and Rubin, R., 2014, March. How video production affects student engagement: An empirical study of MOOC videos. In Proceedings of the first ACM conference on Learning@ scale conference (pp. 41-50). ACM.

[16] Gütl, C., Rizzardini, R.H., Chang, V. and Morales, M., 2014, September. Attrition in MOOC: Lessons learned from drop-out students. In International Workshop on Learning Technology for Education in Cloud (pp. 37-48). Springer, Cham.

[17] Han, T., Xiao, X., Shi, L., Canny, J. and Wang, J., 2015, April. Balancing accuracy and fun: Designing camera based mobile games for implicit heart rate monitoring. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (pp. 847-856). ACM.

[18] Hussain, M.S., Calvo, R.A. and Chen, F., 2013. Automatic cognitive load detection from face, physiology, task performance and fusion during affective interference. Interacting with computers, 26(3), pp.256-268.

[19] Jiang, Y., Baker, R.S., Paquette, L., San Pedro, M. and Heffernan, N.T., 2015, June. Learning, Moment-by-Moment and Over the Long Term. In International Conference on Artificial Intelligence in Education (pp. 654-657). Springer, Cham.

[20] Karsenti, T., 2013. What the research says. International Journal of Technologies in Higher Education, 10(2), pp.23-37.

[21] Kim, J., Guo, P.J., Seaton, D.T., Mitros, P., Gajos, K.Z. and Miller, R.C., 2014, March. Understanding in-video dropouts and interaction peaks in online lecture videos. In Proceedings of the first ACM conference on Learning@ scale conference (pp. 31-40). ACM.

[22] Kovacs, G., 2015, April. QuizCram: A Question-Driven Video Studying Interface. In Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (pp. 133-138). ACM.

[23] Krause, M., Mogalle, M., Pohl, H. and Williams, J.J., 2015, March. A playful game changer: Fostering student retention in online education with social gamification. In Proceedings of the Second (2015) ACM Conference on Learning@ Scale (pp. 95-102). ACM.

[24] Lehman, B. and Graesser, A., 2015, June. To resolve or not to resolve? that is the big question about confusion. In International Conference on Artificial Intelligence in Education (pp. 216-225). Springer, Cham.

[25] McDuff, Daniel Jonathan. Crowdsourcing affective responses for predicting media effectiveness. PhD diss., Massachusetts Institute of Technology, 2014.

[26] McDuff, D., Mahmoud, A., Mavadati, M., Amr, M., Turcot, J. and Kaliouby, R.E., 2016, May. AFFDEX SDK: a cross-platform real-time multi-face expression recognition toolkit. In Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (pp. 3723-3726). ACM.

[27] Mehta, Deepak, The Future Of Massively Open Online Courses (MOOCs), #OnCampus, Forbes.com, March 23, 2017.

[28] Miranda, S., Mangione, G.R., Orciuoli, F., Gaeta, M. and Loia, V., 2013, October. Automatic generation of assessment objects and Remedial Works for MOOCs. In Information Technology Based Higher Education and Training (ITHET), 2013 International Conference on (pp. 1-8). IEEE.

[29] Monkaresi, H., Bosch, N., Calvo, R.A. and 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), pp.15-28.

[30] Oviatt, Sharon. The design of future educational interfaces. Routledge, 2013.

[31] Pham, P. and Wang, J., 2016, October. Adaptive review for mobile MOOC learning via implicit physiological signal sensing. In Proceedings of the 18th ACM International Conference on Multimodal Interaction (pp. 37-44). ACM.

[32] Pham, P. and Wang, J., 2015, June. AttentiveLearner: improving mobile MOOC learning via implicit heart rate tracking. In

International Conference on Artificial Intelligence in Education (pp. 367-376). Springer, Cham.

[33] Pham, P. and Wang, J., 2017, June. AttentiveLearner2: A Multimodal Approach for Improving MOOC Learning on Mobile Devices. In International Conference on Artificial Intelligence in Education (pp. 561-564). Springer, Cham.

[34] Picard, Rosalind. Affective computing. Cambridge: MIT press, 1997.

[35] Raghuveer, Raghuveer, V.R., Tripathy, B.K., Singh, T. and Khanna, S., 2014, December. Reinforcement learning approach towards effective content recommendation in MOOC environments. In MOOC, Innovation and Technology in Education (MITE), 2014 IEEE International Conference on (pp. 285-289). IEEE.

[36] Sauro, J., 2011. A practical guide to the system usability scale: Background, benchmarks & best practices. Measuring Usability LLC.

[37] Szafir, D. and Mutlu, B., 2013, April. ARTFul: adaptive review technology for flipped learning. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1001-1010). ACM.

[38] Van der Sluis, F., Ginn, J. and Van der Zee, T., 2016, April. Explaining Student Behavior at Scale: The influence of video complexity on student dwelling time. In Proceedings of the Third (2016) ACM Conference on Learning@ Scale (pp. 51-60). ACM.

[39] Verkoeijen, P.P., Rikers, R.M. and Özsoy, B., 2008. Distributed rereading can hurt the spacing effect in text memory. Applied Cognitive Psychology, 22(5), pp.685-695.

[40] Weiner, B., 1985. An attributional theory of achievement motivation and emotion. Psychological review, 92(4), p.548.

[41] Xiao, X. and Wang, J., 2016, October. Context and cognitive state triggered interventions for mobile MOOC learning. In Proceedings of the 18th ACM International Conference on Multimodal Interaction (pp. 378-385). ACM.

[42] Xiao, X. and Wang, J., 2015, November. Towards attentive, bi-directional MOOC learning on mobile devices. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (pp. 163-170). ACM.

[43] Xiao, X., Pham, P. and Wang, J., 2017, June. Dynamics of Affective States During MOOC Learning. In International Conference on Artificial Intelligence in Education (pp. 586-589). Springer, Cham.

[44] Xiao, X., Han, T. and Wang, J., 2013, December. LensGesture: augmenting mobile interactions with back-of-device finger gestures. In Proceedings of the 15th ACM on International conference on multimodal interaction (pp. 287-294). ACM.

[45] Yang, D., Sinha, T., Adamson, D. and Rosé, C.P., 2013, December. Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In Proceedings of the 2013 NIPS Data-driven education workshop (Vol. 11, p. 14).