Breaking out of the Lab: Mitigating Mind Wandering with Gaze-Based Attention-Aware Technology in Classrooms

STEPHEN HUTT

University of Pennsylvania, hutts@upenn.edu

KRISTINA KRASICH

Duke University, kkrasich@duke.edu

JAMES R. BROCKMOLE

University of Notre Dame, james.brockmole@nd.edu

SIDNEY K. D'MELLO

University of Colorado, Boulder, sidney.dmello@colorado.edu

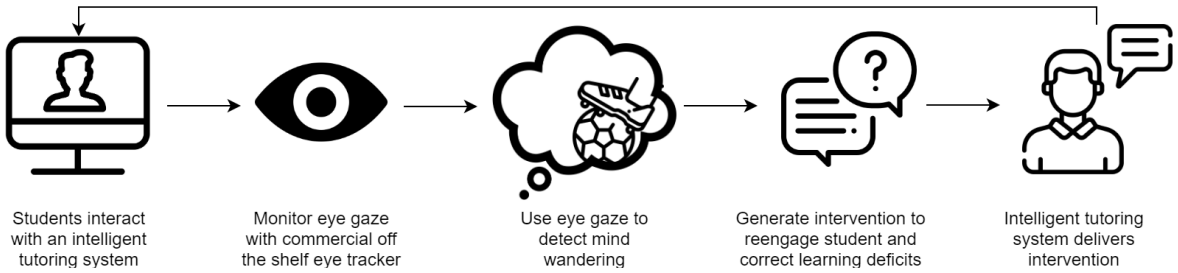

Figure 1. Overview of attention-aware learning technology

We designed and tested an attention-aware learning technology (AALT) that detects and responds to mind wandering (MW), a shift in attention from task-related to task-unrelated thoughts, that is negatively associated with learning. We leveraged an existing gaze-based mind wandering detector that uses commercial off the shelf eye tracking to inform real-time interventions during learning with an Intelligent Tutoring System in real-world classrooms. The intervention strategies, co-designed with students and teachers, consisted of using student names, reiterating content, and asking questions, with the aim to reengage wandering minds and improve learning. After several rounds of iterative refinement, we tested our AALT in two classroom studies with 287 high-school students. We found that interventions successfully reoriented attention, and compared to two control conditions, reduced mind wandering, and improved retention (measured via a delayed assessment) for students with low prior-knowledge who occasionally (but not excessively) mind wandered. We discuss implications for developing gaze-based AALTs for real-world contexts.

CCS CONCEPTS • Human-centered computing~Human computer interaction (HCI) • *Human-centered computing~User studies* • *Applied computing~Interactive learning environments*

Additional Keywords and Phrases: eye-gaze, cyberlearning, intelligent tutoring systems, mind wandering, attention-aware learning

ACM Reference Format:

1 INTRODUCTION

Captivating and maintaining students' attention is challenging—but critical—in any learning environment [1]. Attention facilitates cognitive processes crucial for learning, such as prior knowledge activation and inference generation [41, 68], but it frequently lapses throughout a learning session [61, 77]. In traditional classrooms, teachers can dynamically adapt their instruction to refocus students' attention when it seems to have wandered, for example, by telling a joke, change activities, or suggest a break. However, this real-time adaptivity is currently beyond the scope of most computer-based learning environments, suggesting an opportunity for improving the adaptivity and effectiveness of such technologies. Although researchers have previously suggested the need and utility for attention-aware technologies [63] and, more specifically, attention-aware learning technologies [14], this is particularly relevant today due to increases in computer-based instruction in response to the COVID-19 pandemic.

In the current work, we developed and tested an attention-aware learning technology (AALT) to improve student engagement and learning. We considered one specific kind of attentional lapse called mind wandering (MW) (or zoning out) — defined as an attentional shift from task-related to unrelated thoughts [71]. Research estimates that students spend at least 20-30% of their time mind wandering, either in traditional classrooms [60, 61] or while interacting with learning technologies [17]. Although the *trait-level* tendency to MW has been linked to creative problem solving and prospective planning [48], a meta-analysis of 88 independent samples indicated a negative correlation between *state* (i.e., in the moment) MW and performance across a variety of tasks [58]. In the case of learning with technology, a recent meta-analysis [17] of 25 studies indicated that MW was negatively correlated with learning outcomes (r = -.24). This is unsurprising because when learners mind wandering, they miss out on key concepts [62, 70], have increased difficulty encoding information into memory [67], and fail to comprehend learning content [25, 66]. Thus, there may be benefits to AALTs that address mind wandering in real-time, which is the focus of this work.

Responding to MW entails detection of MW, a challenging task given its covert, internal nature [72]. In the current work, we leverage eye movements, which reveal how the brain processes visual information in real-time [37], for MW detection. Eye-gaze correlates of MW have been identified across a variety of task contexts, including reading [23, 59], lecture viewing [24, 60], suggesting a promising approach for monitoring MW in AALTs [14]. Accordingly, we designed and tested a gaze-based AALT (a biology Intelligent Tutoring System [ITS]) to detect and address MW in real-world classrooms. The ITS tracks student eye movements and uses a pre-trained machine learning model to detect MW in real-time (prior work by Hutt et al. [34]). It then dynamically adapts its instruction to capture and refocus attention to the learning material and to address deficits ostensibly due to mind wandering (current work, Figure 1). In this paper, we describe the design and iterative refinement of our AALT. Further, across two user studies conducted in high school classrooms, we investigate whether the system successfully reengages students' attention, reduces mind wandering, and, consequently, improves learning.

1.1 Related Work

User interfaces that track and respond to attentional states have been explored in a number of domains such as the auto-industry (e.g., monitoring driver fatigue and susceptibility to external distractions - see review [20]), education (e.g., to select adaptive hints in educational games [49]) and adaptive information visualization [8, 74]. In educational environments, Pham and Wang [55, 56] used computer vision techniques to monitor heart rate, which was then used to detect attentional lapses. A widget appeared on the user interface when MW was detected and then disappeared if the student was not MW for over three minutes. These studies show the successful integration of attention-aware adaptation into learning technology but did not examine if the interventions improved student learning. Similarly, AttentiveReview

[54] is a closed-loop system for online learning on mobile phones. It uses video-based photoplethysmography (PPG) to detect a learner's heart rate from a smartphone's camera while viewing MOOC-like lectures on the phone. AttentiveReview ranks the lectures based on its estimates of learners' "perceived difficulty," selecting the most challenging lecture for subsequent review (called adaptive review). In a 32-participant evaluation study, the authors found that the adaptive review condition's learning gains were statistically on par with a full review condition but were achieved in 66.7% less review time. Although this result suggests that AttentiveReview increased learning efficiency, it focused on difficulty rather than attention per se.

Of particular relevance are studies focused on eye gaze in learning technologies, which can be broadly grouped into three categories [9]: (1) offline-analyses of eye gaze to understand attentional processes, (2) modeling of attentional states in real-time, and (3) closed-loop systems that respond to attention in real-time. There has been a wealth of work considering offline analysis of eye movements in educational contexts (see above and [30, 36, 49, 57]). Similarly, real-time modeling of learner attention has become increasingly prevalent [4, 5, 10, 13]. However, closed-loop attentional-aware technologies are still few and far between [13, 28, 69]. In one relevant study, D'Mello and colleagues [13] presented GazeTutor, a multimedia learning environment that used eye gaze to monitor attention and intervene accordingly. If a student appeared to be inattentive (defined below), the tutor would deliver a short phrase (e.g., *"Please pay attention,"* or "*I'm over here, you know.*") designed to redirect the student's attention. An evaluation study indicated that GazeTutor was successful in dynamically reorienting learners' attention and improved learning gains (compared to a control group) in certain contexts. However, GazeTutor characterized inattention as no valid gaze detected for five seconds, which would mischaracterize students who close their eyes to concentrate on the materials (delivered auditorily) or when gaze lapses are due to eye tracking errors. Thus, more objective measures of attentional states are essential.

In this vein, D'Mello et al. [11] showed that eye movement-based real-time MW detection could be leveraged for adaptive intervention during computerized reading. In this study, students read a computerized page-by-page text while their eye movements were recorded and analyzed via a pre-trained and validated MW detector [3]. Here, MW detection was based on more detailed gaze behaviors than in GazeTutor, leveraging general features such as the average length of a fixation and fixation dispersion. If the detector predicted (in real-time) that readers were mind wandering on a page, prior to advancing to the next page, the interface would pose readers with a multiple-choice question on that page's content. If the readers answered incorrectly, they were given the opportunity to re-read the page and then were either presented with the same question again or a different question on the same page. The authors found that this approach corrected deficiencies associated with MW when compared to a control condition.

Mills et al. [46] improved upon several usability and pedagogical limitations with the D'Mello et al. [11] approach. In their study, the trained and validated MW detector [3] predicted mind wandering on larger sections of text spanning multiple pages rather than individual pages. If MW was detected, the readers were asked to generate open-ended self-explanations on the core concept in the section just read. The self-explanations were automatically scored, and readers who received low scores were asked to re-read parts of the text to improve their self-explanation. This approach did not yield improvements to immediate comprehension (compared to a control), but it improved retention as measured by a one-week delayed posttest, suggesting considerable potential.

Both D'Mello et al. [11] and Mills et al. [46] demonstrate the potential for using eye gaze to develop a gaze-based AALT for reading, however both studies occurred in the laboratory using research-grade eye trackers that retail for thousands of dollars, thereby limiting widespread scalability. Recent work has demonstrated the potential for using commercial off-the-shelf (COTS) eye trackers (retailing for approximately 100-200 dollars) to study MW in the laboratory

[36] and, crucially, in authentic real-world environments [35]. We build upon this work to develop and test COTS eye tracking in a closed-loop AALT that addresses MW during interactions with an ITS for use in high school classrooms.

1.2 Current Study and Novelty

The current work presents a prototype for the first fully automated gaze-based attention-aware adaptive ITS for use in classrooms. Building upon previous work from Hutt et al. [34], the ITS tracks students' eye movements to predict the likelihood of MW (prior work) for dynamic real-time interventions (current work).

To our knowledge, this is the first study to detect and combat MW via a closed-loop system in real-world contexts. Eye tracking is theoretically-grounded method for measuring attention, and indeed initial work has utilized eye tracking for attention-aware reading interfaces [11, 46]. However, these endeavors rely on expensive research-grade eye tracking in lab contexts, limiting widespread scalability and ecological validity. We use validated, portable COTS eye trackers that have been used to successfully model attentional states with accuracy comparable to research-grade equipment [27, 36]. Using an eye tracker that retails for approximately 100 USD, our technology can support multiple students and classrooms for the cost of one research-grade eye tracker.

To our knowledge, this work also presents the first gaze-based AALT for use in authentic contexts. Prior work has been conducted in the lab, whether the environment can be strictly controlled (e.g., controlling the lighting or how a user can move), and is free from distractions (e.g., users' phones taken away). Classrooms present a nosier environment where students may turn and whisper to a neighbor or become distracted by others in the room, impacting both attention and eye tracking accuracy. This work designs for both user and use case and, in doing so, shows that a gaze-based AALT can be effective in the real world.

Moreover, prior work on AALTs has focused on computerized reading [18, 46], whereas we consider a rich and varied learning environment, with multiple activities (e.g., lecturing, scaffolded dialogue, concept mapping) and representations (text, audio, media), which increases the complexity of MW detection and the bandwidth of adaptive interventions. Through an extensive design process, we developed interventions to reorient attention and correct any learning deficit attributable to MW. The interventions integrate pedagogically supported approaches from past empirical investigations with structured interviews conducted with students and teachers. Our user-centered approach accounts for imperfect MW detection and delivers "fail-soft" interventions (i.e., those that are not harmful if delivered incorrectly), both of which are novel contributions.

Finally, we evaluated our technology in two large scale experiments with a total of 287 students. We examined the impact of these interventions on mitigating MW in a study with 103 high school students during their regular biology class (Study 1). Following further refinement of the ITS and interventions, we conducted a second study with 184 high school students to investigate learning and retention improvement compared to two control conditions (Study 2). These studies demonstrate the success of both our physical and software implementations

To summarize, our main contribution is the design of the closed-loop AALT that detects and responds to mind wandering and two classroom studies to investigate its effectiveness at reengaging attention and improving student learning. Whereas we leverage an existing ITS [52] with an embedded mind-wandering detector [34], the intervention design, refinement, and real-world usability and efficacy testing are new contributions.

2 MIND WANDERING DETECTION IN THE GURU INTELLIGENT TUTORING SYSTEM (PREVIOUS WORK)

2.1 The Guru ITS

Guru, the ITS used in the current work, is designed to teach biology topics through collaborative conversations in natural language. It was modeled after interactions with expert human tutors [7, 15, 16, 53] and has been shown to be as effective at promoting learning compared to small group tutoring with novice human tutors [51].

Guru's primary interface (see Figure 2) consists of a multimedia panel, a 3D animated agent, and a text response box. The agent speaks (using speech synthesis), gestures, and points using animations. Throughout the dialogue, the tutor gestures to parts of the multimedia panel most relevant to the discussion, and images are slowly revealed as the dialogue advances. For a more detailed description of Guru see [34, 51].

Guru tutorials provided tutoring on introductory biology topics (e.g., osmosis) aligned with state curriculum standards over short sessions (15-40 minutes). Guru begins each tutorial session with a brief introduction to motivate the topic, followed by five phases: Common Ground Building, Intermittent summaries, Concept Maps, Scaffolded Dialogue, and a Cloze task. The two most relevant phases to the current work are described below.

Common-Ground-Building (CGB) Instruction. Biology topics often involve specialized terminology that must be understood before it is advisable to move on to deeper knowledge-building activities. Therefore, Guru begins with a collaborative lecture phase [12], which covers basic information and terminology relevant to the topic with a 3:1 (Tutor:Student) turn ratio.

Scaffolded Dialogue. Students complete a scaffolded dialogue in which Guru uses a Prompt → Feedback → Verification Question → Feedback → Elaboration cycle to cover target concepts in detail. The student model is continually updated throughout this process, prompting additional dialogue where necessary. For example, a student who has not demonstrated knowledge of a certain concept would get additional dialogue on that concept.

2.2 MW Detection in Guru

MW detection has previously been integrated and validated in Guru [34]. As students interact with Guru, their eye gaze is monitored by a COTS eye tracker (the EyeTribe, which retailed for $99[1]), which is affixed just below the screen (see Figure 2). Setup and calibration of the eye tracker are done entirely by the students using previously validated [34] instructions and wizards.

To generate a MW prediction, gaze features are calculated from the previous 30-seconds of eye movements. This window size was motivated by previous work showing that an 18-30 s time window optimized MW detection [34, 36, 39]. The detector first calculates fixations (i.e., points in which gaze is maintained on the same location) and saccades (i.e., the movement of the eyes between fixations) from the raw eye gaze using Open Gaze and Mouse Analyzer (OGAMA) [76]. Next, 57 global eye movement features are calculated from the time series of fixations and saccades to characterize general gaze patterns independent of the displayed content. Pertinent features include the number of fixations, fixation durations, and saccade amplitudes, which are all independent of the content displayed on the screen and thereby more robust to eye tracking errors.

These features are inputted to a previously trained machine-learned model (a Bayesian network implemented with the WEKA data mining package [32]) that outputs a probability of MW. This model was trained using a dataset of eye

[1] The EyeTribe is no longer commercially available, but other COTS eye trackers retail for a similar price and have been shown to have equitable performance [27, 34]

gaze (collected using the same tracker used here) and self-reports of MW collected from 135 high school students as they interacted with Guru in their high school classroom [34]. The model was validated with a further set of 39 high school students in the classroom. In this validation, self-reports were triggered either by the detector or pseudo randomly to evaluate the detector's accuracy [34]. We considered the detector's accuracy (weighted F_1 of 0.51 vs. chance baseline value of.33), to be moderate given the complexity of the problem and designed our interventions to be robust to detection errors (as illustrated below).

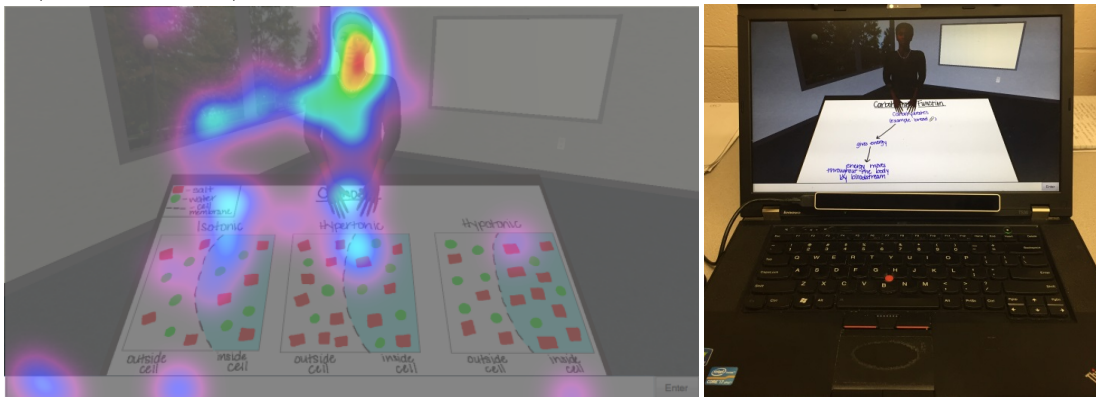

Figure 2. Left: Screenshot of Guru in the CGB phase with overlaying gaze heatmap
Right: Example Eye Tribe setup

3 INTERVENTION DESIGN

Incorporating input from both teachers and students is critical in designing interventions that are effective at reducing MW. Our first task was to obtain input from students and teachers about re-engagement techniques used in classrooms. We did this with structured interviews with 25 high school students, followed by a half-day focus group with three high school biology teachers.

3.1 Structured Interview with Students

We first conducted structured interviews with 25 high school students who had not previously used Guru. Students volunteered to complete one Guru session on a randomly selected topic either during study hall or after class. An experimenter (a high-school research intern) observed students throughout the session. Between each phase (section) of Guru, the session was paused, and students were interviewed with questions related to their experiences in that phase with an emphasis on their levels of attention and engagement (e.g., What would have helped you stay engaged in that last section?). In general, students reported more MW in the CGB phase and the scaffolded dialogue phases, when there was considerable tutor speech compared to the other phases.

At the end of the session, students were asked open-ended questions regarding self- and teacher-related strategies for regaining attention. Approximately a third of students (36%) commented that a teacher simply saying their name helped them pay attention. A further 20% commented that being reminded of an upcoming assessment helped them remain engaged. A representative sample of student responses includes: "*A teacher might tell you to stay focused, or ask you a question when they know you weren't paying attention*"; *"When someone says my name, or a teacher talks to me after class* [about staying engaged]"; "*Being reminded that there is a test coming up"; "Being moved away from friends."*

In summary, the students indicated there were multiple techniques to capture lapsed attention, with teachers using a variety of these throughout a class session. This implied that our approach to responding to the MW within Guru should be also be varied in its approach and potentially include using the student's name or asking a question. It should be noted that some of the suggestions made by students involved physical changes (such as being moved away from a friend) that our software would not be able to do. Such suggestions were not carried forward to future design activities due to being unfeasible in this context.

3.2 Teacher Focus Group

We next held a half-day focus group with three high school biology teachers. All three teachers taught our target population and were from the same large public midwestern high school; one of the teachers had previously worked with the research team. Teachers were shown a demonstration of Guru by a member of the research team and were shown how Guru tracks student progress. Teachers could ask questions or ask for clarifications, which led to an initial discussion of Guru.

We then conducted a series of brainwriting exercises [42] to identify successful remediation tactics that could be computerized to combat MW. Prompts pertained to teachers' everyday classroom experience with engaging students in learning material. Teacher suggestions included: "*Perhaps Guru could call out students by name"; "I think if Guru could ask a question about the immediate content, that would be helpful"; "A complete switch of context might be useful, something like: 'let's come back to this'"; "Adding an incremental summary, everything covered so far in brief, might help counter the effect of mind wandering*."

All three teachers confirmed that calling on a student by name is often effective, as is prompting a student to realize they had zoned out by asking a question. Therefore, the focus group discussions helped validate the structured interviews conducted with students.

3.3 Intervention Design

The overarching goal was to deliver interventions that strike the delicate balance of reengaging students without being abrupt or disruptive, thus interrupting the learning experience [29, 33]. Intervention design must acknowledge that MW detection is imperfect (see [3, 19, 34, 43]). Thus, the design must "fail-soft" in that if delivered incorrectly (due to imperfect MW detection), potentially harmful effects on learning are minimized. Students should not feel like they are being watched or having their privacy violated in any way. Accordingly, interventions were only designed for the Common Ground Building and Scaffolded Dialogue phases of Guru, where the most MW is reported [34, 47]. They were also designed to be "light-touch," as elaborated below.

3.3.1 Intervention Content

Based on our findings documented above, we developed two types of interventions: reiteration and question.

Reiteration. This intervention style consisted of two phrases: an attention redirection phrase and a repeat phrase. The attention redirection phrase aimed to draw the student's attention back to the tutor and emphasize the importance of the learning material. The repeat phrase then restated specific content that was being discussed when MW was detected. The idea was first to reorient attention and second correct any comprehension deficiencies attributed to mind wandering. For example:

> Diffusion involves particles moving from places in the cell where there are a lot of particles, to places where there are fewer of those particles. That's pretty neat! Let's

> go back over that [ATTENTION REDIRECTION]; this will be on the quiz later [EMPHASIZE IMPORTANCE]. Diffusion involves particles … [REPEAT CONTENT]

Question: Rather than repeating content following the attention redirection phrase, the tutor asks a question about the current content being discussed. The tutor then waits for a response and provides feedback as appropriate. To illustrate:

> In fact, the particles spread out naturally and randomly, just by floating around and bouncing off of other molecules! Let's work on this together [ATTENTION REDIRECTION], I have a question for you! Does Facilitated Diffusion require energy? [QUESTION]

Using Students Names. Based upon student feedback, we also included the option to incorporate the student's first name (as entered by the student at the beginning of the session) in the attention redirection phrase for either intervention type. For example, "*Charlie [FIRST NAME], let's work on this some more [ATTENTION REDIRECTION].*" There was a 50% pseudorandom chance of using the student's name in each intervention type described above. This would ensure that the interventions remained varied and did not become stale or boring for students. It should be noted that student names were only temporarily stored while the software was running and were deleted at the end of each Guru session.

3.3.2 Intervention Delivery

As students interact with Guru, MW likelihoods (between 0 and 1) are generated from the MW detector every second. To account for detector inaccuracies, we adopted a nonlinear probabilistic approach in deciding when to intervene. Specifically, a prediction of less than 0.05 resulted in no intervention, and a prediction greater than 0.7 always yielded an intervention (other things considered; see below). In between these bounds, interventions were probabilistic (e.g., a likelihood of 0.45 resulted in a 45% chance of yielding an intervention). After reviewing MW predictions from previous studies [33], these thresholds were selected to align MW predictions with prior empirical research on the incidence of MW with Guru [34, 47].

The final intervention workflow is shown in Figure 3. To maintain the ITS interaction flow and reduce disruptions, except when intended to reengage attention, we placed several restrictions on the intervention mechanism. In particular, an intervention would not be triggered if the student received an intervention within the last 90 seconds, or the ITS was awaiting a response (e.g., if the student was typing). Each intervention was randomly selected without replacement from a set of pre-determined phrases such that the same phrasing was not used twice in one session. Once the intervention phrase was selected, it was then inserted into the tutor's dialogue at the earliest possible moment without interrupting the tutor's speech.

4 TESTING AND REFINEMENT OF FIRST INTERVENTION PROTOTYPE

We conducted several testing and refinement cycles of the first intervention prototype in the lab and the classroom. Laboratory participants were compensated with research credit, while classroom participants were compensated with a $10 gift card. To test whether users noticed the interventions, they were not informed that Guru was responding to MW until they had completed the session. No participant was involved in the testing and refinement process more than once.

4.1 Laboratory Testing – 14 Participants

The prototype was initially tested in the lab with 14 undergraduate students (who had previously not used Guru) for one session each. Sessions were screen recorded and reviewed by an experimenter to evaluate the intervention mechanism. Students were also interviewed regarding their experiences after the session. We found that when asking students if they noticed an intervention, they commented that the interventions could be at times "abrupt" or "interrupting." In contrast, there were some instances in which students commented that they did not notice the interventions at all, an equally concerning issue. As a result, we adjusted attention redirection phrases to improve the learning session's general flow and make the intervention suitably engaging. We also added a longer pause between the session content and the intervention phrase, so that it would appear as a new conversational turn.

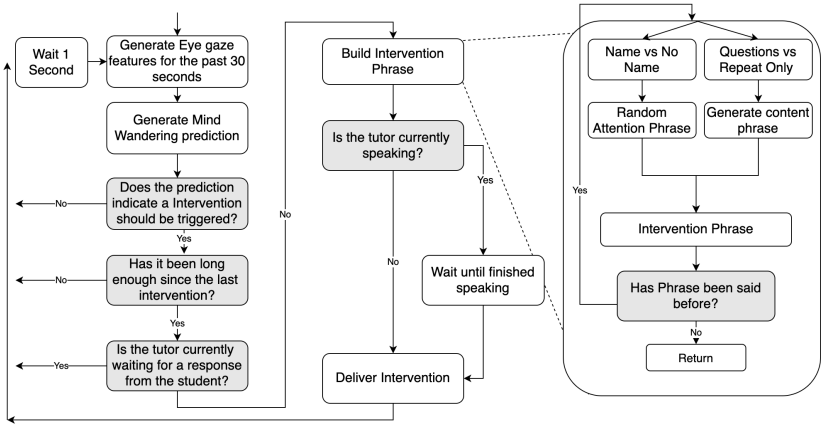

Figure 3. Intervention Delivery Workflow

4.2 Individual Testing in School – 5 Participants

Further testing of the updated intervention was done either after school or in study hall with five high school student volunteers. Students completed one Guru session, which was observed by a researcher who noted critical incidents and student questions. After the session, students were interviewed about their experience, including if they noticed any attempts to reorient their attention (e.g. ".Did you ever notice whether [the tutor] called on you to pay closer attention?"). Students in this round of piloting reported fewer complaints about interruptions, suggesting that intervention delivery had improved; this was then anecdotally confirmed by reviewing the screen recordings. However, students reported that occasionally they found the phrasing of interventions strange, so we further adjusted these to be better suited to high school students.

4.3 Classroom Pilot – 13 Participants

Next, we conducted a pilot study in the same classroom environment intended for the user studies with 13 students who had not previously interacted with Guru. On average, we found students received 1.3 interventions (*SD*=1.48), with six students never receiving an intervention. We were initially concerned by the low number of interventions triggered; however, these students were honors biology students (i.e., more advanced than our target sample), so we expected lower MW rates. With this in mind, we made only minor modifications to the triggering algorithm by lowering the threshold for a guaranteed intervention from 0.7 to 0.6. Students reported noticing interventions but did not feel that they disrupted the software's overall flow or the learning experience.

5 USER STUDY 1

Study 1 was the first test of the affect-aware Guru with mind wandering detection and intervention in an authentic classroom environment (the previous pilot had a very small number of students). The primary goal was to examine whether the intervention mechanism functioned as intended and if there were notable reductions in mind wandering.

5.1 Methods

5.1.1 Participants (Students)

Students were 103 high school seniors at a large, midwestern, public high school who were enrolled in their second high school biology class. The school reports a student population that is 73% White, 9% Black, 6% Asian, and 7% Hispanic. Around 20% were enrolled in the free and reduced lunch program. None of the students had previously used Guru or been involved in the intervention development or testing. Before participating in the study, students provided written assent while their parents provided written consent. Students were compensated with a $10 gift card for their participation in the study.

5.1.2 Procedure

Testing occurred over two days, with each day consisting of a different sample of students. Testing was conducted in students' regular high school classroom. Students each completed two Guru sessions (one on facilitated diffusion and the other on protein function). Each testing session consisted of an introduction to the software, the first Guru session (30 minutes), a short break, and the second Guru session (30 minutes) with topic counterbalanced across the first and second Guru sessions. In either the first or second session (counterbalanced across students) students received interventions based on real-time MW detection as described in Section 2.2; we only analyze these sessions here (the other condition did not yield usable data and is not analyzed further).

Students were each provided with a laptop and eye tracker at their desks (see Figure 2). On-screen instructions with live feedback guided students to establish a seated position that was compatible with eye tracker recommended positions. Calibration was then achieved using a randomized nine-point calibration process. Following calibration, students completed a six-item, multiple-choice pretest to gauge prior knowledge on the assigned biology topic. Pretest questions were randomly selected from a twelve-item, topic-specific question bank, meaning questions varied across students (see Figure 4 for an example). Students then received tutoring from Guru, upon which they completed a posttest to assess learning gains with questions with items randomly selected from the topic-specific question bank but different from the pretest. Students were also asked to self-report their MW on a six-point Likert scale. All procedures were approved by the Institutional Review Board and the principal of the school.

What are two factors that can cause a protein to become deformed?
a) **exposure to chemicals AND heat (correct answer)**
b) exposure to carbohydrates AND other proteins
c) exposure to hormones AND antibodies
d) exposure to water AND oxygen

Figure 4. Example multiple choice question from protein function, the correct answer is shown in bold.

5.2 Results

5.2.1 Frequency of Intervention Delivery

A total of 120 interventions were delivered across the 103 sessions (M = 1.16, SD = 1.61), of which there were 46 sessions with no interventions. This rate of delivered interventions was lower than anticipated based upon previous predicted rates of MW that were validated with the current detector [30]. We hypothesized these differences might, at least in part, be due to differences in the academic level of the student populations. That is, the detector was trained using data from students enrolled in their first high school biology class [34], whereas students in the current study, were fourth-year high school students enrolled in the second biology class. Also, students in the current study took less time (4.6 minutes less on average) to complete the Guru session compared to the students used to train the detector, resulting in less potential for mind wandering and consequently interventions.

Given the observed difference in samples from training to deployment, we next examined whether the detector was still accurate with this new sample. Accordingly, we correlated the mean MW likelihood for every participant from the detector across the entire Guru session with students' self-reported MW following the session. We observed a modest correlation (Spearman *rho* = .28, p=.13), lower, but within the range of what has been previously reported values (*rho* = .4) from a laboratory study with research-grade tracking [22].

5.2.2 Impact of Interventions on reducing MW

To examine whether our interventions reduced MW, we compared the average predicted likelihood of MW in the ten seconds prior to and following an intervention using data from the 57 participants who received at least one intervention (Figure 5). A paired-samples t-test indicated that the predicted likelihood of MW was much greater before (M=.60, SD = .34) than after (M=.10, SD=.24) the intervention, $t(56) = 5.63$, $p<0.01$. These results indicate a reduction in predicted MW following the interventions.

Next, we examined if one type of intervention was more effective at reducing MW than another. Using a linear mixed-effects model (implemented in R with the *lme4* package [2]), we regressed change in MW (post intervention – pre intervention) on the intervention type: (reiteration or question) × use of the student's name (yes or no) interaction, with biology topic as a covariate and participant as an intercept-only random effect. The interaction was not significant (p = .15), nor were any of the main effects (ps > .26). Thus, the different intervention strategies were statistically equivalent in their effectiveness at reducing MW (see Figure 5).

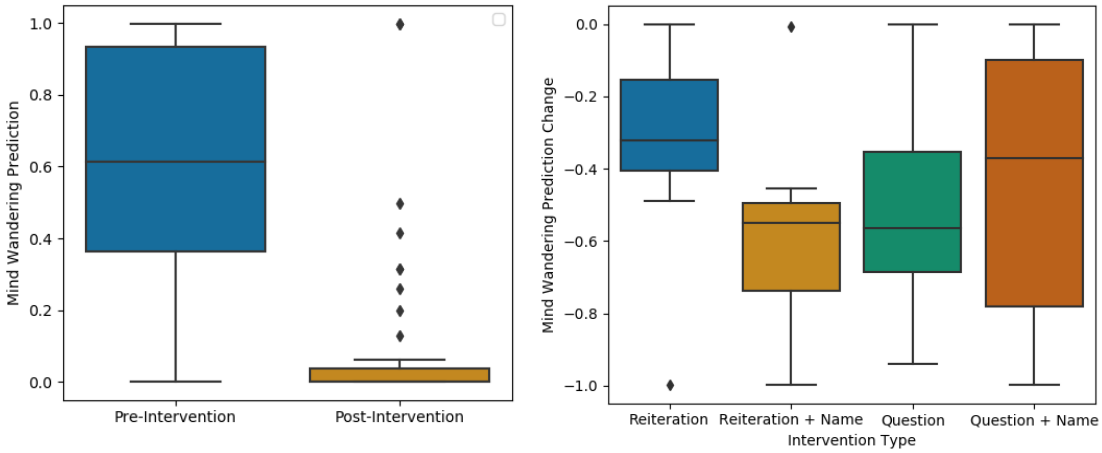

Figure 5. Left: Boxplot of MW predictions before and after interventions for students in the intervention condition
Right: Change in MW prediction (post – pre intervention) for each type of Intervention (negative differences suggest decreased MW)

5.3 Discussion

We conducted an initial user study of the attention-aware Guru ITS in a real-world classroom. We found that the interventions reduced student MW (as measured by our MW detector). However, there was a large proportion of sessions (44%) that received no interventions. Although some of these students may never have been MW, it is unlikely that this was the case for all such students. Instead, it is more likely that the MW detector which triggered the interventions did not gracefully generalize from then training data where novice biology students were used to the current study with advanced students. Learning gains resulting from interventions could not be examined in this study as there was not a suitable comparison group. Instead, we demonstrate the feasibility of AALT in the classroom and show that our interventions reduced student MW, at least in the moments following an intervention. We found no significant differences in the reduction of MW between the types of interventions suggesting both repetition and questions were effective in this context. The use of students name also did not present a significant difference, potentially due to differing social impacts of being called upon in a one on one tutoring environment as opposed to in a classroom setting.

6 FURTHER REFINEMENT TO PRODUCE THE SECOND INTERVENTION PROTOTYPE

We made changes to the prototype based on findings from Study 1. To increase the opportunities to test interventions, we chose to exclusively focus on the Common Ground Building (CGB) instruction phase, as this is where the most MW occurs [34, 47]. To further increase the opportunity for interventions, we updated the learning material to extend the CGB instruction length and match current state standard curricula. Based on the speech rate of the speech synthesis system used by the Guru animated agent and the typical Guru response time data from high school students, we anticipated that the additional material would extend the learning session by approximately four minutes. We also included an additional biology topic (i.e., carbohydrate structure and function), resulting in three Guru CGB tutorials.

To evaluate the impact of these changes on the number of interventions delivered, we conducted a series of simulations using previously observed MW rates from a dataset of high school students interacting with Guru [34]. In each simulation, the Guru session time was generated based on previous high school student data and the additional learning material. For each second of the session, MW probabilities were drawn from a distribution based on data used

to train the MW detector. Simulated probabilities were then inputted to our intervention trigger mechanism (described above) to calculate the total number of interventions, accounting for interactions that might block an intervention (e.g., allowing enough time between interventions.) Simulations indicated that an average of 3.3 (*SD* = 3.0) interventions would be delivered during the modified session.

The refined prototype was then piloted with 24 undergraduate students, who spent an average of 3.2 minutes longer in the CGB session than in previously collected data [36], suggesting that the content length had been appropriately expanded. The number of interventions also correlated positively (rho =.48, p<.01) with a MW survey item completed after the session (i.e., "This activity did not hold my attention at all."), thus reinforcing the validity of our intervention mechanism.

7 USER STUDY 2

Study 1 served as an initial feasibility study for gaze-based engagement interventions in the classroom and demonstrated that interventions could reduce student MW. Having further refined the interventions and tutorial content, we conducted a second user study to examine the impact of interventions on student learning.

7.1 Method

7.1.1 Participants (Students)

Students were 184 freshman and sophomore high school students at the same large public midwestern high school discussed in user Study 1 (see section 5.1.1 for school demographics). Students were all enrolled in their first Biology course. None of the students had used Guru before or been previously involved in this research. Students participated in their regular classroom and were compensated with a $10 gift card. Students provided written assent, and parents provided written consent. Students completed three learning sessions (Protein Function, Carbohydrate Function, and Facilitated Diffusion) in a counterbalanced order. Students were randomly assigned to one of three conditions (described below) for all three Guru sessions in a between-subjects design.

Experimental Condition (n = 107). The experimental condition remained unchanged from Study 1 where students received learning interventions based upon real-time MW detection based on their eye movements.

Active Control (n = 41). Students in the *active control* condition received interventions that were initiated based on a predefined probability distribution based on data used to train the MW detector [34] rather than on real-time MW detection. To illustrate, if 5% of predictions in the previous dataset were between 0.95 and 1, then there is a 5% chance of generating a MW prediction between 0.95 and 1. The intervention delivery mechanism was the same as in the experimental condition, expect based on simulated MW probabilities. Eye movements were tracked in the same manner as the experimental condition, so the user experience was identical. This allows us to disambiguate the effects of interventions that are sensitive to mind wandering (experimental condition) compared to the interventions itself irrespective of mind wandering (active control).

Do Nothing Control (n = 36). Students in the *no-intervention* condition did not receive any learning interventions. This was included to test for any potentially harmful effects of the intervention. We also tracked eye movements in this condition, similar to the above two conditions.

7.1.2 *Procedure*

Each student was tested over two days. On Day 1, students were provided the same equipment as in Study 1 in their regular biology classrooms. During their lesson, students completed three sessions, each on a different biology topic in Guru (i.e., Protein Function, Facilitated diffusion, and Carbohydrate Function) with a mean completion time of 15.54 minutes (*SD*= 3.16 minutes) per topic. As in Study 1, students completed a pre and posttest for each learning session (example shown in Figure 4). At the end of each learning session, students completed items from five subscales of the Intrinsic Motivation Inventory (IMI; Table 2) [44, 64] using a six-point Likert scale (coded as 1-6) As in Study 1, students also answered a question designed to self-report MW. In the interest of time and to avoid fatigue, students answered one randomly selected item per subscale (6 items total) after each session. Due to interruptions (one class was randomly selected for a drug search with sniffer dogs during the experiment) and students being late or having to leave, 44 out of 552 sessions were incomplete. These sessions were not included in later analysis.

The second day of testing consisted of a follow-up assessment to assess students' retention of learned materials. It occurred three weeks after the initial learning session in the same classroom during regular class hours. Students completed a Cloze task [45, 75], where they generated an ideal summary of each topic by filling in missing information (from memory) related to the core concepts (see Figure 6). Retention was measured as the proportion of correctly completed items. Due to student absences, we obtained delay scores for 502 sessions (out of 508 complete sessions).

When a substance is dissolved in water, it is called a _____. When the salute is dissolved in water, the mixture is called a solution. Different solutions can change the water level of a cell. Solutions can have different levels of concentration, this difference is called the concentration gradient. Water molecules move from areas of ____ water concentration to areas of ___ water concentration. In other words, water moves ____ the concentration

Figure 6. Example cloze task (partial view), student responses are shown in blue

7.2 Results

7.2.1 *Number of interventions*

We first examined the number of interventions delivered. Students in the experimental condition received an average of 1.45 (SD = 2.69) interventions, which was on par with students in the active control condition (M = 1.33, SD = 0.49), a Kruskal-Wallis test [40] showed no significant difference between the two distributions (p = .35). Of the 327 tutorial sessions in the intervention condition, 155 of these received 0 interventions. A further examination of these sessions indicated that 90/107 (84%) students in the experimental condition received at least one intervention across their three Guru sessions.

We determined sessions in the experimental condition with no interventions as a separate *intend-to-treat condition*. This adjustment resulted in a mean of 2.77 (SD = 2.87) interventions for the 172 remaining sessions in the intervention condition. For subsequent analyses, we focus on the Do-Nothing control, Active Control, and experimental condition sessions where students received at least one intervention. We also divided the intervention condition sessions into two groups – a low group who received exactly one intervention and a high group who received more than one intervention. Table 1 lists the assignment of students to Guru topics for the various conditions.

Table 1. Alignment of conditions, topics, and Guru sessions

Condition	Module			Total
	Carbohydrate Function	Facilitated Diffusion	Protein Function	
Active Control	36	39	40	115
Do Nothing Control	34	33	33	100
Experimental High (>1 Intervention)	27	35	27	89
Experimental Low (1 intervention)	25	26	32	83
Experimental None (0 Interventions)	56	50	49	155
Total	178	183	181	552

7.2.2 Analysis of Learning

We used the posttest and the delayed retention test after controlling for pretest scores in a regression framework. Figure 7 depicts histograms of these variables. We removed incomplete sessions and sessions where the posttest was the posttest in under 30 seconds as we deemed this insufficient time for the six-item assessment. This resulted in 336 of the 387 sessions retained.

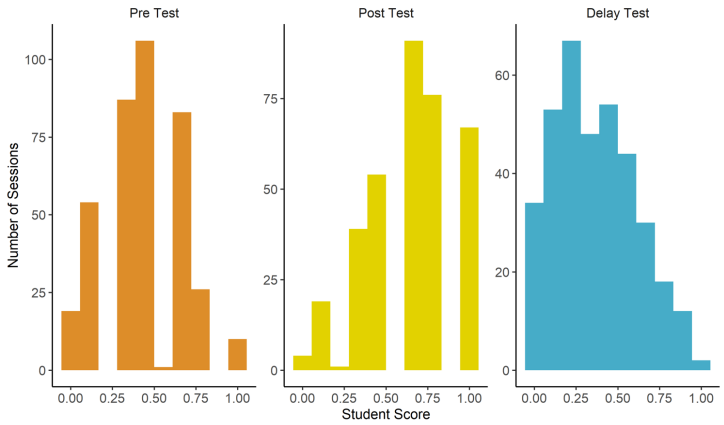

Figure 7. Histograms of learning variables (delay = delayed retention test and value = proportional scores on each assessment

We first examined any differences in pretest scores. Using a linear mixed-effects, we regressed pretest scores on condition and topic with participant as an intercept-only random effect. There were no significant condition differences (p = .87) though prior knowledge significantly varied as a function of the topic (p < .001), so we retained it as a covariate.

Next, we regressed (in two separate models) posttest score and retention (delayed) scores on condition × prior knowledge (as measured by the pretest) interaction with topic and time on task (time spent in Guru) as covariates. There was no significant main effect of condition nor interaction (ps > .45) for posttest scores. However, the condition × prior knowledge interaction was significant (p =.009) for the delayed retention test. Simple slopes analyses

IMI

Finally, we examined items from the IMI questionnaire (See Table 2 for descriptives per condition). We tested for condition differences among the self-report items students completed at the end of the session. For this, we regressed each self-report measure on condition after controlling for pretest, posttest (since the self-reports were completed after the posttest), topic, time spent in Guru, and survey version (there were three different versions – see above). We found

no main effect of condition on self-reported competence (p=.32), perceived value (p=.34), effort/importance (p=.61), interest (p=.95), and felt pressure/tension (p=.18). However, there was a significant effect (p = .002) of condition on self-reported mind wandering. Pairwise comparisons (see Figure 9) indicated lower self-reported mind wandering for both the intervention (low and high) groups compared to the Do-Nothing control groups (ps < .04). However, only the group that received exactly one intervention (i.e., low group) had lower self-reported mind wandering (p = .04) than the Active control group.

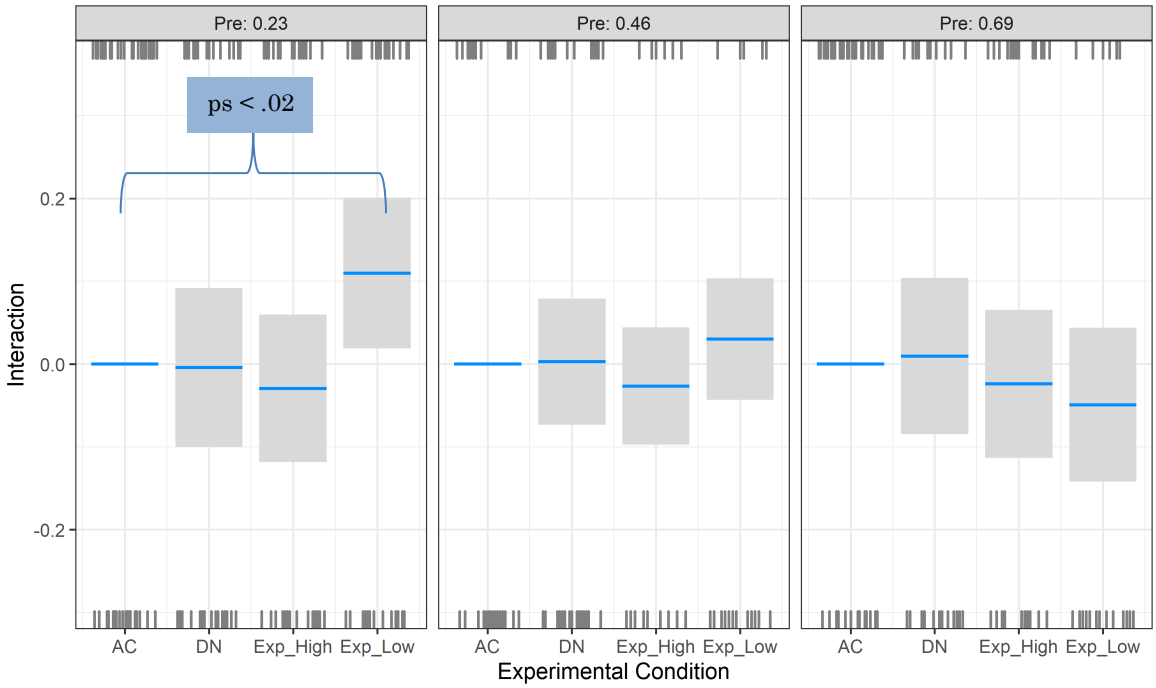

Figure 8. Simple slopes analysis illustrating the interaction between condition and pretest scores (one standard deviation below the mean (0.23), at the mean (0.46), and one standard deviation above the mean (0.69)) for predicting retention. AC = Active-control. DN = Do-nothing control, Exp_High = Intervention group with one or more interventions, Exp_Low = Intervention group with exactly one intervention.

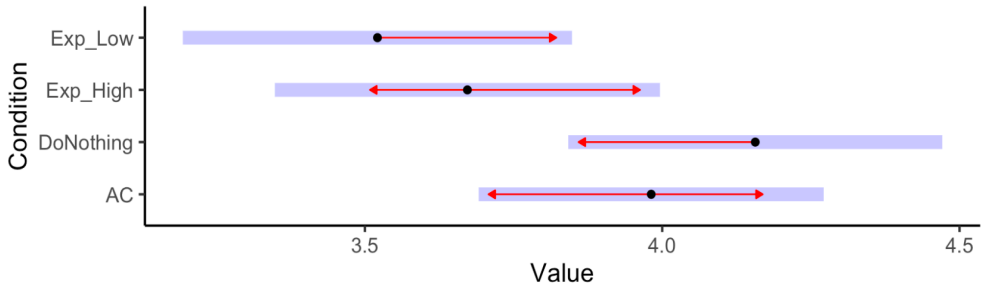

Figure 9. Pairwise comparisons for self-reported mind wandering by condition. AC = Active-control. DoNothing = Do-nothing control, Exp_High = Intervention group with more than one intervention, Exp_Low = Intervention group with exactly one intervention;

Table 2. Example questions given to students following each session and descriptive statistics

Scale	Sample Item	Active Control M (SD)	Do-Nothing Control M (SD)	Exp_High (>1 intervention) M (SD)	Exp_Low (1 intervention) M (SD)
Interest	I thought this was a boring activity [reverse coded]	3.41 (1.36)	3.36 (1.39)	3.24 (1.35)	2.28 (1.43)
Value	I believe this activity could be of some value to me	3.82 (1.35)	3.94 (1.44)	3.48 (1.26)	3.50 (1.27)
Effort /Importance	I put a lot of effort into this	4.03 (1.05)	4.15 (1.29)	3.74 (1.26)	3.76 (1.31)
Perceived Competence	I am satisfied with my performance at this task	3.43 (1.24)	3.21 (1.44)	3.42 (1.34)	3.38 (1.43)
Pressure /Anxiety	I did not feel nervous at all while doing this	4.37 (1.43)	4.30 (1.62)	4.25 (1.57)	4.86 (1.48)
Mind Wandering	My attention drifted towards thoughts unrelated	3.96 (1.42)	4.11 (1.51)	3.71 (1.55)	3.44 (1.61)

7.3 Discussion

We found that our real-time gaze-based interventions successfully reduced (self-reported) MW and promoted retention compared to both control groups. However, the positive effect was only observed for low prior knowledge learners that received only one intervention. It might be the case that the interventions were unnecessary for those with average prior knowledge and higher. Further, requiring more than one intervention might signal a different issue with maintaining attention (e.g., lack of interest), likely not addressed with the light-touch intervention approach considered here. Finally, a substantial number of sessions (48%) in the intervention condition still contained no interventions (14% of students never received a single intervention across all three sessions), suggesting that our detector might still be missing instances of MW. Thus, there might be limits to what can be achieved with imperfect MW detection.

8 GENERAL DISCUSSION

Attention is critical for effective learning [50], but attention-based adaptive instruction is currently beyond the scope of most learning technologies. Here, we developed and tested adaptive interventions to address mind wandering, a form of attentional lapse that negatively correlates with learning [25, 66, 70]. Specifically, we leveraged existing MW detection models and COTS eye trackers to develop an attention-aware ITS with adaptive interventions inspired by teacher remediation strategies for re-orienting attention. After a series of design, testing, and refinement cycles, we evaluated the impact of real-time, attention-aware interventions in two studies (*Ns* of 103 and 184). Both studies took place in high school classrooms, moving AALT out of the laboratory [18, 46] and into the real world.

8.1 Main Findings

We demonstrated the feasibility of an attention-aware ITS in classrooms as well as its efficacy at reducing MW and supporting learning. First, through student feedback, a teacher focus group, and brainwriting exercises [42], we identified remediation tactics that could mimic traditional classroom interactions and successfully reengage learners. We incorporated two types of intervention: reiterating the learning material and asking a content question. User testing

indicated that the interventions had to be carefully embedded to effectively regain lapsed attention while avoiding abrupt interruptions that could disrupt the learning experience. We also noted the necessity of a 'fail soft' approach to avoid negatively impacting the learning experience if the detector incorrectly deemed that the student was mind wandering. Future improvements to MW detection may negate the need for "fail-soft" interventions. However, current MW detection (both the method used here and in the literature) still leaves a considerable margin for error [34, 46].

We showed that these interventions reduced the predicted likelihood of MW (based on the validated MW detector) and that the number of interventions delivered was positively correlated with self-reported MW (Study 1). We further showed (through comparison to an active control and a no-intervention condition) that the attention aware ITS improved long term retention for students with low prior knowledge (as measured by a pretest), but only for those who only occasional mind wandered. (Study 2). Moreover, our studies showed the high usability of our learning technology in the classroom. That is, students calibrated the eye trackers and launched the Guru software with minimal oversight from experimenters. Altogether, our work presents the design and implementation of a closed-loop system for responding to student MW in an ITS.

This work's main application is to support student learning by responding to attentional lapses such as MW. By testing COTS eye trackers in ecologically valid environments, we have taken research on AALTs from the lab and into the real world. Thus, the AALT presented is suitable for both user and (crucially) use case, which has previously been unexplored. Future iterations of attention-aware technologies could be deployed into multiple computer-enabled classrooms simultaneously or in students' homes.

Increased scalability, in turn, affords a wider variety of applications including supporting students involved in at-home or asynchronous learning. For example, beyond responding to MW, other attention-aware technology could encourage students to manage their time effectively by monitoring their attentional states. It could also be used to provide feedback both to developers of learning technology and content developers so that instructional activities and materials could be revised based on what captures students' attention and keeps them engaged.

Finally, this work serves as a proof of concept for monitoring other attentional states beyond MW (e.g., focused attention, alternating attention) in classroom environments to ensure that limited attentional resources are being optimally deployed [14].

8.2 Limitations & Future Work

Like all studies, ours has limitations. For one, there were still fewer attention-aware interventions delivered than expected [34]. Several factors may account for a low intervention count. Although the detector was previously validated, it is known to have inaccuracies. Further, we trained and evaluated our detector on self-reported MW, which, although being a validated measure of such an internal and introspective state [26, 58], requires students to be mindful of their mind wandering and respond honestly. Thus, we are, in some ways, limited by mind wandering detection accuracies. That said, even though Study 2 showed that 47% of sessions did not contain an intervention, only 14% of students never received an intervention across the three sessions. This is consistent with 15% of students who never reported a single instance of mind wandering in a prior study of equivalent length with Guru [34].

Further, both our prior knowledge and post test assessments relied on multiple-choice testing, a method with known limitations [31]. However, well-designed multiple-choice questions are a robust method to assess student learning [6] and are familiar to high school students as they encounter them in standardized tests such as the SAT or ACT. Though multiple-choice tests can potentially lead to guessing, this is easy to detect and correct for. Multiple choice quizzes also offer a time-efficient method to measure knowledge across a topic [21], which was our intended use. Further, although

we considered multiple choice questions suitable for our context, we did include an alternate assessment technique for the delayed test (cloze task).

Another limitation is that only students with low prior knowledge showed improved retention scores. It is unclear whether students with higher prior knowledge require alternate intervention strategies or if there are limited possibilities to improve learning, as they already know the material. Interventions were most successful when only one had to be delivered, implying that the interventions do not support more prolonged MW (ostensibly due to low interest or similar). Future research is needed to design alternative intervention strategies for these students. In general, it is likely that differences in student aptitude, topic interest, difficulties with comprehending the material, and other student-specific variables influence MW [25, 73] and correspondingly require different intervention strategies.

Finally, although COTS eye trackers offer a cheaper, scalable alternative to research-grade equipment, this hardware still comes at a cost, both financial and in terms of setup. In contrast, webcams have become ubiquitous in modern computers, especially laptops and mobile devices. Recent studies have shown increased success using a webcam and corresponding computer vision techniques to monitor eye movements [38]. Though this approach still has a high degree of error [78], the global gaze features used here are somewhat robust to such tracking errors [36, 46], suggesting that future work should consider how webcam technology can increase the accessibility of AALTs.

8.3 Concluding Remarks

The recent introduction of COTS eye trackers has ushered in an exciting time for gaze-based technologies to move beyond the lab and into the real world. We developed and tested a gaze-based attention-aware learning technology to mitigate MW and its effects for use in noisy real-world classroom environments. Our user studies indicated that our intervention approach successfully reoriented attention, reduced MW, and improved learning in certain contexts. They also highlighted several areas of improvement and opportunities for future research aimed at making learning from technology engaging and effective for all.

9 ACKNOWLEDGEMENTS

This research was supported by the National Science Foundation (NSF) (DRL 1235958 and IIS 1523091). Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the NSF. Thanks to fellow lab members for their assistance in the data collection, to the students for their valuable feedback and to our teacher consultants (not named to protect student privacy) for welcoming us into their classrooms.

Figure 1 was designed using resources from Flaticon.com

REFERENCES

[1] Ainley, J. and Luntley, M. 2007. The role of attention in expert classroom practice. *Journal of Mathematics Teacher Education*. 10, 1 (Feb. 2007), 3–22. DOI:https://doi.org/10.1007/s10857-007-9026-z.

[2] Bates, D. et al. 2007. The lme4 package. *October*. (2007).

[3] Bixler, R. and D'Mello, S.K. 2016. Automatic gaze-based user-independent detection of mind wandering during computerized reading. *User Modeling and User-Adapted Interaction*. 26, 1 (2016), 33–68. DOI:https://doi.org/10.1007/s11257-015-9167-1.

[4] Bondareva, D. et al. 2013. Inferring learning from gaze data during interaction with an environment to support self-regulated learning. *International {Conference} on {Artificial} {Intelligence} in {Education}* (Memphis, TN, USA, 2013), 229–238. DOI:https://doi.org/10.1007/978-3-642-39112-5-24.

[5] Brown, L.V.N. and Howard, A.M. 2014. A real- time model to assess student engagement during interaction with intelligent educational agents. *ASEE Annual Conference and Exposition, Conference Proceedings* (2014).

[6] Butler, A.C. 2018. Multiple-choice testing in education: are the best practices for assessment also good for learning? *Journal of Applied*

Research in Memory and Cognition. 3, 7 (2018), 323–331. DOI:https://doi.org/10.1016/j.jarmac.2018.07.002.

[7] Cade, W.L. et al. 2008. Dialogue modes in expert tutoring. *International conference on intelligent tutoring systems* (2008), 470–479. DOI:https://doi.org/10.1007/978-3-540-69132-7-50.

[8] Carenini, G. et al. 2014. Highlighting interventions and user differences: informing adaptive information visualization support. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2014), 1835–1844. DOI:https://doi.org/10.1145/2556288.2557141.

[9] Conati, C. et al. 2013. Eye-tracking for student modelling in intelligent tutoring systems. *Design recommendations for intelligent tutoring systems*. 1, (2013), 227–236.

[10] Conati, C. and Merten, C. 2007. Eye-tracking for user modeling in exploratory learning environments: {an} empirical evaluation. *Knowledge-Based Systems*. 20, 6 (2007), 557–574.

[11] D'Mello, S.K. et al. 2016. Attending to attention: detecting and combating mind wandering during computerized reading. *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (2016), 1661–1669.

[12] D'Mello, S.K. et al. 2010. Collaborative lecturing by human and computer tutors. *International Conference on Intelligent Tutoring Systems* (Berlin, Heidelberg, 2010), 178–187. DOI:https://doi.org/10.1007/978-3-642-13437-1_18.

[13] D'Mello, S.K. et al. 2012. Gaze tutor: a gaze-reactive intelligent tutoring system. *International Journal of human-computer studies*. 70, 5 (May 2012), 377–398. DOI:https://doi.org/10.1016/j.ijhcs.2012.01.004.

[14] D'Mello, S.K. 2016. Giving eyesight to the blind: towards attention-aware aied. *International Journal of Artificial Intelligence in Education*. 26, 2 (2016), 645–659. DOI:https://doi.org/10.1007/s40593-016-0104-1.

[15] D'Mello, S.K. et al. 2010. Mining collaborative patterns in tutorial dialogues. *Journal of Educational Data Mining*. (2010).

[16] D'Mello, S.K. et al. 2010. Monitoring affect states during effortful problem solving activities. *International Journal of Artificial Intelligence in Education*. (2010). DOI:https://doi.org/10.3233/JAI-2010-012.

[17] D'Mello, S.K. 2018. What do we think about when we learn? *Deep Comprehension*. K.K. Mills et al., eds. Routledge. 52–67.

[18] D'Mello, S.K. et al. 2017. Zone out no more: mitigating mind wandering during computerized reading. *Proceedings of the 10th International Conference on Educational Data Mining, EDM 2017* (2017).

[19] DeFalco, J.A. et al. 2018. Detecting and addressing frustration in a serious game for military training. *International Journal of Artificial Intelligence in Education*. 28, 2 (Jun. 2018), 152–193. DOI:https://doi.org/10.1007/s40593-017-0152-1.

[20] Dong, Y. et al. 2011. Driver inattention monitoring system for intelligent vehicles: a review. *IEEE transactions on intelligent transportation systems*. 12, 2 (2011), 596–614. DOI:https://doi.org/10.1109/TITS.2010.2092770.

[21] Downing, S.M. 2002. Assessment of knowledge with written test forms. *International Handbook of Research in Medical Education*. 647–672.

[22] Faber, M. et al. 2018. An automated behavioral measure of mind wandering during computerized reading. *Behavior Research Methods*. 50, 1 (2018), 134–150. DOI:https://doi.org/10.3758/s13428-017-0857-y.

[23] Faber, M. et al. 2018. How the stimulus influences mind wandering in semantically-rich task contexts. *Cognitive Research: Principles and Implications*. (2018). DOI:https://doi.org/10.1186/s41235-018-0129-0.

[24] Faber, M. et al. 2020. The eye-mind wandering link: identifying gaze indices of mind wandering across tasks. *Journal of Experimental Psychology: Human Perception and Performance*. (2020). DOI:https://doi.org/10.1037/xhp0000743.

[25] Feng, S. et al. 2013. Mind wandering while reading easy and difficult texts. *Psychonomic Bulletin & Review*. 20, 3 (2013), 586–592. DOI:https://doi.org/10.3758/s13423-012-0367-y.

[26] Franklin, M.S. et al. 2011. Catching the mind in flight: using behavioral indices to detect mindless reading in real time. *Psychonomic Bulletin & Review*. 18, 5 (Oct. 2011), 992–997. DOI:https://doi.org/10.3758/s13423-011-0109-6.

[27] Gibaldi, A. et al. 2017. Evaluation of the tobii eyex eye tracking controller and matlab toolkit for research. *Behavior Research Methods*. (2017). DOI:https://doi.org/10.3758/s13428-016-0762-9.

[28] Gluck, K.A. et al. 2000. Broader bandwidth in student modeling: what if its were "eye" ts? *International conference on intelligent tutoring systems* (2000), 504–513.

[29] Gobert, J.D. et al. 2015. Operationalizing and detecting disengagement within online science microworlds. *Educational Psychologist*. 50, 1 (2015), 43–57. DOI:https://doi.org/10.1080/00461520.2014.999919.

[30] Graesser, A.C. et al. 2005. Question asking and eye tracking during cognitive disequilibrium: {comprehending} illustrated texts on devices when the devices break down. *Memory & Cognition*. 33, 7 (2005), 1235–1247. DOI:https://doi.org/10.3758/BF03193225.

[31] Haladyna, T.M. et al. 2002. A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*.

[32] Hall, M. et al. 2009. The weka data mining software: an update. *SIGKDD Explorations*. 11, 1 (Nov. 2009), 10–18. DOI:https://doi.org/10.1145/1656274.1656278.

[33] Hassenzahl, M. and Tractinsky, N. 2006. User experience - a research agenda. *Behaviour and Information Technology*. (2006). DOI:https://doi.org/10.1080/01449290500330331.

[34] Hutt, S. et al. 2019. Automated gaze-based mind wandering detection during computerized learning in classrooms. *User Modeling and User-Adapted Interaction*. 29, 4 (Sep. 2019), 821–867. DOI:https://doi.org/10.1007/s11257-019-09228-5.

[35] Hutt, S. et al. 2017. "Out of the fr-eye-ing pan": towards gaze-based models of attention during learning with technology in the classroom. *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization* (New York, NY, USA, 2017), 94–103. DOI:https://doi.org/10.1145/3079628.3079669.

[36] Hutt, S. et al. 2016. The eyes have it: gaze-based detection of mind wandering during learning with an intelligent tutoring system. *The 9th International Conference on Educational Data Mining* (Raleigh, NC, USA, NC, USA, 2016), 86–93.

[37] Just, M.A. and Carpenter, P. 1976. Eye fixations and cognitive processes. *Cognitive Psychology*. 8, 4 (1976), 441–480. DOI:https://doi.org/10.1016/0010-0285(76)90015-3.

[38] Kar, A. and Corcoran, P. 2017. A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms. *IEEE Access*.

[39] Krasich, K. et al. 2018. Gaze-based signatures of mind wandering during real-world scene processing. *Journal of Experimental Psychology: General*. 147, 8 (2018), 1111. DOI:https://doi.org/10.1037/xge0000411.

[40] Kruskal, W.H. and Wallis, W.A. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*. (1952). DOI:https://doi.org/10.1080/01621459.1952.10483441.

[41] Linnenbrink, E.A. 2007. The role of affect in student learning: a multi-dimensional approach to considering the interaction of affect, motivation, and engagement. *Emotion in Education*. R. Pekrun, ed. Elsevier. 107–124.

[42] Litcanu, M. et al. 2015. Brain-writing vs. brainstorming case study for power engineering education. *Procedia - Social and Behavioral Sciences*. (2015). DOI:https://doi.org/10.1016/j.sbspro.2015.04.452.

[43] Loboda, T.D. 2014. *Study and detection of mindless reading*. University of Pittsburgh.

[44] McAuley, E.D. et al. 1989. Psychometric properties of the intrinsic motivation inventoiy in a competitive sport setting: a confirmatory factor analysis. *Research Quarterly for Exercise and Sport*. (1989). DOI:https://doi.org/10.1080/02701367.1989.10607413.

[45] McCray, G. and Brunfaut, T. 2018. Investigating the construct measured by banked gap-fill items: evidence from eye-tracking. *Language Testing*. (2018). DOI:https://doi.org/10.1177/0265532216677105.

[46] Mills, C. et al. 2020. Eye-mind reader: an intelligent reading interface that promotes long-term comprehension by detecting and responding to mind wandering. *Human-Computer Interaction*. 00, 00 (2020), 1–27. DOI:https://doi.org/10.1080/07370024.2020.1716762.

[47] Mills, C. et al. 2015. Mind wandering during learning with an intelligent tutoring system. *Artificial Intelligence in Education* (Madrid, Spain, Spain, Jun. 2015), 267–276. DOI:https://doi.org/10.1007/978-3-319-19773-9_27.

[48] Mooneyham, B.W. and Schooler, J.W. 2013. The costs and benefits of mind-wandering: a review. *Canadian Journal of Experimental Psychology*. 67, 1 (Mar. 2013), 11–18. DOI:https://doi.org/10.1037/a0031569.

[49] Muir, M. and Conati, C. 2012. An analysis of attention to student–adaptive hints in an educational game. *International Conference on Intelligent Tutoring Systems* (2012), 112–122.

[50] Olney, A.M. et al. 2015. Attention in educational contexts: the role of the learning task in guiding attention. *The Handbook of Attention*. J. Fawcett et al., eds. MIT Press.

[51] Olney, A.M. et al. 2012. Guru: a computer tutor that models expert human tutors. *Intelligent Tutoring Systems* (Chania, Crete, Greece, Jun. 2012), 256–261. DOI:https://doi.org/10.1007/978-3-642-30950-2_32.

[52] Olney, A.M. et al. 2012. Guru. *Cross-Disciplinary Advances in Applied Natural Language Processing*.

[53] Olney, A.M. et al. 2010. Tutorial dialog in natural language. *Studies in Computational Intelligence*. R. Nkambou et al., eds. Springer Berlin Heidelberg. 181–206.

[54] Pham, P. and Wang, J. 2018. Adaptive review for mobile mooc learning via multimodal physiological signal sensing - a longitudinal study. *ICMI 2018 - Proceedings of the 2018 International Conference on Multimodal Interaction*. (2018), 63–72. DOI:https://doi.org/10.1145/3242969.3243002.

[55] Pham, P. and Wang, J. 2015. AttentiveLearner: improving mobile mooc learning via implicit heart rate tracking. *Artificial Intelligence in Education* (Madrid, Spain, 2015), 367–376. DOI:https://doi.org/10.1007/978-3-319-19773-9_37.

[56] Pham, P. and Wang, J. 2017. AttentiveLearner2: a multimodal approach for improving mooc learning on mobile devices. *International Conference on Artificial Intelligence in Education* (2017), 561–564. DOI:https://doi.org/10.1007/978-3-319-61425-0_64.

[57] Ponce, H.R. and Mayer, R.E. 2014. Qualitatively different cognitive processing during online reading primed by different study activities. *Computers in Human Behavior*. 30, (Jan. 2014), 121–130. DOI:https://doi.org/10.1016/j.chb.2013.07.054.

[58] Randall, J.G. et al. 2014. Mind-wandering, cognition, and performance: a theory-driven meta-analysis of attention regulation. *Psychological Bulletin*. 140, 6 (Nov. 2014), 1411–1431. DOI:https://doi.org/10.1037/a0037428.

[59] Reichle, E.D. et al. 2010. Eye movements during mindless reading. *Psychological Science*. 21, 9 (Sep. 2010), 1300–1310. DOI:https://doi.org/10.1177/0956797610378686.

[60] Risko, E.F. et al. 2013. Everyday attention: mind wandering and computer use during lectures. *Computers & Education*. 68, (2013), 275–283. DOI:https://doi.org/10.1016/j.compedu.2013.05.001.

[61] Risko, E.F. et al. 2012. Everyday attention: variation in mind wandering and memory in a lecture. *Applied Cognitive Psychology*. 26, 2 (2012), 234–242. DOI:https://doi.org/10.1002/acp.1814.

[62] Robertson, I.H. et al. 1997. "Oops!": performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. *Neuropsychologia*. 35, 6 (Jun. 1997), 747–758. DOI:https://doi.org/10.1016/S0028-3932(97)00015-8.

[63] Roda, C. and Thomas, J. 2006. Attention aware systems: {theories}, applications, and research agenda. *Computers in Human Behavior*. 22, 4 (2006), 557–587.

[64] Ryan, R.M. 1982. Control and information in the intrapersonal sphere: an extension of cognitive evaluation theory. *Journal of Personality and*

Social Psychology. (1982). DOI:https://doi.org/10.1037/0022-3514.43.3.450.

[65] Schielzeth, H. 2010. Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution*. (2010). DOI:https://doi.org/10.1111/j.2041-210x.2010.00012.x.

[66] Schooler, J.W. et al. 2004. Zoning out while reading: evidence for dissociations between experience and metaconsciousness. *Thinking and seeing: Visual metacognition in adults and children*. MIT Press. 203–226.

[67] Seibert, P.S. and Ellis, H.C. 1991. Irrelevant thoughts, emotional mood states, and cognitive task performance. *Memory & Cognition*. 19, 5 (Sep. 1991), 507–513. DOI:https://doi.org/10.3758/BF03199574.

[68] Shernoff, D.J. et al. 2003. Student engagement in high school classrooms from the perspective of flow theory. *School Psychology Quarterly*. 18, 2 (2003), 158–176. DOI:https://doi.org/10.1521/scpq.18.2.158.21860.

[69] Sibert, J.L. et al. 2000. The reading assistant: eye gaze triggered auditory prompting for reading remediation. *Proceedings of the 13th Annual ACM Symposium on User Interface Software and Technology* (New York, NY, USA, 2000), 101–107. DOI:https://doi.org/10.1145/354401.354418.

[70] Smallwood, J. et al. 2008. When attention matters: the curious incident of the wandering mind. *Memory & Cognition*. 36, 6 (Sep. 2008), 1144–1150. DOI:https://doi.org/10.3758/MC.36.6.1144.

[71] Smallwood, J. and Schooler, J.W. 2006. The restless mind. *Psychological bulletin*. 132, 6 (Nov. 2006), 946–958. DOI:https://doi.org/10.1037/0033-2909.132.6.946.

[72] Smallwood, J. and Schooler, J.W. 2015. The science of mind wandering: empirically navigating the stream of consciousness. *Annual Review of Psychology*. 66, (2015), 487–518. DOI:https://doi.org/10.1146/annurev-psych-010814-015331.

[73] Soemer, A. and Schiefele, U. 2019. Text difficulty, topic interest, and mind wandering during reading. *Learning and Instruction*. (2019). DOI:https://doi.org/10.1016/j.learninstruc.2018.12.006.

[74] Steichen, B. et al. 2014. Te, te, hi, hi: eye gaze sequence analysis for informing user-adaptive information visualizations. *International Conference on User Modeling, Adaptation, and Personalization* (2014), 183–194. DOI:https://doi.org/10.1007/978-3-319-08786-3_16.

[75] Taylor, W.L. 1953. "Cloze procedure": a new tool for measuring readability. *Journalism Quarterly*. (1953). DOI:https://doi.org/10.1177/107769905303000401.

[76] Voßkühler, A. et al. 2008. OGAMA (open gaze and mouse analyzer): open-source software designed to analyze eye and mouse movements in slideshow study designs. *Behavior Research Methods*. 40, 4 (Nov. 2008), 1150–1162. DOI:https://doi.org/10.3758/BRM.40.4.1150.

[77] Wilson, K. and Korn, J.H. 2007. Attention during lectures: beyond ten minutes. *Teaching of Psychology*. (2007). DOI:https://doi.org/10.1080/00986280701291291.

[78] Zhang, X. et al. 2019. MPIIGaze: real-world dataset and deep appearance-based gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 41, 1 (2019), 162–175. DOI:https://doi.org/10.1109/TPAMI.2017.2778103.