Automatic Detection of Mind Wandering from Video in the Lab and in the Classroom

Nigel Bosch and Sidney K. D'Mello

Abstract—We report two studies that used facial features to automatically detect mind wandering, a ubiquitous phenomenon whereby attention drifts from the current task to unrelated thoughts. In a laboratory study, university students (N = 152) read a scientific text, whereas in a classroom study high school students (N = 135) learned biology from an intelligent tutoring system. Mind wandering was measured using validated self-report methods. In the lab, we recorded face videos and analyzed these at six levels of granularity: (1) upper-body movement; (2) head pose; (3) facial textures; (4) facial action units (AUs); (5) co-occurring AUs; and (6) temporal dynamics of AUs. Due to privacy constraints, videos were not recorded in the classroom. Instead, we extracted head pose, AUs, and AU co-occurrences in real-time. Machine learning models, consisting of support vector machines (SVM) and deep neural networks, achieved F_1 scores of .478 and .414 (25.4% and 20.9% above-chance improvements, both with SVMs) for detecting mind wandering in the lab and classroom, respectively. The lab-based detectors achieved 8.4% improvement over the previous state-of-the-art; no comparison is available for classroom detectors. We discuss how the detectors can integrate into intelligent interfaces to increase engagement and learning by responding to wandering minds.

Index Terms—Affective computing, computer vision, educational technology, human–computer interaction

1 Introduction

Most of us can recall a time when we realized our attention had drifted away from thinking about what we were trying to do towards something completely unrelated. For example, we might be reading a book or news article and suddenly realize that we have no idea what we were reading. Or we might find ourselves attending a lecture but have no recollection of what the speaker just said. Such lapses in attention, known as mind wandering [1], are ubiquitous experiences. For example, one large-scale study that used experience sampling to track mind wandering of 5,000 people in 86 countries found that it occurred 46.9% of the time during day-to-day life [2]. Mind wandering is not merely incidental; recent meta-analyses have confirmed that it is negatively related to performance across a variety of tasks [3], [4]. Here, our goal is to develop automated methods to detect mind wandering to support a variety of applications aimed at improving task performance.

1.1 What is Mind Wandering?

At its core, mind wandering is an attentional shift away from the processing of task-related information to the processing of task-irrelevant thoughts or ideas [1], [5]–[12]. By task-related we mean thoughts that support the primary task objective. For example, during reading, inferences or memory retrievals that go beyond the textual content would not be considered mind wandering as long as they are related to the content, whereas reflecting on how boring the text is would. These shifts in the locus of attention usually occur without intention or even awareness [1], [9] but people can also intentionally go off task [13]. Mind wandering is related to, but not the same as, boredom [14] and aligns with the attentional subcomponent of the cognitive component of tripartite (affective, cognitive, and behavioral [15]) models of engagement [16].

There are multiple hypotheses regarding the cognitive mechanisms underlying mind wandering (reviewed in [17]). According to the *executive-resource hypothesis* [10], when a task does not sufficiently consume all of one's attentional resources, unused resources are directed to task-unrelated thoughts, leading the mind to wander. In contrast, the *control-failure hypothesis* posits that mind wandering occurs when executive control fails to suppress task-unrelated thoughts [11], [18]. Despite these differences, the basic idea is that both task-related and task-unrelated thoughts compete for consciousness, a limited resource, and mind wandering occurs when task-unrelated thoughts win the competition of consciousness [19].

There are many antecedents of mind wandering (e.g., current concerns and prospective thoughts, aspects of the task stimulus, environmental distractions, introspection, semantic and autobiographical memory retrievals; see [20], [21]). It is also more likely to occur when a person is in a negative mood [22], [23] and among those diagnosed with dysphoria (depression) [24] or attention-deficit/hyperactivity disorder [25]. Importantly, in semantically-rich tasks contexts, like reading, the stimulus itself is often a source of mind wandering due to the automaticity of memory associations (see [21]).

1.2 Current Study

We explore video-based detection of mind wandering as a step towards intelligent technologies that sense and re-

- *Nigel Bosch is with the School of Information Sciences and the Department of Educational Psychology, University of Illinois at Urbana-Champaign, Champaign, IL 61820. E-mail: pnb@illinois.edu*
- *Sidney K. D'Mello is with the Institute of Cognitive Science and the Department of Computer Science, University of Colorado Boulder, Boulder, CO 80309. E-mail: sidney.dmello@colorado.edu*

spond to users' mental states. We focus on mind wandering detection during learning with technology, due its high incidence and negative consequences in this context. In particular, mind wandering is frequent during routine learning activities like computerized reading and video lecture viewing [10], [26], occurring between 20% to 40% of the time [4]. And although mind wandering does have some benefits [27], such as the association between trait day dreaming and creative problem solving [28], mind wandering *during* learning is consistently negatively related to learning outcomes (e.g., [5], [26] and recent meta-analysis in [4]).

There is considerable potential for intelligent learning environments (e.g., intelligent tutoring systems, e-textbooks, massive open online courses or MOOCs) to improve engagement and learning by automatically detecting and adapting the learning environment when minds wander [29]–[32]. For example, the technology might ask the student to take a short quiz when mind wandering is detected [33], encourage re-reading [34], change topics, or even suggest taking a break. Alternatively, instructors and instructional designers might be given feedback about incidence of students' mind wandering to identify course materials that could be made more engaging. Such strategies necessitate methods for automatic mind wandering detection, which is the focus of this paper.

1.3. Related Work

There has been considerable work on automated engagement detection in general [35]–[45] including research in the context of learning environments (see recent review [16]). However, these previous studies are different from mind wandering detection in that they either conceptualize engagement as a holistic construct (e.g., [35], [43], [45]), or focus on different aspects of engagement, such as behavioral engagement (e.g., going off-task) [38], [40] or affective states such as interest [46], curiosity [47], or boredom [48], [49]. In contrast, mind wandering is most closely related to attentional disengagement, which is related to boredom [14]. Further, whereas most previous work focuses on overt *appearances* of disengagement [35], [37], mind wandering reflects a covert state of inattention [31], making it particularly challenging to detect.

To illustrate, the left column in Fig. 1 depicts examples of facial expressions preceding mind wandering reports (i.e., when people catch themselves mind wandering – see Section 2.1) whereas the right columns depicts cases where people did not report mind wandering. The person on the bottom-left has her eyes closed and subsequently reported mind wandering, whereas the person on the bottom-right appears to be bored or uninterested due to a prolonged yawn but did not report mind wandering. Consider the top row – here the people may appear to be engaged in both cases (subjectively speaking), but reported mind wandering for only the example on the left. Similarly, the middle row depicts two people who appear to be intently focused with eyes directed toward the screen. Here, the person on the left reported mind wandering while the one on the right did not. As these examples illustrate, facial indicators of mind wandering are nonobvious (to the extent that we can rely on self-reports as discussed in Section 2.1). Thus, research has mainly focused on alternate modalities for mind wandering detection, as reviewed below.

Fig. 1. Examples of facial expressions for positive (left column) and negative mind wandering (right column) cases.

Before proceeding, it is useful to point out that in all the studies reviewed below, and indeed in almost all of research in this area, mind wandering is measured via self-reports, either using thought probes (e.g., "were you zoned out [or attending] at the time of the probe?") or relying on self-caught instances of mind wandering (e.g., "press the Z key every time you catch yourself zoning out") – see Section 2.1 for a methodological discussion.

Mind Wandering Detection from Eye Gaze. In a pioneering reading study, Smilek et al. [50] found that participants blinked more frequently and fixated (focused their eyes on one spot) less frequently while mind wandering compared to normal reading. Thus, tracking the location of eye gaze, ostensibly in tandem with the words being read, should be diagnostic of mind wandering [51]. Consequently, machine learning methods applied to eye gaze data have proven effective for automatic mind wandering detection, achieving above-chance accuracies ranging from 28% to 45% [51]–[55] when validated in a person-independent manner. Unfortunately, these studies were conducted in laboratory contexts and utilized high-quality gaze trackers that cost tens of thousands of dollars, raising substantial scalability concerns.

Hutt et al. [56] addressed this problem by using consumer-off-the-shelf (COTS) gaze tracking in real-world classrooms, achieving 46% above-chance improvements.

However, they used a $99 USD[1] eye tracker called the EyeTribe which is no longer available after the company was acquired. A similarly priced gaze tracker, the Tobii 4C, requires an additional license ($2000 USD when we last enquired) for research usage. Furthermore, most schools have very limited budgets, and even purchasing such relatively inexpensive hardware is untenable at a large scale.

Mind Wandering Detection from Physiology. Physiological features (e.g., heart rate, skin conductance) have also been the basis of some mind wandering detection research [57], [58]. Blanchard et al. [57] utilized the Affectiva Q wrist-mounted sensor to measure physiology during reading, achieving a 22% above-chance mind wandering detection accuracy. This and similar sensors are still prohibitively expensive (about $1,700 USD[2]) for classroom use. Current COTS alternatives (e.g., Fitbit HR, $100 USD[3]) typically do not include the physiological channels utilized in costly research-grade physiological devices (e.g., Empatica, Shimmer), do not provide the same fine-grained sampling rates, and might still be prohibitively expensive for classroom use at scale.

An alternative is to obtain physiological signals indirectly from video. In particular, in a recent study [58], participants watched online lectures while placing their fingers over the camera lens of a smart phone with the flash on. Heart rate was measured from changes in color due to blood pumping through the finger (photoplethysmography). They achieved a 22% above-chance accuracy for mind wandering detection via heart rate extracted from smart phone cameras. Though innovative, it is not clear how well their method works beyond mobile applications, or whether it would be effective outside a laboratory setting when finger placement is harder to control, and battery life is of central concern.

Mind Wandering Detection from Reading and Textual Features. Some researchers have adopted approaches to mind wandering detection based on reading activities (keypresses) alone. In one of the earliest studies, participants read a text one word at a time [59] and were classified as having mind wandered when they spent too little or too much time on difficult sections of the text, as determined by word length, number of syllables, and word familiarity. Despite achieving a 45% above-chance accuracy, this method is limited by the threshold-based approach and the unnaturalness of word-by-word reading.

Mills and D'Mello [60] addressed these limitations by using machine learning to detect mind wandering from reading times and textual features (e.g., number of words, text difficulty) in more naturalistic reading paradigms. They achieved a 20.7% above-chance accuracy with person-independent validation. Though promising in terms of scalability, an obvious limitation is that the detector cannot be applied to non-reading contexts.

Mind Wandering Detection from Facial Features. Most similar to the present research are two of our own studies on mind wandering detection from video in the lab. In the first study [61], we recorded videos of participants' faces as they watched a narrative film for approximately 35 minutes. We extracted facial action units (AUs) with FACET, a commercial version of the Computer Expression Recognition Toolbox (CERT) [62], and body movement using a motion tracking algorithm [63]. We trained a variety of classifiers including support vector machines, logistic regression, naïve Bayes, and others to detect mind wandering from the video features. The best-performing model achieved a person-independent mind wandering F_1 score of .390 – a 13% above-chance improvement.

In a subsequent study [64], we analyzed the generalizability of this method across task contexts. One set of participants watched a narrative film, while a separate set of participants read a scientific text. The model trained on narrative film data achieved a 25% above-chance accuracy and generalized to the scientific text reading task (21% above-chance accuracy). The model trained on scientific text data also achieved a 25% above-chance accuracy, and after tuning the mind wandering prediction threshold, also generalized to the narrative film watching task (22% above-chance accuracy).

These studies study demonstrated the potential for video-based mind wandering detection and their generalizability, but used basic facial features and achieved only low to moderate accuracy, which we improve on here.

1.4 Novelty of Current Study

Researchers have demonstrated the feasibility of automatic mind wandering detection, but with some drawbacks (see Section 1.3). To address these limitations, we propose mind wandering detection based on facial and movement features derived from video. This offers several advantages over previous work. First, cameras are almost universally present on laptops and mobile computing devices used in schools, or can be purchased quite cheaply (for under $10 USD[4]). Second, cameras require little to no expertise to set up and require no calibration compared to gaze trackers and some physiological sensors. Third, facial features are not strictly dependent on the task at hand and should generalize across domains. In contrast, gaze features are more dependent on the stimulus (e.g., fixation durations, scan paths, etc. are different for reading compared to scene viewing [65]), and are less likely to generalize.

We also extend previous work [61], [64] on detecting mind wandering from video in the following five ways:

Feature Engineering. Whereas previous work exclusively relied on basic descriptives (mean, standard deviation, max) of AUs and body movement, we propose a novel combination of features extracted at six levels of complexity: (1) gross body movement; (2) head pose; (3) facial texture patches; (4) individual AUs; (5) co-occurring AU pairs; and (6) temporal dynamics of AUs. We hypothesize that such a multifaceted analysis is needed since mind wandering is a visually subtle phenomenon (as illustrated in Fig. 1) and its overt behavioral cues are unknown.

Comparison and Fusion of Feature Types. We compare individual models trained using different feature to identify which feature sets capture facial cues that communi-

[1] Price as of May 2016
[2] Empatica E4 REV2 price as of March 2018
[3] Price as of March 2018
[4] JESWELL USB2 webcam price as of March 2018

cate mind wandering. Furthermore, we show that a combination of these different feature sets improves detection accuracy over previous approaches that relied on feature sets (1-2) and (4) from above.

Classifiers. We also improve on previous work by considering more complex classification algorithms, including SVMs with a range of hyperparameters, and deep neural networks with varying structures.

Feature Selection. We introduce a novel feature selection method for datasets with a large number of dimensions and non-linear feature–label relationships. Such datasets are commonly encountered in affective computing applications, where there may be many input features but few instances – thus necessitating feature selection.

Classroom Context. Finally, we also explore video-based mind wandering detection in an authentic classroom environment, where participants interacted with a computerized biology tutor. The classroom environment is especially challenging due to privacy concerns (no videos could be recorded), thereby incurring the added constraint of real-time feature extraction. Additionally, all processing had to be performed on budget hardware already available in the school classroom, which required a simplification of the feature set. Our results indicate that our approach was successful despite these challenging constraints.

2 Study 1: Self-Caught Mind Wandering Detection During Reading in the Lab

We reanalyzed video data previously reported in [64] to enable comparisons of the proposed approach with previous work. The data itself was collected as part of a larger study – see [66] for full details.

2.1 Data Collection

Participants (152 university students) read the introductory chapter of *Soap Bubbles: Their Colors and the Forces that Mould Them* by C.V. Boys [67]. The text is about the physical behaviors of soap bubbles, how surface tension enables bubble formation, and how chemical composition affects bubble formation. We used this text because it is likely to be unfamiliar to most participants but is written to be understandable without prior knowledge of the topic.

The text was presented on 57 screens (called pages) with about 114 words per page. Participants used the right arrow key to advance to the next page. Videos of participants' faces were recorded with a Logitech C270 webcam ($20 USD[1]) at 12.5 frames per second. Of the 152 participants, 10 were removed due to video recording errors and three were removed because they did not sign a data release agreement, leaving 139 participants in the dataset.

Participants used pre-designated keys to report whenever they caught themselves zoning out – a colloquial term for mind wandering. These served as "ground-truth" labels for supervised machine learning. Zoning out was defined as: *At some points during reading, you may realize that you have no idea what you just read. Not only were you not thinking about what you are actually reading, you were thinking about something else altogether. This is called "zoning out"*. Participants were further instructed to distinguish between two types of zone outs – task-related interferences vs. task-unrelated thoughts – as part of a larger study. However, both these types of zone outs were grouped because they are related, and multiclass detection was infeasible given the dataset size.

We used the self-caught method here vs. the probe-caught method (Study 2) because we were interested in tracking mind wandering without task disruptions and were focused on mind wandering with meta-awareness [1] (i.e., people are consciously aware that they are mind wandering).

It is important to emphasize a few points about this method to track mind wandering. First, the method relies on self-reports because mind wandering is an inherently internal phenomenon, which requires conscious awareness for reporting [20]. At this time, there are no reliable neurophysiological or behavioral markers that can accurately substitute for the self-report methodology [20]. Second, self-reports of mind wandering have been objectively linked to a host of theoretically-grounded behavioral and physiological signals [6], [8], [57]–[60], [68]–[78], providing convergent validity for this approach. Self-reports also consistently correlate with objective outcome measures, which provides evidence for their predictive validity [3]. Finally, our reliance on self-reports to measure mind wandering is consistent with the state of the art in the psychological and neuroscience literatures [20].

2.2 Extracting Video Clips

There were a total of 2,577 mind wandering reports across 7,923 pages of text (about one report every 3 pages). On average, each participant provided 18.5 reports ($SD = 13.5$) As shown in Fig. 2, the number of reports was quite variable across participants, which makes person-independent mind wandering detection quite challenging.

Participants reported mind wandering an average of 16 seconds into the page. Accordingly, we extracted video clips in 10s windows leading up to each mind wandering report; these corresponded to positive instances of mind wandering. We used 10s as a compromise between having longer, potentially more informative clips, while maximizing the number of clips that could be extracted. Of the 2,577 clips, 1,339 clips overlapped across pages and were discarded because of the concern that the action of leaning forward and looking at the keyboard to find the page-turn key might have influenced facial feature tracking.

We also added a 4s buffer before the mind wandering report to ensure that clips did not capture the movements associated with the self-report key press. We chose a 4s buffer length based on a pilot study where four raters made judgments on whether the keypress was visible in 540 randomly-selected video clips with buffer lengths ranging from 0-6s. Raters were instructed to report "if there is apparent hand or eye movement at the end of the clip as participants look and reach for the MW key." Two raters initially coded 250 clips with 0s-4s buffers. They reported apparent hand movements in 73% of clips with a 0s buffer

[1] Price as of March 2018

(eye movements in 93%), down to 4% hand movements and 5% eye movements for 4s clips. We increased the buffer lengths to 5s and 6s, and obtained ratings from the same raters and two new raters, finding no further decrease in apparent hand or eye movements with longer buffers. Thus, we proceeded with a 4s buffer length.

A further 207 clips were removed because the face could not be automatically detected for at least 1 second of the clip, which was our minimum threshold for usable data. A real-time application of our methods could also discard such clips, so removing them does not harm validity. In total, there were thus 1,031 usable mind wandering clips of which 64% were task-unrelated mind wandering reports. These served as positive instances for the classifiers,

Negative instances were extracted from periods of time between mind wandering reports (see Fig. 3). We divided each video into 14s instances (10s window of data + 4s offset to avoid including page turn movements) and removed any instances that coincided with page turn events. We also removed any negative mind wandering instances that fell within a 30s period before each mind wandering report, because the participant might have been mind wandering but had not yet realized or reported it. The duration of mind wandering is an open question [20], but is hypothesized to not exceed 20s (see [79]); the 30s buffer was taken out of an abundance of caution.

We randomly selected 2,406 negative mind wandering instances from the remaining instances to obtain a 30% mind wandering rate, which is consistent with previous research on the incidence of mind wandering during learning, especially during reading (see meta-analysis in [4]). The dataset comprised a total of 3,437 instances (1,031 positive mind wandering).

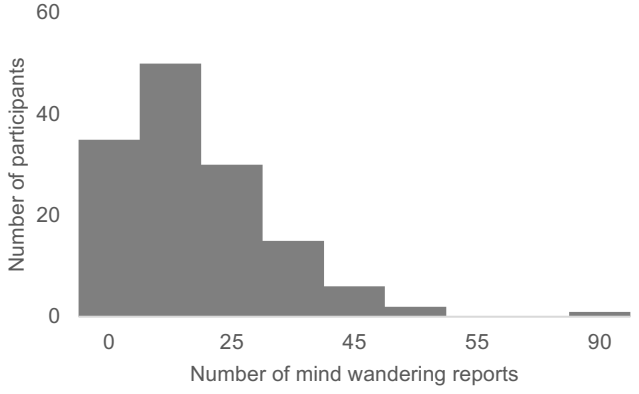

Fig. 2. Histogram of the number of self-caught mind wandering reports made by participants.

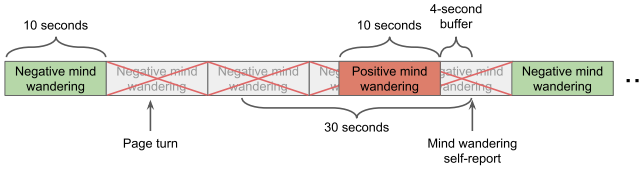

Fig. 3. Instance extraction scheme illustrating how we selected positive and negative instances of mind wandering. We eliminated instances that overlapped with page turn events, because body and head movements due to page turn actions are tangential to mind wandering in general. We selected negative instances of mind wandering that were at least 30 seconds before mind wandering self-reports.

2.3 Feature Extraction

We extracted features at five levels of granularity, ranging from a simple measure of upper-body motion to complex patterns in AU temporal dynamics.

Upper-body Movement Features. We used a validated motion silhouetting method [63], where each video frame is compared to a continuously-updated background image formed by the weighted average of the previous four frames. Gross body movement was estimated as the proportion of pixels that changed compared to the background motion silhouette (Fig. 4A). This movement estimation method also serves as an accurate proxy for pressure-sensitive posture sensors [63]. We extracted the following statistical features from the body movement time series in each 10s clip: mean, median, standard deviation, minimum, maximum, and range.

Head Pose Features. We utilized head pose features as a proxy for gaze direction, motivated by the link between eye gaze and mind wandering [79], [80]. Specifically, we extracted head yaw (looking to the side), pitch (looking up or down), and roll (tilting to the side; Fig. 4B), summarizing each with mean, median, standard deviation, minimum, maximum, and range across the 10s clips – yielding 18 head pose features in total.

Local Binary Pattern (LBP) Texture Features. We extracted texture patch features with local binary patterns [81], which have been shown to be effective for engagement classification [35], [37]. LBP features capture texture patterns, which are indicative of changes in facial expressions changes. For example, texture patches near the mouth change during smiles as wrinkles appear on the skin, lips widen, and teeth become visible.

LBP features were computed following the *uniform, rotation invariant* method [81]. Features were computed for individual pixels in a patch (see below) by measuring brightness in a ring around that pixel. Pixels brighter than the central pixel were coded as 1 while dimmer pixels were coded as 0, producing an eight-digit binary pattern for the pixel (e.g., 00001111 – see Fig. 4C). The method then counts the frequency of the various patterns in the patch. Uniform LBP features are those with one consecutive area of brightness (e.g., 01110000 but not 01010100). All non-uniform patterns were grouped together before counting pattern frequencies. Rotation-invariant LBP features are those that were equivalent after bit-shifting so that orientation of the pattern did not matter (e.g., 1110000 and 00011100 were counted together, since both have three consecutive bright pixels). All rotations of the same pattern were grouped to yield 10 patterns in all: a non-uniform pattern, a sequence of all 0's, and 8 possible sequences of consecutive 1's.

We extracted LBP features from fifteen 16 × 16 pixel patches from both eyes and the center of the mouth, automatically located with OpenFace [82]. We selected eye regions to capture events such as blinking and general eye movement patterns (e.g., horizontal saccades should be indicative of normal reading), which have both been linked to mind wandering [50], [80]. The mouth regions were chosen to capture movements such as yawning that would result in texture changes (e.g., as the teeth became visible).

We extracted ten LBP features from each patch in each

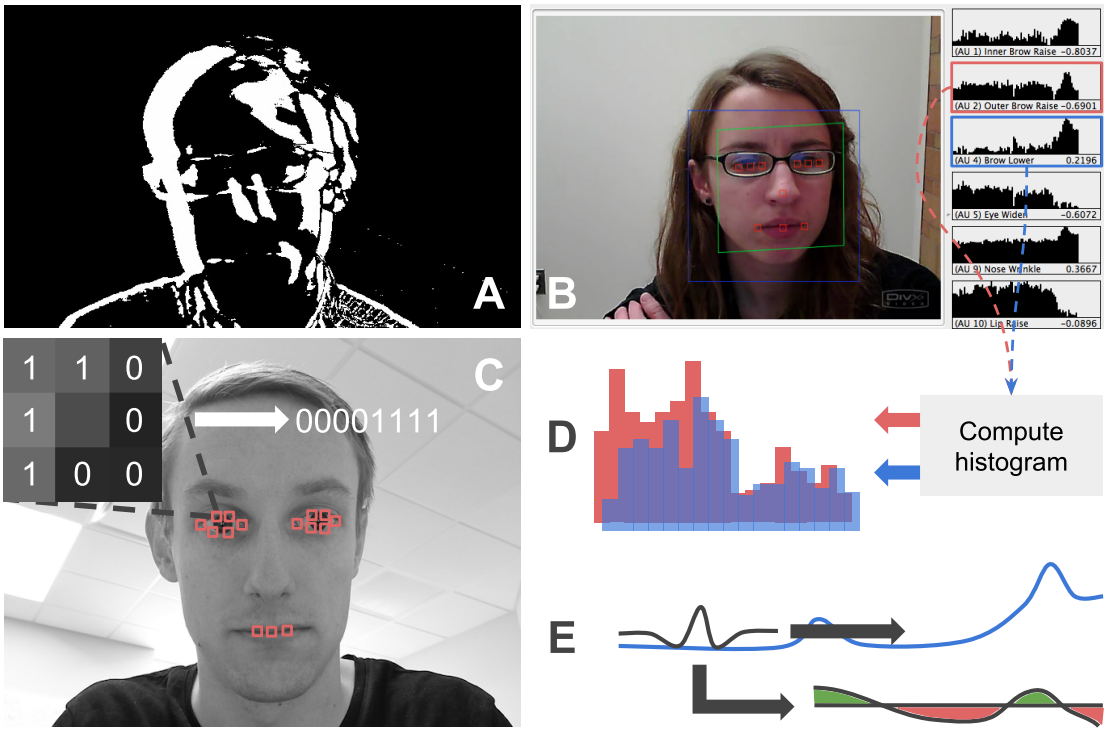

Fig. 4. Feature extraction examples illustrating: A) upper-body motion, B) estimates of action units (AUs) and head pose provided by EmotientSDK, C) local binary patterns extracted from key areas of the face, D) Jensen-Shannon divergence measuring similarity between pairs of AU estimates, and E) counting positive and negative responses of a Gabor filter convolved across an AU estimate.

frame, and aggregated over frames in each clip with minimum, maximum, mean, median, range, and standard deviation functionals to obtain a total of 900 LBP features (15 patches × 10 features/patch × 6 functionals/feature).

Basic Action Unit (AU) Features. Facial action units (AUs) represent specific muscle activations; the AUs we considered were AU1 (inner brow raiser), AU2 (outer brow raiser), AU4 (brow lowerer), AU5 (upper lid raiser), AU6 (cheek raiser), AU7 (lid tightener), AU9 (nose wrinkler), AU10 (upper lip raiser), AU12 (lip corner puller), AU14 (dimpler), AU15 (lip corner depressor), AU17 (chin raiser), AU18 (lip puckerer), AU20 (lip stretcher), AU23 (lip tightener), AU24 (lip pressor), AU25 (lips part), AU26 (jaw drop), and AU28 (lip suck). We detected AUs with EmotientSDK, an updated commercial version of the Computer Expression Recognition Toolbox (CERT) [62]. CERT has been previously validated against human annotations of facial expressions on thousands of video frames [36], [83]. It recognizes AUs by extracting the responses of 72 two-dimensional Gabor filters and uses support vector machines (SVMs) for AU classification. AU intensity is estimated by measuring the distance from the decision boundary of the SVM [84]. Fig. 4B illustrates time series of example AUs. We extracted the mean, median, standard deviation, minimum, maximum, and range across the 10s clips, resulting in 114 AU features (6 functionals × 19 AUs).

Co-occurring AU Features. We captured co-occurrence relationships between AUs to model more complex expressions. For example, co-occurring muscle movements near both the mouth and eyes when smiling can indicate genuine smiles compared to smiles involving the mouth only [85]. We estimated AU co-occurrences based on the similarity between their distributions within each clip using Jensen-Shannon divergence (JSD) [86], which is an extension of Kullback-Leibler divergence (KLD) [87]. KLD (Equation 1) measures the information lost by using a prior distribution Q to approximate a posterior distribution P, given probability density functions p and q for P and Q respectively. JSD (Equation 2) is a modification of KLD that is symmetric, which allows measurement of symmetric relationships between AUs (e.g., co-occurrence of eyebrow + mouth movements is equivalent to mouth + eyebrow movements). JSD was chosen over other measures (e.g., correlation-based measures) because it captures non-linear relationships [86]. Furthermore, JSD measures expressions that consist of multiple AUs activating in the same clip, even if they do not activate at exactly the same moment. For example, JSD features can measure a mouth movement that is accompanied by an eyebrow movement within the same clip. We computed a total of $(19 \times [19 - 1] \ / \ 2) = 171$ JSD features for each AU pair.

$$KLD(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx \quad (1)$$

$$JSD(P||Q) = \frac{1}{2} KLD(P||\frac{1}{2}[P+Q]) + \frac{1}{2} KLD(Q||\frac{1}{2}[P+Q]) \quad (2)$$

AU Temporal Dynamics Features. We captured changes in AUs over time to model facial expression dynamics that might be obscured by mean aggregation as with the basic AU features. We applied one-dimensional Gabor filters to AU time series using an existing method [88]. Gabor filters capture responses in specific frequencies

and can thus distinguish between static facial expressions (such as an open mouth) and dynamic facial expressions (such as a yawn) if tuned to the right frequency [88].

Gabor filters consist of a cosine wave multiplied by a Gaussian envelope. From a filter wavelength λ and sample frequency f (i.e., frames per second), a frequency multiplier k scales sample indices to the appropriate domain (Equation 3). Then a filter G is defined with the same width w as the number of frames in the video (Equation 4).

$$k = \frac{2\pi}{\lambda f} \tag{3}$$

$$G[t] = \cos\left(\left(t - \frac{w}{2}\right)k\right) e^{-\frac{1}{8}\left(t-\frac{w}{2}\right)^2 k^2} \tag{4}$$

Filters were convolved across the frame-by-frame AU time series with missing values (failed face tracking) linearly interpolated. Filter responses indicate changes in AUs that occurred over a period of time similar to the period of the cosine wave. Following the method of [88], we squared the filter response to emphasize larger values and counted regions of both positive and negative responses. We applied a bank of 8 filters with periods ranging from 1 to 12 seconds (Fig. 5), because the specific duration of any facial expressions associated with mind wandering is unknown but similar periods have been effective in previous research [88]. Periods much shorter than 1 second would be unlikely to work for our videos since they were recorded at 12.5 frames per second; filters would have few samples from which to recognize short periods (e.g., just 5 frames for a period of 0.4 seconds).

The temporal filter features were counts of positive and negative responses for each filter (i.e., the number of times the filter produced a positive or negative response within a clip). We then grouped counts into bins (number ranges) according to size of response (area under the squared filter response curve), ranging from -6 to 6 in accordance with previous research [88]. With 19 AUs, 8 filter wavelengths, and 12 bins, there were 1,824 temporal filter features.

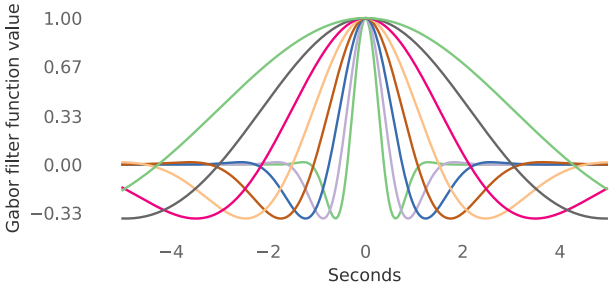

Fig. 5. Gabor filters that were convolved over AU intensity time series.

2.4 Supervised Classification

We build models with two commonly-used classifiers: SVMs and feed-forward deep neural networks (DNNs). We used SVMs because of their flexibility and general efficacy in related computer vision research [89], [90]. Furthermore, they are well-suited to the high-level features we extracted and the relatively small size of the dataset. We also explored DNNs given their efficacy [91], [92] – though their full potential is typically only realized with large amounts of data. For the same reason, we did not consider convolutional neural networks (CNNs). Pre-trained CNNs could be applied to extract feature maps from individual video frames, but would yield tens of thousands of dimensions when applied across clips with up to 125 frames per clip. We thus restricted analyses to the relatively low-dimensional feature types described above.

We trained individual SVM and DNN models for each feature set (motion, head pose, LBP textures, AUs, co-occurring AUs, and temporal AU dynamics) and also considered fusion of feature sets so that the predictive power of different types of facial features could be compared (e.g., static versus dynamic facial expressions).

We used person-independent four-fold cross validation [93]. By person-independent, we mean that all data from a participant was either in the training or testing data, but never both, thereby increasing the likelihood of generalization to new participants (at least within similar populations and interaction contexts). We used data from three of the four folds (data from 75% of participants) for training, while models were tested on the remaining fold (25% of participants). We further split training data to select features, weigh instances, and select hyperparameters with two-fold nested cross-validation (see Fig. 6). These procedures were only applied to the training data.

Feature Selection. For DNN models, we added L_1 regularization to the first layer, thereby minimizing the influence of ineffectual features. For SVM models, we applied feature selection to reduce the dimensionality of the feature space. We initially experimented with forward feature selection (FFS) [94], given that model-free alternatives such as RELIEF-F and correlation-based feature selection (CFS) do not capture the same nonlinear patterns in data that our classifiers do [95], [96]. However, FFS was computationally impractical due to the large number of LBP features and hyperparameter combinations (described below).

Thus, we developed a new two-step variation of FFS (test-correlate feature selection; TCFS) as a compromise between the model-specific advantages of forward feature selection and the computational simplicity of model-free methods. In TCFS, we trained an SVM on each feature and then ranked features based on the accuracy of these individual feature models, as measured by the area under the receiver operating characteristic curve estimated from a single point (minimum proper curve [97]). We eliminated any feature that was correlated (Spearman's *rho* > .6) with a better ranked feature. The final models were trained on up to 50 (if there were that many) of the highest-ranked remaining features. The 50-feature maximum was informed by recommendations that the square root of the number of instances per training fold ($3437 \times 3/4 = 2578$ in our case) is an appropriate conservative limit on the number of features [98].

Instance Weighting and Hyperparameter Tuning. Initial experiments with unweighted instances yielded models that exclusively predicted the majority class. Thus, we weighted training instances such that the sum of weights for positive and negative mind wandering instances were equal. SVM models fit the decision boundary according to

these weights, setting the decision boundary further from higher weight (minority class) instances. Similarly, DNN models made larger parameter adjustments for minority class instances.

We trained SVM models with radial basis function (RBF) kernels, which require a regularization hyperparameter C and a support vector radius of influence hyperparameter γ. We varied C and γ via grid search; values of C varied from 10^{-2} to 10^{2} and values of γ varied from 10^{-5} to 10^{2} by powers of 10. We selected hyperparameters for each feature set individually because each captures different behavioral expressions and has different distributions.

We also varied DNN hyperparameters via grid search, including the number of hidden layers (1, 2, 4, or 8), number of neurons in each hidden layer (4, 8, 16, or 32), dropout [99] applied before each hidden layer (0%, 25%, 50%, or 75%), L_1 regularization applied to the input layer (0, .001, .01, or .1), and learning rate for the Adam [100] optimizer (.01, .001, or .0001). The DNN decision threshold was initially set to a typical default of .5, but this resulted in very few positive mind wandering predictions (e.g., < 5% for four of the six individual feature set models) Thus, we instead chose DNN decision thresholds to match the SVM predicted rate of mind wandering as closely as possible (since DNNs produced continuous predictions while SVMs did not) to enable unbiased model comparisons.

Fusion Methods. We considered three methods to fuse the individual feature sets. For feature-level fusion, we concatenated the selected features from each set, performed another round of feature selection to reduce the feature set size, and trained a new model. For majority voting, we classified an instance as mind wandering if at least three of the six individual SVM models classified it as such, or if the sum of DNN prediction probabilities exceeded a threshold tuned to match SVM prediction rates. We also trained a Classification and Regression Tree (CART) model on the predictions of the individual feature set models [101]. We cross-validated fusion models with the same training and testing folds as the individual models, thus preserving person-independence.

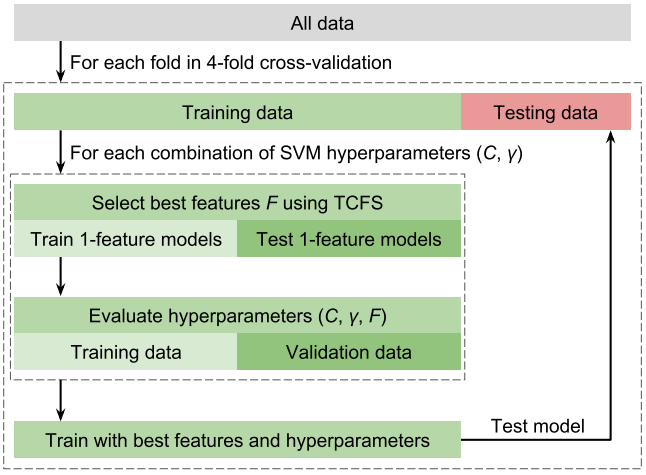

Fig. 6. Illustration of SVM model training procedure showing which portions of data we trained and tested on at each step

2.5 Study 1: Mind Wandering in the Lab Results

2.5.1 Classification Results

We measured accuracy primarily with the F_1 score of mind wandering. F_1 is the harmonic mean of precision (proportion of instances classified as mind wandering that were truly mind wandering) and recall (proportion of true mind wandering instances that were classified as mind wandering). As a baseline, chance-level mind wandering F_1 (.300) was defined as the mind wandering F_1 obtained by randomly assigning positive mind wandering labels to 30% of the instances (i.e., the base rate) and negative to the rest. We also computed area under the receiver operating characteristic curve (AUC), where .500 represents chance level and 1 represents perfect classification. For SVMs we utilized the distance of each instance from the separating hyperplane as a measure of confidence (which is required to calculate AUC) while DNNs naturally yield continuous confidence predictions. The results in Table 1 indicate that detection accuracy was modest, but better than chance in terms of both F_1 and AUC (for SVM models more so than DNNs).

TABLE 1
MIND WANDERING (MW) DETECTION RESULTS IN THE LAB

	MW F_1		MW Precision		MW Recall		AUC		
Feature Set	**SVM**	**DNN**	**SVM**	**DNN**	**SVM**	**DNN**	**SVM**	**DNN**	**Predicted MW Rate**
Body Motion	.429	.356	.355	.295	.541	.449	.565	.495	.457
Head Pose	.330	.412	.269	.335	.429	.534	.557	.514	.478
LBP Textures	.458	.421	.361	.332	.626	.575	.600	.552	.520
Basic AUs	.430	.352	.358	.293	.536	.439	.576	.476	.449
Co-occurring AUs	.380	.353	.339	.315	.433	.402	.545	.525	.383
Temporal AUs	.389	.356	.312	.285	.518	.473	.513	.474	.498
Feature-level Fusion	.362	.366	.303	.306	.448	.453	.499	.506	.443
Majority Vote	.454	.420	.368	.324	.545	.598	.580	.518	.445
CART Fusion	.478	.376	.360	.283	.711	.562	.603	.481	.593

We compared models using mixed-effects logistic regression to predict agreement between the model outputs and the self-reports (1 for agree; 0 for disagree). We included participant as an intercept-only random effect for all comparisons, due to the repeated and nested structure of the data – one or more instances nested within a participant. We also included participant-level predicted mind wandering rates as a fixed effect covariate, because predicted mind wandering rates were correlated with model accuracy (Pearson *r*s between -.158 and -.409). For this reason, comparisons might not align with F_1 scores in the table since those do not adjust for prediction rates. We used the *lme4* [102] package in R [103] for model fitting, the *car* package [104] for significance testing, and the *emmeans* package [105] for pairwise comparisons.

We first compared SVMs versus DNNs across all six individual feature sets by regressing accuracy on the classifier type (two-level categorical variable for SVM or DNN) × self-reported mind wandering label (1 or 0) interaction term, with feature set included as a fixed effect. The classifier type × label interaction term allows us to examine model accuracy for positive vs. negative instances of (self-reported) mind wandering, the former being of interest here. The results indicated a significant interaction term, χ^2 (1) = 5.11, $p < .05$, which suggests that relative model accuracies varied for positive vs. negative instances of mind wandering. Focusing on the positive instances, estimated marginal means comparison showed that SVMs yielded statistically better results than DNNs, on average.

Focusing on SVM models only, we compared individual feature sets by regressing correctness on the feature set (six-level categorical variable for the individual feature sets) × label interaction term. The interaction was significant (χ^2 (5) = 183, $p < .001$), so we conducted pairwise comparisons between feature sets with false discovery rate adjusted for 15 comparisons. The results yielded the following overall pattern for positive mind wandering instances: LBP Textures > [Basic AUs = Body Motion = Temporal AUs] > [Co-occurring AUs = Head Pose]. Notably, Basic AUs and Temporal AUs features were more effective than Co-occurring AUs – perhaps because Basic and Temporal AUs capture simpler, first-order expressions of a single facial muscle, while Co-occurring AUs capture subsets of these facial expressions, which might have been too sparse given available data.

Finally, comparing the best individual feature set (LBP Textures) to SVM fusion models yielded a significant model × interaction term (χ^2 (3) = 281, $p < .001$). Comparisons for positive mind wandering instances indicated the models were statistically ranked as follows: CART Fusion > [Majority Vote = LBP Textures] > Feature-level Fusion. Thus, results indicate that the CART model was most accurate, though its tendency to predict high levels of mind wandering might be undesirable in some applications, in which case the Majority Vote model may be a better choice.

2.5.2 Comparison Across Mind Wandering Types

Students could report mind wandering incidents that were either task-related or task-unrelated thoughts. We replicated individual SVM models described above (Section 2.4) as a three-class classification task (task-unrelated thought, task-related interference; not mind wandering). We computed accuracy by combining predictions of either mind wandering type and re-computing F_1, so that we could directly compare accuracy to the binary model. However, F_1 scores were ≈ .300 (chance level) with the exception of Body Motion (F_1 = .336), which was not higher than the binary classification F_1 (.429). This is likely due to the fact that the class imbalance for the binary classification gets much more severe for the three-way classification.

TABLE 2
COMPARISON OF RECALL FOR TASK-RELATED AND TASK-UNRELATED TYPES OF MIND WANDERING

Feature Set	Task-related recall	Task-unrelated recall
Body Motion	.542	.541
Head Pose	.439	.423
LBP Textures	.574	.655
Basic AUs	.501	.556
Co-occurring AUs	.423	.438
Temporal AUs	.507	.524
Mean	.477	.506

To further examine differences across mind wandering types, we inspected the proportions of each mind wandering type that were correctly classified (recall) by the binary classifiers. We first compared recall across each mind wandering type for each feature channel (Table 2). Overall, recall was similar across the two types of mind wandering. Next, focusing on the positive mind wandering instances only, we regressed model correctness (correct or incorrect) on the mind wandering type and feature set interaction. Neither the mind wandering type main effect nor the feature set × mind wandering type interaction were significant, indicating that models were similarly accurate for both types of mind wandering.

2.5.3 Comparison Across Genders

Students (39% male) reported their gender following the text reading portion of the study. We compared mind wandering reports and classification accuracies across genders, regressing agreement on the gender × mind wandering label interaction term, again including feature set as a fixed effect. The interaction term was significant (χ^2 (1) = 73, $p <$ .001), and pairwise comparisons revealed that the individual feature set models were significantly more accurate for male than female students – despite the fact that we controlled for individual mind wandering rates in this comparison. This difference was primarily due to higher recall for male students (.797 versus .665), especially since precision was higher for the female students (.326 for females; .386 for males).

2.5.4 Including Undetectable Face Instances

We removed 207 positive mind wandering instances because no face could be detected in the video clips (Section 2.2). These could also be removed in a real-time application of our detectors, but gaps in predictions might need to be

filled in – for example, to create uniform time series predictions. We thus examined the influence of making random predictions for these instances, with additional negative mind wandering instances sampled to maintain the 30% mind wandering base rate. We set the randomly-predicted mind wandering rate equal to that of the best model (CART Fusion; .593 predicted rate), appended the random predictions to the list of CART Fusion predictions, and recomputed accuracy to simulate a scenario where the CART Fusion predictions are supplemented with random predictions when needed. Given the small number of unusable positive instances, this approach has minimal detriment to accuracy – the newly calculated F_1 was .466, precision was .351, and recall was .694 (versus .478, .360, and .711, respectively; Table 1).

2.6 Comparison to Human Observers

Some visual perception tasks are relatively easy for humans but difficult for computers (e.g., recognizing faces, following gaze directions). To assess the difficulty of recognizing mind wandering from facial expressions, we recruited human observers from Amazon's Mechanical Turk's [106] crowdsourcing platform to each code a random subset of 100 video clips (30 of which corresponded to self-reported mind wandering) for mind wandering. We recruited nine different human observers per clip and used majority voting to determine the final observer mind wandering label for each clip.

Observer achieved mind wandering precision, recall, and F_1 scores of .333, .467, and .389, respectively. On the same subset of 100 clips, the CART decision-level fusion model yielded precision, recall, and F_1 scores of .421, .800, and .552 with a predicted mind wandering rate of 57% compared to 42% for human observers. Accuracy varied considerably across observation rounds (Fig. 7), though none exceeded the accuracy of the CART model (F_1 of .552 on this sample).

Despite the small sample size, this result illustrates the difficulty of the task for humans. It also highlights the potential for automatic mind wandering detectors to outperform humans, though more formal validation with a larger sample is needed.

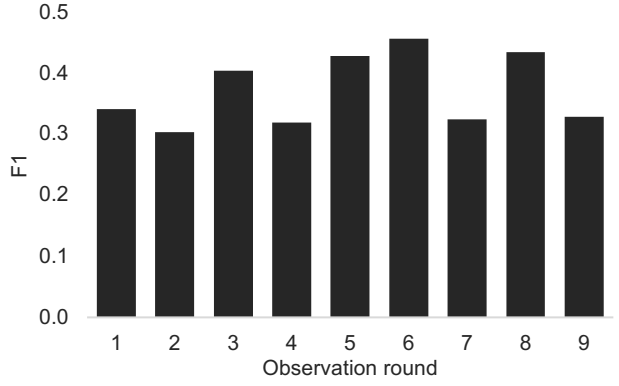

Fig. 7. Observer accuracy (F_1) for all nine rounds of human mind wandering ratings

3 Study 2: Probe-Caught Mind Wandering Detection with an Intelligent Tutoring System in the Classroom

We followed up on the lab study with a classroom study, using a different participant sample (high-school students), a different method to obtain mind wandering reports (probe-caught) and with a more interactive task: learning from an intelligent tutoring system called Guru [107]. We also used a different method to extract facial features as elaborated below. These methodological differences were due to practical constraints, but provide an opportunity to test core components of our approach in a vastly different context. Finally, participant-level demographics were not available for these data, so we could not compare model accuracy by gender as in Study 1.

3.1 Guru Tutor

Guru is an intelligent tutoring system designed to teach biology topics (e.g., osmosis; protein function) aligned with state curriculum standards. It engages students in one-on-one collaborative conversations in natural language [107]. It was modeled after interactions with expert human tutors and has been shown to be effective at promoting learning at levels compared to small group human tutoring [107].

Guru utilizes an animated pedagogical agent that references a multimedia workspace (see Fig. 8). The tutor communicates via synthesized speech and gestures, while students communicate by typing their responses, which are analyzed using natural language processing. Guru maintains a dynamic model of student progress (called a student model [108]), which it uses to adapt instruction to individual students.

A topic in Guru involves interrelated concepts and facts, which are covered in 15- to 40-minute tutoring sessions. Guru begins with an introduction to motivate the topic, which is then followed by a five-phase tutorial session (see [107] for details of each phase).

3.2 Data Collection

Data were collected from 135 (41% male) high-school freshmen and sophomores enrolled in an introductory biology course. Students provided written assent to participate, while their parents provided written consent. The study was approved by the university institutional review board, and by the high school's principal. Students were given a $10 gift card for participating.

The study occurred over the course of two days in students' regular biology classroom with students sharing a desk (Fig. 9). There were seven class periods per day, with enrollment ranging from 14 to 30 students per class. Students used a school-provided laptop to interact with Guru, which we equipped with an inexpensive (Logitech C270) external webcam. The cameras for the two students at each desk were connected to a third laptop, which was used solely for facial feature processing and was synchronized to the Guru laptops via an internet time server.

Upon providing assent, students were introduced to the study, followed by a 30-minute Guru learning session on one topic, a short break, and another 30-minute Guru session on a different topic.

We used the probe-caught method [9] to monitor mind wandering during the two Guru sessions. Specifically, we defined mind wandering to students before their first interaction with Guru, provided instructions on how to report mind wandering to the probes, and also administered a brief quiz to verify their understanding.

Thought probes occurred pseudorandomly every 90-120 seconds. The 90-120s time range was selected based on previous research which tracked mind wandering during interactions with Guru [109]. The probes automatically paused the tutoring session. If the tutor was speaking at the time the probe was to be triggered, the probe was delayed until the tutor finished speaking. The probe consisted of an auditory beep along with an opaque overlay on screen, instructing the participant to press the "N" key if they were not mind wandering, "I" if they were intentionally (deliberately) mind wandering, or "U" if they were unintentionally (spontaneously) mind wandering. Here, we do not differentiate between intentional and unintentional mind wandering in order to maximize the number of instances for machine learning.

Participants encountered an average of 12 probes over the course of each session; they reported mind wandering for an average of 27.6% (SD = 23.5%). There was considerable variability in the mind wandering distribution across participants as noted in Fig. 10.

As expected, the classroom environment was much less controlled than the lab environment. Students interrupted and distracted each other, left to go to the bathroom, and occasionally even used their cellphones. Due to computer failures (e.g., power supply failure, unexpected software updates), data from 10 students were unusable, resulting in data from 125 students.

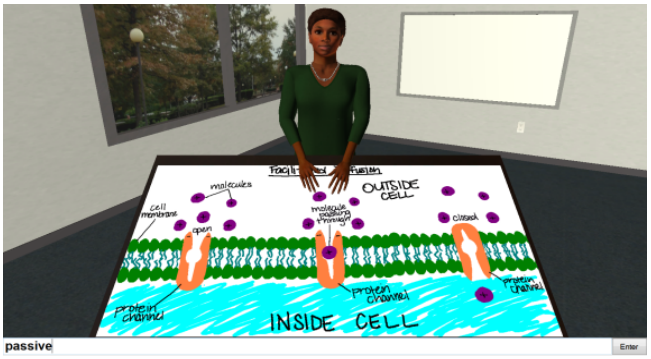

Fig. 8. Screenshot of Guru in the CGB phase.

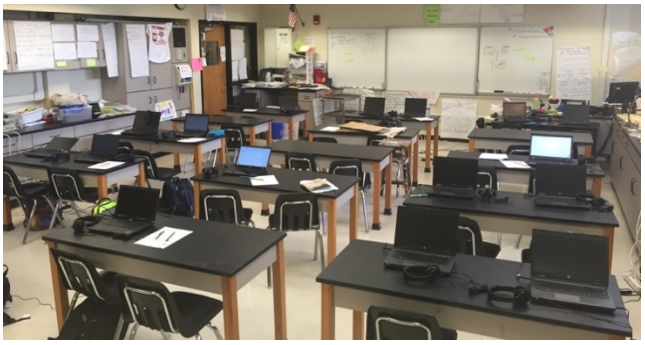
Fig. 9. Example classroom layout.

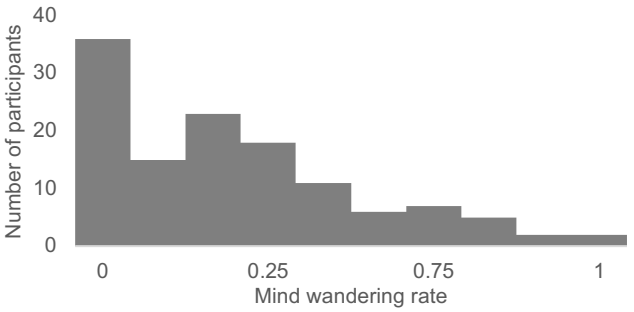

Fig. 10. Histogram of mind wandering rates.

3.3 Automatic Mind Wandering Detection

Real-Time Feature Extraction. Due to privacy considerations, videos of students could not be recorded for later feature extraction and analysis. Therefore, features were extracted in real-time. We could not extract features with EmotientSDK, as we did in the lab study, due to licensing constraints. Instead, we extracted AUs and head pose with OpenFace [82]. The feature extraction frame rate was variable because of external computational resource demands (e.g., system processes) and varying demands of the feature extraction process itself (e.g., when face tracking is lost the entire image must be searched to rediscover the face – a computationally expensive process). For this reason, frame rate was also relatively low (mean = 4.6 frames per second) compared to the lab study (exactly 12.5 frames per second). Additionally, temporal filter features could not be extracted from AU estimates because of the variable timing and sparsity of frames. Body motion and LBP features were also not extracted since they add additional computational complexity. Thus, we extracted head pose and AU features real-time, and calculated AU co-occurrence features (JSD features) offline.

Instance Extraction. We extracted 2,888 instances, each 10s long, from the 125 students. We discarded 502 instances because they contained fewer than 5 frames of data (approximately 1s), leaving 2,386 instances (25.9% positive mind wandering instances, 62.5% of which were unintentional).

Supervised Classification. As in the lab study, we trained SVM and DNN classifiers for the individual channels (Basic AUs, Co-occurring AUs, and Head Pose only) using the exact same cross-validation, feature selection, instance weighting, and hyperparameter tuning procedures from Study 1. We also trained similar feature-level fusion, decision-level fusion (CART), and majority vote models as in Study 1.

3.4 Results

Mind wandering detection was accurate above chance level (up to F_1 = .414, 20.9% above chance of .259 for the feature-level fusion model; Table 3). AUC results indicated a possible advantage for SVMs over DNNs, which we followed up via statistical comparisons of accuracy in the same manner as Study 1 (see Section 2.5.1 Classification Results) to rigorously investigate this possibility. We found a similar trend toward SVM models outperforming DNNs overall for the individual feature sets, but it was not significant (p = .154). Follow-up analysis of pairwise comparisons for the individual feature sets for SVMs (as in Study 1) revealed the statistical ordering: Basic AUs > Co-occurring AUs > Head Pose. Interestingly, the results show that the feature-level fusion model had the highest overall F_1, exceeding the individual feature sets and even the decision-level fusion methods (statistical ordering: Feature-level Fusion > CART Fusion > Basic AUs > Majority Vote. Once again, however, it is worth noting that the most accurate model predicts mind wandering at a high rate, so other models may be preferable if this is of concern.

3.4.1 Comparison Across Mind Wandering Types

We further analyzed each feature set, comparing recall for intentional and unintentional mind wandering report types (Table 4). We focused on the SVM models, given their higher accuracies. Similar to Study 1, there were no significant differences in detector accuracies between the two types of mind wandering included in Study 2. Recall was similar as well: .369 for intentional versus .345 for unintentional (Table 4) mind wandering.

3.4.2 Including Undetectable Face Instances

There were 502 instances in which the face could not be detected for at least 1 second (see Section 3.3). As in Study 1, we generated random mind wandering predictions with the same predicted rate as the feature-level fusion SVM (highest F_1 model). We found that F_1 was only slightly diminished at .396 (recall was .545, and precision was .311) with these instances included compared to them being excluded (F_1 = .414, recall = .573, precision = .324).

TABLE 4
COMPARISON OF DETECTOR RECALL FOR INTENTIONAL VERSUS UNINTENTIONAL MIND WANDERING CASES

Feature Set	Intentional recall	Unintentional recall
Basic AUs	.517	.453
Co-occurring AUs	.310	.339
Head Pose	.280	.244
Mean	.369	.345

4 DISCUSSION

We review the main findings, discuss limitations, and point to opportunities for future research.

4.1 Main Findings

Automatic mind wandering detection is a challenging problem, especially given the lack of prototypical mind wandering facial expressions (Fig. 1), variance in mind wandering reports across participants (in fact, 25% of participants reported no mind wandering at all in Study 1 [Fig. 2] and 29% in Study 2 [Fig. 10]), and the difficulty of the task for human observers (Section 2.6). Despite these challenges, we found that automatic computer vision methods detected mind wandering at better than chance-levels in both a laboratory reading context (decision-level fusion F_1 = .478 versus .300 chance) and in a noisy biology classroom with real-time feature extraction (feature-level fusion F_1 = .414 versus .259 chance).

Although these results reflect a modest improvement over chance level predictions (25.4% for the lab study and 20.9% for the classroom study), the models surpassed previous state-of-the-art face-based mind wandering during reading in a lab context. Specifically, mind wandering detection accuracy using all six feature sets was F_1 = .478, compared to F_1 = .441 previously reported using basic AU and head pose features [64] – an improvement of 8.4%. This finding demonstrates a slight advantage for considering multiple feature sets when detecting subtle facial expressions associated with mind wandering. In fact, LBP features (a feature set not previously considered for mind wandering detection) were the most accurate at the task. Additionally, our analysis of decision-level and feature-level fusion models in both studies showed a statistically

TABLE 3
OVERVIEW OF MIND WANDERING (MW) DETECTION RESULTS IN THE CLASSROOM

	MW F_1		MW Precision		MW Recall		AUC		
Feature Set	**SVM**	**DNN**	**SVM**	**DNN**	**SVM**	**DNN**	**SVM**	**DNN**	**Predicted MW Rate**
Basic AUs	.379	.340	.314	.282	.477	.429	.568	.528	.394
Co-occurring AUs	.320	.285	.312	.278	.328	.293	.536	.510	.273
Head Pose	.248	.272	.240	.263	.257	.282	.511	.502	.277
Feature-level Fusion	.414	.331	.324	.259	.573	.458	.584	.487	.458
Majority Vote	.326	.292	.313	.280	.341	.306	.547	.499	.283
CART Fusion	.385	.311	.304	.246	.523	.422	.566	.530	.444

significant increase in accuracy compared to the best individual feature set, though with high mind wandering prediction rates.

These results showed a notable advantage for SVM models versus DNNs in both studies. In Study 1, the best SVM model yielded a 9.8% improvement over the best DNN model; the improvement was 11.2% for Study 2. The SVM advantage in these studies comes despite a thorough hyperparameter search for DNNs. DNNs may have suffered from the relatively small size (3,437 instances in Study 1 and 2,386 in Study 2) and high-level feature sets, compared to machine learning problems where DNNs typically excel: millions of instances and low-level features.

We also investigated model accuracies for two different types of mind wandering in each study. In Study 1, we found no statistical difference in accuracy for task-related and task-unrelated mind wandering instances. Similarly, Study 2 showed no difference in accuracy for intentional versus unintentional mind wandering, indicating that mind wandering is not easier to detect across these types, and may be associated with similar facial expressions.

There were some instances where facial expressions could not be detected (207 in Study 1 and 502 in Study 2). However, for some applications it may be necessary to make predictions for *all* instances. We therefore computed accuracy with random predictions set to match the classifier prediction rate for these instances, and found that model accuracy was not drastically diminished in Study 1 (F_1 decreased from .478 to .466) nor in Study 2 (F_1 changed from .414 to .396), thus indicating models could be deployed in such applications without notable decreases in accuracy.

4.2 Limitations and Future Work

The most notable limitation of the current paper is the modest accuracy achieved for mind wandering detection. However, this is expected since the problem of mind wandering detection has been shown to be exceedingly difficult with other modalities as well, such as eye gaze and physiology. That said, gaze-based mind wandering detectors do appear to outperform video-based detectors (see [110]) on the same data. Thus, future work should strive to improve these results through additional feature engineering methods and additional deep learning methods which have been successful in other domains.

Additionally, the features that could be extracted in the classroom environment were limited by the processing power of the computers available. While this is a realistic constraint that must be dealt with, future work with increased processing power for real-time feature extraction will be necessary for determining performance upper limits in this context.

It is also possible that facial expressions of mind wandering differ in contexts with more or less social pressure to appear engaged. Participants read alone in Study 1, but were still aware (at least initially) that they were being recorded, which might have increased self-regulatory behaviors. Thus, one possible avenue for future work is to compare facial expressions in contexts where individuals do not know they are being observed. Similarly, participants' interest in reading or learning about a topic might influence their rate of mind wandering [109], [111]–[113]. Our results also indicated that models in Study 1 were significantly more accurate for male than female students, an effect driven by differences in recall, which implies that demographic differences are worthy of further exploration.

Another limit pertains to our instance extraction scheme in Study 1, which required discarding a large number of clips (1,339) because they contained a page turn event. This process limits the mind wandering detectors to function only in situations with no such events. However, this limitation is necessary to avoid the possibility of models simply detecting movement associated with page-turn keypresses (a trivial task), which in turn might be related to mind wandering, but only in a highly task-specific way. On the other hand, data discarded due to undetectable facial features might be improved upon in future work, by imputing missing values (e.g., with a Kalman filter).

4.3 Applications

The mind wandering detection approach reported here represent the first automatic face-based mind wandering detection in a laboratory and in a classroom. The results we presented indicate that mind wandering can be detected at levels above chance – though far from perfectly. Although more research is needed to ascertain a plausible upper-bound for mind wandering detection accuracy, the current level of accuracy is likely sufficient to support fail-soft, probabilistic interventions that utilize these detectors in computerized learning environments. For example, a computerized reading environment with an automated mind wandering detector could recommend a break if repeated mind wandering is detected. Similarly, brief test questions could be inserted into a learning session for occasional instances of detected mind wandering. The learning environment can also sense mind wandering passively and provide class-level aggregates (by leveraging the principle of aggregation to improve reliability of noisy signals [114]) of students' attentional levels to teachers of instruction designers to guide pedagogy. Thus, the next critical step is to use the detectors in these and other ways in order to provide a more enjoyable, efficient, and effective learning experience for all students.

Acknowledgment

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 NSF.

References

[1] J. W. Schooler, J. Smallwood, K. Christoff, T. C. Handy, E. D. Reichle, and M. A. Sayette, "Meta-awareness, perceptual decoupling and the wandering mind," *Trends in Cognitive Sciences*, vol. 15, no. 7, pp. 319–326, Jul. 2011.

[2] M. A. Killingsworth and D. T. Gilbert, "A wandering mind is an unhappy mind," *Science*, vol. 330, no. 6006, pp. 932–932, Nov. 2010.

[3] J. G. Randall, F. L. Oswald, and M. E. Beier, "Mind-wandering, cognition, and performance: A theory-driven meta-analysis of

attention regulation.," *Psychological bulletin*, vol. 140, no. 6, pp. 1411–1431, 2014.

[4] S. K. D'Mello, "What do we think about when we learn?," in *Understanding Deep Learning, Educational Technologies and Deep Learning, and Assessing Deep Learning*, K. Millis, J. Magliano, D. Long, and K. Wiemer, Eds. Routledge/Taylor and Francis, in press.

[5] J. Smallwood, D. J. Fishman, and J. W. Schooler, "Counting the cost of an absent mind: Mind wandering as an underrecognized influence on educational performance," *Psychonomic Bulletin & Review*, vol. 14, no. 2, pp. 230–236, Apr. 2007.

[6] J. Smallwood, E. Beach, J. W. Schooler, and T. C. Handy, "Going AWOL in the brain: Mind wandering reduces cortical analysis of external events," *Journal of Cognitive Neuroscience*, vol. 20, no. 3, pp. 458–469, 2008.

[7] J. Smallwood, M. McSpadden, and J. W. Schooler, "When attention matters: The curious incident of the wandering mind," *Memory & Cognition*, vol. 36, no. 6, pp. 1144–1150, Sep. 2008.

[8] J. C. McVay and M. J. Kane, "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*, vol. 35, no. 1, pp. 196–204, 2009.

[9] L. M. Giambra, "A laboratory method for investigating influences on switching attention to task-unrelated imagery and thought," *Consciousness and Cognition*, vol. 4, no. 1, pp. 1–21, Mar. 1995.

[10] J. Smallwood and J. W. Schooler, "The restless mind," *Psychological Bulletin*, vol. 132, no. 6, pp. 946–958, 2006.

[11] J. C. McVay and M. J. Kane, "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*, vol. 38, no. 3, pp. 525–549, May 2012.

[12] J. C. McVay, M. J. Kane, and T. R. Kwapil, "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 & Review*, vol. 16, no. 5, pp. 857–863, Oct. 2009.

[13] P. Seli, E. F. Risko, and D. Smilek, "On the necessity of distinguishing between unintentional and intentional mind wandering," *Psychological Science*, vol. 27, no. 5, pp. 685–691, May 2016.

[14] J. D. Eastwood, A. Frischen, M. J. Fenske, and D. Smilek, "The unengaged mind: Defining boredom in terms of attention," *Perspectives on Psychological Science*, vol. 7, no. 5, pp. 482–495, Sep. 2012.

[15] J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris, "School engagement: Potential of the concept, state of the evidence," *Review of Educational Research*, vol. 74, no. 1, pp. 59–109, 2004.

[16] S. K. D'Mello, E. Dieterle, and A. L. Duckworth, "Advanced, analytic, automated (AAA) measurement of engagement during learning," *Educational Psychologist*, vol. 52, no. 2, pp. 104–123, Apr. 2017.

[17] J. Smallwood, "Distinguishing how from why the mind wanders: A process–occurrence framework for self-generated mental activity.," *Psychological bulletin*, vol. 139, no. 3, pp. 519–535, 2013.

[18] J. C. McVay and M. J. Kane, "Does mind wandering reflect executive function or executive failure? Comment on Smallwood and Schooler (2006) and Watkins (2008)," *Psychological Bulletin*, vol. 136, no. 2, pp. 188–197, 2010.

[19] J. Smallwood, "Why the global availability of mind wandering necessitates resource competition: Reply to McVay and Kane (2010).," *Psychological Bulletin*, vol. 136, no. 2, pp. 202–207, 2010.

[20] J. Smallwood and J. W. Schooler, "The science of mind wandering: Empirically navigating the stream of consciousness," *Annual Review of Psychology*, vol. 66, pp. 487–518, 2015.

[21] M. Faber and S. K. D'Mello, "How the stimulus influences mind wandering in semantically rich task contexts," *Cognitive Research: Principles and Implications*, vol. 3, no. 35, pp. 1–14, 2018.

[22] J. Smallwood, A. Fitzgerald, L. K. Miles, and L. H. Phillips, "Shifting moods, wandering minds: Negative moods lead the mind to wander," *Emotion*, vol. 9, no. 2, pp. 271–276, 2009.

[23] G. L. Poerio, P. Totterdell, and E. Miles, "Mind-wandering and negative mood: Does one thing really lead to another?," *Consciousness and Cognition*, vol. 22, no. 4, pp. 1412–1421, Dec. 2013.

[24] J. Smallwood, R. C. O'Connor, M. V. Sudbery, and M. Obonsawin, "Mind-wandering and dysphoria," *Cognition and Emotion*, vol. 21, no. 4, pp. 816–842, Jun. 2007.

[25] P. Seli, J. Smallwood, J. A. Cheyne, and D. Smilek, "On the relation of mind wandering and ADHD symptomatology," *Psychonomic Bulletin & Review*, vol. 22, no. 3, pp. 629–636, Jun. 2015.

[26] E. F. Risko, N. Anderson, A. Sarwal, M. Engelhardt, and A. Kingstone, "Everyday attention: Variation in mind wandering and memory in a lecture," *Applied Cognitive Psychology*, vol. 26, no. 2, pp. 234–242, Mar. 2012.

[27] B. W. Mooneyham and J. W. Schooler, "The costs and benefits of mind-wandering: A review," *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, vol. 67, no. 1, pp. 11–18, 2013.

[28] B. Baird, J. Smallwood, M. D. Mrazek, J. W. Y. Kam, M. S. Franklin, and J. W. Schooler, "Inspired by distraction: Mind wandering facilitates creative incubation," *Psychological Science*, vol. 23, no. 10, pp. 1117–1122, Oct. 2012.

[29] C. Roda and J. Thomas, "Attention aware systems: Theories, applications, and research agenda," *Computers in Human Behavior*, vol. 22, no. 4, pp. 557–587, Jul. 2006.

[30] S. K. D'Mello and A. C. Graesser, "Feeling, thinking, and computing with affect-aware learning technologies," in *Handbook of Affective Computing*, R. A. Calvo, S. K. D'Mello, J. Gratch, and A. Kappas, Eds. New York, NY: Oxford University Press, 2015, pp. 419–434.

[31] S. K. D'Mello, "Giving eyesight to the blind: Towards attention-aware AIED," *International Journal of Artificial Intelligence in Education*, vol. 26, no. 2, pp. 645–659, Jun. 2016.

[32] D. N. Rapp, "The value of attention aware systems in educational settings," *Computers in Human Behavior*, vol. 22, no. 4, pp. 603–614, Jul. 2006.

[33] K. K. Szpunar, N. Y. Khan, and D. L. Schacter, "Interpolated memory tests reduce mind wandering and improve learning of online lectures," *PNAS*, vol. 110, no. 16, pp. 6313–6317, Apr. 2013.

[34] S. K. D'Mello, C. Mills, R. Bixler, and N. Bosch, "Zone out no more: Mitigating mind wandering during computerized reading," in *Proceedings of the 10th International Conference on Educational Data Mining (EDM 2017)*, 2017, pp. 8–15.

[35] H. Monkaresi, N. Bosch, R. A. Calvo, and S. K. D'Mello, "Automated detection of engagement using video-based estimation of facial expressions and heart rate," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 15–28, 2017.

[36] J. F. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, and J. C. Lester, "Automatically recognizing facial expression: Predicting engagement and frustration," in *Proceedings of the 6th International Conference on Educational Data Mining*, 2013.

[37] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagement from facial expressions," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 86–98, Jan. 2014.

[38] J. E. Beck, "Engagement tracing: Using response times to model student disengagement," in *Proceedings of the 12th International Conference on Artificial Intelligence in Education (AIED 2015)*, 2005, pp. 88–95.

[39] J. F. Grafsgaard, J. B. Wiggins, A. K. Vail, K. E. Boyer, E. N. Wiebe, and J. C. Lester, "The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring," in *Proceedings of the 16th International Conference on Multimodal Interaction*, 2014, pp. 42–49.

[40] C. Mills, N. Bosch, A. Graesser, and S. K. D'Mello, "To quit or not to quit: Predicting future behavioral disengagement from reading patterns," in *Proceedings of the 12th International Conference on Intelligent Tutoring Systems (ITS 2014)*, Cham, CH, 2014, pp. 19–28.

[41] J. A. Walonoski and N. T. Heffernan, "Detection and analysis of off-task gaming behavior in intelligent tutoring systems," in *Proceedings of the 8th International Conference on Intelligent Tutoring Systems (ITS 2006)*, 2006, pp. 382–391.

[42] S. Cetintas, L. Si, Y. P. P. Xin, and C. Hord, "Automatic detection of off-task behaviors in intelligent tutoring systems with machine learning techniques," *IEEE Transactions on Learning Technologies*, vol. 3, no. 3, pp. 228–236, Jul. 2010.

[43] J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P. W. McOwan, and A. Paiva, "Automatic analysis of affective postures and body motion to detect engagement with a game companion," in *6th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2011)*, 2011, pp. 305–311.

[44] M. Raca, L. Kidzinski, and P. Dillenbourg, "Translating head motion into attention - Towards processing of student's body-language," in *Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015)*, 2015, pp. 320–326.

[45] N. Bosch, S. K. D'Mello, J. Ocumpaugh, R. S. Baker, and V. Shute, "Using video to automatically detect learner affect in computer-enabled classrooms," *ACM Transactions on Interactive Intelligent*

Systems (TiiS), vol. 6, no. 2, pp. 17:1-17:26, 2016.
[46] M. Yeasin, B. Bullot, and R. Sharma, "From facial expression to level of interest: A spatio-temporal approach," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004)*, 2004, vol. 2, pp. 922–927.
[47] N. Jaques, C. Conati, J. M. Harley, and R. Azevedo, "Predicting affect from gaze data during interaction with an intelligent tutoring system," in *Proceedings of the 12th International Conference on Intelligent Tutoring Systems (ITS 2014)*, 2014, pp. 29–38.
[48] S. K. D'Mello, S. D. Craig, A. Witherspoon, B. McDaniel, and A. Graesser, "Automatic detection of learner's affect from conversational cues," *User Modeling and User-Adapted Interaction*, vol. 18, no. 1–2, pp. 45–80, 2008.
[49] R. A. Calvo and S. K. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 18–37, Jan. 2010.
[50] D. Smilek, J. S. A. Carriere, and J. A. Cheyne, "Out of mind, out of sight: Eye blinking as indicator and embodiment of mind wandering," *Psychological Science*, vol. 21, no. 6, pp. 786–789, 2010.
[51] T. D. Loboda, "Study and detection of mindless reading," University of Pittsburgh, 2014.
[52] R. Bixler and S. K. D'Mello, "Automatic gaze-based detection of mind wandering with metacognitive awareness," in *Proceedings of the 23rd International Conference on User Modeling, Adaptation and Personalization (UMAP 2015)*, 2015, pp. 31–43.
[53] S. K. D'Mello, K. Kopp, R. Bixler, and N. Bosch, "Attending to attention: Detecting and combating mind wandering during computerized reading," in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, New York, NY, 2016, pp. 1661–1669.
[54] S. K. D'Mello, J. Cobian, and M. Hunter, "Automatic gaze-based detection of mind wandering during reading," in *Proceedings of the 6th International Conference on Educational Data Mining*, 2013.
[55] R. Bixler and S. K. D'Mello, "Toward fully automated person-independent detection of mind wandering," in *User Modeling, Adaptation, and Personalization*, 2014, pp. 37–48.
[56] S. Hutt, C. Mills, N. Bosch, K. Krasich, J. Brockmole, and S. K. D'Mello, "Out of the fr-"eye"-ing pan: Towards gaze-based models of attention during learning with technology in the classroom," in *Proceedings of the 2017 Conference on User Modeling, Adaptation, and Personalization (UMAP 2017)*, New York, NY, 2017, pp. 94–103.
[57] N. Blanchard, R. Bixler, T. Joyce, and S. K. D'Mello, "Automated physiological-based detection of mind wandering during learning," in *Proceedings of the 12th International Conference on Intelligent Tutoring Systems (ITS 2014)*, 2014, pp. 55–60.
[58] P. Pham and J. Wang, "AttentiveLearner: Improving mobile MOOC learning via implicit heart rate tracking," in *Artificial Intelligence in Education*, 2015, pp. 367–376.
[59] M. S. Franklin, J. Smallwood, and J. W. Schooler, "Catching the mind in flight: Using behavioral indices to detect mindless reading in real time," *Psychonomic Bulletin & Review*, vol. 18, no. 5, pp. 992–997, Oct. 2011.
[60] C. Mills and S. K. D'Mello, "Toward a real-time (day) dreamcatcher: Sensor-free detection of mind wandering during online reading," in *Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015)*, 2015, pp. 69–76.
[61] A. Stewart, N. Bosch, H. Chen, P. J. Donnelly, and S. K. D'Mello, "Face forward: Detecting mind wandering from video during narrative film comprehension," in *Proceedings of the 18th International Conference on Artificial Intelligence in Education (AIED 2017)*, Cham, CH, 2017, pp. 359–370.
[62] G. Littlewort *et al.*, "The computer expression recognition toolbox (CERT)," in *2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011)*, 2011, pp. 298–305.
[63] J. K. Westlund, S. K. D'Mello, and A. M. Olney, "Motion Tracker: Camera-based monitoring of bodily movements using motion silhouettes," *PLoS ONE*, vol. 10, no. 6, Jun. 2015.
[64] A. Stewart, N. Bosch, and S. K. D'Mello, "Generalizability of face-based mind wandering detection across task contexts," in *Proceedings of the 10th International Conference on Educational Data Mining (EDM 2017)*, 2017, pp. 88–95.
[65] K. Rayner, "Eye movements in reading and information processing: 20 years of research.," *Psychological Bulletin*, vol. 124, no. 3, pp. 372–422, 1998.
[66] K. Kopp, S. K. D'Mello, and C. Mills, "Influencing the occurrence of mind wandering while reading," *Consciousness and Cognition*, vol. 34, pp. 52–62, Jul. 2015.
[67] C. V. Boys, *Soap-bubbles, and the forces which mould them*. London, England: Society for Promoting Christian Knowledge, 1890.
[68] P. Seli, J. S. A. Carriere, D. R. Thomson, J. A. Cheyne, K. A. E. Martens, and D. Smilek, "Restless mind, restless body," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 40, no. 3, pp. 660–668, 2014.
[69] J. Drummond and D. Litman, "In the zone: Towards detecting student zoning out using supervised machine learning," in *Intelligent Tutoring Systems*, 2010, pp. 306–308.
[70] K. Christoff, A. M. Gordon, J. Smallwood, R. Smith, and J. W. Schooler, "Experience sampling during fMRI reveals default network and executive system contributions to mind wandering," *Proceedings of the National Academy of Sciences*, vol. 106, no. 21, pp. 8719–8724, May 2009.
[71] M. Mittner, W. Boekel, A. M. Tucker, B. M. Turner, A. Heathcote, and B. U. Forstmann, "When the brain takes a break: A model-based analysis of mind wandering," *Journal of Neuroscience*, vol. 34, no. 49, pp. 16286–16295, Dec. 2014.
[72] R. G. O'Connell, P. M. Dockree, I. H. Robertson, M. A. Bellgrove, J. J. Foxe, and S. P. Kelly, "Uncovering the neural signature of lapsing attention: Electrophysiological signals predict errors up to 20 s before they occur," *Journal of Neuroscience*, vol. 29, no. 26, pp. 8604–8611, Jul. 2009.
[73] D. H. Weissman, K. C. Roberts, K. M. Visscher, and M. G. Woldorff, "The neural bases of momentary lapses in attention," *Nature Neuroscience*, vol. 9, no. 7, pp. 971–978, Jun. 2006.
[74] T. Foulsham, J. Farley, and A. Kingstone, "Mind wandering in sentence reading: Decoupling the link between mind and eye.," *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, vol. 67, no. 1, p. 51, 2013.
[75] D. J. Frank, B. Nara, M. Zavagnin, D. R. Touron, and M. J. Kane, "Validating older adults' reports of less mind-wandering: An examination of eye movements and dispositional influences.," *Psychology and Aging*, vol. 30, no. 2, p. 266, 2015.
[76] E. D. Reichle, A. E. Reineberg, and J. W. Schooler, "Eye movements during mindless reading," *Psychological Science*, vol. 21, no. 9, pp. 1300–1310, Sep. 2010.
[77] S. Uzzaman and S. Joordens, "The eyes know what you are thinking: Eye movements as an objective measure of mind wandering," *Consciousness and Cognition*, vol. 20, no. 4, pp. 1882–1886, Dec. 2011.
[78] M. S. Franklin, J. M. Broadway, M. D. Mrazek, J. Smallwood, and J. W. Schooler, "Window to the wandering mind: Pupillometry of spontaneous thought while reading," *The Quarterly Journal of Experimental Psychology*, vol. 66, no. 12, pp. 2289–2294, Dec. 2013.
[79] K. Krasich, R. McManus, S. Hutt, M. Faber, S. K. D'Mello, and J. Brockmole, "Gaze-based signatures of mind wandering during real-world scene processing," *Journal of Experimental Psychology: General*, in press.
[80] M. Faber, R. Bixler, and S. K. D'Mello, "An automated behavioral measure of mind wandering during computerized reading," *Behavior Research Methods*, vol. 50, no. 1, pp. 134–150, 2018.
[81] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
[82] T. Baltrušaitis, P. Robinson, and L. P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1–10.
[83] G. C. Littlewort, M. S. Bartlett, L. P. Salamanca, and J. Reilly, "Automated measurement of children's facial expressions during problem solving tasks," in *Proceedings of the 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG 2011)*, 2011, pp. 30–35.
[84] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *Journal of Multimedia*, vol. 1, no. 6, pp. 22–35, 2006.
[85] D. S. Messinger, A. Fogel, and K. L. Dickson, "All smiles are positive, but some smiles are more positive than others," *Developmental Psychology*, vol. 37, no. 5, pp. 642–653, 2001.
[86] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, Jan. 1991.
[87] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, Mar. 1951.

[88] M. S. Bartlett, G. C. Littlewort, M. G. Frank, and K. Lee, "Automatic decoding of facial movements reveals deceptive pain expressions," *Current Biology*, vol. 24, no. 7, pp. 738–743, Mar. 2014.
[89] I. Buciu, C. Kotropoulos, and I. Pitas, "ICA and Gabor representation for facial expression recognition," in *Proceedings of the 2003 International Conference on Image Processing (ICIP 2003)*, 2003, vol. II, pp. 855–858.
[90] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on Local Binary Patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, May 2009.
[91] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.
[92] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 1097–1105.
[93] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995.
[94] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, Mar. 2003.
[95] I. Kononenko, "Estimating attributes: Analysis and extensions of RELIEF," in *European Conference on Machine Learning (ECML 94)*, F. Bergadano and L. D. Raedt, Eds. Berlin Heidelberg: Springer, 1994, pp. 171–182.
[96] M. A. Hall, "Correlation-based feature selection for machine learning," PhD Thesis, Waikato University, New Zealand, 1999.
[97] S. Parodi, V. Pistoia, and M. Muselli, "Not proper ROC curves as new tool for the analysis of differentially expressed genes in microarray experiments," *BMC Bioinformatics*, vol. 9, no. 410, Oct. 2008.
[98] J. Hua, Z. Xiong, J. Lowey, E. Suh, and E. R. Dougherty, "Optimal number of features as a function of sample size for various classification rules," *Bioinformatics*, vol. 21, no. 8, pp. 1509–1515, Apr. 2005.
[99] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
[100] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference for Learning Representations (ICLR 2015)*, San Diego, CA, 2015.
[101] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and regression trees," Belmont, CA, 1984.
[102] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *arXiv:1406.5823 [stat]*, Jun. 2014.
[103] R. Core Team, "R: A language and environment for statistical computing," 2013.
[104] J. Fox, M. Friendly, and S. Weisberg, "Hypothesis tests for multivariate linear models using the car package," *The R Journal*, vol. 5, no. 1, pp. 39–52, 2013.
[105] R. Lenth, "Emmeans: Estimated marginal means, aka least-squares means," *R package*, 2017.
[106] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing user studies with Mechanical Turk," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2008, pp. 453–456.
[107] A. Olney *et al.*, "Guru: A computer tutor that models expert human tutors," in *Proceedings of the 11th International Conference on Intelligent Tutoring Systems (ITS 2012)*, 2012, pp. 256–261.
[108] R. Sottilare, A. Graesser, X. Hu, and H. Holden, *Design recommendations for intelligent tutoring systems: Volume 1 - Learner modeling*. U.S. Army Research Laboratory, 2013.
[109] C. Mills, S. K. D'Mello, N. Bosch, and A. Olney, "Mind wandering during learning with an intelligent tutoring system," in *Proceedings of the 17th International Conference on Artificial Intelligence in Education (AIED 2015)*, Cham, CH, 2015, pp. 267–276.
[110] S. Hutt *et al.*, "Gaze-based models of mind wandering in classrooms," *User Modeling and User-Adapted Interaction*, in review.
[111] N. Unsworth and B. D. McMillan, "Mind wandering and reading comprehension: Examining the roles of working memory capacity, interest, motivation, and topic experience," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 39, no. 3, pp. 832–842, 2013.
[112] A. Grodsky and L. M. Giambra, "The consistency across vigilance and reading tasks of individual differences in the occurrence of task-unrelated and task-related images and thoughts," *Imagination, Cognition and Personality*, vol. 10, no. 1, pp. 39–52, Sep. 1990.
[113] S. I. Lindquist and J. P. McLean, "Daydreaming and its correlates in an educational environment," *Learning and Individual Differences*, vol. 21, no. 2, pp. 158–167, Apr. 2011.
[114] R. Rosenthal, "Conducting judgment studies: Some methodological issues," in *The New Handbook of Methods in Nonverbal Behavior Research*, J. A. Harrigan, R. Rosenthal, and K. R. Scherer, Eds. New York, NY: Oxford University Press, 2005, pp. 199–234.

Nigel Bosch is an Assistant Professor in the School of Information Sciences and the Department of Educational Psychology at the University of Illinois at Urbana-Champaign. His research interests involve affect detection from video-based signals in learning contexts and applying affect detection as a means to study the affective component of learning. In 2017 he received his PhD in Computer Science from the University of Notre Dame, where he was advised by Sidney D'Mello.

Sidney D'Mello (PhD in Computer Science) is an Associate Professor in the Institute of Cognitive Science and Department of Computer Science at the University of Colorado Boulder. He is interested in the dynamic interplay between cognition and emotion while individuals and groups engage in complex real-world tasks. D'Mello has co-edited seven books and published over 250 journal papers, book chapters, and conference proceedings (13 of these have received awards). His work has been funded by numerous grants and he serves(d) as associate editor or on the editorial boards of ten journals.