Adaptive Review for Mobile MOOC Learning via Implicit Physiological Signal Sensing
Phuong Pham\textsuperscript{1,2} and Jingtao Wang\textsuperscript{1,2}
\textsuperscript{1}Department of Computer Science
\textsuperscript{2}Learning Research and Development Center (LRDC)
University of Pittsburgh, PA, USA
\{phuongpham, jingtaow\}@cs.pitt.edu

ABSTRACT
Massive Open Online Courses (MOOCs) have the potential to enable high quality knowledge dissemination in large scale at low cost. However, today’s MOOCs also suffer from low engagement, uni-directional information flow, and lack of personalization. In this paper, we propose AttentiveReview, an effective intervention technology for mobile MOOC learning. AttentiveReview infers a learner’s perceived difficulty levels of the corresponding learning materials via implicit photoplethysmography (PPG) sensing on unmodified smartphones. AttentiveReview also recommends personalized review sessions through a user-independent model. In a 32-participant user study, we found that: 1) AttentiveReview significantly improved information recall (+14.6\%) and learning gain (+17.4\%) when compared with the no review condition; 2) AttentiveReview also achieved comparable performances at significantly less time when compared with the full review condition; 3) As an end-to-end mobile tutoring system, the benefits of AttentiveReview outweigh side-effects from false positives and false negatives. Overall, we show that it is feasible to improve mobile MOOC learning by recommending review materials adaptively from rich but noisy physiological signals.

CCS Concepts
\begin{itemize}
  \item Human-centered computing→Ubiquitous and mobile computing→Ubiquitous and mobile computing systems and tools
\end{itemize}

Keywords
MOOC; Heart Rate; Intelligent Tutoring System; Physiological Signal; Affective Computing; Mobile Interface.

1. INTRODUCTION
Massive Open Online Courses (MOOCs) are changing the landscape of how people learn. MOOCs can provide high quality educational contents in large scale at low cost. More importantly, MOOC learners have control over what, when and how they learn. As the result, the number of registered MOOC learners reached 35 million in December 2015 [26]. A major portion of learners has reported both career benefits (72\%) and educational benefits (61\%) after finishing MOOCs [3].

Tutorial videos are the primary learning resource on today’s MOOC platforms. Different from hour-long lecture videos recorded in traditional classrooms, the video length in MOOCs is usually brief, typically 3-15 minutes, for increased engagement [11]. Such short video clips are also ideal to watch on mobile devices. Indeed, major MOOC providers, such as Coursera, edX, Khan Academy, and Udacity, have released their mobile apps to offer “on the go” learning environments.

Despite the great potential, today’s MOOCs also suffer from low retention rates (e.g. 7%-9\% in Coursera [22]), low engagement [14][15], lack of student-instructor interaction [10][21], and little personalization [2][20][24]. Such “one-size-fits-all” dilemma in MOOCs was caused, at least in part, by three major reasons. First, the restrictive affordance of prerecorded video. The static and asynchronous nature of tutorial videos increase the cost of offering personalized learning experiences; Second, insufficient monitoring and modeling of the learning process. Different from traditional classrooms, instructors in MOOCs no longer have access to important cues such as raised hands or facial expressions at the time of teaching. Also different from traditional intelligent tutoring systems (ITS), only limited profile and behavioral information, such as self-reported background [24] and clickstream analysis [5][11][14], are available for researchers to understand the learning process. As a result, today’s modeling efforts in MOOCs focus primarily on providing the aggregated learning analytics to instructors and stakeholders rather than offering personalized interventions to learners; Third, increased costs in designing fine-grained in-lecture assessments. Although it...
is possible to integrate the assessment process [21] into the video watching interface, designing evaluative questions for each learning topic in today’s MOOC offerings could be expensive and time consuming.

In this paper, we propose AttentiveReview (Figure 1), an intelligent intervention technology for mobile MOOC learning. AttentiveReview uses on-lens finger gestures for video control, i.e. covering the back camera lens to play a lecture video, uncovering the camera lens to pause the video. Throughout a learning session, AttentiveReview infers learners’ perceived difficulty levels during educational presentations and recommends lecture topics that learners might benefit most from reviewing by analyzing the photoplethysmography (PPG) waveforms captured implicitly from fingertip transparency changes. AttentiveReview builds upon and extends findings of AttentiveLearner [23][30]. While AttentiveLearner focused on demonstrating the feasibility of using implicit PPG signals to detect mind wandering [23], boredom and confusion [30] in mobile MOOC learning via post-hoc analysis, this research shows both the feasibility and efficacy of adaptive content review in an interactive system.

The contributions are threefold:

- We present the design, evaluation, and direct benefits of AttentiveReview, an adaptive review interface for mobile MOOC learning in a 32-participant study.
- We show that AttentiveReview, based on inferring the perceived difficulty levels from implicit PPG signal sensing, can improve both information recall and the learning gain when compared with the no review condition. AttentiveReview can also achieve comparable performances with significantly less time when compared with a full review condition.
- Through a series of fine-grained analysis on experimental results, we show that the benefits of AttentiveReview outweigh side-effects from false positives and false negatives.

2. RELATED WORK

2.1 Understanding and Improving MOOCs

Given the opportunities and challenges that MOOCs face, researchers have made significant efforts 1) to understand how students learn in MOOCs [4] [11][14][25]; and 2) to improve today’s MOOC interfaces [10][15][21] and delivering strategies [2][20][24].

Researchers have leveraged clickstream analysis to understand learners’ in-video [11][14][25], in-forum [4], and in-course [5] behaviors in MOOCs. In-video behaviors, e.g. playing, pausing or rewinding a video, have revealed five scenarios causing low engagement [14]. Sluis and colleagues [25] discovered a polynomial relationship between learner’s dwelling time and video complexity. In-forum behaviors, such as posting, viewing, and replying, show that most students do not actively discuss [4] in discussion forums offered by today’s MOOCs. In-course behaviors, e.g. liking a document or sharing a link, have been used to understand what contributes to a student’s correct response, and what helps a learner get the certificate [5]. Although clickstream analysis can reveal informative insights from existing activity logs, it works better for disclosing the aggregated trend from thousands of learners for revisions of course content in the future, rather than providing personalized scaffolding for individual learners.

Researchers have invented new interfaces and interaction techniques to improve the usability, engagement, and retention of MOOCs. L.IVE by Monserrat et al. [21] showed the benefits of overlaying and integrating discussions and assessments on top of the video watching interface. Glassman and colleagues [10] created Mudslide, a rich feedback collection and visualization interface, to facilitate the information flow from students to instructors. Krause et al. [15] recently explored the use of social gamification to improve the retention rate in MOOCs.

Personalization has also been explored in MOOCs to improve both retention and learning outcomes. Researchers have explored the use of learners’ browsing behavior [2], learning objectives [24], and assessment performance [20] to provide more relevant learning materials. However, behavioral data in MOOCs are usually sparse, e.g. there could be only one mouse click in a typical 8-min video watching session. As a result, multiple learning sessions from a large number of learners are necessary at the bootstrapping stage. At the same time, most personalization techniques rely on learners’ performance on assessment questions, which may not always be available if learners skip the quizzes or there are no assessment questions for a specific learning topic. In comparison, AttentiveReview is an adaptive tutoring system for mobile MOOC learning. AttentiveReview leverages fine-grained PPG signals implicitly collected from tutorial videos, rather than clickstreams or results of quizzes, to infer learners’ perceived difficulty level in each topic and adaptively suggests the optimal review material.

2.2 Physiological Signals in Education

Physiological signals, e.g. skin conductance (a.k.a. EDA or GSR) [29], electroencephalogram (EEG) [27], functional near-infrared spectroscopy (fNIRS) [1], heart rate/PPG [13][18][23][30], eye gaze [7], and facial expressions [29], can be collected to infer learners’ cognitive and affective states during the learning process. While cognitive states (e.g. attention, memory workload) directly influence learning outcomes, affective states (e.g. confusion, boredom, and frustration) also have indirect impacts on learning [6]. Affective AutoTutor [8] used conversational cues, gross body language, and facial features to detect learner’s boredom, confusion, and frustration. GazeTutor by D’Mello et al. [7] explored the use of eye gazes to detect attentional disengagement. ARTFul [27] leveraged EEG to infer learner’s attention. The Wayang Tutor [29] successfully used a combination of GSR, facial expression, mouse pressure, and learning posture to infer learners’ affective states, e.g. motivation, frustration, and engagement, in learning.

Difficult task could affect learners’ engagement [1] and performance [17]. A difficult task could overload learners’ cognitive capacity. This overload can further induce negative emotions, e.g. anxiety and anger. Afergan et al. [1] was able to differentiate perceived difficulty levels from fNIRS signals. Lyu et al. [18] discovered a decrease in several heart rate variability components in ECG signals, such as high frequency and mean interbeat, for increased task difficulties. McDuff et al. [19] found that the high frequency component of heart rate variability in PPG signals was significantly reduced in the high difficulty condition.

One common limitation with most of current affect/cognitive state aware tutoring systems [1][8][7][18][27][29] is the requirement of dedicated sensors, such as wristbands, gaze trackers, and EEG headsets. The cost, availability, and portability of such sensors have become a major obstacle preventing the wide adoption of such technologies in real world scenarios, especially in MOOCs.
In contrast, our proposed system runs on unmodified smartphones and uses the built-in camera as an implicit PPG sensor, hence eliminates the need for dedicated physiological signal sensors.

AttentiveLearner [23][30] is the most relevant project to this research. Pham et al. demonstrated the feasibility of detecting mind wandering [23]. Xiao and colleagues [30] systematically studied the usability of AttentiveLearner and proposed the use of aggregated learning events and extreme learning events to improve the prediction accuracy of learners’ affective states. Despite the promising outcomes, existing efforts of AttentiveLearner have been focusing on the feasibility of inferring and presenting learners’ cognitive and affective states for instructors. No student-side intervention was studied to improve the learning outcomes of mobile MOOCs.

AttentiveReview builds upon and improves AttentiveLearner by providing direct benefit to learners via in-situ adaptation and feedback. In particular, AttentiveReview suggests appropriate review materials based on learners’ perceived difficulty levels in each learning topic. To the best of our knowledge, AttentiveReview is the first affect-aware learning system that provides direct, measurable benefits for mobile MOOC learners.

3. DESIGN OF ATTENTIVERVIEW

AttentiveReview is an intelligent tutoring system optimized for mobile MOOC learning. Similar to popular mobile MOOC clients such as Coursera, EdX, and Khan Academy, users of AttentiveReview can browse, stream, and watch tutorial videos on their mobile phones. Figure 2 shows the primary video watching interface of AttentiveReview.

AttentiveReview recommends lecture topics learners might benefit most from by inferring the perceived difficulty levels of a learner through sensing and analyzing the photoplethysmography (PPG) waveforms of learners implicitly. Similar to AttentiveLearner [23][30], AttentiveReview naturally blends the PPG sensing process into the video consuming process via a tangible video control channel.

AttentiveReview consists of four major components: 1) a tangible video control channel; 2) an implicit PPG waveform sensing module; 3) on-screen video play and PPG visualization interface; and 4) algorithms for supporting adaptive review.

3.1 Tangible Video Control

AttentiveReview recommends lecture topics learners might benefit most from by inferring the perceived difficulty levels of a learner through sensing and analyzing the photoplethysmography (PPG) waveforms of learners implicitly. Similar to AttentiveLearner [23][30], AttentiveReview naturally blends the PPG sensing process into the video consuming process via a tangible video control channel.

AttentiveReview builds upon and improves AttentiveLearner by providing direct benefit to learners via in-situ adaptation and feedback. In particular, AttentiveReview suggests appropriate review materials based on learners’ perceived difficulty levels in each learning topic. To the best of our knowledge, AttentiveReview is the first affect-aware learning system that provides direct, measurable benefits for mobile MOOC learners.

3.2 Implicit PPG Sensing

AttentiveReview converts the built-in back camera of unmodified smartphones as an implicit PPG sensor. In every cardiac cycle, the heart pumps blood to capillary vessels and changes the transparency of the corresponding human body parts, including the fingertips. These transparency changes correlate directly with heart beats and can be detected by the covered back camera. The implicit PPG signals have been used to infer users’ cognitive and affect states [23][30].

3.3 Video Playback Interface

The primary user interface of AttentiveReview is similar to that of a mobile video player, with additional widgets for visualizing the camera preview window, the attention indicator and the PPG preview window (Figure 2). The camera preview window shows a live video stream from the back camera. The attention indicator shows a learner’s instant heart rate and finger-covering state. The PPG indicator shows the live waveform of a learner’s PPG signals. A learner can move both the attention indicator and the PPG indicator by drag-and-drop, and can also switch each indicator off via a single finger tap.

3.4 Adaptive Review Algorithm

Our adaptive algorithm bases on the pedagogical principle that reviewing relevant and appropriate tutorial materials will improve learning [16][27].

AttentiveReview extracts both temporal domain and frequency domain features from a learner’ PPG waveforms collected from the learning process. AttentiveReview then uses a ranking SVM algorithm to determine learners’ perceived difficulty in each learning topic, and suggests the learner to review the most difficult topic.

AttentiveReview uses LivePulse [12], a heuristic peak counting algorithm, to identify normal-to-normal (NN) inter-beat intervals from PPG signals. The NN interval signal is smoothed and then resampled to 20Hz.

As shown in Figure 3, for each topic in a tutorial video, AttentiveReview skips the first and the last 30 seconds of PPG waveforms to minimize noise and carry-over effects. We extract 8 dimensions of heart rate variability (HRV) related features from both the truncated topic window (i.e. global features) and a 15-second non-overlapping sliding window within a topic (i.e. local features). The definitions of the features are: 1) AVNN (average heart rate); 2) SDNN (temporal standard deviations of heart beats); 3) pNN10 (percentage of adjacent heart beats with a
4. USER STUDY

We conducted a user study to further understand AttentiveReview. We had two major goals for this study. First, we would like to evaluate the usability of AttentiveReview in actual mobile MOOC learning sessions. Second, we would like to investigate the feasibility and efficacy of improving the learning outcome of MOOCs through personalized review recommendations via implicit physiological signal sensing.

4.1 Experimental Design

The study consisted of four main phases: 1) background survey and pre-test; 2) MOOC learning; 3) MOOC reviewing; 4) post-test and closing survey (Figure 4). To simulate the facts that reviews in the real world do not immediately follow the learning sessions, we scheduled a two-minute relaxation session between the learning session and the review session, and between the review session and the post-test session. Learners played Candy Crush Saga, a popular mobile game on major mobile platforms, during relaxation sessions.

We designed four different interventions for the MOOC reviewing phase. First, no review. Learners do not complete any review; Second, full review. Learners go through all topics of a lecture video in this condition; Third, adaptive review. Learners review the most difficult topic inferred by our algorithm. Fourth, counter adaptive review. Learners review the easiest topic inferred by our algorithm. We included the counter adaptive condition to differentiate whether the reviewing action or the reviewing content could have an impact on learning outcomes. In the no review condition, the participants listened to the song "Twelve Variations on "Ah vous dirai-je, Maman"" by Wolfgang Amadeus Mozart for the duration of the review session. Otherwise, the participants watched the corresponding review video generated by our algorithm.

To illustrate the fact that majority of MOOCs today are offered in English, but learners include both native English speakers and English as Second language speakers, we recruited both native participants (EL1) and non-native participants (EL2) to evaluate the impact of native language on the outcomes of reviewing.

In summary, our study used a 4x2 between-subjects design. The independent variables were review condition (no review, full review, adaptive, and counter adaptive) and participant’s native language (EL1 and EL2). The type and order of the independent variables have been counter-balanced via Latin-square patterns. The dependent variables were participant’s performance on information recall (Recall) and learning gain (Learning). Recall was measured by the accuracy (# correct answers / # questions) of a learner in the post-test. Learning gain was measured as the percentage accuracy difference between the post-test and the pre-test. In the end, each participant completed a closing survey including 1) subjective ratings about AttentiveReview’s usability, and 2) perceived difficulty ranking of each topic in descending order (most perceived difficult = Rank 1; least perceived difficult = Rank 3). The perceived difficulty rankings would be used as the ground truth for AttentiveReview in sections 5.4.2 and 5.4.3.

4.2 Learning Material

To avoid the interference with participants’ background knowledge, we chose law, an area that was unfamiliar to all the participants, as the learning topic. In the study, participants watched a lecture on law consisting three topics: criminal laws, human rights, and surveillance laws. The corresponding tutorial videos were from Coursera. The criminal law was taught by Professor Stephen Morse from University of Pennsylvania. The topic of human rights was taught by Professor Laurence R. Helfer of Duke University. The surveillance laws’ topic was taught by

---

1 https://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html
lawyer Jonathan Mayer at Stanford University. Each topic lasted for 8 minutes, leading to a 24-minute lecture video.

We created 9 multiple-choice questions for the pre-test (3 questions per topic) and 24 multiple-choice questions for the post-test (8 questions per topic). Samples questions include, “What kind of disputes does Tort law deal with?” (criminal laws) and “When did the bulk domestic email surveillance program end?” (surveillance laws).

4.3 Participants and Apparatus
There were 32 subjects (9 females; 16 native English speakers and 16 non-native English speakers) from a local university participating in our study. Each of the four review conditions was balanced with four EL1 participants and four EL2 participants. The average age was 23.6 (σ = 4.2). No participant reported any exposure to any of the three law topics prior to our study.

Our experiment was completed on a Nexus 5 smartphone with a 4.95 inch, 1920 x 1080pixel display, 2.26 GHz quad-core processor, running Android 5.0.

5. RESULTS

5.1 Subjective Feedback
Participants reported positive experiences with AttentiveReview in general (Figure 5). To be specific, participants found AttentiveReview was easy to use 4.30 (σ = 0.81) and responsive 4.42 (σ = 0.5) on a five-point Likert scale. Detailed comments include ‘[AttentiveReview was] Easy to play and pause’, ‘I like that I can control the app without touching the screen’, ‘Using the covering back camera strategy is very natural’.

Figure 5. Subject feedback on a five-point Likert scale.

Although participants were positive about using AttentiveReview for their future learning activities 3.82 (σ =1.05), they also reported usability problems on comfort 3.73 (σ =1.09), primarily caused by the heat emitted by the camera flash: ‘Sometimes my finger got hot’, ‘Light got hot after a while which was slightly uncomfortable’. It was worth noticing that participants watched a 24-minute tutorial video, which was longer than lecture clips used in AttentiveLearner [30] and most MOOC platforms. We hope smartphone manufacturers could take the heat emission ratio of the flash light into account when choosing camera optical assemblies in the future. It was also possible to capture PPG signals without turning on the flash light or only turning it on intermittently at a higher signal-noise-ratio (SNR). We plan to investigate such alternative settings in follow-up studies.

5.2 Signal Quality
Figure 6 illustrates the PPG signal quality captured by AttentiveReview from eight participants during our study. We adopted the same signal quality metric as Xiao and Wang[30]. That is, in a 5-second normal-to-normal (NN) inter-beat signal window, the signal in the window was rated high quality if at least 80% of the NN intervals were within the ± 25% range of the window’s median. On average, 93.76% (σ = 4.39%) of the PPG signals from this study were in high quality. This finding confirmed the reliability of PPG signals captured by AttentiveReview.

Figure 6. Sample PPG signal quality of eight participants.

Figure 7 shows the HRV spectrograms (normalized amplitude) from the perceived least difficult topic (top row) and the perceived most difficult topic (second row) of six participants. The HRV spectrograms were plotted by calculating the power spectral density from NN intervals [19]. Each topic used a one-minute sliding window with one second increment to calculate power spectral density. The high frequency (HF) power was reduced under the stress condition [18][19]. The spectrograms show similar observations for those learning difficult topics. The low HF power in the bottom row indicated corresponding learners were under higher cognitive workload than those learning least difficult topics.

Figure 7: Heart rate variability spectrogram (LF and HF) of six participants (P3, P9, P14, P19, P25, P27) in their perceived least difficult topic (top row) and perceived most difficult topic (second row).

5.3 Learning Outcome
Figure 8 illustrates the learning outcomes (Left: learning recall; Right: learning gain) by review conditions (None: no review; Counter: counter adaptive review; Adapt: adaptive review; Full: full review) on information recall (Recall) and learning gain (Learning). Only significant p values are reported.

Figure 8. Learning outcomes (Left: learning recall; Right: learning gain) by review conditions (None: no review; Counter: counter adaptive review; Adapt: adaptive review; Full: full review) on information recall (Recall) and learning gain (Learning). Only significant p values are reported.
Using a two-way ANOVA analysis, we found significant main effects on both Recall (F(2,24) = 7.91, p < .001) and Learning (F(2,24) = 4.04, p < .05) among the four review conditions. The average Recall scores were 54.17% (σ = 5.89%), 55.73% (σ = 10.89%), 68.75% (σ = 9.18%), and 72.39% (σ = 8.01%) in the no review, counter adaptive review, adaptive review, and full review conditions respectively. Learning scores were 19.44% (σ = 16.85%), 26.56% (σ = 11.32%), 36.81% (σ = 17.2%), and 43.23% (σ = 19.89%) in the no review, counter adaptive review, adaptive review, and full review conditions.

We also used pairwise mean comparison (t-tests) with Bonferroni correction to better understand the relative performances (Figure 8). There were two major findings from Figure 8.

First, participants taking any review performed better or at least equally well in both Recall and Learning than participants having no review. In particular, participants taking full review performed significantly better than participants taking no review in both Recall (t(14) = 15.64, p < .001) and Learning (t(14) = 10.22, p < .005). Similarly, adaptive review had significantly better performances in both Recall (t(14) = 10.01, p < .005) and Learning (t(14)= 5.45, p < .005) than no review. In the case of counter adaptive review, even watching easy learning materials for one extra time did not hurt performances in Recall (t(14) = 0.11, p = .738) or Learning (t(14)= 0.92, p = .348) when compared with no review. This result confirmed that reviewing in general and AttentiveReview, in particular, can improve learning outcome. At the same time, this result showed that adaptive review was a low risk intervention even if the prediction was imperfect.

Second, adaptive review was more efficient in time than full review under comparable performance. There were no significant differences between the adaptive condition and the full review condition in both Recall (t(14) = 0.63, p = .437) and Learning (t(14) = 0.75, p = .396). In other words, AttentiveReview was able to achieve equivalent cognitive learning performance as a full review, with 66.7% less review time. These results show the efficacy of the adaptation algorithm in AttentiveReview.

There was a significant main effect on Learning (t(30)= 5.04, p = .034) between learners’ native language. EL1 scored significantly higher on Learning 37.41% (σ = 20.77%) than EL2 25.61% (σ = 13.71%). The different on Recall between EL1 and EL2 was not significant (p = .529).

5.4 Detecting Perceived Difficulty

5.4.1 Perceived Difficulty vs. Learning Recall

We also investigated the relationship of the four review conditions and self-reported difficulty levels on learning recall. Two-way ANOVA analysis confirmed a significant main effect on Recall (F(2,84) = 0.59, p < .001) among different perceived difficulty levels. However, there was no significant main effect of review conditions on Recall (F(3,84) = 0.79, p = .499). We found that conducting perceived difficulty ratings after the review and post-test was responsible for this outcome. A participant’s perceived difficulty levels of three topics would have been changed after the participant reviewed one (adaptive or counter adaptive) or three (full review) topics and/or took the post-test. Thus, the reported perceived difficulty levels of each topic may not match the initial perceived difficulty levels when watching the lesson.

Among all the four review conditions, no review was not affected by this confounder because participants in the no review condition did not watch any review content. One-way ANOVA analysis in no review showed a significant main effect of perceived difficulty on Recall (F(2,21) = 3.96, p < .05). Further pairwise mean comparison (t-tests) with Bonferroni correction on perceived difficulty levels show a significant difference in Recall between the easiest topic (Rank 1) and other ranks, i.e. Rank 2 (t(21) = 5.55, p < .05), and Rank 3 (t(21)=6.31, p < .05). However, there was no significant difference in Recall between the most difficult topic (Rank 3) and the second most difficult topic (Rank 2) (t(21) = 0.03, p = .877). There were two implications in these findings. First, the perceived difficulty levels were highly correlated with Recall and can be a good indicator for review content. Second, both the most difficult topic and the second most difficult topic were beneficial for review.

5.4.2 Model Performance

The confounding factor in 5.4.1 has minimum influence on our ranking algorithm for three reasons: 1) our model is user-independent and was trained on 8 participants from a pilot study; 2) none of the participants knew their performances in the pre-and post-test; and 3) both learning topic and post-test questions focused on understanding and recall rather than deep inference. We quantified our algorithm’s ranking performance on perceived difficulty levels (i.e. whether the model’s ranking outputs match participants’ subjective ratings) in 2 settings: all four review conditions (All-conditions) and no review only (No-Review). In addition to the strict ranking criterion (Strict-Ranking), we also evaluated AttentiveReview in a relaxed ranking condition (Relaxed-Ranking). Since our results in section 5.4.1 showed that both the most difficulty and the second most difficult topic can benefit learners, the relaxed ranking criterion only marks the model prediction incorrect when the recommended topic was neither the most difficult nor the second most difficult topic. This criterion can be considered a variant of Precision@N, which had been widely used in information retrieval community. We used random guessing and participants’ agreement (Fleiss’ kappa) as baselines.

### Table 1. Perceived difficulty ranking performance.

<table>
<thead>
<tr>
<th></th>
<th>Strict-Ranking</th>
<th>Relaxed-Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All-conditions</td>
<td>No-Review</td>
</tr>
<tr>
<td></td>
<td>All-conditions</td>
<td>No-Review</td>
</tr>
<tr>
<td>AttentiveReview</td>
<td>Accuracy</td>
<td>38.5%</td>
</tr>
<tr>
<td></td>
<td>Kappa</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.83</td>
</tr>
<tr>
<td>Random predictor</td>
<td>Accuracy</td>
<td>33.3%</td>
</tr>
<tr>
<td></td>
<td>Kappa</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.12</td>
</tr>
</tbody>
</table>

Table 1 shows the performance of AttentiveReview on the ranking of perceived difficulty levels. In the Strict-Ranking criterion, AttentiveReview only performed better than baselines in the No-Review condition, where no confounding affects participants’ subjective ratings. However, in the Relaxed-Ranking condition, AttentiveReview outperformed all baselines in both All-conditions and No-Review.

5.4.3 Benefits of Review

Although we found that adaptive review could improve both information recall and learning gain in section 5.3, is it really
from the adaptive presentation of confusing topics to learners? We defined and used a fine-grained measurement metric, Review Efficacy Index (REI), to study the impact of review recommendations on learning outcomes. REI is defined as the percentage of learners having Recall scores from the easiest topic higher than Recall scores from the most difficult topic. Let $e_i$ and $d_i$ be the total Recall score of the easiest topic’s questions and the total Recall score of most difficult topic’s questions perceived by the $i^{th}$ participant, we define $l_i$ as follow:

$$l_i(e_i, d_i) = \begin{cases} 1, & \text{if } e_i > d_i \\ 0, & \text{otherwise} \end{cases}$$

Then, we can compute the REI of a review condition $r$ as:

$$REI_r = \sum_{i \in r} l_i(e_i, d_i) / \text{#participant in } r$$

where $r \in \{\text{No, Adapt, Full}\}$

A large value of REI means most of the learners do better in the easiest topic’s questions compared to the most difficult topic’s question when receiving a particular review condition, and vice versa. Note that by modifying $l_i$, REI can be extended to estimate other types of effectiveness, such as comparing one learning topic versus top $k$ topics.

![Figure 9. The REI metric by review conditions.](image)

Figure 9 shows the REI of 3 review conditions. We can see that adaptive review indeed increased learners’ performance on difficult topics when compared with no review. To be specific, the REI of the no review condition, where participants did not watch any review material, was the highest (88%) because learners are likely to score higher in the easiest topic than in the most difficult topic. In the adaptive review condition, the REI decreased to 63%. This means the performance of the most difficult topic were improved when learners got exposed to difficult topics in the review sessions. We attribute the lowest REI (38%) in full review to two reasons. First, the predicted perceived difficulty levels from AttentiveReview was not perfect. So a full review was more beneficial to capture all the difficult topics. Second, there was a ceiling effect for reviewing easy topics, so the benefits in terms of the number of correct answers from difficult topics were higher when compared with those from easy topics.

6. DISCUSSIONS

Not surprisingly, we found the mastery of the English language did have a significant impact on the learning outcomes. In particular, EL1 (with average scores = 37.4%, $\sigma = 20.8\%$) performed significantly better than EL2 (25.6%, $\sigma = 13.7\%$) overall. Additional contrast tests on Learning showed a significant difference between EL1 and EL2 under counter adaptive ($p = .32$) or no review condition ($p = .44$). These results suggest that EL2 may encounter higher challenges in reviewing activities when compared with native English speakers. It would be beneficial to go through the review session at a slower speed, shorter duration, together with highlighted screen captions for learners whose first language were not English.

We also ran a regression analysis to evaluate the importance of features used in AttentiveReview. Despite the fact that AttentiveReview used a ranking SVM, regression analysis could also predict the regression score of each topic based on the perceived difficulty ranking from participants. To avoid confounding factors from reviewing activities and post-tests, we only analyzed data from the no review condition. The mean normal-to-normal (AVNN) feature played an important role in predicting the perceived difficulty levels where its global feature had a significant impact ($p = .047$) and its local feature had a marginal impact ($p = .083$). This finding is consistent with findings in [30], where the authors also found significant changes in AVNN were good indicators of confusion. In addition, the global median absolute deviation (MAD) feature and the local standard deviation of normal-to-normal (SDNN) feature had a significant impact with $p = .044$ and $p = .046$, respectively. However, the MW feature did not have a significant impact ($p = .399$) on the model’s performance. The combination of all experimental features accounted for 84.9% of the variability in perceived difficulty levels.

In addition to using a supervised machine learning algorithm to predict perceived difficulty from PPG signals, it is also possible derive review commendations from deterministic metrics such as the attention index feature [27] from EEG signals directly. It would be interesting to explore whether two such drastically different strategies can complement each other to achieve better learning outcome. Further, for which kinds of situations/students is one strategy more effective than the other?

Despite the encouraging results of AttentiveReview, our research still has at least three major limitations. First, although our model is user-independent, it was trained with data from an 8-participant pilot study on the same learning topics. While this is already a major improvement when compared with user-dependent models, AttentiveReview still requires course-dependent training and is not completely plug-and-play. We plan to study the feasibility of course-independent adaptive review in the near future. Second, our current algorithm only supports review recommendations at the learning topic level. Although most of today’s MOOCs have already organized tutorial videos per learning topics, it may not always be the case. Since a learning topic typically lasts 3 – 15 minutes, our algorithm is not capable of providing ultra fine-grained predictions at sub-minute levels. Third, the intervention in AttentiveReview, i.e. recommending appropriate review materials, happens after finishing multiple learning topics due to the ranking SVM algorithm we chose. It will be necessary to explore complementary interventions that can be activated during the learning process. We are particularly interested in exploring whether it is possible to improve the learning outcome without increasing the total time spent in learning and reviewing.

7. CONCLUSIONS AND FUTURE WORK

We presented AttentiveReview, a novel intelligent tutoring system and algorithm for mobile MOOC learning. We explained in detail how AttentiveReview could collect and use PPG signals implicitly to improve mobile MOOC learning via adaptive review on
unmodified smartphones. Through a 32-participant user study and follow-up analyses, we found AttentiveReview was intuitive and responsive to use. We also found that AttentiveReview can capture learner’s PPG signals in high quality and effectively recommend review materials that improved learners’ information recall and learning outcome. The proposed adaptive approach is simple to implement and can be integrated into today’s major MOOC platforms easily. Overall, AttentiveReview demonstrated the feasibility and efficacy of building an end-to-end, affect-aware intelligent tutoring system on today’s unmodified smartphones.

We have three specific goals in the near future. First, AttentiveReview has been successfully integrated into OpenEdX, an open-source MOOC platform that powered edx.org, one of the major MOOC providers nowadays. We plan to conduct large-scale, longitudinal studies in real MOOCs or flipped classrooms to investigate the use of implicit PPG sensing and adaptive review to improve both teaching and learning. Second, we are testing the use of the front camera as a facial expression sensor and we are integrating state-of-the-art facial expression analysis algorithms, such as the Affectiva SDK, to provide new complementary signals to understand and improve MOOC learning. Third, we plan to investigate the trade-offs between the duration of PPG sensing and the prediction accuracies. We hypothesize that it may not be necessary to track learners’ physiological signals throughout the learning sessions.

8. ACKNOWLEDGMENTS
We thank Jesse Davis, Carrie Demmans Epp, Xiangmin Fan, Longhao Li, Yanjin Long, and Xiang Xiao for the help and support. We also thank the anonymous reviewers for the insightful and constructive feedback.

This research is in part supported by an RDF from the Learning Research and Development Center (LRDC) at the University of Pittsburgh.

9. REFERENCES