Out of the Fr-"Eye"-ing Pan

Towards Gaze-Based Models of Attention during Learning with Technology in the Classroom

Stephen Hutt[1], Caitlin Mills[2], Nigel Bosch[3], Kristina Krasich[1], James Brockmole[1], Sidney D'Mello[1]
[1]University of Notre Dame, [2]University of British Columbia, [3]University of Illinois at Urbana-Champaign
{shutt|sdmello}@nd.edu

ABSTRACT

Attention is critical to learning. Hence, advanced learning technologies should benefit from mechanisms to monitor and respond to learners' attentional states. We study the feasibility of integrating commercial off-the-shelf (COTS) eye trackers to monitor attention during interactions with a learning technology called GuruTutor. We tested our implementation on 135 students in a noisy computer-enabled high school classroom and were able to collect a median 95% valid eye gaze data in 85% of the sessions where gaze data was successfully recorded. Machine learning methods were employed to develop automated detectors of mind wandering (MW) – a phenomenon involving a shift in attention from task-related to task-unrelated thoughts that is negatively correlated with performance. Our student-independent, gaze-based models could detect MW with an accuracy (F_1 of MW = 0.59) significantly greater than chance (F_1 of MW = 0.24). Predicted rates of mind wandering were negatively related to posttest performance, providing evidence for the predictive validity of the detector. We discuss next steps towards developing gaze-based, attention-aware, learning technologies that can be deployed in noisy, real-world environments.

Author Keywords

eye-gaze; cyberlearning; intelligent tutoring systems; mind wandering; attention-aware learning

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous; I.2.6 Artificial Intelligence: Learning; K.3.1 Computers and Education: Computer Uses in Education

1 INTRODUCTION

Imagine you are tutoring a student in cell biology only to realize that the student has completely "zoned out." Although the plan is for the two of you to collaboratively explain osmosis, the student's attention has drifted to unrelated thoughts of lunch, the football game, or an upcoming vacation. You might try to momentarily reorient his or her attention by asking a probing question. However, if attentional focus continues to wane, you realize that you must adapt your instruction to better engage the student. You might re-engage attention by switching topics or even suggesting a break, thereby giving the student an opportunity to recharge. This form of dynamic adaptivity was only possible because you had the ability to continually monitor your student's levels of attentional focus, to detect when their attention, and to adapt your instruction to address attentional lapses as they occurred.

The attention-awareness capabilities exhibited in the example are beyond the radar of current educational technologies that are largely unaware of users' attentional states. It is important that we address this gap because it is widely acknowledged that attention is crucial for effective learning. Cognitive processes such as prior knowledge activation, inference generation and comprehension all demand attentional resources [23, 31, 54]. Students who are unable to sustain attentional focus are more likely to engage in self-distracting and other unproductive behaviors [19], which leads to superficial understanding as opposed to deep comprehension.

Accordingly, our goal is to develop learning technologies that model a user's attentional state and can respond accordingly as a means to improve attentional focus and learning outcomes [16]. As an initial step in this direction, we focus on mind wandering (MW), the attentional shift from task-related processing towards internal task-unrelated thoughts [57]. In the context of learning, both lab and field studies have consistently reported MW rates in the 20%-50% range [39, 48, 49, 61, 62]. Although individual differences in trait-level MW have been shown to be positively correlated with creative problem solving and prospective planning [37], a recent meta-analysis of 88 independent samples indicated a negative correlation between MW and performance across a variety of tasks [44]. MW negatively impacts a learner's ability to attend to external events [50, 56], to encode information into memory [53], and to comprehend learning materials [18, 52, 56]. Hence, MW is generally found to be detrimental to learning outcomes.

MW is related to other forms of disengagement, such as boredom, behavioral disengagement, and off-task behaviors [2, 3, 32, 36, 65], it is inherently distinct because it involves internal thoughts rather than overt expressive behaviors. This raises two challenges. First, while other disengaged behaviors often involve detectable behavioral markers (e.g., yawns signaling boredom), MW is an internal state that can be difficult to distinguish from being on-task [57]. Secondly, because MW can occur outside of conscious awareness the onset and duration of MW remains an open question [58].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
UMAP '17, July 09-12, 2017, Bratislava, Slovakia © 2017 Association for Computing Machinery.ACM ISBN 978-1-4503-4635-1/17/07…$15.00
http://dx.doi.org/10.1145/3079628.3079669

Despite these challenges, there has been some progress toward automatic detection of MW (discussed further in Related Works section). Eye tracking is an attractive technique for detecting attentional states like MW due to decades of scientific evidence in support of an *eye-mind link* that suggests a tight coupling between attentional states and eye movements [13, 27, 46]. Until recently, the cost of research-grade eye trackers has limited the applicability of eye tracking in real-world environments at scale. However the recent introduction of consumer off-the-shelf (COTS) eye trackers (retailing for $100 to $150) has ushered forth an exciting era by affording the application of decades of lab-based research on eye gaze, attention, and learning to real-world classrooms, thereby affording new discoveries about how students learn, and designing innovations to sustain attention during learning.

1.1 Novelty

There are three novel aspects to this work. First, it is currently unknown whether COTS eye trackers can be implemented with sufficient fidelity in noisy classroom settings so as to afford collection of actionable gaze data. We address this challenge by tracking gaze while high-school students learn biology, as part of their biology class, with GuruTutor (or Guru) [40, 41], an intelligent tutoring system (ITS) with conversational dialogues (see Figure 1). We show, for the first time, that it is feasible to use COTS eye trackers to collect valid data from entire classes of students in the real-world context of an uncontrolled classroom environment.

Second, we demonstrate that the gaze data collected is of sufficient fidelity to detect a critical form of attentional lapses, specifically MW. Previous work has shown that MW can be detected using eye tracking in Guru [28] (discussed further in Related Work section), but this was done using data collected in a very controlled lab environment. We investigate how detectors developed using similar supervised machine learning methods perform on data collected in a more noisy and complex environment.

Third, the previous study on gaze-based MW detection with Guru [28] used *global* gaze features. These features encode general eye movements (e.g., number of gaze fixations) and are independent of what is displayed on the screen. In the context of Guru, eye gaze on specific areas of interest might be of importance for MW detection due to the dynamically changing visual display (see Figure 1). Accordingly, we investigate whether there are added advantages to utilizing a new set of *locality* features that are sensitive to gaze on specific locations on the screen.

1.2 Related Work

The idea of attention-aware user interfaces was proposed almost a decade ago [51], including for education contexts [45]. Prior to this, [22] discussed the use of eye tracking to increase the bandwidth of information available to an ITS in an aptly titled paper "Broader Bandwidth in Student Modeling: What if ITS Were "Eye" TS?" Similarly, [1] followed up on some of these ideas by demonstrating how particular beneficial instructional strategies could only be launched via a real-time analysis of eye gaze. Most of the recent work on leveraging eye gaze to increase the bandwidth of learner models has been pioneered by Conati and colleagues [8, 11, 12, 29, 30, 38] .

Conati et al. [11] provide an excellent review of much of the existing work in this area. We can group the research into three categories: (1) offline-analyses of eye gaze to understand attentional processes, (2) modeling of attentional states, and (3) closed-loop systems that respond to attention in real-time. Offline-analysis of eye movements has enjoyed considerable attention in cognitive psychology, and educational psychology for several decades (e.g. [24, 26, 33, 38, 43]). However, online models of learner attention are just beginning to emerge (e.g. [5–8, 12, 17, 30]). Closed-loop attention-aware systems are few and far between (for a more or less exhaustive list see [15, 22, 55, 64]).

MW detection is related to attentional state modeling as both entail identifying the focus of a user's attention. However, MW is inherently different from other forms of attention (i.e., fatigue, distractibility, object-of focus) because the eyes might be fixated on the appropriate external stimulus, but very little is being processed. To date, MW has rarely been considered as an aspect of a user's state that warrants detection and corrective action (but see recent work by [14, 42]). As such, automated approaches to detect MW in near real-time are in their infancy [17, 20].

Eye movements offer a promising methodology for automatically detecting MW due to well-known relationships between visual attention and the locus of eye gaze. For example, MW has been associated with longer fixation durations [47] as well as more blinking during reading [59]. Researchers have recently leveraged these relationships to build gaze-based MW detectors during reading [4, 6]. In these studies, MW was measured via pseudo-random thought probes that were interspersed during the reading process. Supervised classification models were successful in discriminating between "yes" and "no" responses to the probes using eye gaze features and were validated in a manner that generalized to new students.

Gaze-based MW detection has also been applied to more complex visual stimuli, such as film viewing [34]. In this study, participants viewed a 32.5-minute film and reported when they caught themselves MW. Supervised learning models were built using both global and locality gaze features (defined above). Locality features were superior in terms of predicting MW, ostensibly due to their sensitivity to the dynamic content being displayed on the screen.

Of particular relevance is a previous lab study on detecting MW during learning with Guru, the same ITS we explore here [28]. Students' eye gaze was tracked with a Tobii EyeX (another COTS eye tracker) as they completed a 30-40 minute learning session with Guru. Students reported MW by responding to pseudo-random thought probes throughout the session. A variety of supervised classification models were trained to detect MW from global gaze features alone, achieving person-independent accuracies that were substantially greater than chance.

All current work on gaze-based MW detection, such as the reading studies [4, 6], the film study [34] and ITS study [28] have been limited to using training data collected in the laboratory. Laboratory environments have the advantage of relatively consistent lighting and freedom from distractions from other students, cell phones, ambient sounds, and numerous other

factors. In contrast, we consider the possibility of building MW detectors from eye-gaze data collected in the noisy real-world context of a computer-enabled classroom. In contrast to the lab, students in our study were subject to all the usual distractions of a high school classroom, which makes the data far noisier.

2 IMPLEMENTATION

Our implementation involves integrating eye tracking into an ITS called GuruTutor.

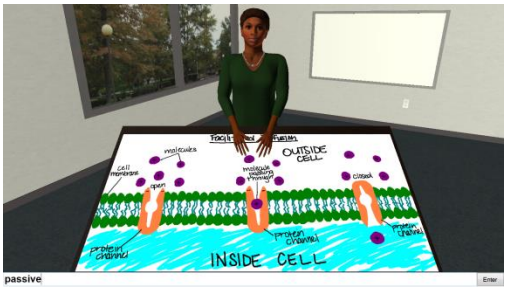

Figure 1. Screenshot of Guru in the CGB phase

2.1 Guru Tutor

GuruTutor (Guru) is an ITS designed to teach biology topics through collaborative conversations in natural language. It was modeled after interactions with expert human tutors and has been shown to be effective at promoting learning and retention at levels similar to human tutors [40].

Guru engages the student through natural language conversations, using an animated tutor agent that references a multimedia workspace (see Figure 1). The tutor communicates via synthesized speech and gestures, while students communicate by typing responses, which are analyzed using natural language processing techniques. Guru maintains a student model [60] throughout the session, which it uses to tailor instruction to individual students.

Guru teaches introductory biology topics (e.g., osmosis; protein function) in line with state curriculum standards in short sessions, typically lasting 15 to 40 minutes. Each topic involves interrelated concepts and facts. Guru begins with a basic introduction to motivate the topic, which is then followed by a five-phase session that develops students' understanding of the topic. The five phases are described below. **Common-Ground-Building (CGB) Instruction.** Biology lessons often involve specialized terminology that needs to be understood before it is possible to move on to deeper knowledge building activities. Therefore, Guru begins with a collaborative lecture phase that covers basic information and terminology relevant to the topic. **Intermittent Summaries (Summary).** Following CGB, students construct their own natural language summaries of the material covered in CGB. These summaries are automatically analyzed to determine which concepts require further tutoring in the remainder of the session. **Concept Maps**. For the target concepts, students complete skeleton concept maps, node-link structures that are automatically generated from text (see Figure 2). **Scaffolded Dialogue.** Next, students complete a scaffolded natural language dialogue in which Guru uses a Prompt → Feedback → Verification Question → Feedback → Elaboration cycle to cover target concepts. If a student shows difficulty mastering particular concepts, a second Concept Maps phase is initiated followed by an additional Scaffolded Dialogue phase. **Cloze Task.** The session concludes with a cloze task requiring students to complete an ideal summary of the topic by filling in missing information by retrieving it from memory

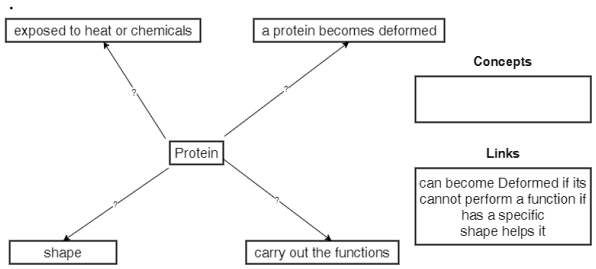

Figure 2. Example Concept Map

2.2 Integrating Eye Tracking in Guru

Our first task was to integrate eye tracking into Guru. In line with the goals of the project, we chose a COTS eye-tracker called the EyeTribe that retails for $99. The eye tracker was affixed to a laptop computer just below the screen.

Our goal was to facilitate eye tracker setup and calibration by the students themselves. This was accomplished via on-screen instructions that included a mixture of images, interactive tools, and text directions. The instructions first guided students on positioning using live feedback, followed by information on the calibration process itself. This is followed by a nine-point calibration process, where nine points appear on the screen in turn and students fixate on each until it disappears.

2.3 Iterative Testing & Refinement

We completed several testing and refinement cycles to improve the implementation to be as user friendly and autonomous as possible. Laboratory participants were compensated with research credit, while classroom participants were compensated with a $10 gift card. Students provided written assent while their parents provided written consent prior to participating in the study, which was approved by the University's Institutional Review Board and the principal of the school.

Lab Testing

The software was initially tested in the lab on individual students. Undergraduate students who had not used the software before were asked to follow the calibration instructions and complete a session on one biology topic with Guru. The setup process was observed and pain-points were noted. The students were then interviewed about their experience with the system. Insights gleaned from this testing were used to improve the clarity of the on-screen instructions and increase the level of feedback that users receive during the eye tracker calibration process.

Individual Testing in School – 9 Participants

Initial testing of the implementation was done in after-school sessions with high-school student volunteers. Students

completed the eye tracking setup along with one Guru session. Each student was observed by a researcher, who noted incidents and recorded student questions. After the session, students were interviewed about the software, including how easy it was to use, how well they understood what they needed to do, and whether they understood why they were doing each step. This informed our development of the software and streamlined the on-screen instructions, providing additional help as needed.

Small Group Testing in School – 7 Participants

As a step towards testing with entire classes of students, we tested the implementation with a group of seven student volunteers after school. Students were given instructions as a group and then interviewed individually once they had completed the session. This allowed us to identify issues with scaling of the software that might arise when working with full classes of students. As a result, we further improved the instructions and addressed other technological challenges.

Classroom Pilot – 35 Participants

The final stage of the iterative development process was a classroom pilot using the specific classroom used for main data collection. We piloted with two class periods during students' regular biology classes. A key observation at this stage was the range of times it took students to complete a Guru session, which had not been as apparent in previous iterations. Students finished the session up to twenty minutes apart, which poses challenges as these students could be sources of distraction for others.

With respect to usability, the overall conclusion was that the students could independently complete the setup and calibration process via the on-screen instructions. In other words, they could use the Guru implementation with minimal guidance from the researchers and the resultant eye gaze was deemed sufficiently valid for larger-scale data collection. One final development was a seating position check and potential recalibration of the eye tracker halfway through the session, in case head position had changed considerably.

2.4 Main Data Collection in Classroom

Students were 9th and 10th graders enrolled in a Biology 1 class. We collected data from 135 students (41% male) over the course of two school days in students' regular biology classroom. Students were compensated with a $10 gift card for their participation in the study.

Students had biology class on alternating days, so the two days of data collection included different students. Each class period consisted of an introduction to the software, 30 minutes of completing a biology session using Guru, a short break, and another 30-minute Guru session on a different biology topic. The following topics were included in the study: Protein Function, Carbohydrate Function, Osmosis, Interphase, Facilitated Diffusion and Biochemical Catalysts, with students randomly assigned a topic (except that they could not get the same topic for both sessions). The classroom layout remained unchanged from the setup used for standard teaching, with the addition of two laptops per desk. The laptops were provided by the high school. Each laptop had an eye tracker affixed below the screen. Class sizes ranged from 14 to 30 students based on regular enrollment. Two researchers were present during data collection to answer procedural questions from students and address any hardware or software issues they encountered.

2.5 Eye Tracking Validity

The majority of students were able to use the software, including eye tracker calibration, without any intervention from the experimenters. However, running the software for a full day presented new challenges. Over the course of the two days, there was the potential to collect 270 sessions as each of the 135 students completed two sessions. The software was completely successful (students able to run through a Guru session and we collected gaze data with no issues) for 85% of the sessions. The following is a breakdown of the causes for the 15% missing sessions: (1) Hardware failure: some of the computers had incorrect drivers for the USB 3.0 ports, preventing the functionality of the eye tracker. (2) Background processes: several computers attempted to automatically update during the session, causing an increased load on the processor which caused the software to occasionally crash. (3) Calibration failure: students who failed calibration three times continued without eye tracking.

The eye tracker records a validity for each sample based on number of eyes tracked and the quality of the tracking. We considered a valid sample to include at least one eye tracked. Figure 3 shows a histogram of percent of valid gaze points per session. Of the 85% of sessions where eye tracking was collected, we observed a median validity rate of 95% per session (mean was 89%). We consider this promising given the difficulties presented by the relatively unconstrained classroom environment, where students were free to fidget, look around the room, and even occasionally laid their heads on the table as they interacted with Guru. If we enforce a stricter validity threshold of both eyes tracked, mean validity drops to 71%, median to 75%, still promising scores.

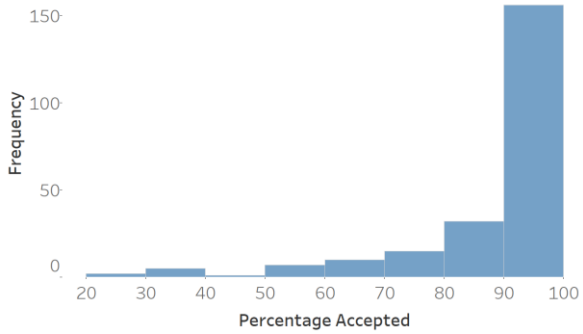

Figure 3. Histogram showing gaze validity rate per session where eye tracking was recorded

Figure 4 shows an example heatmap from the CGB phase for one participant, illustrating gaze concentration. We note the largest concentration of gaze on the tutor's face and upper body, followed by the multimedia panel, and the response box (on the bottom). Visualizing several such heatmaps served as a good initial check for the quality of eye gaze. Our overall conclusion was that we were able to track eye gaze with a reasonable

accuracy when small groups of students used a COTS eye trackers in a noisy real-world environment.

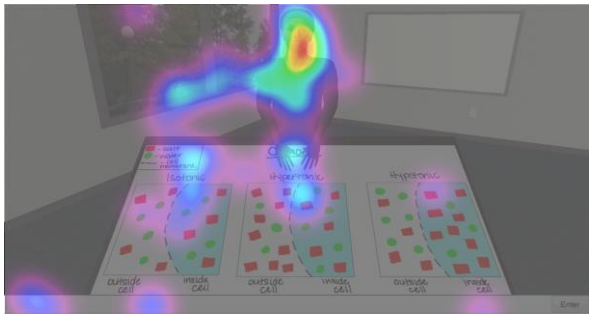

Figure 4. Heatmap overlay showing participants eye gaze in CCB phase. Red indicates high concentration of fixations, purple low concentration of fixations

3 MIND WANDERING DETECTION

Our next step was to leverage the eye gaze data to build automated mind wandering (MW) detectors. We adopted a supervised learning approach, which required labeled data, collected using thought probes.

3.1 Thought Probes

Mind wandering was measured during learning with Guru using auditory thought probes, which is a standard approach in the literature [56]. MW was first defined to the participants. Instructions and MW reporting procedure were extensively tested and refined in the preliminary studies described above. Participants were required to demonstrate understanding of how to respond to the thought probes (via multiple choice questions and feedback) before proceeding.

Participants were probed at pseudo-random intervals with probes occurring every 90-120 seconds, based on previously observed MW rates in Guru [35]. The probes automatically paused the tutoring session. If the tutor was speaking at the time the probe was to be triggered, the probe was delayed until the tutor finished speaking. The probe consisted of an auditory beep along with an opaque overlay on screen, instructing the participant to press the "N" key if they were not mind wandering, "I" if they were intentionally (deliberately) mind wandering, or "U" if they were unintentionally (spontaneously) mind wandering. In this study, we do not differentiate between intentional and unintentional mind wandering, so both "I" and "U" responses were considered MW. Participants encountered an average of 12 probes over the course of each session with a mean MW rate of 28% (SD = 24%, Min = 0%, Max = 100%).

It is important to emphasize a few points about the method used to track MW. First, this method relies on self-reports because MW is an inherently conscious phenomenon, which requires self-awareness for reporting [58]. Second, self-reports of MW have been objectively linked to patterns in pupillometry [21], eye-gaze [47], and task performance [44], providing validity for this approach. However, at this time, there are no reliable neurophysiological or behavioral markers that can accurately substitute for the self-report methodology [58]. Indeed, this is the very reason we set out to build objective gaze-based MW detectors. The limits of thought probes are considered further in the Discussion section. Our use of thought-probes to measure MW is consistent with the state of the art in the psychological and neuroscience literatures [58].

3.2 Feature Engineering

We calculated features from 30-second windows (window size was based on previous work [4, 28]) preceding each auditory probe. We investigated two types of gaze features: global gaze (from previous work [28]) as well as a new set of locality features. Global gaze features focus on general gaze patterns and are independent of the content on the screen, whereas locality features encode where gaze is fixated. We also considered a set of context features to encode information from the session.

Global Gaze Features. Eye movements were measured by fixations (i.e., points in which gaze was maintained on the same location) and saccades (i.e., the movement of the eyes between fixations). We calculated fixations and saccades from the raw eye gaze data using the Open Gaze and Mouse Analyzer (OGAMA) [63]. We considered six general measures across the 30-second window (bolded in Table 1), from which we computed the number, mean, median, minimum, maximum, standard deviation, range, kurtosis, and skew of the distributions, yielding 54 features. We also included three other features (see Table 1), yielding a total of 57 global gaze features.

Table 1. Eye-gaze features. Bolded cell indicates that nine descriptives (e.g., mean) were used as features (see Text)

Feature	Description
Fixation Duration	Elapsed time in ms of fixation
Saccade Duration	Elapsed time in ms of saccade
Saccade Length	Distance of saccade in pixels
Saccade Angle Absolute	Angle in degrees between the x-axis and the saccade
Saccade Angle Relative	Angle of the saccade relative to previous gaze point.
Saccade Velocity	Saccade Length / Saccade Duration
Fixation Dispersion	Root mean square of the distances of each fixation to the average fixation position
Horizontal Saccade Proportion	Proportion of saccades with relative angles <= 30 degrees above or below the horizontal axis
Fixation Saccade Ratio	Ratio of fixation duration to saccade duration

Locality Gaze Features. In contrast to the global features, the locality features were based on locality of gaze. Specifically, a 10 × 8 grid was overlaid on the screen. Each cell represented a feature and was assigned a weight proportional to the number of gaze fixations on that corresponding location (see Figure 5). In addition to these 80 locality features, we included an additional "out of bounds" feature that encoded the proportion of fixations that were off the screen bounds.

Context Features. The gaze features were complemented by eight features that provide a snapshot of the student-tutor interaction. One feature was the assigned biology *topic*. A second encoded participants' *pretest* scores. The next three features represented participants' progress within Guru, such as the *current phase* of the session (e.g., cloze, concept map), the amount of elapsed *time into the session*, and the amount of elapsed *time into the current phase*. The last three features focused on participants' performance within Guru, measured as the proportion of *positive*, *neutral*, and *negative* feedback received.

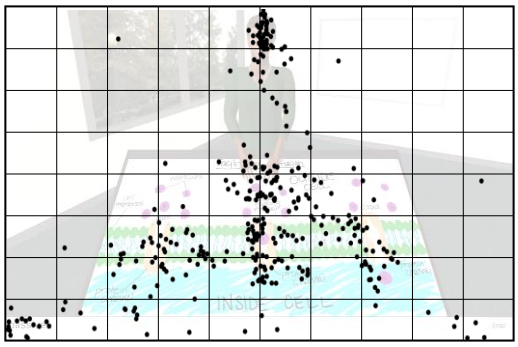

Figure 5. Example grid used for locality features, the count of fixations in each cell becomes a feature

3.3 Classification Models and Validation

We focused on Bayesian Networks because they yielded the best performance compared to several other standard classifiers on this task in our previous work [28]. We used the default implementation from the Weka data mining package [25].

In total, there were 2,720 probes during the Guru sessions. Of those, 386 were discarded due to insufficient eye gaze data (< 1 fixation) in the respective window to compute any of the global features, ostensibly due to students looking away from the screen, chatting with a neighbor, or closing their eyes. The remaining 2,334 instances were used across all feature sets to ensure a fair comparison. Features that could not be computed (e.g. distribution features when there is only one fixation) were treated by the models as missing data and values were imputed based on the training set.

We validated the models with a leave-several-participants-out cross-validation scheme. For each fold, instances from a random 66% of the participants were assigned to a training set and the instances of the remaining 33% participants were assigned to a test set. This process ensures that no instances of any individual participant could appear in both the training and test sets within a fold. This process was repeated for 15 folds and the results were accumulated before computing accuracy metrics.

Students reported MW in 23% of the 2,334 instances, giving a substantial data skew. Class imbalance poses a challenge as supervised learning methods tend to bias predications towards the majority class label. To compensate for this concern, we used the SMOTE algorithm [9] to create synthetic instances of the minority class by interpolating feature values between an instance and its randomly chosen nearest neighbors until the classes were equated. SMOTE was *only applied on the training sets;* the original class distributions were maintained in the testing sets in order to ensure validity of the results.

3.4 Results

The classification results are shown in Table 2. Because our intention is to detect instances of MW, we focus on the precision, recall, and F_1 score of the MW class as our key metric. This is a strict evaluation criterion as the base rate of MW is only 23% in our data. For comparison, a chance-level baseline was created by *randomly* assigning the MW label to 23% of the instances and computing accuracy accordingly.

Table 2. MW detection results for school data

Feature Set	F_1 MW	Precision MW	Recall MW
Global	0.59	0.55	0.65
Locality	0.59	0.51	0.70
Context (Cntxt)	0.49	0.58	0.43
Global + Locality	0.46	0.51	0.41
Global + Context	0.53	0.51	0.53
Locality + Context	0.49	0.59	0.42
Global + Locality + Cntxt	0.44	0.53	0.38
Chance	*0.24*	*0.22*	*0.26*

The results indicated that: (1) all models substantially outperformed the chance-baseline; (2) both global and locality models had similar F_1 MW scores, but slightly different precision and recall scores; (3) the combined global + locality model had (surprisingly) lower performance than either feature set alone; and (4) adding context to the individual models did not result in any improvement; if anything it reduced classification accuracy.

Proportionalized confusion matrices for the gaze-based models are shown in Table 3. We note that the errors for global and locality models were skewed towards false positives (vs. misses), which would explain the higher recall with respect to the Global + Locality model, we saw a higher proportion of misses, which would explain its lower recall score.

Locality features relate to spatial location of gaze, however, each phase of Guru had different screen content (e.g., Figure 1 vs. Figure 2). To examine if this caused bias against locality features, we compared global vs. locality models for the Common Ground Building phase - the only phase with enough data to build phase-specific models. The number of available instances was reduced to 1,259 (from 2,334) and MW rate increased to 30%. Classification results are shown in Table 4, where we note no substantial differences compared to the phase-independent models shown in Table 2, assuaging concerns of bias.

To further explore the validity of our detector we investigated whether predicted MW was related to posttest performance in the same way reported MW was. Participant-level *reported* MW rate was negatively correlated with posttest score (*rho* = -.189 p = .058) while *predicted* MW was also negatively correlated with posttest for detectors built with both the Global (*rho* = -.112, p = .269) and Locality (*rho* = -.177, p = .076) feature sets.

Table 3. Confusion matrices for gaze-based models

Actual	**Predicted**	
Global	*MW*	*Not MW*
MW	0.65 (hit)	0.35 (miss)
Not MW	0.52 (false pos.)	0.48 (correct rej.)
Locality	*MW*	*Not MW*
MW	0.70 (hit)	0.30 (miss)
Not MW	0.56 (false pos.)	0.44 (correct rej.)
Global + Locality	*MW*	*Not MW*
MW	0.41 (hit)	0.59 (miss)
Not MW	0.31 (false pos.)	0.69 (correct rej.)

Table 4. Models built for CGB phase

Feature Set	**F_1 of MW**	**Precision of MW**	**Recall of MW**
Global	0.59	0.55	0.64
Local	0.61	0.58	0.65
Global + Local	0.44	0.54	0.37

Feature Analysis

We compared the global gaze features across instances of MW versus not MW to characterize the differences in gaze during MW. Cohen's *d*, an effect size measure, was used to assess the direction and magnitude of the differences between the two classes [10]. For each class (MW and Not MW) the average for each feature across instances was computed. Cohen's *d* was computed by calculating the difference of each feature across MW and Not MW divided by the pooled standard deviation. Positive *d* values for a feature indicate higher values for instances of MW compared to instances of Not MW. Twenty (out of 57) of the effect sizes observed are consistent with small effects, using the convention that .2, .5, and .8 for small, medium, and large effects respectfully [10] the remaining effect sizes were less than .2 suggesting that no one feature dominated, but that a combination was needed for MW detection.

To establish which features contributed most to MW detection, the ten largest effect sizes were ranked in terms of their absolute Cohen's *d*. Fewer fixations (d = -.41) and saccades (d = -.40) (which are by definition highly correlated) were found for MW, and fixations were more dispersed (d = .23). Differences in median and mean saccade velocity (d = -.33, d = -.32 respectively), range of saccade angles (d = -.26), mean saccade duration (d = .24), maximum saccade angle (d = -.24), maximum and median saccade duration (d = .23, d = .22 respectively) suggest that saccades were slower, longer and covered a smaller range of angles during MW. These findings are consistent with previous work on eye gaze surrounding MW in reading, which also found number of fixations to be predictive [4], highlighting consistent differences in eye gaze features across learning tasks. These effects suggest that during MW, students focus on fewer points on the screen for a longer time. In addition, the effects for saccade duration and fixation dispersion suggest that these points are likely to be more spread around the screen rather than focusing in on information or visual stimulus such as diagrams.

4. GENERAL DISCUSSION

It is widely acknowledged that attention is necessary for learning [39]. An attention-aware learning technology [16, 42] that can monitor and react to a student's attentional state could assuage the cost of attentional failures (like MW), thereby improving learning. However, until now, the high cost of eye trackers (which are the most robust method to track visual attention) has relegated these technologies to the confines of the lab. We addressed this issue in the current paper by studying the feasibility of using COTS eye trackers in a real world classroom environment.

4.1 Main Findings

We have shown that, although the classroom provides a noisier environment than the lab, it is still feasible to collect valid eye tracking data with COTS eye trackers. Further, to maintain ecological validity, students were relatively unconstrained and independent in our study. We did give initial guidance with respect to seating position for calibration, however students were free to fidget, move, and behave as they would in a classroom. Despite this, we were able to achieve a median gaze validity of 95%. This is for the students where the gaze was collected at all. We were unable to collect data for 15% of sessions, however, this was primarily for reasons beyond our control (e.g., hardware issues with school computers and auto update).

Validity, however, does not imply usefulness. To address this, we built person-independent MW detectors based on the gaze data collected in the classroom. Our main finding was that our models were moderately accurate at detecting MW in a person-independent fashion despite the numerous challenges involved, such as class imbalance, noisy gaze data, and unrestricted movements. Importantly, our MW F_1 score of .59 was higher than the previous score of .49, achieved in a lab study with the same learning environment [28], although the comparison should be taken with a modicum of caution since the two studies differed along multiple factors (e.g., student population, type of eye tracker). Nevertheless, these results are encouraging as a detector with similar accuracy was used to successfully trigger interventions that improved learning gains in the context of reading [14]. We also extended the previous work that only investigated global gaze features, by exploring locality features as well as a combination of the two. This did not yield any performance improvements over the global features alone. One possibility is that the global features are sufficient for this task. However, it is more likely that the locality features considered here were too simplistic and benefits may be gained by refining them (see Future Work). In analysis of features we observed consistent differences with previous work in eye gaze and MW [4], most notably that when MW, students are more likely to have fewer, more spread out fixations than when not MW.

4.2 Applications

The key application of this work is to develop an attention-aware version of Guru that detects and combats MW in real-time. Such a system has a number of paths to pursue to re-engage students when MW is detected. One immediate effect of MW is that a student fails to attend to a unit of information or event because

they are consumed by internal, off-task thoughts. To combat this, one approach may be to simply repeat the missed information (e.g., "John, let me repeat that…") or to direct the user's attention to an area of the screen that may help them (e.g., "John, you might want to look at the image showing the enzyme breakdown…"). A more involved approach might be to ask the student a content question (e.g., "John, what happens to an enzyme when it is subjected to heat?") or ask the student to self-explain a concept. Additional measures might be needed if MW persists despite interventions. One option is to simply change to a new activity. Guru might even suggest changing topics or offering a choice for what students would prefer to do next. If all else fails, Guru might even suggest that the student take a break.

It is important to consider that the aforementioned interventions rely on MW detection, which is inherently imperfect. The detector may issue a false alarm, suggesting that a student is MW when they are not, or it could miss that a student is MW. In our view, MW detection does not need to be perfect as long as there is a modicum of accuracy. Imperfect detection can be addressed with a probabilistic approach, where the detector outputs a MW likelihood that is then used to determine whether an intervention is triggered (i.e., if the likelihood of MW is 70%, then there is a 70% chance of an intervention). The interventions should also be designed to "fail-soft" in that there are no harmful effects to learning if delivered incorrectly.

Beyond MW detection and response, COTS eye tracking in the classroom opens doors to several potential applications. One involves monitoring attentional states beyond MW (e.g., focused attention, alternating attention) so as to ensure that limited attentional resources are being optimally deployed [16]. Another application is alternate interaction methods that use eye-gaze as input, keeping learning novel and interesting. A further application is large-scale user testing of new learning technologies in the classroom. Student eye-gaze could also be used as a feedback tool to teachers, who can revise instruction/materials based on what captures students' attention.

4.3 Limitations

There were several limitations of this work. Our system was designed to include a low-cost eye tracker so that it may scale to classrooms. However, COTS eye trackers have a low sampling-rate, limiting their accuracy compared to research-grade eye trackers. Further, factors beyond our control, such as incorrect USB drivers on a school-owned computer, meant that for some of the sessions, no eye tracking was collected at all.

With regard to MW detection, we are limited by the features used in the supervised learning models. We used a small subset of gaze features and did not model any temporal gaze patterns. For example, if a participant had multiple fixations in one area, were these concentrated or distributed across time? In addition, we only considered a small number of contextual features.

At this time, locality features are not related directly to content and do not use Areas of Interest (AOI's). Guru's display is not fixed and changes throughout the tutorial phases. To further investigate locality features would require separate AOIs and corresponding models per phase. Because students spend different amounts of time in each phase, there were not enough instances to build phase-specific models.

A further limitation relates to the use of thought probes, which require users to be mindful of their MW and respond honestly. Although this method has been previously validated [21, 44, 47] there is no clear alternative to track a highly internal state like MW outside of potentially measuring brain activity in an fMRI scanner. One futuristic possibility is to combine self-reports and wearable electroencephalography (EEG) as a means of collecting more accurate MW responses, but it is unclear if this can be done in the wild.

4.4 Future Work

The results discussed here invite several possibilities for improvement that we will address as future work. First, we will explore a more refined set of locality features for MW detection. Example locality features involve fixations on various parts of the display, such as the tutor agent, aspects of the multimedia panel, the response box, and so on. When images are present, we will analyze image-specific gaze fixations, such as proportion of fixations on images, number of components fixated on, and fixation durations on different components. Guru uses a slow-reveal animation, where image components appear as they referenced in the session. This affords computing of animation-based locality features that measure gaze latencies to different image components as they are revealed.

We also plan to integrate our detectors into Guru to detect MW in real-time. Here, the MW probes will be triggered based on the detector's real-time probabilistic assessment of MW instead of the pseudo-random probing. Alignment between students' reports and the detector's estimates will be used to evaluate the detector's real-time MW accuracy when applied to new students. The detectors will be refined based on the outcome of these studies. The refined detector will then be used to deliver interventions (as noted above), leading to an attention-aware version of Guru.

4.5 Conclusion

The recent introduction of COTS eye trackers has ushered in an exciting time for gaze-based technologies that assist learning in the classroom. We have shown that valid and actionable eye-gaze data can be collected in an unconstrained manner despite the noisy real-world classroom environment. Our findings suggest that it might finally be possible to apply decades of lab-based research on eye gaze, attention, and learning to classrooms, thereby affording new discoveries about how students learn while designing new interfaces to sustain attention during learning.

5. ACKNOWLEDGMENTS

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 consultant (not named to protect student privacy) for welcoming us into their classroom.

REFERENCES

[1] Anderson, J.R. 2002. Spanning seven orders of magnitude: A challenge for cognitive modeling. *Cognitive Science*. 26, 1 (2002), 85–112.

[2] Arroyo, I. et al. 2007. Repairing disengagement with non-invasive interventions. *Proceedings of the 2007 Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work* (Amsterdam, The Netherlands, 2007), 195–202.

[3] Baker, R.S.J. d. 2007. Modeling and understanding students' off-task behavior in intelligent tutoring systems. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2007), 1059–1068.

[4] Bixler, R. and D'Mello, S. 2016. Automatic gaze-based user-independent detection of mind wandering during computerized reading. *User Modeling and User-Adapted Interaction*. 26, 1 (2016), 33–68.

[5] Bixler, R. and D'Mello, S.K. 2015. Automatic gaze-based detection of mind wandering with metacognitive awareness. *User Modeling, Adaptation and Personalization* (Dublin, Ireland, 2015), 31–43.

[6] Bixler, R. and D'Mello, S.K. 2014. Toward fully automated person-independent detection of mind wandering. *User Modeling, Adaptation, and Personalization* (Aalborg, Denmark, 2014), 37–48.

[7] Blanchard, N. et al. 2014. Automated physiological-based detection of mind wandering during learning. *Intelligent Tutoring Systems* (Switzerland, 2014), 55–60.

[8] 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.

[9] Chawla, N.V. et al. 2002. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Int. Res.* 16, 1 (Jun. 2002), 321–357.

[10] Cohen, J. 2013. *Statistical power analysis for the behavioral sciences*. Taylor & Francis.

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

[12] 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.

[13] Deubel, H. and Schneider, W.X. 1996. Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision research*. 36, 12 (1996), 1827–1837.

[14] 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.

[15] D'Mello, S.K. et al. 2012. Gaze tutor: A gaze-reactive intelligent tutoring system. *Int. J. Hum.-Comput. Stud.* 70, 5 (May 2012), 377–398.

[16] 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.

[17] Drummond, J. and Litman, D. 2010. In the zone: Towards detecting student zoning out using supervised machine learning. *Intelligent Tutoring Systems* (Pittsburgh, PA, USA, 2010), 306–308.

[18] Feng, S. et al. 2013. Mind wandering while reading easy and difficult texts. *Psychonomic Bulletin & Review*. 20, 3 (2013), 586–592.

[19] Forbes-Riley, K. and Litman, D. 2011. When does disengagement correlate with learning in spoken dialog computer tutoring? *Artificial Intelligence in Education* (Auckland, New Zealand, Jul. 2011), 81–89.

[20] 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.

[21] Franklin, M.S. et al. 2013. Window to the wandering mind: pupillometry of spontaneous thought while reading. *The Quarterly Journal of Experimental Psychology*. 66, 12 (2013), 2289–2294.

[22] 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.

[23] Graesser, A.C. et al. 2007. Inference generation and cohesion in the construction of situation models: Some connections with computational linguistics. *Higher level language processes in the brain: Inference and comprehension processes*. (2007), 289–310.

[24] 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.

[25] Hall, M. et al. 2009. The WEKA data mining software: An update. *SIGKDD Explor. Newsl.* 11, 1 (Nov. 2009), 10–18.

[26] Hegarty, M. and Just, M.A. 1993. Constructing mental models of machines from text and diagrams. *Journal of Memory and Language*. 32, 6 (Dec. 1993), 717–742.

[27] Hoffman, J.E. and Subramaniam, B. 1995. The role of visual attention in saccadic eye movements. *Perception & psychophysics*. 57, 6 (1995), 787–795.

[28] 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, 2016), 86–93.

[29] Jaques, N. et al. 2014. Predicting affect from gaze data during interaction with an intelligent tutoring system. *Intelligent Tutoring Systems:*, (Honolulu, HI, USA, Jun. 2014), 29–38.

[30] Kardan, S. and Conati, C. 2012. Exploring gaze data for determining user learning with an interactive simulation. *User Modeling, Adaptation, and Personalization* (Montreal, Canada, Jul. 2012), 126–138.

[31] 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.

[32] M. Cocea and S. Weibelzahl 2011. Disengagement detection in online learning: validation studies and perspectives. *IEEE Transactions on Learning Technologies*. 4, 2 (Jun. 2011), 114–124.

[33] Mathews, M. et al. 2012. Do your eyes give it away? Using eye tracking data to understand students' attitudes towards open student model representations. *Intelligent Tutoring Systems* (Chania, Crete, Greece, Jun. 2012), 422–427.

[34] Mills, C. et al. 2016. Automatic gaze-based detection of mind wandering during film viewing. *The 9th International*

Conference on Educational Data Mining. (Raleigh, North Carolina, 2016).

[35] Mills, C. et al. 2015. Mind wandering during learning with an intelligent tutoring system. *Artificial Intelligence in Education* (Madrid, Spain, Jun. 2015), 267–276.

[36] Mills, C. et al. 2014. To quit or not to quit: predicting future behavioral disengagement from reading patterns. *Intelligent Tutoring Systems* (Honolulu, HI, USA, Jun. 2014), 19–28.

[37] 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.

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

[39] 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.

[40] 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.

[41] Person, N.K. et al. 2012. Interactive concept maps and learning outcomes in Guru. *Florida Artificial Intelligence Research Society Conference* (Marco Island, FL, USA, May 2012), 456–461.

[42] 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.

[43] 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.

[44] 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.

[45] Rapp, D.N. 2006. The value of attention aware systems in educational settings. *Computers in Human Behavior Special issue: Attention aware systems*. 22, 4 (Jul. 2006), 603–614.

[46] Rayner, K. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*. 124, 3 (Nov. 1998), 372–422.

[47] Reichle, E.D. et al. 2010. Eye movements during mindless reading. *Psychol Sci*. 21, 9 (Sep. 2010), 1300–1310.

[48] Risko, E.F. et al. 2013. Everyday attention: Mind wandering and computer use during lectures. *Computers & Education*. 68, (2013), 275–283.

[49] 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.

[50] 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.

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

[52] 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.

[53] Seibert, P.S. and Ellis, H.C. 1991. Irrelevant thoughts, emotional mood states, and cognitive task performance. *Mem Cognit*. 19, 5 (Sep. 1991), 507–513.

[54] Shernoff, D.J. et al. 2014. Student engagement in high school classrooms from the perspective of flow theory. *Applications of Flow in Human Development and Education: The Collected Works of Mihaly Csikszentmihalyi*. M. Csikszentmihalyi, ed. Springer Netherlands. 475–494.

[55] 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.

[56] Smallwood, J. et al. 2008. When attention matters: the curious incident of the wandering mind. *Memory & Cognition*. 36, 6 (Sep. 2008), 1144–1150.

[57] Smallwood, J. and Schooler, J.W. 2006. The restless mind. *Psychological Bulletin*. 132, 6 (Nov. 2006), 946–958.

[58] 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.

[59] Smilek, D. et al. 2010. Out of mind, out of sight: Eye blinking as indicator and embodiment of mind wandering. *Psychological Science*. (Apr. 2010).

[60] Sottilare, R.A. et al. 2013. *Design recommendations for intelligent tutoring systems: Volume 1-learner modeling*. US Army Research Laboratory.

[61] Szpunar, K.K. et al. 2013. Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proc. Natl. Acad. Sci. U.S.A.* 110, 16 (Apr. 2013), 6313–6317.

[62] Szpunar, K.K. et al. 2013. Mind wandering and education: from the classroom to online learning. *Front Psychol*. 4, (2013), 495.

[63] Vosskuhler, A. et al. 2008. OGAMA (Open Gaze and Mouse Analyzer): open-source software designed to analyze eye and mouse movements in slideshow study designs. *Behav Res Methods*. 40, 4 (Nov. 2008), 1150–1162.

[64] Wang, H. et al. 2006. Empathic tutoring software agents using real-time eye tracking. *Proceedings of the 2006 Symposium on Eye Tracking Research &Amp; Applications* (New York, NY, USA, 2006), 73–78.

[65] Wixon, M. et al. 2012. WTF? Detecting students who are conducting inquiry without thinking fastidiously. *User Modeling, Adaptation, and Personalization* (Montreal, Canada, Jul. 2012), 286–296.