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, a multimodal intelligent tutoring system optimized for MOOC learning on unmodified mobile devices. AttentiveReview 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 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.

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.

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 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. Coetsee 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. Raghuvane 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.

2.3.1 Triple Stream Signal Sensing

AttentiveReview 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.

2.3.2 Triple Stream Signal Sensing

AttentiveReview 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 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 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 utilizes the front camera to capture a learner’s facial expressions while watching lecture videos. As a result, AttentiveReview enables the automatic collection of both PPG signals and facial expressions implicitly during learning sessions.
AttentiveReview\textsuperscript{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\textsuperscript{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\textsuperscript{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 \times 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.
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 multiple times 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) 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

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 was easy to learn and enjoyable to use. Besides the SUS, we also collected subjective feedback from participants. In general, participants like AttentiveReview 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]bout 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 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. 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.](image)

5.2.2 Review Effectiveness

In general, using AttentiveReview 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 low 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 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 [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.
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.

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 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.

Since AttentiveReview 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.](image)

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).

<table>
<thead>
<tr>
<th>Modality</th>
<th>Curiosity</th>
<th>Weekly Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clickstream</td>
<td>1.96 (±1.72)</td>
<td>5.41% (±0.03)</td>
</tr>
<tr>
<td>PPG</td>
<td>1.87 (±1.63)</td>
<td>5.21% (±0.03)</td>
</tr>
<tr>
<td>Facial</td>
<td>1.93 (±1.69)</td>
<td>5.24% (±0.03)</td>
</tr>
<tr>
<td>Fusion</td>
<td>1.86 (±1.62)</td>
<td>5.41% (±0.03)</td>
</tr>
</tbody>
</table>

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 intends 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


