Automated Gaze-Based Mind Wandering Detection during Computerized Learning in Classrooms

Stephen Hutt, *University of Colorado, Boulder*

Kristina Krasich, *University of Notre Dame*

Caitlin Mills, *University of British Columbia*

Nigel Bosch, *University of Illinois at Urbana-Champaign*

Shelby White, *Indiana University-Purdue University Indianapolis*

James R. Brockmole, *University of Notre Dame*

Sidney K. D'Mello, *University of Colorado, Boulder*

Abstract

We investigate the use of commercial off-the-shelf (COTS) eye-trackers to automatically detect mind wandering - a phenomenon involving a shift in attention from task-related to task-unrelated thoughts – during computerized learning. Study 1 (N = 135 high-school students) tested the feasibility of COTS eye tracking while students learn biology with an intelligent tutoring system (ITS) called GuruTutor in their classroom. We could successfully track eye gaze in 75% (both eyes tracked) and 95% (one eye tracked) of the cases for 85% of the sessions where gaze was successfully recorded. In Study 2, we used this data to build automated student-independent detectors of mind wandering, obtaining accuracies (mind wandering $F_1 = 0.59$) substantially better than chance ($F_1 = 0.24$). Study 3 investigated context-generalizability of mind wandering detectors, finding that models trained on data collected in a controlled laboratory more successfully generalized to the classroom than the reverse. Study 4 investigated gaze- and video- based mind wandering detection, finding that gaze-based detection was superior and multimodal detection yielded an improvement in limited circumstances. We tested live mind wandering detection on a new sample of 39 students in Study 5 and found that detection accuracy (mind wandering $F_1 = 0.40$) was considerably above chance (F1 = 0.24), albeit lower than offline detection accuracy from Study 1 ($F_1 = 0.59$), a finding attributable to handling of missing data. We discuss our next steps towards developing gaze-based attention-aware learning technologies to increase engagement and learning by combating mind wandering in classroom contexts.

1. Introduction

The number of students utilizing computer-based learning has soared in the past few years. For instance, more than a quarter of students in higher education in the United States are enrolled in at least one online course (Allen & Seaman, 2016). Computer-based learning is hailed as resistant to time, location, and situation barriers (Bates, 2005) and is a cost-effective alternative to traditional learning environments (Twigg, 2003). Yet, the impoverished student-instructor interaction in computer-based learning leaves much to be desired. While a human tutor can dynamically adapt instruction to better engage students (Ainley & Luntley, 2007), this is largely beyond the scope of current educational technologies. For example, a tutor who notices that a student appears disengaged may attempt to reengage the student by asking him or her a question. This level of adaptivity is only possible if the tutor can monitor the student's level of attentional focus. Current educational technologies are largely unable to assess a student's attentional state and therefore cannot provide such dynamic attention-aware instruction.

It is imperative that we address this deficiency, as it is widely acknowledged that attention is crucial for effective learning (Berliner, 1990; Olney, Risko, D'Mello, & Graesser, 2015; Shernoff, Csikszentmihalyi, Schneider, & Shernoff, 2003; Smallwood, Fishman, & Schooler, 2007; Szpunar, Khan, & Schacter, 2013). Students who are unable to sustain attentional focus are more likely to engage in self-distracting and other unproductive behaviors (Forbes-Riley & Litman, 2011), which leads to superficial understanding as opposed to deep comprehension. Accordingly, our goal is to

This is the author's manuscript of the article published in final edited form as:

Hutt, S., Krasich, K., Mills, C., Bosch, N., White, S., Brockmole, J. R., & D'Mello, S. K. (2019). Automated gaze-based mind wandering detection during computerized learning in classrooms. User Modeling and User-Adapted Interaction, 29(4), 821–867. https://doi.org/10.1007/s11257-019-09228-5

develop attention-aware learning technologies that can sense and respond to students' attentional states as a means to improve attentional focus, engagement, and learning (D'Mello, 2016).

As an initial step in this direction, we focus on one kind of attentional lapse called mind wandering. Mind wandering is defined as an attentional shift from task-related processing towards internal task-unrelated thoughts (Smallwood & Schooler, 2006) (more detail in section 2.1). Mind wandering is quite frequent during learning, occurring 20%-40% of the time (D'Mello, 2018; Olney et al., 2015; Risko, Anderson, Sarwal, Engelhardt, & Kingstone, 2012; Risko, Buchanan, Medimorec, & Kingstone, 2013; Szpunar, Khan, et al., 2013; Szpunar, Moulton, & Schacter, 2013). Although the *trait-level* tendency to mind wander is positively associated with creative problem solving and prospective planning (Mooneyham & Schooler, 2013), a meta-analysis of 88 independent samples indicated a negative correlation between *state* mind wandering and performance across a variety of tasks (Randall, Oswald, & Beier, 2014). Further, the magnitude of the negative correlation increases for more complex tasks, such as learning. To this point, a recent meta-analysis (D'Mello, 2018) of 25 studies that tracked mind wandering across a range of digital learning environments indicated that mind wandering is negatively correlated with learning outcomes (r = -.24). This is unsurprising because when learners mind wandering, they miss out on key concepts (Robertson, Manly, Andrade, Baddeley, & Yiend, 1997; Smallwood, McSpadden, & Schooler, 2008), have increased difficulty encoding information into memory (Seibert & Ellis, 1991), and fail to comprehend learning content (Feng, D'Mello, & Graesser, 2013; Jonathan W. Schooler, Reichle, & Halpern, 2004). Thus, there may be benefits to attention-aware technologies that address mind wandering in real-time.

In order to address mind wandering in real-time we must first be able to detect when a student is mind wandering. However, whereas mind wandering is related to other forms of disengagement, such as boredom, behavioral disengagement, and off-task behaviors, it is inherently distinct because it primarily involves internal thoughts. This raises two challenges for detecting it. First, while other disengaged behaviors often involve detectable behavioral markers (e.g., yawns signaling boredom), mind wandering is an internal state (Smallwood & Schooler, 2006) with fewer overt signals that we know of. Second, mind wandering can occur outside of conscious awareness (Smallwood & Schooler, 2015), making it difficult to precisely measure in the first place.

Despite these challenges, there has been some progress toward automatic detection of mind wandering (see Section 2.3). Eye tracking is an attractive method for this purpose due to decades of evidence in support of a tight coupling between attention and eye movements—the so-called "eye-mind" link (Deubel & Schneider, 1996; Hoffman & Subramaniam, 1995; Just & Carpenter, 1976; Rayner, 1998) (see Section 2.2). Eye tracking has been used as a research tool for over a century (Buswell, 1936, 1937; Dodge, 1900; Huey, 1898, 1908; Javel, 1878; Yarbus, 1967) as well as for several real-world applications, such as military training in flight simulations (Weibel, Fouse, Emmenegger, Kimmich, & Hutchins, 2012), target identification (Hild, Kühnle, & Beyerer, 2016), and to help surgeons critically analyze their surgical skills (Ahmidi et al., 2010). Although, these applications were designed for use outside the laboratory, they typically use research-grade eye trackers that cost thousands of dollars, thereby limiting widespread scalability. Some work has also employed cellphone cameras to track eye movements in smartphone applications (Krafka et al., 2016), though many technical hurdles (e.g. tracking the gaze of users wearing glasses) need to be overcome before it is suitable for real-world use.

Fortunately, the recent availability of consumer off-the-shelf (COTS) eye trackers (retailing for hundreds of dollars) has ushered forth an exciting era by enabling scalable "in the wild" gaze-based research and applications (e.g., Maurer, Krischkowsky, & Tscheligi, 2017; Navarro & Sundstedt, 2017; Zhang, Chong, Müller, Bulling, & Gellersen, 2015). In the domain of learning, it presents new opportunities to explore attention during learning and to design learning technologies that improve engagement and learning by monitoring students' attentional states. Accordingly, we develop mind wandering detectors from eye-gaze data collected in the real-world context of a computer-enabled classroom, taking us one step closer to scalable attention-aware systems in the wild.

1.1 Current Study & Novelty

Based on the literature review (see section 2), there has been an uptake of research on automated mind wandering detection. However, all of the studies have focused on mind wandering detection in a laboratory environment. Laboratory environments have the advantage of relatively consistent lighting (important for some sensors) and freedom from distractions from other students (as students are usually individually tested), cell phones, ambient sounds, and numerous other factors. In contrast, we build upon these laboratory investigations to develop mind wandering detectors from eye-gaze data collected in the real-world context of a computer-enabled classroom.

It should be emphasized that mind wandering, like any other psychological construct, must be

operationalized in order to be investigated. In this work we use self-reports through thought probes (i.e., asking a person if they are currently mind wandering) as our measure of mind wandering because it is inherently a conscious phenomenon (Smallwood & Schooler, 2015) and self-reports are the most effective method to access conscious content (Ericsson & Simon, 1980). Although this methodology has been previously validated (Franklin, Broadway, Mrazek, Smallwood, & Schooler, 2013; Randall et al., 2014; Reichle, Reineberg, & Schooler, 2010) it relies on students being aware of their mind wandering and responding honestly (discussed further in limitations in section 8.3). Thus, our mind wandering detector is constrained with respect to our use of self-reports to operationalize mind wandering.

We make five contributions[1]. First, it is currently unknown whether COTS eye trackers can be implemented with sufficient fidelity in noisy classroom settings. We address this challenge by tracking eye gaze while high school students learn biology as part of their regular classes with GuruTutor (Olney et al., 2012), a dialogue-based intelligent tutoring system (discussed in Section 3.1). We show that it is feasible to use COTS eye trackers to collect valid data in a classroom environment.

Second, we demonstrate that the aforementioned eye gaze data collected in the classroom is of sufficient quality to automatically detect mind wandering in a student-independent fashion. We also experiment with different types of feature sets, including global (general) eye movements, locality (content sensitive) eye movements, and contextual features from GuruTutor. We find equitable performance between global and locality features. Because global features require less precise eye gaze, this affords more tolerance to eye tracking errors.

Generalization of detection is of particular interest here. With the exception of Stewart et al. (Stewart, Bosch, & D'Mello, 2017), much of the work in automated MW detection has focused on particular environment or context (Bixler & D'Mello, 2016; Blanchard, Bixler, Joyce, & D'Mello, 2014; Stewart, Bosch, Chen, Donnelly, & D'Mello, 2017). Taking a machine learnt model trained in one situation and applying it to another situation, be that a different context, feature space, or different classification, is a complex problem across in machine learning (Pan & Yang, 2010). Accordingly, in our third study, we investigate the differences between data collected in the laboratory versus data collected 'in the wild.' Using cross-training methods (i.e., building models on lab data and testing on classroom data), we show that lab-based mind wandering detectors are transferable to the classroom environment, but not vice versa – a finding with interesting implications.

Fourth, given the success of facial features for detecting mind wandering (Stewart, Bosch, Chen, et al., 2017), we compared face- and gaze- based mind wandering detection, finding a strong advantage for gaze. We also build multimodal models trained using gaze data and facial feature data to investigate how fusion methods can be used to improve mind wandering detection, especially in cases where one of the streams is missing, a common occurrence in a classroom environment.

Finally, we implement our gaze based mind wandering detector in GuruTutor and show that the detector trained on previously collected data can then be used in real-time on a new sample of classroom students. This is an important step because it demonstrates that the detector can be deployed in the real-world.

To our knowledge, this is the first study to investigate COTS eye tracking for mind wandering detection in the wild, the generalizability of gaze-based models across laboratory and classroom contexts, a multimodal gaze+video combination to detect mind wandering under real-world constraints, and real-time "online" mind wandering detection. Taken together, the models we develop will play a critical role in the first fully-automated attention-aware intelligent tutoring system for use in classrooms.

2. Background and Related Work

2.1 What is Mind Wandering?

At its core, mind wandering is an attentional shift away from the processing of external, task-related information to the processing of internal, task-irrelevant thoughts or ideas (Giambra, 1995; McVay & Kane, 2009, 2012; McVay, Kane, & Kwapil, 2009; J W Schooler et al., 2011; Smallwood, Beach, Schooler, & Handy, 2008; Smallwood et al., 2007; Smallwood, McSpadden, et al., 2008; Smallwood & Schooler, 2006). These shifts in the locus of attention usually occur without intention or even awareness (Giambra, 1995; J W Schooler et al., 2011). There are multiple hypotheses regarding the cognitive mechanisms underlying mind wandering (Smallwood & Schooler, 2015). According to

[1] Study 1 and 2 have been previously published in (Hutt, Mills, et al., 2017)

the *executive-resource hypothesis* (Smallwood & Schooler, 2006), when a task does not sufficiently consume all of one's attentional resources, unused resources are automatically allocated to task-unrelated thoughts and mind wandering occurs. In contrast, the *control-failure hypothesis* posits that mind wandering occurs when executive control fails to suppress task-unrelated thoughts (McVay & Kane, 2010, 2012). Despite these differences, the common argument is that both task-related and task-unrelated thoughts compete for attention, a limited resource, and mind wandering occurs when attentional resources are diverted to task-unrelated thoughts.

Furthermore, different types of mind wandering can also be identified based on the content of task-unrelated thoughts (Faber & D'Mello, in press.; Stawarczyk, Majerus, Maj, Van der Linden, & D'Argembeau, 2011). For example, thoughts related to feelings pertaining to a task (e.g., "This is so boring.") are different from completely unrelated thoughts (e.g., "I wonder what they will serve for lunch today."). Recent research also suggests that aspects of the learning materials themselves (e.g., encountering the word water in a chemistry course) can trigger mind wandering due to the associative nature of memory (e.g., water triggers a memory of a summer at the beach) (Faber & D'Mello, in press).

2.2 Why eye gaze reflect mind wandering?

It is generally assumed that attention is focused on where the eyes are fixated (e.g., Hoffman & Subramaniam, 1995; Yonetani et al., 2012). Eye gaze is considered a real-time index of the information-processing priorities of the visual system because physiological and cognitive limitations on vision, attention, and memory require the eyes to shift from location to location to construct a comprehensive representation of the external world. Furthermore, visual information is only acquired during *fixations*, periods when the eye remains relatively stable, whereas visual input is suppressed during *saccades*—ballistic movements of the eyes between fixation points (e.g., Campbell & Wurtz, 1978; Matin, 1974; Zuber & Stark, 1966). Therefore, ongoing task goals are often best served when central fixation is allocated to the most important visual information within the environment, and, thus eye movements should reflect where visual attention is allocated.

Although eye movements are used as signatures of attention, a growing body of research has observed changes in eye movements when people are *not* attending to visual tasks, such as during mind wandering. For instance, self-reports of mind wandering during reading are associated with fewer and longer fixations, greater variability in fixation patterns, and more frequent eye blinks (Faber, Bixler, & D'Mello, 2017; Reichle et al., 2010; Smilek, Carriere, & Cheyne, 2010; Uzzaman & Joordens, 2011). Mind wandering during visual scene processing has been associated with associated with fewer and longer fixations, greater fixation dispersion, and more frequent eye blinks (Krasich et al., 2018) Thus, there appears to be a shift in the sampling strategy of the visual system during mind wandering in that fewer regions are sampled and incoming information is restricted via blinks (Gawne & Martin, 2000). This is consistent with accounts showing reduced cortical processing of external information during mind wandering (Baird, Smallwood, Lutz, & Schooler, 2014; Barron, Riby, Greer, & Smallwood, 2011; Kam et al., 2011; Smallwood, Beach, et al., 2008). Collectively, this research demonstrates that as mind wandering becomes prioritized and task-related processing is deprioritized (Csifcsák & Mittner, 2017), certain aspects of gaze behavior change. In essence, mind wandering is a form of "looking without seeing" because the eyes might fixate on the appropriate external stimulus, but very little is processed (Baird et al., 2014). This is the key insight that motivates gaze-based mind wandering detection as reviewed below.

2.3. Previous work on mind wandering detection

Automated approaches to detect mind wandering are growing over the last decade. Here, we distinguish between non-gaze and gaze-based mind wandering detection.

Non-gaze mind wandering detection. There has been sporadic work in modalities, such as reading time and textual features (Franklin, Smallwood, & Schooler, 2011; Mills & D'Mello, 2015), prosody (Drummond & Litman, 2010), facial features (Stewart, Bosch, Chen, et al., 2017), and peripheral (Blanchard et al., 2014; Pham & Wang, 2015) and central physiology (Mittner et al., 2014). In particular, researchers have built MW detectors based on interaction information readily available in log files collected during the reading (e.g., reading time, complexity of the text). For example, (Mills & D'Mello, 2015), attempted to classify whether students were MW while reading a screen of text using reading behaviors and features of the text, such as text difficulty. Similarly, (Franklin et al., 2011) also attempted to predict MW during reading using textual features, such as word familiarity, difficulty, and reading time. However, rather than using supervised machine

learning, they used a set of researcher-defined thresholds to ascertain if participants were "mindlessly reading" based on difficulty and reading time.

Facial features have also been used for mind wandering detection. Stewart et al. (Stewart, Bosch, Chen, et al., 2017) recorded videos of participants' faces as they watched a narrative film, from which they extracted facial action unit (AU) and body motion features. A combination of these feature sets proved moderately successful at predicting MW. MW detectors trained on facial features (AU and body motion features have also been shown to generalize between contexts (Stewart, Bosch, & D'Mello, 2017). This study used two datasets, one in which participants watched a narrative film, and another in which a separate set of participants read a scientific text. Models were trained on each domain and were first each shown to be successful in their own domain. In addition, the model trained on the narrative film model performed above chance when applied to the scientific text reading data, and vice versa after adjusting the prediction threshold.

Gaze-based mind wandering detection. Researchers have recently leveraged aforementioned relationships between gaze and attention to build gaze-based mind wandering detectors during reading (Bixler & D'Mello, 2014, 2016). In these studies, mind wandering was measured via pseudo-random thought probes triggered during computerized reading, which required participants to report whether they were mind wandering or attentive to their task in the moments prior to the probe (called probe-caught mind wandering). Supervised classification models were successful in discriminating between "yes" and "no" responses to the thought probes using global eye gaze features (e.g., number of gaze fixations; fixation duration) and were validated in a manner that generalized to new students. Importantly, gaze-based mind wandering estimates correlated with learning outcomes, thereby providing evidence for the predictive validity of the detector (Bixler & D'Mello, 2016).

A recent study (Faber et al., 2017) used similar methods to build gaze-based mind wandering detectors during reading. However, participants reported when they caught themselves mind wandering (called self-caught mind wandering) rather than responding to a thought probe. The researchers used a combination of supervised classification models trained on global eye gaze features and probabilistic prediction to address missing data to successfully detect mind wandering. Relatedly, Loboda (2014) showed that eye gaze was predictive of self-caught and probe-caught mind wandering in reading. He also showed that predictive features varied between the two methods, suggesting that the different context (e.g. whether the person has meta-awareness of the mind wandering or not) impacts the nature of eye movements.

Compared to the scale, variability, and density of visual information within many computerized learning environments, reading involves a relatively sparse visual context and prescribes specific scan patterns. Beyond reading, (Mills, Bixler, Wang, & D'Mello, 2016) applied gaze-based mind wandering detection to narrative film viewing. In this study, participants viewed a 32.5-minute film and reported when they caught themselves mind wandering. Supervised learning models were built using both global and content-sensitive local gaze features, with the latter being more accurate, presumably due to their sensitivity to the dynamic content being displayed on the screen.

All of the above studies have used research-grade eye trackers and were done in the lab. Hence, of particular relevance to this work is a lab study that detected mind wandering with COTS eye trackers during learning (Hutt, Mills, White, Donnelly, & D'Mello, 2016). This study tracked students' eye gaze with a Tobii EyeX as students completed a 30-40 minute learning session with GuruTutor (Olney et al., 2012)the same intelligent tutoring system (ITS) we use in the current research. . Students reported mind wandering by responding to pseudo-random thought probes throughout the session. A variety of supervised classification models were trained to detect mind wandering from global gaze features alone, achieving person-independent accuracies that were substantially greater than chance. This work served as a proof of concept that COTS eye trackers can be sufficient for mind wandering detection in a lab context. Here, we extend this work by investigating mind wandering detection in a classroom environment.

We note that this is not the first time eyetracking has been utilized outside of the laboratory. For example, eyetracking has been used to improve interaction for military training in flight simulations (Weibel et al., 2012) and for target identification (Hild et al., 2016). In another setting, eyetracking software has helped surgeons to critically analyze their surgical skills (Ahmidi et al., 2010). Although, these applications were designed for use outside the laboratory, they typically use research-grade eye trackers that cost thousands of dollars, thereby limiting widespread scalability. Non-research grade equipment has shown some potential as well. One example includes using eye-gaze to interact with public displays (e.g., changing the page) (Zhang et al., 2015), though the range of interactions was limited and the system was susceptible to head movements over a certain threshold. Video game designers have also embraced using eye gaze to augment more traditional interaction techniques (Maurer et al., 2017; Navarro & Sundstedt, 2017). Recent developments have

also employed cellphone cameras to track eye movements in smartphone applications (Krafka et al., 2016). However, this work is still in an early phase and many technical hurdles (e.g. tracking the gaze of users wearing glasses) need to be overcome before it is suitable for real-world use. Nevertheless, these and other recent works suggest the promise of using cost-efficient eye tracking as a mode of interaction in user-centered technologies with an eye towards wider scalability.

3 Study 1: Feasibility of COTS Eye tracking in Classrooms

3.1 Motivation

Eye tracking has been shown to be an effective modality for mind wandering detection using both research grade eye tracking (Bixler & D'Mello, 2016) as well as COTS eye tracking (Hutt et al., 2016) in controlled lab environments, often with an experimenter to guide participants with seating position, calibration, etc. For real-world use, students should be able to calibrate the eye trackers themselves after receiving basic instruction and the resultant gaze data should be of sufficient fidelity for user modeling. Study 1 explores the feasibility of collecting valid eye tracking data in a high school classroom using COTS eye trackers integrated with an ITS called GuruTutor. We also instructed students to respond to thought-probes of mind wandering, which were interspersed during the learning sessions as discussed in detail in Section 4.2.1.

3.2 Method

3.2.1 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 (Cade, Copeland, Person, & D'Mello, 2008; D'Mello, Lehman, & Person, 2010; D'Mello, Olney, & Person, 2010; Olney, Graesser, & Person, 2010) and has been shown to be as effective at promoting learning compared to small group tutoring with novice human tutors (Olney et al., 2012).

Guru's primary interface (see Figure 1) 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 images on the multimedia panel most relevant to the discussion, and images are slowly revealed as the dialogue advances.

Guru maintains a student knowledge model of various concepts (Sottilare, Graesser, Hu, & Holden, 2013), which it uses to tailor instruction to individual students. Student typed input is mapped to a speech act category (e.g., Answer, Question, Affirmative, etc.) using regular expressions and a decision tree learned from a labeled tutoring corpus (Rasor, Olney, & D'Mello, 2011). Guru uses the speech act category and multiple models of dialogue context to decide what to do next. For example, an affirmative response in the context of a verification question (e.g. "Do you understand?") is interpreted as a content-based answer, while an affirmative in the context of a statement like "Are you ready to begin?" is not. Guru uses a general model of dialogue (e.g., feedback, questions, and motivational dialogue) and specific models representing the *mode* of the tutoring session (Cade et al., 2008), including Lecture and Scaffolding (see below). The mode models contain specific logic for answer assessment, feedback delivery (positive, neutral, or negative), and student model maintenance.

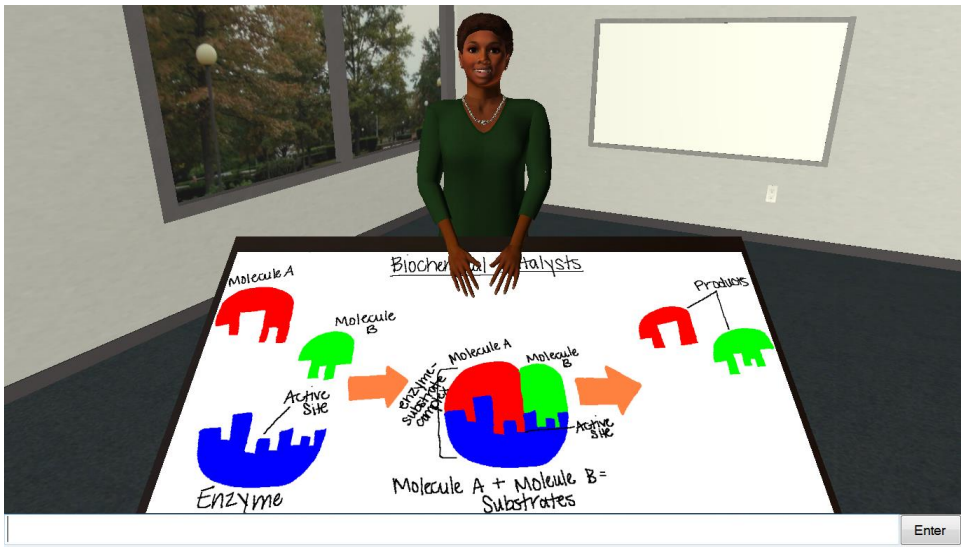

Figure 1. Screenshot of Guru in the CGB phase

Guru tutorials focus on introductory biology topics (e.g., osmosis; protein function) in line with state curriculum standards in short sessions, typically lasting between 15 to 40 minutes. Guru begins each tutorial session with a basic introduction to motivate the topic, which is then followed by the following five-phase session that develops students' understanding of the topic:

1. **Common-Ground-Building (CGB) Instruction.** Biology topics often involve specialized terminology that needs to be understood before it is advisable to move on to deeper knowledge building activities. Therefore, Guru begins with a collaborative lecture phase (D'Mello et al., 2010) which covers basic information and terminology relevant to the topic. The lectures are collaborative with a 3:1 (Tutor:Student) turn ratio in which the tutor asks concept-completion questions (e.g., "Enzymes are a type of what?"), verification questions (e.g., "Is connective tissue made up of proteins?"), or comprehension-gauging questions (e.g., "Is this making sense so far?").
2. **Intermittent Summaries (Summary).** Following CGB, students construct their own natural language summaries of the material covered in CGB. These summaries are automatically analyzed to develop the initial student model, which determines which concepts require further tutoring in the remainder of the session.
3. **Concept Maps**. For the target concepts, students complete skeleton concept maps, node-link structures that are automatically generated from text (see Figure 2; (Person, Olney, D'Mello, & Lehman, 2012)).
4. **Scaffolded Dialogue.** Students then complete a scaffolded dialogue in which Guru uses a Prompt → Feedback → Verification Question → Feedback → Elaboration cycle to cover target concepts in detail. An example of this would be the following:

   Tutor: What breaks the carbohydrate down?
   Student: Food
   Tutor: Not quite, does breaking the carbon bonds break the carbohydrate down?
   Student: Yes
   Tutor: Bingo! Breaking the carbon bonds within a carbohydrate breaks down the carbohydrate and releaces energy

   The student model is continually updated during the concept mapping and scaffolded dialogue phases. If a student shows difficulty mastering particular concepts, a second Concept Maps phase is initiated followed by an additional Scaffolded Dialogue phase.
5. **Cloze Task.** The session concludes with a cloze task requiring students to complete an ideal summary of the topic, filling in missing information in the summary by retrieving it from memory (see Figure 3).

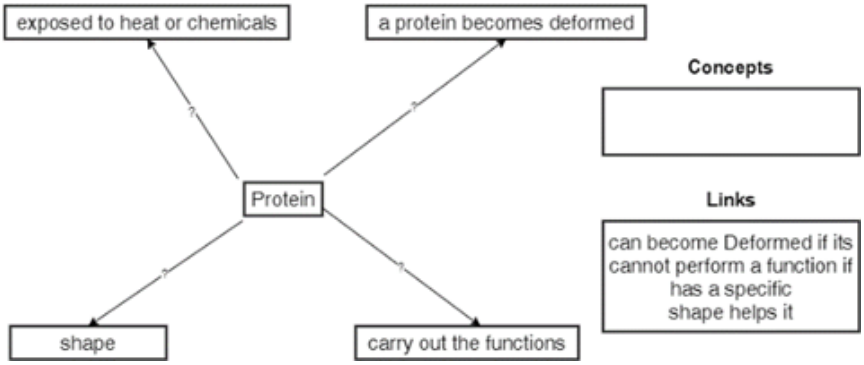

Figure 2. Example Concept Map

I'm done Show items I need help

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 _____. Water is allowed to move freely through the cell _____, and this movement of water is called osmosis. Depending on the level of water concentration inside and outside of the cell, the cell can be in an isotonic, hypotonic, or hypertonic solution. If a cell is in an isotonic solution, then the solution concentration inside of the cell is _____ __ the solution concentration outside of the cell. When the solute concentration is _____, the cell is in homeostasis. If a cell is in a hypertonic solution, then there is ____ of a solute outside of the cell than in the cell. When the solute concentration is unbalanced like this, water rushes ___ __ the cell, causing a water ____ that may lead to the cell shriveling up. If a cell is in a hypotonic solution, then there is ____ of a solute outside the cell than in the cell. In this case, the imbalance of solutes causes water to rush ____ the cell. The cell can potentially burst from the _____ water rushing into the cell. For plant cells, the _____ water from a hypotonic solution _____ the cells and allows the plant to stand up _____.

Figure 3. Example Cloze task, blue text denotes student answer

3.2.2 Integrating Sensors in Guru

Our first task was to integrate eye tracking into Guru. We chose a COTS eye-tracker called the EyeTribe that retailed for $99. The eye tracker connects to a computer via USB 3.0 and was affixed to a laptop as shown in in Figure 4. This eye tracker is no longer available for retail, but similar results have been obtained with an alternate COTS tracker (see Study 3).

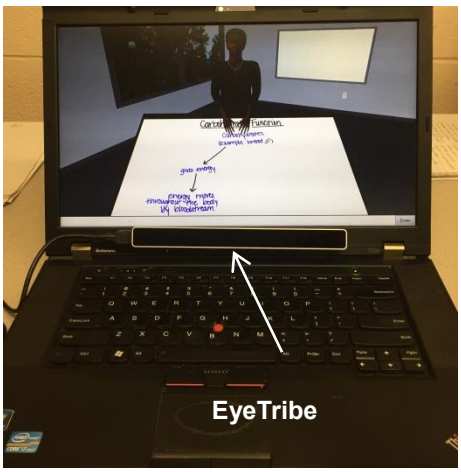

Figure 4. Example EyeTribe setup

One important 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 to position themselves within the range of the tracker, followed by information on the calibration process itself. In particular, students were shown a window with two cartoon eyes that represent their eye gaze (see Figure 5). A red background signified an incorrect position, whereas a green background denoted an acceptable

position. This was followed by a nine-point calibration process, where points appeared on the screen sequentially and students fixated on each until it disappeared.

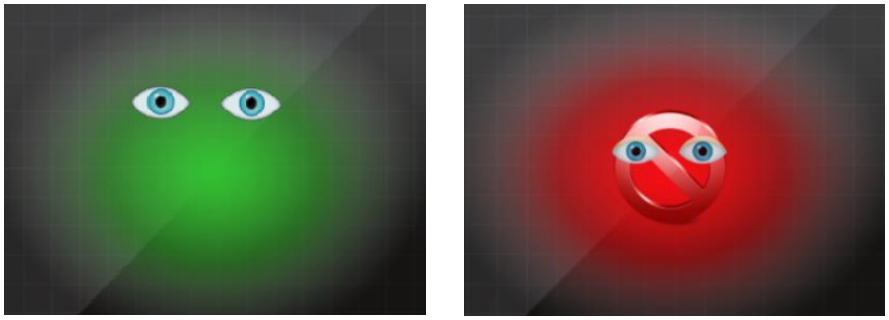

Figure 5. Head position positive feedback (left) and negative feedback (right).

In addition to eye tracking, we also collected facial feature data. Due to privacy concerns we could not record videos of students. Instead facial features were extracted live using OpenFace (Baltrusaitis, Robinson, & Morency, 2016). Facial feature extraction placed no additional requirement on the student beyond the correct positioning of the camera which was done using guided feedback provided at the beginning of the session.

3.2.3 Iterative Testing & Refinement

We completed several testing and refinement cycles to ensure that the entire implementation was as user friendly and autonomous as possible. Laboratory participants were compensated with research credit, while classroom participants were compensated with a $10 gift card.

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 tutorial session on one biology topic with Guru. Researchers observed the setup process to identify pain-points (e.g., unclear seating position instruction). 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 students received during the eye tracker calibration process.

Individual Testing in School – 9 Students. Initial testing of the implementation was conducted in after-school sessions with high-school student volunteers. Students completed the eye tracking setup and one Guru learning session. Each student was observed by a researcher, who noted critical 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 during each step. This informed our development of the software and streamlined the on-screen instructions, adding expanded instruction as needed.

Small Group Testing in School – 7 Students. As a step towards testing with entire classes of students, we tested the implementation with a small group (seven) of 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 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 Students. Finally, we conducted a classroom pilot using the same classroom as the main data collection (illustrated in Figure 6). We piloted with two class periods during students' regular biology classes. A key observation was the range of completion times with students finishing up to 20 minutes apart. This poses challenges as students who finished early could be sources of distraction for others. As a result, we enforced a time limit on the Guru session, auto advancing students to the posttest after 25 minutes.

Figure 6. Example classroom layout

We concluded that the usability of the system was considerably improved after these four rounds of testing and iterative refinement. Students were able to independently complete the setup, calibration process, and tutoring session via the on-screen instructions. In other words, they could use Guru with minimal guidance from the researchers and the resultant eye gaze was deemed sufficiently valid for larger-scale data collection. Figure 7 shows the final software workflow which includes the potential for a seating position check and recalibrating the eye tracker halfway through the session in case head position had changed considerably. In addition to this workflow we implemented a timout fuction that would automatically advance students to the post test after 25 minutes of working with Guru. This was done to ensure that all students could experience two topics per session.

3.2.4 Main Data Collection in Classroom

We collected data from 135 (41% male) freshmen and sophomore high-school students enrolled in a Biology 1 course over two sequential school days in students' regular biology classroom (see Figure 6). 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 high school.

Each class period consisted of an introduction to the software, 30 minutes of a biology session using Guru, a short break, then another 30-minute Guru session on a different biology topic. We used the following topics: Protein Function, Carbohydrate Function, Osmosis, Interphase, Facilitated Diffusion and Biochemical Catalysts, with students randomly assigned to a topic except that they could not receive the same topic for both sessions.

Each session began with a multiple-choice pretest and concluded with a multiple-choice posttest on the tutorial topic. As in the classroom pilot study, students were automatically advanced to the posttest at the 25-minute mark in the session. An example multiple choice item from the protein function topic is shown below (Figure 8).

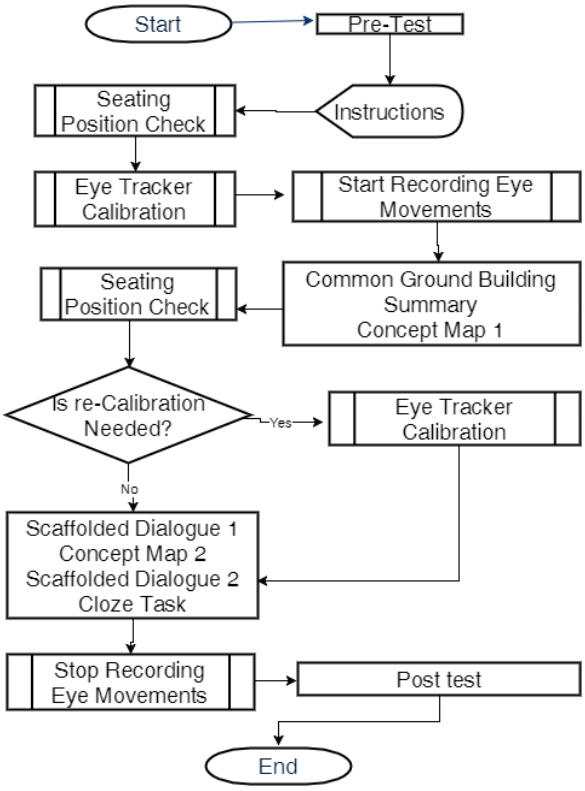

Figure 7. Final software flow diagram

What are two factors that can cause a protein to become deformed?

a) **exposure to chemicals AND heat**
b) exposure to carbohydrates AND other proteins
c) exposure to hormones AND antibodies
d) exposure to water AND oxygen

Figure 8. Example multiple choice question from protein function, correct answer shown in bold.

Class sizes ranged from 14 to 30 students based on regular enrollment. The classroom layout was the same as the setup used for regular instruction with the addition of two school-provided laptops per desk. Each laptop had an eye tracker affixed below the screen (see Figure 6 and Figure 9). A third laptop was present on each desk (not shown in Figure 6) in order to record facial features for a secondary study (discussed further in Section 6).

Students received instructions onscreen throughout the study. Specifically at the time of calibration, students were simply reminded to sit comfortably and instructed to follow the dot onscreen with their eyes. Two researchers were present during data collection to answer procedural questions and address any technical issues students encountered. In the event of calibration failure, the researchers would provide guidance on tracker positioning. The students' teacher remained at the back of the classroom and did not interact with the students throughout the study.

3.3 Results

The majority of students were able to use the software, including eye tracker calibration, without any intervention from the experimenters. There was a potential to collect 270 sessions as each of the 135 students completed two sessions. We obtained gaze data for 85% of these sessions with the following 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, causing problems with the eye tracker; (2) background processes—several computers attempted to automatically update during the session, causing an increased load on the processor, which in turn caused the software to occasionally crash; (3) calibration failure—students who failed to successfully calibrate after three attempts continued without eye tracking. Due to the unforseen nature of these errors, they were not appropraitely logged

by the software. Hence, exact proportions for each of these three failures are not availble, requiring us to instead rely on anecdotal reports and notes of the experimenter.

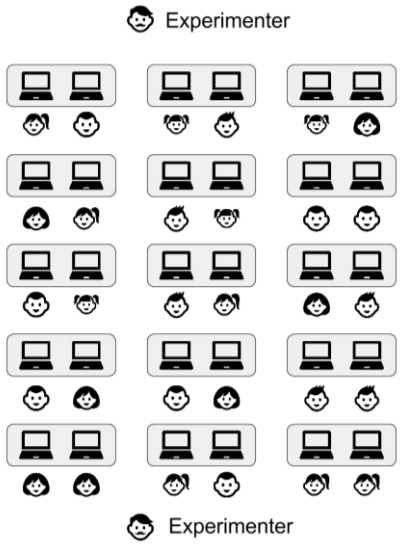

Figure 9. Classroom layout for data collection

We considered a valid sample to include at least one eye tracked with "high" quality as determined by the EyeTribe API. Figure 10 shows a histogram of percent of valid gaze points per session. We observed a median session-level validity rate of 95% for the sessions with gaze data. The median session-level validity dropped to 75% if we enforced a more stringent criterion that both eyes were tracked.

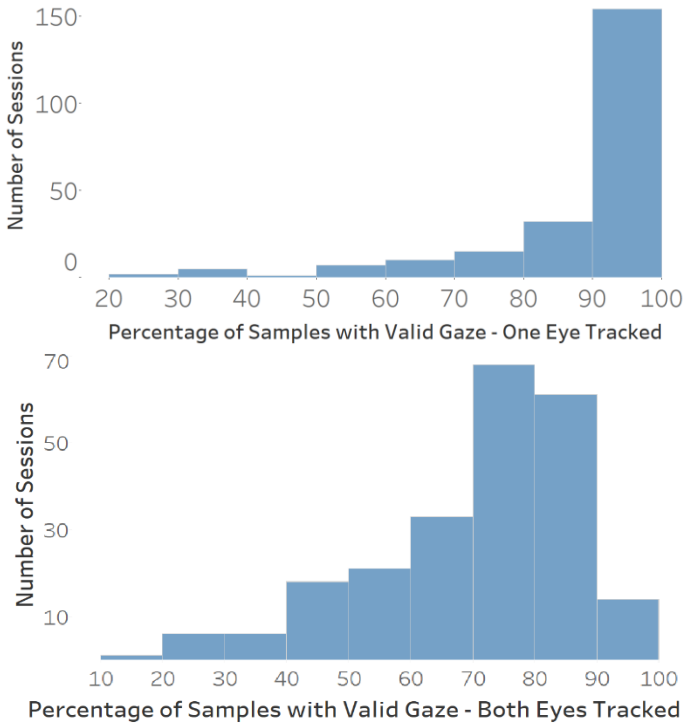

Figure 10. Histograms showing gaze validity rate per session where eye tracking was recorded

We also visualized the data as a first pass quality check. Figure 11 shows an example heatmap of one student's eye gaze during the CGB phase. As expected, we note the largest concentration of gaze on the tutor's face and upper body, followed by the multimedia panel, and the response box (at the bottom). This is in line with how active each element is in that the tutor moves and changes the

most throughout the session.

Having explored both tracker validity and visualizations for each participant, our overall conclusion was that we were able to track eye gaze with reasonable accuracy when entire classes of students used COTS eye trackers in a noisy real-world environment.

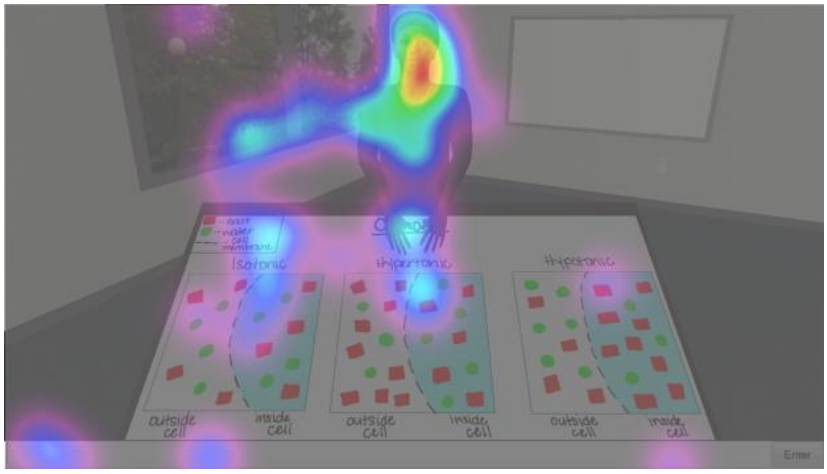

Figure 11. Heatmap overlay showing a student's eye gaze in CCB phase. Red indicates high concentration of fixations, purple low concentration of fixations

3.4 Discussion

Whereas the laboratory affords controlled data collection, the classroom presents a far nosier environment where students are free to turn and whisper to a neighbor or may become distracted by other students in the room. Further, although students were given initial guidance as to seating position for calibration, they were relatively unconstrained throughout the two tutorial sessions. Despite this, we were able to achieve a median gaze validity ranging from 75% (both eyes tracked) to 95% (one-eye tracked criterion) for the sessions where gaze was collected at all. Give the differences in validity rate we considered monocular data in subsequent analyses.

Although we were unable to collect data for 15% of sessions, these failures were primarily for reasons beyond our control (e.g., hardware issues with school computers and automatic system updates) that were not present in initial pilot studies. Overall, we considered these results to be adequate given the difficulties presented by the relatively unconstrained classroom environment.

We should note that eye tracking is not perfect. The Eye Tribe is designed to be tolerant to minor movements and we did not restrict movement other than requiring students to remain seated. During the tutoring session students were free to fidget, look around the room, and even occasionally lay their heads on the desk. As students look away tracking is lost, and their eyes must be re-detected when their focus returns to the screen. The redetection is supported by the EyeTribe but this series of events adds to the noise and inaccuracy of the tracker. Nevertheless, we considered the benefit of a 'real world' environment and a low cost eye tracker to outweigh tracking errors introduced into the dataset.

4 Study 2: Mind Wandering Detection (Offline Model)

4.1 Motivation

Study 1 indicated that COTS eye trackers could provide valid data in a noisy real-world environment. Our next step was to build automated mind wandering detectors using the data collected in the classroom.

4.2 Methods

We adopted a supervised learning approach using eye gaze and interaction data from the main data collection described in Study 1 (section 3.2.4). Thought probes (described below) interspersed during the Guru sessions were used to measure mind wandering.

4.2.1 Thought Probes to Measure Mind Wandering

We measured mind wandering during learning with Guru using visual & auditory thought probes, which is a standard approach in the literature (Smallwood, McSpadden, et al., 2008). Mind wandering was first defined to the students as follows:

> "While you're learning from the tutor, we want to know when you zone out. Zoning out is when you realize that you are no longer paying attention to what you're supposed to be doing. You probably experience it everyday! For example, instead of thinking about the biology, you may be thinking about something else altogether (maybe thinking about lunch or what you might be doing after school).
>
> In addition, we are also interested in what type of zone outs occur while you are learning from the biology tutor.
>
> Zone outs can occur either because you INTENTIONALLY decided to think about unrelated things or because your thoughts UNINTENTIONALLY drifted away despite your best intentions to stay focused.
>
> So when we ask you if you are zoning out, want you to distinguish between these two types of zone outs when you respond."

The instructions and the mind wandering reporting procedure were extensively tested and refined in the preliminary studies described above. Students were required to demonstrate understanding of how to respond to the thought probes (via multiple choice questions and feedback) before proceeding.

Students were probed at pseudo-random intervals with probes occurring every 90-120 seconds, based on previously observed mind wandering rates in Guru (Mills, D'Mello, Bosch, & Olney, 2015). The tutoring session paused when a probe was to be delivered (see Figure 12). 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 translucent overlay on screen, instructing the student 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 (Seli, Risko, & Smilek, 2016). In this study, we do not differentiate between intentional and unintentional mind wandering, so both "I" and "U" responses were considered as mind wandering. Students encountered an average of 12 probes over the course of each session with a mean mind wandering rate of 28% (SD = 24%, min = 0%, max = 100%).

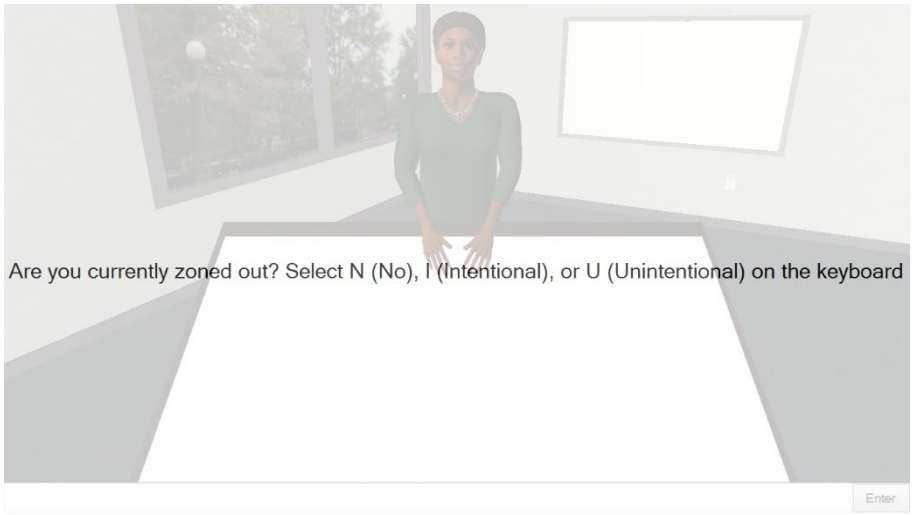

Figure 12. Example probe during Guru session

It is important to emphasize a few points about this probing method. First, it relies on self-reports because mind wandering is an inherently conscious phenomenon, which requires self-awareness for reporting (Franklin et al., 2013). Second, self-reports of mind wandering have been objectively

linked to patterns in pupillometry (Reichle et al., 2010), eye-gaze (Randall et al., 2014), and task performance (Smallwood & Schooler, 2015), providing validity for this approach (also see Section 8.3). That said, it is possible that the probing method may have re-oriented the students attention, which is why we limited the number of probes a student could receive. It was made clear to all students that their responses from the probes would not be shared with their teacher and they were encouraged to always answer honestly. At this time, there are no reliable neurophysiological or behavioral markers that can accurately substitute for the self-report methodology (Smallwood & Schooler, 2015). The limits of thought probes are considered further in the Discussion section, but as it currently stands, our use of thought-probes is consistent with the state of the art in the psychological and neuroscience literatures (Smallwood & Schooler, 2015).

4.2.2 Feature Engineering

We calculated features from 30-second windows (based on previous work (Hutt et al., 2016)) preceding each thought probe. Due to the lower validity binocular gaze recording (see section 3.3), we use monocular gaze throughout this study. We investigated two types of gaze features: global gaze (from previous work (Hutt et al., 2016)) 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 relative to specific scene content. We also encoded interaction data from the Guru session to obtain a set of context features (e.g., topic covered during the session).

Global Gaze Features. Eye gaze is measured as fixations (i.e., points in which gaze is maintained on the same location) and saccades (i.e., the movement of the eyes between fixations). We first fixation filtered the raw eye gaze data using Open Gaze and Mouse Analyzer (OGAMA) (Vosskuhler, Nordmeier, Kuchinke, & Jacobs, 2008). We considered six general measures of eye gaze 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 fixation dispersion, horizontal saccade proportion, and fixation saccade ratio (see Table 1), for a total of 57 global gaze features.

Table 1. Eye-gaze features. Bolded cell indicates that nine descriptives (e.g., mean, range) 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. Whereas the global features emphasize *how* eye's move, the locality features are based on *where* gaze is fixated on the screen. We computed these by overlaying a 10 × 8 grid on the screen. Each grid cell represented a feature and was assigned a weight proportional to the number of fixations on that corresponding cell (see Figure 13). In addition to these 80 locality features, we included an "out of bounds" feature that encoded the proportion of fixations that were off the screen bounds. We chose this approach instead of an Area of Interest approach (AOI) in an attempt to improve generalizability across Guru topics because the media panel is updated per topic.

Context Features. The gaze features were complemented by eight features that provide a snapshot of the current student-tutor interaction. One feature was the assigned biology *topic*. A second encoded students' *pretest* scores. The next three features represented students' 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 measured focused on students' performance within Guru, measured as the proportion of *positive*, *neutral*, and *negative* feedback received.

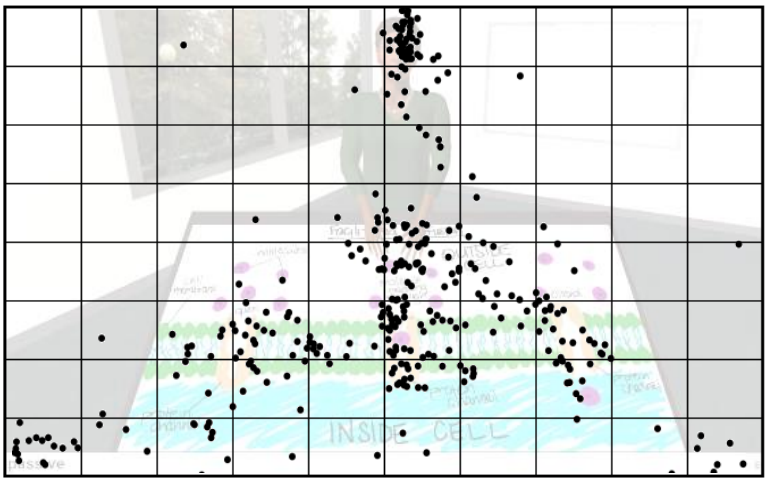

Figure 13. Example grid used for locality features - the proportion of fixations in each cell is a feature

4.2.3 Supervised Classification and Validation

We focused on Bayesian Networks for classification because they yielded the best performance compared to several other standard classifiers on this task in our previous work (Hutt et al., 2016). We used the default implementation from the Weka data mining package (Hall et al., 2009).

In total, there were 2,720 thought probes during the Guru sessions. Of those, 386 were discarded due to insufficient eye gaze data (< 1 fixation in the 30s window) to compute several of the global features. 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 as missing values and were imputed based on mean values in the training set.

We validated the models with a leave-several-students-out cross-validation scheme. For each fold, instances from a random 67% of the students were assigned to a training set and the instances of the remaining 33% students were assigned to a test set. This process ensures that no instances of any individual student could appear in both the training and test sets within a fold. This process was repeated for 15 iterations and the results were accumulated before computing accuracy metrics.

Students reported mind wandering for 23% of the 2,334 instances, resulting in substantial data skew. Class imbalance poses a challenge because supervised learning methods tend to bias predications towards the majority class. To compensate for this concern, we used the SMOTE algorithm (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) 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 used to oversample the minority class and was *only applied on the training sets;* the original class distributions were maintained in the test sets in order to ensure validity of the results.

4.3 Results

Because our intention is to detect instances of mind wandering, we focus on the precision, recall, and F_1 score of the mind wandering class as our key metrics. For comparison, a chance-level baseline was created by *randomly* assigning the mind wandering label to 23% of the instances (the mind wandering base rate) and computing accuracy metrics accordingly.

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

Proportionalized confusion matrices 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. In contrast the global+locality model had a higher proportion of misses, which would explain its lower recall score.

We statistically compared the individual models using mixed-effects linear regression (due to the repeated and nested structure of the data—one or more instances nested within a participant) with participant as an intercept-only random effect. We regressed agreement between model prediction and ground truth (i.e., accuracy) on model type with session as a random intercept. To account for differences in predicted MW rate across models, we included it as a fixed effect covariate. We used the lme4 package in R for model fitting, and the emmeans package for pairwise comparisons. Additionally, *p*-value adjustment for multiple comparisons was performed using the false-discovery rate method.

There was a significant main effect of model type ($p < .05$), which we probed with a series series of planned comparisons. We found no significant difference ($p = .445$) between global and locality classifiers with both being significantly ($ps < 0.001$) better than the context classifier. We then considered the effect of combining global and locality features and observed that both unimodal classifiers (global and locality) were significantly ($ps < .055$) better than the combined model. We also explored the effect of adding context to the individual classifiers and found either no significant improvement ($p = .16$ for global*)* or a reduction in accuracy ($p < .001$). Finally, the unimodal classifiers (global, locality, context) were significantly better than the trimodal classifier ($ps < .01$).

Table 2. Mind wandering (MW) detection results for classroom data

Feature Set	Predicted MW Rate	F_1 MW	Precision MW	Recall MW
Global	0.52	0.59	0.55	0.65
Locality	0.55	0.59	0.51	0.70
Context	0.35	0.49	0.58	0.43
Global + Locality	0.36	0.46	0.51	0.41
Global + Context	0.45	0.53	0.51	0.53
Locality + Context	0.34	0.49	0.59	0.42
Global + Locality + Context	0.33	0.44	0.53	0.38
Chance	0.23	0.24	0.22	0.26

Results for CGB Phase. 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 reduced the effectiveness of locality features, we compared global vs. locality models for the Common Ground Building phase - the only Guru phase with enough data to build phase-specific models. The number of available instances was reduced to 1,259 (from 2,334) and mind wandering rate increased to 30%. Classification results are shown in Table 4, where we note a small (.02 mind wandering F1) improvement for the locality model. There were no other substantial differences compared to the phase-independent models shown in Table 2, assuaging concerns of bias.

Predictive validity. To further explore the validity of our detector, we investigated whether predicted mind wandering was related to posttest performance similar to self-reported mind wandering. As expected, the student-level self-*reported* mind wandering rate was negatively correlated with posttest scores (Spearman's $rho = -.189$, $p = 0.058$) as was *predicted* mind wandering for both the global ($rho = -.112$, $p = 0.26$) and locality ($rho = -.177$, $p = 0.076$) models. None of these correlations were significant (including the ground truth correlation) at the $p < .05$ level, ostensibly due to the small sample size. Nevertheless, it is encouraging that the predicted MW rate of the locality model correlate with learning to the same extent as the self-reported MW rate.

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 Common Ground Building phase, comparison values shown in parentheses.

Feature Set	**Predicted MW Rate**	**F_1 MW**	**Precision MW**	**Recall MW**
Global	0.57	0.59 (0.59)	0.55 (0.55)	0.64 (0.65)
Local	0.55	0.61 (0.59)	0.58 (0.51)	0.65 (0.70)
Global + Local	0.37	0.44 (0.46)	0.54 (0.51)	0.37 (0.41)

4.3.1 Feature Analysis

In order to explore how gaze features were related to mind wandering, we compared the global gaze features across positive vs. negative instances of mind wandering. We only considered global gaze features as these are more interpretable in terms of understanding general gaze patterns. Cohen's *d*, an effect size measure, was used to assess the direction and magnitude of the differences between the two classes (Cohen, 2013). Positive *d* values for a feature indicate higher values for positive instances of mind wandering. To better understand how eye gaze relates to mind wandering detection, we ranked the ten largest effect sizes in terms of their absolute Cohen's *d* (shown in Table 5). We observe that during mind wandering, 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 specific regions. These effects mimic those observed in other scene viewing tasks (Krasich et al., in press) suggesting that they are relatively context-free.

Table 5. Ten largest effect sizes in terms of absolute Cohen's *d*

Feature	**Cohen's *d***
Number of Fixations	-.41
Number of Saccades	-.40
Median Saccade Velocity	-.33
Mean Saccade velocity	-.32
Range of Saccade Angles	-.26
Mean Saccade Duration	.24
Maximum Saccade Angle	-.24
Fixation Dispersion	.23
Maximum Saccade Duration	.23
Median Saccade Duration	.22

4.4 Discussion

We used the data from Study 1 to build student-independent gaze based mind wandering detectors. Despite the challenges involved in real-world data collection, we achieved mind wandering

detection rates (mind wandering F1 of .59) that substantially outperformed chance (mind wandering F1 = .24). We also extended previous work that only investigated content independent (global) features by exploring content sensitive (locality) features as well as a combination of the two. Content-sensitive features have been shown to be successful in other learning environments (Hutt, Hardey, et al., 2017; Mills et al., 2016) whereas in reading global features were sufficient (Bixler & D'Mello, 2016). Locality features did not yield any performance boost compared to global features, but did achieve similar results. In fact, predictions from the two feature sets were not too strongly correlated (*rho = 0.25)* suggesting they encode overlapping but non-redundant information..

A combination of global and locality features also did not provide a boost over global features alone, possibly because the locality features we considered were too simplistic for this task. A further explanation is that the Bayesian network was affected by mutual information between the two feature sets. To address this, we regressed each of the global features using the locality features as inputs. The average R^2 was *.25* (SD = 0.12, Min = .06, Max = 0.66), suggesting that on average 25% of the variance in global features is explained by local features. Although it seems there is some mutual information between the feature sets, the overlap is perhaps too low to suggest that this is the only reason for the reduced performance of the two models.

Finally, the models' estimates of mind wandering correlated with posttest scores at similar levels as self-reported mind wandering (particularly for the locality models), providing evidence of their predictive validity. Thus, though modest in accuracy, our mind wandering detector appears to have adequate validity.

5 Study 3: Cross-Training Detectors between the Lab and Classroom

5.1 Motivation

Studies 1 and 2 focused on mind wandering detection in the classroom. However, learning occurs in many contexts, raising the question of whether mind wandering detectors generalize across contexts? Thus, Study 3 sought to address how models built from lab data compare to those built classroom data, whether a lab-model generalizes to a classroom-model, and vice versa.

5.2 Methods

5.2.1 Lab Data Summary

Lab data were individually collected from 153 undergraduate students (one at a time) who received course credit for their participation (this dataset is an extension of data used in Hutt et al., 2016). After providing informed consent, students were seated at a desk in front of a 15-inch laptop connected to a Tobii EyeX eye tracker (another COTS eye tracker). We used the Tobii EyeX (with the appropriate liscencing to record data) because the EyeTribes were not available at that time. Students completed one session with Guru on one of the same topics used in the classroom study. Mind wandering was monitored using the same thought-probe method. In total, there were 2,066 mind wandering probes, with a mind wandering rate of 23% (the classroom yielded an equivalent rate of 23%).

5.2.2 Cross-Training Method

We considered three datasets: lab data, classroom data, and the lab and school data combined. The combined dataset had a total of 4,400 instances, 2,334 from the classroom and 2066 from the lab. We considered various combinations of training and testing data as showing in Table 5. As before, the validation method ensured that a participant could be in wither the training and testing set but never both. As was done previously, each process was repeated over 15 iterations. We only focused on global eye gaze as these features are more generalizable, less affected by eye tracker accuracy, and resulted in equitable performance compared to locality features (see above).

The two studies used different eye trackers, but both produced a similar quality of data (~95% average valid samples per session). However, the trackers had differing sampling rates, 30Hz for the EyeTribe and an average of 60Hz for the EyeX (note the EyeX sampling rate was not constant and varied slightly throughout the sessions). For both trackers, we used the same fixation filtering algorithm, however, due to the differing sampling rates we had to adjust the parameters by tracker. Specifically the number of samples required to constitute a fixation was set to three for the EyeTribe and six for the EyeX. We applied z-score standardization by dataset to further mitigate any

differences between the two trackers.

5.3 Results

Cross-training results are shown in Table 6. We first note that school-trained detectors outperformed lab-trained detectors (F_1 of .59 versus .44), although the findings could be confounded by differences in eye gaze trackers and participants (see Discussion). Importantly, detectors trained on data collected in the controlled laboratory environment were transferable to the more complex school environment (F_1 of .44 versus .43). The reverse was not as successful as models trained on the school data did not generalize well to the lab (F_1 of .59 versus .33), though there was still improvement over chance (F_1 = .23). It is possible that the different trackers had an effect here, but that would not explain the asymmetric relationship between the two cross-trained modes. The combined lab+classroom model generalized across data collected in both environments.

As in Study 2, we compared the individual models using mixed-effects linear regression. For the Lab+Classroom model, we observed no significant difference between testing in the classroom or the lab (p = .32) . For the model trained on the lab, we note that after accounting for MW prediction rate, testing on the lab was significantly (p < .001) better than testing on the classroom data (despite the similar F_1 scores). The reverse was true for models trained on classroom data as these yielded higher accuracies when tested in the classroom vs. the lab (p < .001).

Table 6. Mind wandering F_1 scores for cross training, predicted mind wandering rates shown in parenthesis

Training Set	**Testing Set**		
	Lab + Classroom	Lab	Classroom
Lab + Classroom	0.50 (0.49)	0.44 (0.36)	0.58 (0.55)
Lab	-	*0.44 (0.42)*	0.43 (0.58)
Classroom	-	0.33 (0.30)	*0.59 (0.52)*
Chance	0.24 (0.23)	0.23 (0.23)	0.24 (0.23)

To further explore the asymmetric relationship, we examined the confusion matrices for the two cross trained models (shown in Table 7). The model trained in the lab and tested in the classroom followed a pattern (hit rate) similar to the model trained and tested on classroom data. However, the model trained on the classroom data and tested on the lab data had a low hit rate, whilst maintaining a high correct rejection rate, ostensibly because it may be underpredicting mind wandering compared to the other models. We attempted to mitigate the underprediction by altering the decision threshold used by the model. The model calculates a likelihood of mind wandering for each instance and we previously considered instances with likelihoods greater than .5 as mind wandering. We experimented with altering this threshold as a way to adjust the predicted mind wandering rate, however, the asymmetric relationship in the results was maintained.

Table 7. Confusion matrices for cross trained models

Actual	**Predicted**	
Train Classroom, Test Classroom	*MW*	*Not MW*
MW	0.65 (hit)	0.35 (miss)
Not MW	0.52 (false pos.)	0.48 (correct rej.)
Train Lab, Test Classroom	*MW*	*Not MW*
MW	0.82 (hit)	0.18 (miss)
Not MW	0.65 (false pos.)	0.35 (correct rej.)
Train Classroom Test Lab	*MW*	*Not MW*
MW	0.39 (hit)	0.61 (miss)
Not MW	0.28 (false pos.)	0.72 (correct rej.)

5.4 Discussion

We investigated generalizability of the mind wandering detectors across contexts. Although the stimuli (Guru topics, instruction framework etc.) were very similar between the classroom and the lab studies, the students, environment, and eye tracker were very different. Despite these challenges, the models largely generalized across these two contexts when trained on combined data from both contexts. There was also an asymmetry in the results in that the lab model yielded equitable performance when applied to the classroom, but accuracy of the classroom model was reduced by almost half when applied to the lab data.

We also found that the classroom model yielded overall better results than the lab model. Though this result could be attributed to multiple factors (e.g., eye tracker, students, environment), the fact remains that while it may be appropriate to start out in the lab, in this case, real-world data provided better detection accuracy.

6 Study 4: Multimodal mind wandering Detection from Gaze and Video

6.1 Motivation

Studies 1-3 indicated that eye gaze is a valid, albeit imperfect, channel to detect mind wandering. In Study 4, we attempt to improve detection accuracy, by considering a multimodal approach combining our gaze based detectors with detectors trained on facial features extracted from video. Previous work on video-based mind wandering detection (Stewart, Bosch, Chen, et al., 2017; Stewart, Bosch, & D'Mello, 2017) focused on facial features collected in the laboratory. In contrast, (Bosch & D'Mello, in preparation) developed a video-based mind wandering detector using the same classroom data used to develop our gaze-based mind wandering detector (see Studies 1-2).

6.2 Methods

6.2.1 Facial Feature Models

The development of the facial feature models is outlined briefly here, for a more detailed description see (Bosch & D'Mello, in review). Facial features were extracted in real-time (simultaneously while gaze was recorded) using OpenFace (Baltrusaitis et al., 2016). Real-time processing was done due to privacy considerations, which precluded recording videos of students for offline feature extraction. The live feature extraction placed no additional requirement on the student (e.g. calibration) beyond positioning the camera correctly.

OpenFace provides intensity estimates of 14 (at the time of the study) action units (AUs), which represent specific facial muscle activations (Ekman & Friesen, 1978). AUs were extracted per frame as fast as real-time processing would allow (mean frames per second = 4.6). We also captured co-occurrence relationships between AUs, which can be important for distinguishing facial expressions. For example, a connection between muscle movements around the mouth and eyes when smiling can be a telling of genuine smiles, versus smiles involving the mouth only (Messinger, Fogel, & Dickson, 2001). Co-occurrence relationships between all pairs of AUs ($n = 14$) were computed via the Jensen-Shannon divergence (JSD) (Lin, 1991) an extension of Kullback-Leibler divergence (KLD) for measuring similarity between two probability distributions (Kullback & Leibler, 1951)

We used a 10-second window length for facial features based on previous work (Stewart, Bosch, & D'Mello, 2017) showing that a shorter window led to more accurate MW detectors for this modality. There were 2,888 10-second-long instances extracted from 135 students, of which 502 were discarded because they contained fewer than 4 frames of data (approximately 1 second). As features were extracted in real time, missing data occurred when the face could not be automatically detected, for example, if the face was no longer in frame or if the student had turned the side. The automatic facial recognition is tolerant to small movements as well as multiple faces in frame.

For each of the remaining 2,386 instances, AUs were aggregated across frames to obtain the mean, median, standard deviation, minimum, maximum, and range of each AU estimate within the video clip. This resulted in 84 AU features and 91 JSD features (all possible unordered pairs of 14 AUs).

A support vector machine (SVM) classifier was then trained on each feature set using a student-level k-fold cross validation scheme. The SVM models used a radial base kernel and hyperparameters were trained using a cross-validated grid search. The SVM's were used here instead of a Bayesian approach because they have been shown to be successful in this domain (Stewart, Bosch, & D'Mello, 2017).

6.2.2 Decision Fusion

The modalities operate at different time scales (10s for facial features and 30s for gaze features) and there are cases where data was only available from a single modality. Therefore, we used decision-level fusion to combine the two modalities instead of feature-level fusion. We combined the gaze-based models (Global Gaze and Locality Gaze) developed in Study 2 with the facial feature models (Face AU, Face JSD) to develop a multimodal detector by aligning the instances of the two modalities by end-point (i.e., time of the MW probe). We compared the four individual predictions to an unweighted classifier that simply averaged the four individual binary predictions and to a weighted average classifier that weighted the individual predictions by the mind wandering F_1 score of each individual classifier. Averages that exceeded a 0.5 threshold were taken as mind wandering.

6.3 Results

Results of the feature fusion classifiers are shown in Table 8. There was an overlap of 1,743 instances with no missing data across the two channels (instances that contained both valid gaze and face data). To ensure a fair comparison, this subset was used in initial fusion experiments. The results, shown in Table 8, indicate that neither of the fusion approaches outperformed the best single-channel models. As in Study 2, we compared the individual models using mixed-effects linear regression. We first compared within modality, observing no significant difference between the two gaze models (p = .417) or the two face models (p =. 683). We then compared the modalities to each other and observed that both gaze classifiers were significantly (ps < .001) better than both face classifiers. Finally, we compared the two gaze classifiers to the two fusion classifiers and found no significant difference (ps > .505).

Table 8. Results of multimodal feature fusion for complete data (1,734 instances,)

Classifier	Predicted MW Rate	F_1 MW	Precision MW	Recall MW
Individual classifiers				
Gaze Global	0.46	0.45	0.33	0.71
Gaze Locality	0.47	0.49	0.41	0.62
Face AU	0.35	0.31	0.25	0.4
Face JSD	0.29	0.3	0.28	0.32
Fused classifiers				
Unweighted Average	0.42	0.44	0.33	0.64
Weighted Average	0.33	0.45	0.37	0.57
Chance	*0.22*	0.23	0.22	0.23

Next, we considered the 2,558 instances with either valid gaze or valid face data, by simply ignoring missing predictions when computing the averages. For example, if only gaze predictions were present, then only these two values were averaged instead of four. Results of this classification are shown in Table 9. We again conducted a statistical analysis (following the same method as above) and noted that similarly, no significant difference between the gaze classifiers (p = .145), which were significantly better than the face classifiers (ps < .001). We also found that the fusion classifiers were significantly better than the global gaze classifier (p = .02), but not the local gaze classifier (p = .531), likely because of the differing MW prediction rates.

6.4 Discussion

We compared gaze- and face- based mind wandering detectors and found that the gaze models yielded the highest accuracies. We also showed that one solution to the missing data problem inherent in real-world contexts is to combine two modalities. It also should be noted that the improvement of the multimodal classifier over gaze classifiers was minimal so adding additional modalities may well over complicate MW detection and increase points of failure.

Table 9. Results of multimodal feature fusion with missing data included (2,558 instances)

Classifier	Predicted MW Rate	F_1 MW	Precision MW	Recall MW
Individual classifiers				
Gaze Global	0.48	0.47	0.37	0.61
Gaze Local	0.37	0.47	0.43	0.54
Face AU	0.38	0.35	0.32	0.39
Face JSD	0.26	0.29	0.32	0.27
Fused classifiers				
Average	0.38	0.5	0.42	0.63
Weighted Average	0.29	0.49	0.46	0.53
Chance	0.25	0.24	0.24	0.24

7. Study 5: Real-Time Mind Wandering Detection in Classrooms (Online model)

7.1 Motivation

Having provided evidence that eye gaze collected in the classroom could be used to detect mind wandering in Studies 2-4, we next considered if these models could be used to detect mind wandering "live" as students are using Guru. This is a critical step to verify that the models can be used to trigger real-time mind wandering interventions, an important goal of this work. In this initial exploration of live detection, we considered only gaze for simplicity and although there was an improvement in accuracy for the gaze+face multimodal model, the improvement was quite small (mind wandering F_1 .50 vs. .47). We also focused on global features because they are more generalizable, more robust to noise than locality features, and yielded better performance when missing data was included (see Table 9), which is an important requirement for a real-time system.

7.2 Methods

7.2.1 Development

We integrated the mind wandering detector from Study 2 into Guru to enable real-time mind wandering detection. In this initial implementation, mind wandering predictions were generated in non-overlapping 30s windows; overlapping windows are considered in Section 7.4. We reconfigured Guru to deliver two types of probes: (1) a probe triggered by the detector (triggered probe) and (2) a pseudorandom probe as used previously (random probe). Triggered probes occurred based upon the mind wandering detector's estimates with a 0.5 classification threshold, whereas random probes occurred as before (every 90-120 seconds). Students would receive only one probe (triggered or random) in a 90 second window. Pseudocode for the probing algorithm is shown in Figure 14.

A lab-based user test with seven college students indicated that the detector triggered only one probe each for six students; the seventh received 12 probes. Due to this imbalance, and to account for detector error, we replaced the binary 0.5 cutoff with a nonlinear probabilistic approach. Specifically, a prediction of less than 0.25 resulted in no probe, a prediction greater than 0.75 always yielded a probe. In between these bounds, probing was probabilistic (e.g., a prediction of 0.45 resulted in a 45% chance of yielding a probe whereas a prediction of 0.65 garnered a 65% chance).

7.2.2 Classroom Testing

Using a similar setup to Study 1 (Section 3.4), we tested Guru with real-time mind wandering detection in two classes. Participants were a new set of 39 high school sophomores, who were enrolled in their first high school biology class with the same teacher and in the same classroom as Study 1. Each student completed two Guru sessions in an 80-minute class period (with the exception of two students who were late to class and therefore only completed one session). Students were randomly assigned to two topics from among the following: Carbohydrate Function, Facilitated Diffusion, Interphase, Protein Function, or Osmosis.

```
//Run every 30 seconds
Function CheckMW(Gaze)
    instance = Generate30SecondInstance(Gaze)
    prediction = Detector(inst)
    LogPrediction(pred)
    T = timeSinceLastProbe()
    if T < 90 Seconds then
        break
    end
    if prediction < 0.5 then
        break
    else
        TriggerProbe()
        ResetTimer()
    end
```

```
//Run every 90-120 Seconds
Function RandomProbing()
    T = timeSinceLastProbe()
    if T < 90 Seconds then
        break
    else
        TriggerProbe()
        ResetTimer()
    end
```

Figure 14. Pseudocode for probing algorithms

7.3 Results

Although the equipment used for the study was unchanged from Studies 1 and 2 (Study 4 continued to use the EyeTribe), software and protocol were adapted to address some of the technical issues that had previously arisen such as computer crashes due to windows updates. As a result of these modifications, our software was successful in recording gaze data for 93.4% of sessions.

There were 351 total probes triggered by the detector; a further 350 probes were triggered pseudo-randomly. The observed mind wandering rate across *all* probe responses was 27%, on par with the rates found in Study 2 (23%).

7.3.1. Live mind wandering detection results

We treated a positive mind wandering response to a *random probe* as a miss, a negative mind wandering response to a *triggered probe* as a false positive. There were six cases in which the detector predicted that a student was mind wandering, however, the constraint of requiring 90 seconds between consecutive probes prevented these from occurring. However, a random probe was triggered within the next 10s (as the time constraint had then expired), and the student answered that they were mind wandering. We counted those cases as correct predictions of mind wandering and are reflected in the confusion matrix shown in Table 10.

Table 10. Confusion matrix for live detector (predicted mind wandering rate 0.50)

Actual	**Predicted**	
	MW	*Not MW*
MW	0.55 (hit)	0.45 (miss)
Not MW	0.48 (false pos.)	0.52 (correct rej.)

We obtained a mind wandering F_1 of 0.40, precision of 0.32, and recall of 0.55 (chance baseline values 0.24, 0.26, 0.22, respectively). Although we outperform chance, there was still a reduction from results obtained in Study 2 (mind wandering $F_1 = 0.59$). This could be attributed to how missing data was handled. In Study 2, cases with insufficient data (< 1 fixation in the window) were excluded, but all cases were considered for the online model. For example, a student may be mind wandering but the detector was unable to generate a prediction if valid gaze data was not recorded; this would be considered to be a miss. We examine how many instances did not contain enough valid eye gaze to generate a prediction. The detector attempted to make a prediction every 30 seconds, but there were 84 instances in which there was not enough data to make a prediction (mean per session = 1.10, SD = 1.62 min = 0, max = 7).

7.3.2. Follow-up analyses

Locality model. We investigated how the locality model would have performed compared to the current global model by using the locality model trained in Study 2 to generate (offline) predictions for the same instances as the global model. The results, shown in Table 11, suggest equitable performance compared to the global models with a mind wandering F_1 of 0.42, precision of 0.37, and recall of 0.53. Again, we observed a drop-off from the results obtained in Study 2 (mind wandering F_1 = 0.59). The two models also agreed on 74% of all instances - the locality model agreed with the global model's mind wandering and not mind wandering predictions for 63% and 82% of the instances, respectively.

Table 11. Confusion matrix for Locality Model (predicted mind wandering rate 0.38)

Actual	**Predicted**	
	MW	*Not MW*
MW	0.45 (hit)	0.54 (miss)
Not MW	0.33 (false pos.)	0.66 (correct rej.)

Overlapping Windows. In this study, we generated a "live" mind wandering prediction every 30 seconds to be consistent with the offline model from Study 2. However, to improve temporal granularity of the predictions, we conducted a post-hoc analysis by generating a mind wandering prediction every one second using a 30-second sliding window and the global gaze features. This is particularly important if the goal is to address mind wandering when it occurs as a 30s delay is likely too excessive.

Visual examination of the predictions across time suggested three representative patterns. Figure 15 (a) shows a representative time series of the sliding window predictions for 19 of the 71 sessions with no triggered probes, whereas Figure 15 (b) represents a session with a high reported mind wandering rate (46%). The difference among the two time series is readily apparent with Figure 15a containing low and flat mind wandering predictions, but Figure 15b shows multiple spikes, indicating MW predicitons.. Finally, Figure 15 (c) represents a session with an average reported mind wandering rate (20%). For this session, the detector was mainly flat like Figure 15a with a few bursts of high mind wandering predictions (like Figure 15b). Though Figure 15c represents a session with average reported MW rate, the timing of predicted MW varies between participants since not all occurred at the end of the session.

Correlations with learning. We computed an average mind wandering rate for each session by averaging the sliding window predictions. Whereas self-reported mind wandering was largely uncorrelated with posttest scores (Spearman's rho = -0.04), the predicted mind wandering rate showed a similar negative correlation (rho = -0.20) as in Study 2 and previous research (D'Mello, in press).

7.4 Discussion

We have demonstrated that offline detectors trained in Study 2 can be used for real-time mind wandering detection albeit with a drop in performance. We partly attribute this performance change to differences in how missing data was handled. In future refinements, we will investigate additional methods to handle missing data in real-time. One such approach is to consider multimodal mind wandering detection when a second modality could substitute when one is missing and vice versa as discussed in Study 4.

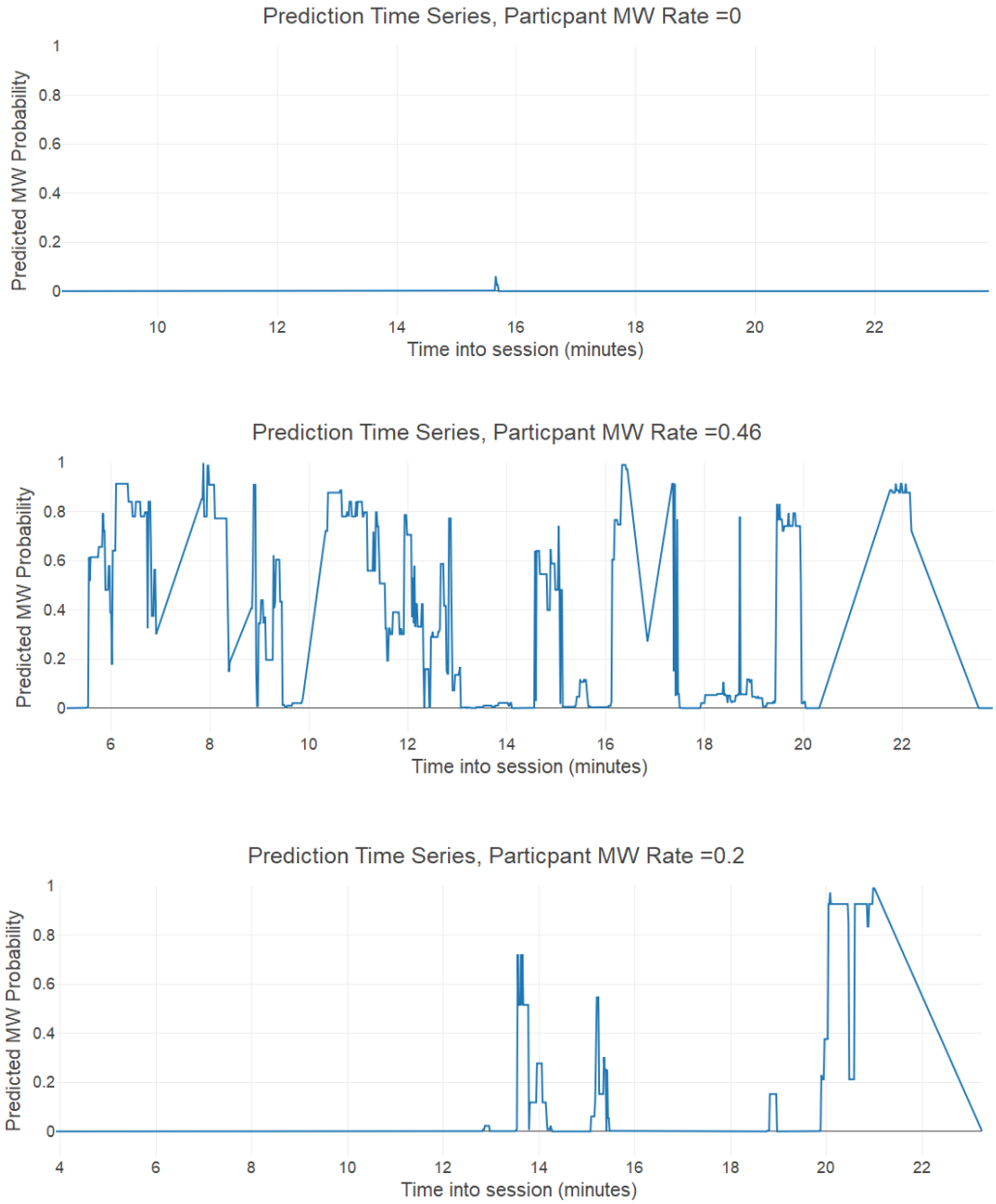

Figure 15. Mind wandering predictions for overlapping 30 second windows

8 General Discussion

It is widely acknowledged that attention is necessary for learning (D'Mello, 2016; Pham & Wang, 2015). An attention-aware learning technology (Olney et al., 2015) which can monitor and react to a student's attentional state could assuage the cost of attentional failures (like mind wandering), thereby improving engagement and learning. However, until very recently, the high cost of eye trackers (which are the most robust method to track visual attention) has relegated eye tracking technology to the confines of the lab. We addressed this issue by studying the feasibility of using COTS eye trackers to obtain valid gaze data in a noisy classroom environment (Studies 1) and to use this data to build automatic mind wandering detectors (Study 2). We then extended this work by comparing models trained on lab vs. classroom data as well as built cross-trained models (Study 3); compared gaze-based models to and face-based models and multimodal combinations of eye gaze and facial features for mind wandering detection in classrooms (Study 4); and investigated real-time mind wandering detection in classrooms (Study 5). Our main findings are summarized below, followed by a discussion of applications, limitations, and future work.

8.1 Main Findings

Despite the fact that the classroom is a complex and noisy environment, we show that is feasible to collect valid eye tracking data with COTS eye trackers (Study 1). Furthermore, we maintained realistic classroom conditions in that students were relatively unconstrained throughout the study. And other than receiving initial guidance as to seating position for eye tracker calibration, they also independently completed the calibration. Despite these lack of constraints, we achieved median gaze validity scores of 75% (both eyes tracked) and 95% (one eye tracked) for the sessions (85%) where gaze data was collected. Although we were only able to collect data for 85% of the sessions in Study 1, this was primarily for reasons beyond our control (e.g., hardware issues with school computers, automatic system updates, failure to calibrate). Success rates improved to 93% in Study 5.

Validity, however, does not imply usefulness. To address this, we built student-independent mind wandering detectors from the eye gaze data and self-reported mind wandering collected in the classroom (Study 2). We achieved moderately accurate mind wandering detection rates despite challenges of class imbalance, noisy gaze data, and unrestricted movements. The F_1 mind wandering score of 0.59 was higher than the previous score of 0.49 achieved in a lab study with GuruTutor (Hutt et al., 2016), though this comparison is tentative due to several differences between the two studies. We also extended this previous work that only investigated global gaze features by exploring locality features as well as a combination of the two. These enhancements did not yield accuracy improvements, not did the inclusion of a simple set of contextual features. One possibility is that the global features are sufficient for this task. However, it is more likely that the locality and contextual features were too simplistic and benefits may be gained by refining them (see Future Work).

We found that mind wandering detectors trained on laboratory data were also transferable to an "in the wild" environment, despite several differences among the two samples (Study 3). The reverse, however, was not true. The implication is that while it might be sufficient to start out in a controlled lab environment, this is no substitute for collecting real-world data (and in our case the difference was quite profound). Combining data from both sources led to equitable performance on each source though.

We explored decision fusion between gaze- and video- based mind wandering detectors (Study 4). One finding was that the gaze-models substantially outperformed the video-based detectors. More importantly, when considering cases where both modalities had valid data, we did not observe any significant improvement of the multimodal model over the best individual models. However, when cases with missing data were included, the combination of the gaze and facial detectors improved overall mind wandering detection. Interestingly we note that the fusion classifiers showed a significant difference to the global gaze model but not to the locality model.

Importantly, we provided evidence for the feasibility of real-time mind wandering detection (Study 5) using the gaze-based model. Although model performance decreased during live detection vs an offline model (mind wandering $F_1 = 0.59$ to 0.40) the detector still outperformed chance (0.24) and the difference can partly be attributable to differences in how missing data was handled (also see above). Specifically, In Study 2, we only considered the 2,334 instances with valid gaze data, achieving a mind wandering F_1 score of 0.59 for the global gaze model (Table 2). When we reduced the number of instances to 1,734 for which there was both face and gaze data, we obtained a reduction in classifier performance ($F_1 = 0.46$ for the global model – Table 8). When we increased the number of instances to the 2,558 with *either* face or gaze data we also noted a decrease in performance ($F_1 = 0.48$ for the global model). Thus, it is important to devise better methods to

address cases with missing data especially since they are unlikely to be missing at random.

Finally, we note that detected mind wandering rates negatively correlated with learning at rates comparable (Study 2) or better than self-reports (Study 5). This provides critical evidence regarding its predictive validity.

8.2 Applications

One important application of this work is to develop an attention-aware version of Guru that detects *and combats* mind wandering in real time in an effort to reengage students' attention toward learning materials. Such a system has a number of paths to pursue. One immediate effect of mind wandering is that students may fail 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 student's attention to an area of the screen that may help them (e.g., "Mary, you might want to look at the image showing the enzyme breakdown…"). A more involved approach might be to ask the student a content specific question (e.g., "Santiago, what happens to an enzyme when it is subjected to heat?") or ask the student to self-explain a concept (e.g., "Kiara, why don't you summarize what you just learned"). Additional measures might be needed if mind wandering persists despite these interventions. One option is to simply change to a new activity (e.g., quitting the lecture and moving to concept mapping). Guru might even suggest changing topics or offering students a choice of what 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 mind wandering detection, which is inherently imperfect. In our view, mind wandering 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's mind wandering likelihood is used to determine whether an intervention is triggered (i.e., if the likelihood of mind wandering is 70%, then there is a 70% chance of an intervention), similar to the approach used in Study 5 for real-time mind wandering detection. The interventions should also be designed to be "fail-soft" in that there are no harmful effects if delivered incorrectly and the examples above were designed with this principle in mind.

Beyond immediate intervention, the mind wandering detector could also passively monitor mind wandering rates, tagging content with high mind wandering rates for retesting or restudy or providing reports to teachers and instructional designers on which sections or activities were associated with high mind wandering rates (as considerations for redesign). It could also be used as a feedback tool for students, for example, for use as an objective measure of attention for those interested in improving attentional focus via mindfulness training (Zoogman, Goldberg, Hoyt, & Miller, 2015), meditation, or some other strategy. Similarly, it can be used as tool to promote metacognitive reflection, where students monitor their own attention levels to identify periods where there are most attentive.

COTS eye tracking in the classroom opens doors to several additional applications beyond mind wandering detection and responding. One involves monitoring attentional states beyond mind wandering (e.g., focused attention, alternating attention) to ensure that limited attentional resources are being optimally deployed (D'Mello, 2016). Another includes 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 and sustains students' attention. Indeed, there are numerous potential applications afforded by scalable tracking of eye gaze in real-world environments.

8.3 Limitations & Future Work

There were several limitations of this work. Our system was designed to include a low-cost eye tracker so that it may scale to large numbers of students. However, COTS eye trackers have a lower sampling-rate and are less accurate compared to research-grade eye trackers. Regrettably, the specific eye tracker we used is no-longer available after the company was acquired, but alternatives are available (Tobii 4C; GazePoint). There is also some cost associated with these alternatives, so deployment in underfunded schools is not a guarantee, thereby creating equity issues. Eye tracking with inexpensive web-cams are a promising alternative to consider (Papoutsaki et al., 2016; Sewell & Komogortsev, 2010).

In a related vein, the lab- and school- studies used different eye trackers (due to shipping delays) and different participants (due to convenience), which limits what can be concluded from a direct comparison of models built from the two data sets. That said, the fact that differences among the two were asymmetric (lab models generalized to school but not vice versa), suggests that other factors beyond type of tracker and sample are at play. Future work is needed to resolve the difference

among the two approaches.

With regard to mind wandering 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 patterns in eye gaze. For example, if a student had multiple fixations in one area, were these concentrated or distributed across time? In addition, we only considered a small number of contextual features and our locality features were quite primitive. Future work should explore a more refined set of locality features for mind wandering detection – for example, AOI (area of interest) features that capture 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 can analyze image-specific gaze fixations, such as proportion of fixations on images, number of image components fixated on, and fixation durations on different components (e.g., objects, labels, and arrows). Guru uses a slow-reveal animation, where image components slowly appear as they are being referenced throughout the session. This affords computing of animation-based locality features that measure gaze latencies to different image components as they are slowly revealed. A further possibility would be to explore temporal gaze features, for example how the fixations and saccades interact with the displayed content over time.

Furthermore, because we extracted facial features in real-time and discarded the video frames to protect privacy, the features were limited in scope (e.g., facial textures could not be extracted). A richer set of facial features (Bosch et al., 2015) may make this channel more competitive with eye-gaze. Future work should also extract a richer se of contextual features from Guru interaction data to augment the enhanced set of face and gaze features and integrate them using more advanced multimodal fusion models (D'Mello, Bosch, & Chen, in press) than the simplistic decision-fusion approaches considered here.

A further limitation relates to the use of thought probes, which require users to be mindful of their mind wandering and respond honestly. Although this methodology has been previously validated (Franklin et al., 2013; Randall et al., 2014; Reichle et al., 2010) it is still limited due to the reliance on self-reports. Unfortunately, there is no clear alternative to track a highly internal state like mind wandering outside of measuring brain activity in an fMRI scanner, which is also limited in many respects. One futuristic possibility is to combine self-reports and wearable electroencephalography (EEG) as a means of collecting more accurate mind wandering responses, but it is unclear if this can be done in the wild – though one previous study shows some promise (Girn et al., 2017) .

In addition to the aforementioned improvement to mind wandering detection, it is important to close the loop by developing mind wandering interventions. This presents two challenges to explore: what kind of intervention should be delivered and given the imperfect nature of mind wandering detection, what mechanism should be in place for triggering said intervention? We are in the process of addressing these challenges akin to the strategies discussed above (Section 7.1). Having completed initial design activities with teachers and students, we have implemented the first attention-aware version of Guru. Upon completing multiple rounds of iterative refinement, we will summatively test the technology by randomly assigning students to the attention-aware Guru with the mind wandering interventions enabled (experimental group), disabled (business as usual control group), or with interventions triggered based on historic mind wandering distributions rather than a student's current mind wandering likelihoods (active control group). Whether the attention-aware Guru increases engagement and learning compared to the controls awaits future technology and development.

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

References

Ahmidi, N., Hager, G. D., Ishii, L., Fichtinger, G., Gallia, G. L., & Ishii, M. (2010). Surgical Task and Skill Classification from Eye Tracking and Tool Motion in Minimally Invasive Surgery. In T. Jiang, N. Navab, J. P. W. Pluim, & M. A. Viergever (Eds.), *13th International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 295–302). Beijing, China: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-15711-0_37

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

Allen, I. E., & Seaman, J. (2016). Online Report Card: Tracking Online Education in the United States. *Babson Survey Research Group*.

Baird, B., Smallwood, J., Lutz, A., & Schooler, J. W. (2014). The decoupled mind: mind-wandering disrupts cortical phase-locking to perceptual events. *Journal of Cognitive Neuroscience*, *26*(11), 2596–2607. https://doi.org/10.1162/jocn_a_00656

Baltrusaitis, T., Robinson, P., & Morency, L. P. (2016). OpenFace: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016* (pp. 1–10). https://doi.org/10.1109/WACV.2016.7477553

Barron, E., Riby, L. M., Greer, J., & Smallwood, J. (2011). Absorbed in Thought: The Effect of Mind Wandering on the Processing of Relevant and Irrelevant Events. *Psychological Science*, *22*(5), 596–601. https://doi.org/10.1177/0956797611404083

Bates, A. T. (2005). *Technology, e-learning and distance education*. Routledge.

Berliner, D. C. (1990). What's all the fuss about instructional time? In *The Nature of Time in School. Theoretical Concepts, Practitioners Perceptions* (pp. 3–35). Teachers College Press. https://doi.org/10.1108/EUM0000000002517

Bixler, R., & D'Mello, S. K. (2014). Toward Fully Automated Person-Independent Detection of Mind Wandering. In V. Dimitrova, T. Kuflik, D. Chin, F. Ricci, P. Dolog, & G.-J. Houben (Eds.), *User Modeling Adaptation and Personalization* (pp. 37–48). Aalborg, Denmark, Denmark: Springer. https://doi.org/10.1007/978-3-319-08786-3_4

Bixler, R., & 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), 33–68. https://doi.org/10.1007/s11257-015-9167-1

Blanchard, N., Bixler, R., Joyce, T., & D'Mello, S. K. (2014). Automated Physiological Based Detection of Mind Wandering during Learning. In S. Trausan-Matu, K. Boyer, M. Crosby, & K. Panourgia (Eds.), *Intelligent Tutoring Systems* (Vol. 8474, pp. 55–60). Switzerland: Springer International Publishing.

Bosch, N., & D'Mello, S. K. (in review). Detecting Mind Wandering from Video in the Lab and in the Classroom, *IEEE Transactions on Affective Computing*.

Bosch, N., D'Mello, S. K., Baker, R. S. J. d., Ocumpaugh, J., Shute, V., Ventura, M., … Zhao, W. (2015). Automatic Detection of Learning-Centered Affective States in the Wild. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (pp. 379–388). New York, NY, USA: ACM. https://doi.org/10.1145/2678025.2701397

Buswell, G. T. (1936). How People Look at Pictures. *Psychological Bulletin*, *33*(2), 142–143. https://doi.org/10.1037/h0053409

Buswell, G. T. (1937). How adults read. *Supplementary Educational Monographs*, *45*, 158. https://doi.org/10.1086/614288

Campbell, F. W., & Wurtz, R. H. (1978). Saccadic omission: Why we do not see a grey-out during a saccadic eye movement. *Vision Research*, *18*(10), 1297–1303. https://doi.org/10.1016/0042-6989(78)90219-5

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*(1), 321–357. https://doi.org/10.1613/jair.953

Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences*. Taylor & Francis.

Csifcsák, G., & Mittner, M. (2017). Linking brain networks and behavioral variability to different types of mind-wandering. *Proceedings of the National Academy of Sciences*, *114*(30), E6031–E6032. https://doi.org/10.1073/pnas.1705108114

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

D'Mello, S. K. (2018). What do we Think About When we Learn? In K. K. Mills, D. Long, J. Magliano, & K. Wierner (Eds.), *Deep Comprehension* (pp. 52–67). Routledge.

D'Mello, S. K., Bosch, N., & Chen, H. (2018). Multimodal-Multisensor Affect Detection. In S. Oviatt, P. Cohen, & A. Krueger (Eds.), *The Handbook of Multimodal-Multisensor Interfaces.* (pp. 167–202). ACM Books/Morgan Claypool. https://doi.org/10.1145/3107990.3107998

D'Mello, S. K., Hays, P., Williams, C., Cade, W., Brown, J., & Olney, A. (2010). Collaborative Lecturing by Human and Computer Tutors. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Intelligent Tutoring Systems* (pp. 178–187). Berlin, Heidelberg: Springer Berlin Heidelberg.

https://doi.org/10.1007/978-3-642-13437-1_18
Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, *36*(12), 1827–1837. https://doi.org/10.1016/0042-6989(95)00294-4
Dodge, R. (1900). Visual perception during eye movement. *Psychological Review*, *7*(5), 454–465. https://doi.org/10.1037/h0067215
Drummond, J., & Litman, D. (2010). In the Zone: Towards Detecting Student Zoning Out Using Supervised Machine Learning. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Intelligent Tutoring Systems* (pp. 306–308). Pittsburgh, PA, USA: Springer. https://doi.org/10.1007/978-3-642-13437-1_53
Ekman, P., & Friesen, W. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto: Consulting Psychologists Press.
Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*(3), 215–251. https://doi.org/10.1037/0033-295X.87.3.215
Faber, M., Bixler, R., & D'Mello, S. K. (2017). An automated behavioral measure of mind wandering during computerized reading. *Behavior Research Methods*, *50*(1), 134–150. https://doi.org/10.3758/s13428-017-0857-y
Faber, M., & D'Mello, S. K. (in press). How the stimulus influences mind wandering in semantically-rich task contexts. *Cognitive Research: Principles and Implications*.
Feng, S., D'Mello, S. K., & Graesser, A. C. (2013). Mind wandering while reading easy and difficult texts. *Psychonomic Bulletin & Review*, *20*(3), 586–592. https://doi.org/10.3758/s13423-012-0367-y
Forbes-Riley, K., & Litman, D. (2011). When does disengagement correlate with learning in spoken dialog computer tutoring? In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Artificial Intelligence in Education* (pp. 81–89). Auckland, New Zealand,: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-21869-9_13
Franklin, M. S., Broadway, J. M., Mrazek, M. D., Smallwood, J., & Schooler, J. W. (2013). Window to the wandering mind: pupillometry of spontaneous thought while reading. *The Quarterly Journal of Experimental Psychology*, *66*(12), 2289–2294. https://doi.org/10.1080/17470218.2013.858170
Franklin, M. S., Smallwood, J., & Schooler, J. W. (2011). Catching the mind in flight: using behavioral indices to detect mindless reading in real time. *Psychonomic Bulletin & Review*, *18*(5), 992–997. https://doi.org/10.3758/s13423-011-0109-6
Gawne, T. J., & Martin, J. M. (2000). Activity of primate V1 cortical neurons during blinks. *Journal of Neurophysiology*, *84*(5), 2691–2694. https://doi.org/10.1152/jn.2000.84.5.2691
Giambra, L. M. (1995). A laboratory method for investigating influences on switching attention to task-unrelated imagery and thought. *Conscious Cogn*, *4*(1), 1–21. https://doi.org/10.1006/ccog.1995.1001
Girn, M., Mills, C., Laycock, E., Ellamil, M., Ward, L., & Christoff, K. (2017). Neural Dynamics of Spontaneous Thought: An Electroencephalographic Study. In D. D. Schmorrow & C. M. Fidopiastis (Eds.), *Augmented Cognition. Neurocognition and Machine Learning: 11th International Conference*. https://doi.org/10.1007/978-3-319-58628-1_3
Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations*, *11*(1), 10–18. https://doi.org/10.1145/1656274.1656278
Hild, J., Kühnle, C., & Beyerer, J. (2016). Gaze-based Moving Target Acquisition in Real-time Full Motion Video. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications* (pp. 241–244). New York, NY, USA: ACM. https://doi.org/10.1145/2857491.2857525
Hoffman, J. E., & Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception & Psychophysics*, *57*(6), 787–795. https://doi.org/10.3758/BF03206794
Huey, E. B. (1898). Preliminary Experiments in the Physiology and Psychology of Reading. *The American Journal of Psychology*, *9*(4), 575. https://doi.org/10.2307/1412192
Huey, E. B. (1908). *The Psychology and Pedagogy of Reading: With a Review of the History of Reading and Writing and of Methods, Texts, and Hygiene in Reading*. The Macmillan company.
Hutt, S., Hardey, J., Bixler, R., Stewart, A., Risko, E., & D'Mello, S. K. (2017). Gaze-based Detection of Mind Wandering during Lecture Viewing. In H. Xiangen, B. Tiffany, H. Arnon, & L. Paquette (Eds.), *Proceedings of the 10th International Conference on Educational Data*

Mining (EDM 2017) (pp. 226–231).

Hutt, S., Mills, C., Bosch, N., Krasich, K., Brockmole, J., & D'Mello, S. K. (2017). "Out of the Fr-Eye-ing Pan": Towards Gaze-Based Models of Attention During Learning with Technology in the Classroom. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization* (pp. 94–103). New York, NY, USA: ACM. https://doi.org/10.1145/3079628.3079669

Hutt, S., Mills, C., White, S., Donnelly, P. J., & D'Mello, S. K. (2016). The eyes have it: gaze-based detection of mind wandering during learning with an intelligent tutoring system. In T. Barnes, M. Chi, & M. Feng (Eds.), *The 9th International Conference on Educational Data Mining* (pp. 86–93). Raleigh, NC, USA.

Irwin, D. E., & Carlson-Radvansky, L. A. (1996). Cognitive suppression during saccadic eye movements. *Psychological Science*, *7*(2), 83–88. https://doi.org/10.1111/j.1467-9280.1996.tb00334.x

Javel, É. (1878). Essai sur la physiologie de la lecture. *Annales d'Oculistique*, *80*, 61–73.

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

Kam, J. W. Y., Dao, E., Farley, J., Fitzpatrick, K., Smallwood, J., Schooler, J. W., & Handy, T. C. (2011). Slow Fluctuations in Attentional Control of Sensory Cortex. *Journal of Cognitive Neuroscience*, *23*(2), 460–470. https://doi.org/10.1162/jocn.2010.21443

Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., & Torralba, A. (2016). Eye tracking for Everyone. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2176–2184). Las Vegas, NV, USA. https://doi.org/10.1109/CVPR.2016.239

Krasich, K., McManus, R., Hutt, S., Faber, M., D'Mello, S. K., & Brockmole, J. (2018). Gaze-Based Signatures of Mind Wandering During Real-World Scene Processing. *Journal of Experimental Psychology: General*, *147*(8), 1111. https://doi.org/10.1037/xge0000411

Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. https://doi.org/10.1214/aoms/1177729694

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, *37*(1), 145–151. https://doi.org/10.1109/18.61115

Loboda, T. D. (2014). *Study and Detection of Mindless Reading*. Retrieved from http://d-scholarship.pitt.edu/21734/

Matin, E. (1974). Saccadic suppression: A review and an analysis. *Psychological Bulletin*, *81*(12), 899–917. https://doi.org/10.1037/h0037368

Maurer, B., Krischkowsky, A., & Tscheligi, M. (2017). Exploring Gaze and Hand Gestures for Non-Verbal In-Game Communication. In *Extended Abstracts Publication of the Annual Symposium on Computer-Human Interaction in Play* (pp. 315–322). ACM. https://doi.org/10.1145/3130859.3131296

McVay, J. C., & Kane, M. J. (2009). Conducting the Train of Thought: Working Memory Capacity, Goal Neglect, and Mind Wandering in an Executive-Control Task. *Journal of Experimental Psychology: Learning Memory and Cognition*, *35*(1), 196–204. https://doi.org/10.1037/a0014104

McVay, J. C., & Kane, M. J. (2010). Does mind wandering reflect executive function or executive failure? Comment on Smallwood and Schooler (2006) and Watkins (2008). *Psychological Bulletin*, *136*(2), 188–197. https://doi.org/10.1037/a0018298

McVay, J. C., & Kane, M. J. (2012). Drifting from slow to "D'oh!": Working memory capacity and mind wandering predict extreme reaction times and executive control errors. *Journal of Experimental Psychology: Learning Memory and Cognition*, *38*(3), 525–549. https://doi.org/10.1037/a0025896

McVay, J. C., Kane, M. J., & Kwapil, T. R. (2009). Tracking the train of thought from the laboratory into everyday life: an experience-sampling study of mind wandering across controlled and ecological contexts. *Psychonomic Bulletin and Review*, *16*(5), 857–863. https://doi.org/10.3758/PBR.16.5.857

Messinger, D., Fogel, A., & Dickson, K. L. (2001). All Smiles Are Positive, but Some Smiles Are More Positive Than Others. *Developmental Psychology*, *37*(5), 642–653. https://doi.org/10.1037/0012-1649.37.5.642

Mills, C., Bixler, R., Wang, X., & D'Mello, S. K. (2016). Automatic gaze-based detection of mind wandering during film viewing. In T. Barnes, M. Chi, & M. Feng (Eds.), *The 9th International Conference on Educational Data Mining.* (pp. 30–37). Raleigh, North Carolina.

Mills, C., & D'Mello, S. K. (2015). Toward a Real-time (Day) Dreamcatcher: Sensor-Free Detection of Mind Wandering During Online Reading. In O. C. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, … M. Desmarais (Eds.), *Proceedings of the 8th International Conference on Educational Data Mining.* (pp. 69–76).

Mills, C., D'Mello, S. K., Bosch, N., & Olney, A. M. (2015). Mind Wandering During Learning with an Intelligent Tutoring System. In C. Conati, N. Heffernan, A. Mitrovic, & F. M. Verdejo (Eds.), *Artificial Intelligence in Education* (pp. 267–276). Madrid, Spain, Spain: Springer International Publishing. https://doi.org/10.1007/978-3-319-19773-9_27

Mittner, M., Boekel, W., Tucker, A. M., Turner, B. M., Heathcote, A., & Forstmann, B. U. (2014). When the Brain Takes a Break: A Model-Based Analysis of Mind Wandering. *The Journal of Neuroscience*, *34*(49), 16286–16295. https://doi.org/10.1523/JNEUROSCI.2062-14.2014

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

Navarro, D., & Sundstedt, V. (2017). Simplifying game mechanics: gaze as an implicit interaction method. In *SIGGRAPH Asia 2017 Technical Briefs* (p. 4). ACM. https://doi.org/10.1145/3145749.3149446

Olney, A. M., D'Mello, S. K., Person, N., Cade, W., Hays, P., Williams, C., … Graesser, A. (2012). Guru: A Computer Tutor That Models Expert Human Tutors. In S. Cerri, W. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Intelligent Tutoring Systems* (pp. 256–261). Chania, Crete, Greece: Springer. https://doi.org/10.1007/978-3-642-30950-2_32

Olney, A. M., Risko, E. F., D'Mello, S. K., & Graesser, A. C. (2015). Attention in Educational Contexts: The role of the learning Task in Guiding Attention. In J. Fawcett, E. F. Risko, & A. Kingstone (Eds.), *The Handbook of Attention*. MIT Press.

Pan, S. J., & Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). WebGazer : Scalable Webcam Eye Tracking Using User Interactions. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (pp. 3839–3845).

Person, N. K., Olney, A., D'Mello, S. K., & Lehman, B. (2012). Interactive Concept Maps and Learning Outcomes in Guru. In *Florida Association for Institutional Research Conference* (pp. 456–461). Marco Island, FL, USA.

Pham, P., & Wang, J. (2015). AttentiveLearner: improving mobile MOOC learning via implicit heart rate tracking. In C. Conati, N. Heffernan, A. Mitrovic, & M. F. Verdejo (Eds.), *Artificial Intelligence in Education* (pp. 367–376). Madrid, Spain: Springer International Publishing. https://doi.org/10.1007/978-3-319-19773-9_37

Randall, J. G., Oswald, F. L., & Beier, M. E. (2014). Mind-wandering, cognition, and performance: a theory-driven meta-analysis of attention regulation. *Psychological Bulletin*, *140*(6), 1411–1431. https://doi.org/10.1037/a0037428

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422. https://doi.org/10.1037/0033-2909.124.3.372

Reichle, E. D., Reineberg, A. E., & Schooler, J. W. (2010). Eye movements during mindless reading. *Psychological Science*, *21*(9), 1300–1310. https://doi.org/10.1177/0956797610378686

Risko, E. F., Anderson, N., Sarwal, A., Engelhardt, M., & Kingstone, A. (2012). Everyday attention: Variation in mind wandering and memory in a lecture. *Applied Cognitive Psychology*, *26*(2), 234–242. https://doi.org/10.1002/acp.1814

Risko, E. F., Buchanan, D., Medimorec, S., & Kingstone, A. (2013). Everyday attention: Mind wandering and computer use during lectures. *Computers & Education*, *68*, 275–283. https://doi.org/10.1016/j.compedu.2013.05.001

Robertson, I. H., Manly, T., Andrade, J., Baddeley, B. T., & Yiend, J. (1997). "Oops!": performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. *Neuropsychologia*, *35*(6), 747–758. https://doi.org/10.1016/S0028-3932(97)00015-8

Schooler, J. W., Reichle, E. D., & Halpern, D. V. (2004). Zoning Out while Reading: Evidence for Dissociations between Experience and Metaconsciousness. In *Thinking and seeing: Visual metacognition in adults and children* (pp. 203–226). Cambridge, MA, US, MA, US: MIT Press.

Schooler, J. W., Smallwood, J., Christoff, K., Handy, T. C., Reichle, E. D., & Sayette, M. A. (2011). Meta-awareness, perceptual decoupling and the wandering mind. *Trends in Cognitive Sciences*, *15*(7), 319–326. https://doi.org/10.1016/j.tics.2011.05.006

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

Seli, P., Risko, E. F., & Smilek, D. (2016). On the Necessity of Distinguishing Between Unintentional and Intentional Mind Wandering. *Psychological Science*, *27*(5), 685–691. https://doi.org/10.1177/0956797616634068

Sewell, W., & Komogortsev, O. (2010). Real-time eye gaze tracking with an unmodified commodity webcam employing a neural network. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems - CHI EA '10* (p. 3739). https://doi.org/10.1145/1753846.1754048

Shernoff, D. J., Csikszentmihalyi, M., Schneider, B., & Shernoff, E. S. (2003). Student engagement in high school classrooms from the perspective of flow theory. *School Psychology Quarterly*, *18*(2), 158–176. https://doi.org/10.1521/scpq.18.2.158.21860

Smallwood, J., Beach, E., Schooler, J. W., & Handy, T. C. (2008). Going AWOL in the brain: mind wandering reduces cortical analysis of external events. *Journal of Cognitive Neuroscience*, *20*(3), 458–469. https://doi.org/10.1162/jocn.2008.20037

Smallwood, J., Fishman, D. J., & Schooler, J. W. (2007). Counting the cost of an absent mind: mind wandering as an underrecognized influence on educational performance. *Psychological Bulletin & Review*, *14*(2), 230–236. https://doi.org/10.3758/BF03194057

Smallwood, J., McSpadden, M., & Schooler, J. W. (2008). When attention matters: the curious incident of the wandering mind. *Memory & Cognition*, *36*(6), 1144–1150. https://doi.org/10.3758/MC.36.6.1144

Smallwood, J., & Schooler, J. W. (2006). The restless mind. *Psychological Bulletin*, *132*(6), 946–958. https://doi.org/10.1037/0033-2909.132.6.946

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

Smilek, D., Carriere, J. S. A., & Cheyne, J. A. (2010). Out of mind, out of sight: Eye blinking as indicator and embodiment of mind wandering. *Psychological Science*. https://doi.org/10.1177/0956797610368063

Sottilare, R. A., Graesser, A., Hu, X., & Holden, H. (2013). *Design recommendations for intelligent tutoring systems - Volume 1:Learner modeling* (Vol. 1). US Army Research Laboratory.

Stawarczyk, D., Majerus, S., Maj, M., Van der Linden, M., & D'Argembeau, A. (2011). Mind-wandering: Phenomenology and function as assessed with a novel experience sampling method. *Acta Psychologica*, *136*(3), 370–381. https://doi.org/10.1016/j.actpsy.2011.01.002

Stewart, A., Bosch, N., Chen, H., Donnelly, P., & D'Mello, S. K. (2017). Face Forward: Detecting Mind Wandering from Video During Narrative Film Comprehension. In E. André, R. Baker, X. Hu, M. M. T. Rodrigo, & B. du Boulay (Eds.), *Artificial Intelligence in Education* (pp. 359–370). Springer International Publishing. https://doi.org/10.1007/978-3-319-61425-0_30

Stewart, A., Bosch, N., & D'Mello, S. K. (2017). Generalizability of Face-Based Mind Wandering Detection Across Task Contexts. In *Proceedings of the 10th International Conference on Educational Data Mining (EDM 2017)* (pp. 88–95). Wuhan, China.

Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences*, *110*(16), 6313–6317. https://doi.org/10.1073/pnas.1221764110

Szpunar, K. K., Moulton, S. T., & Schacter, D. L. (2013). Mind wandering and education: from the classroom to online learning. *Frontiers in Psychology*, *4*, 495. https://doi.org/10.3389/fpsyg.2013.00495

Twigg, C. A. (2003). Models for online learning. *Educause Review*, *38*, 28–38.

Uzzaman, S., & Joordens, S. (2011). The eyes know what you are thinking: eye movements as an objective measure of mind wandering. *Consciousness and Cognition*, *20*(4), 1882–1886. https://doi.org/10.1016/j.concog.2011.09.010

Vosskuhler, A., Nordmeier, V., Kuchinke, L., & Jacobs, A. M. (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), 1150–1162. https://doi.org/10.3758/BRM.40.4.1150

Weibel, N., Fouse, A., Emmenegger, C., Kimmich, S., & Hutchins, E. (2012). Let's look at the cockpit: exploring mobile eye-tracking for observational research on the flight deck. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (pp. 107–114). ACM. https://doi.org/10.1145/2168556.2168573

Yarbus, A. L. (1967). *Eye Movements and Vision*. New York.

Yonetani, R., Kawashima, H., & Matsuyama, T. (2012). Multi-mode Saliency Dynamics Model for Analyzing Gaze and Attention. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (pp. 115–122). New York, NY, USA: ACM. https://doi.org/10.1145/2168556.2168574

Zhang, Y., Chong, M. K., Müller, J., Bulling, A., & Gellersen, H. (2015). Eye tracking for public displays in the wild. *Personal and Ubiquitous Computing*, *19*(5), 967–981. https://doi.org/10.1007/s00779-015-0866-8

Zoogman, S., Goldberg, S. B., Hoyt, W. T., & Miller, L. (2015). Mindfulness Interventions with Youth: A Meta-Analysis. *Mindfulness*, *6*(2), 290–302. https://doi.org/10.1007/s12671-013-0260-4

Zuber, B. L., & Stark, L. (1966). Saccadic suppression: Elevation of visual threshold associated with saccadic eye movements. *Experimental Neurology*, *16*(1), 65–79. https://doi.org/10.1016/0014-4886(66)90087-2