

Can Computers Outperform Humans in Detecting User Zone-Outs? Implications for Intelligent Interfaces

Nigel Bosch

School of Information Sciences and Department of Educational Psychology,
University of Illinois at Urbana-Champaign, Champaign, IL, USA, pnb@illinois.edu

Sidney K. D’Mello

Department of Computer Science and Institute of Cognitive Science,
University of Colorado Boulder, Boulder, CO, USA, sidney.dmello@colorado.edu

ABSTRACT

The ability to identify whether a user is “zoning out” (mind wandering) from video has many applications (e.g., distance learning, high-stakes vigilance tasks). However, it remains unknown how well humans can perform this task, how they compare to automatic computerized approaches, and how a fusion of the two might improve accuracy. We analyzed videos of users’ faces and upper bodies recorded 10s prior to self-reported mind wandering (i.e., ground truth) while they engaged in a computerized reading task. We found that a state-of-the-art machine learning model had comparable accuracy to aggregated judgments of nine untrained human observers (area under receiver operating characteristic curve [AUC] = .598 versus .589). A fusion of the two (AUC = .644) outperformed each, presumably because each focused on complementary cues. Further, adding more humans beyond 3-4 observers yielded diminishing returns. We discuss implications of human-computer fusion as a means to improve accuracy in complex tasks.

CCS CONCEPTS

• **Human-centered computing~Empirical studies in HCI** • *Computing methodologies~Computer vision* • Applied computing~Computer-assisted instruction

KEYWORDS

Mind wandering, Human-machine comparison, Facial expression recognition, Attention-aware interfaces

1 Introduction

Mind wandering, or “zoning out”, occurs when a person (in our case, the user of a computer interface) involuntarily shifts their attentional focus from the task at hand to task-unrelated thoughts [17,84]¹. These thoughts might involve memories of the past, current or future concerns, introspection, and other thoughts [37,85]. For example, when asked to report the contents of their thoughts during a mind wandering episode and the preceding trigger, a participant in a study [37] reported: “[Reading the word soap]_{TRIGGER} [reminded me that] [I need to give my dog a bath and make an appointment for him to get his nails clipped]_{CONTENT}”). Similarly, a user might be overwhelmed with thoughts of the future, such as worries over an upcoming exam, and struggle to keep attention focused on the current task.

Mind wandering is estimated to occur as much as 50% of the time as people engage in everyday activities [53]. It is also frequent during human–computer interactions, such as electronic textbook (e-text) reading, video watching, automobile driving, classification tasks, and others [43,50,61,73,84,98]. Although the trait to mind wander has been associated with positive outcomes like creativity and planning for the future [1,63], meta-analyses indicate that it is consistently negatively related to performance on a variety of information processing lab- and real-world- tasks [24,71] that require sustained attentional focus. Mind wandering is

¹ Although we study unintentional attentional shifts in this paper, there is debate about whether mind wandering can also include intentional shifts of attentional focus [81].

similarly negatively related to performance on HCI relevant-tasks, such as computer use in classrooms, note-taking during lectures, watching online videos, and others [50,74,91,98].

Accordingly, recognizing if a user is mind wandering has widespread applications in human-computer interactions. In the following examples, mind wandering might be measured by human observers (example 1 and 2) or an automated computer vision algorithm (examples 3 and 4). In some cases either human observation or automated methods are feasible (examples 5 and 6), as is a combination of the two.

- 1) Instructors of an online classes assesses how engaging their lectures are by determining whether their students are zoning out
- 2) A supervisor at an airport is monitoring the baggage screeners to determine whether they are focused or have zoned out.
- 3) Students studying for an exam use a computerized mind wandering detection method to help determine which content to revisit based on when their attention may have waned
- 4) Air traffic control software measures mind wandering over time to inform a policy of taking breaks to avoid waning vigilance
- 5) An e-book publisher improves a draft manuscript by revising sections of the text that induce mind wandering in users
- 6) A movie studio compares alternative cuts of a film in A/B tests to determine whether viewers mind wander more during one of the cuts

Both human observers and computerized methods can be used to detect mind wandering. To get a sense of the task, consider Figure 1, recorded while users were reading text on a computer screen. Are they deeply engaged, superficially engaged, or have they zoned out? Most people are correct in identifying that person A is diligently reading while the person B has zoned out. What about C and D? Does the fact that D is yawning indicate that they are disengaged or just fatigued? And what can be concluded from the somewhat unusual gesture depicted by C? It appears that most people are confused by the displays in C and D, guessing zoned out about half the time and attentively reading for the other half.

As these examples illustrate, mind wandering detection can be quite a challenging task. Whereas computerized approaches can use specialized sensing such as eye trackers and physiological sensing [9,11,50,68], human observations typically focus on visual cues [64,96]. This raises the question of whether humans are capable of detecting users' mind wandering from visual cues, and how computerized approaches stack up when limited to the same visual cues. Would computers be more or less accurate than human observers, who have finely evolved abilities to read social cues? Is a combination of observer and computer predictions more accurate than either individually?

Of course, we do not expect there are canonical, one-to-one mappings between users' experiences of mind wandering and their facial expressions, but we do expect that context-specific expressions of mind wandering may arise during human-computer interaction tasks. For example, users reading text on a computer screen may blink more often while mind wandering [46,86]. These expressions can perhaps be assessed, with a modicum of accuracy, by observers (including other users, as in example 1 above) and automated methods. We thus seek to examine context-specific computer models for detecting mind wandering and compare them to human observers' perceptions, which are informed by the context, but not necessarily limited to it given their enhanced social perception capabilities compared to computers.

1.1 Research Questions

In the present study we compare a computer algorithm to untrained human observers on the task of deciding whether users depicted in short video clips (similar to Figure 1) are mind wandering or not. We address three research questions (RQs), each of which has implications for HCI research or practice.

RQ1. How accurately can human observers detect mind wandering from videos of users engaged in computer-mediated tasks and how does their accuracy compare to automatic methods?

Implications for HCI: The question of how machines stack up to human observers is significant because computer systems that perform measurement of complex psychological constructs like emotion, social communication, psychopathologies, and similar are on the rise [6,13,14,26,30]. However, in the case of mind

wandering, it is unclear how accurate we might expect automated methods to be, and whether collaboration between computers and humans can improve accuracy. Comparing the accuracy of computers to observers on the same task also provides a useful measure of task difficulty and can serve to establish approximate guidelines on the accuracy of automated systems, which can then inform how they can be used.

RQ2. Does a fusion of human- and machine- predictions outperform either one independently?

Implications for HCI: A human-in-the-loop fusion approach would be valuable if it improves accuracy or if equivalent accuracy can be obtained at a lower cost. Fusion could occur in multiple ways, such as joint prediction for all cases, or a two-step approach in which observers verify machine predictions for specific cases (e.g., positive predictions or ambiguous cases). The latter possibility offers a compromise between accuracy and automation to reduce the burden of human observation.

RQ3. What visual cues do observers utilize to make their decisions, and how are those cues complementary to those used by the machine?

Implications for HCI: Discovering the cues observers and machines utilize can increase transparency by informing users how and why the machine offers a complementary perspective to humans and why a fusion approach might be effective. It may also inform future automatic methods by highlighting different types of information that could be extracted from videos.

1.2 Novelty & Contribution

The present study is concerned with comparing computers to observers on the task of detecting mind wandering from video, as well as investigating a hybrid method that combines observer and computer judgments. Our research is novel in several ways. While previous research has explored automatic mind wandering detection from various sources of data, including video [11,87,88], this is the first in-depth study of how well humans perform the same task (RQ1). These analyses contribute to a growing body of research into the specific situations where machines do comparatively well (or poorly) compared to observers [5,48,52,77,99,100]. Here, we find that the observers and computers outperform each other on different components of the task, suggesting that the two attend to different aspects of the videos or make different tradeoffs. The difficulty faced by observers also places some reasonable bounds on what computers might eventually achieve. This study is also the first to measure and compare fusion of human and machine judgments for mind wandering detection (RQ2), where we find that fusion does indeed improve accuracy, further implying that observer and machine judgments utilize the same data somewhat differently. To delve into this matter, we used natural language processing to analyze textual justifications provided by observers (RQ3), which revealed insights into specific cues that humans rely on, biases they might bring with them. These analyses also suggested some of the behavioral correlates of mind wandering that can be perceived by humans versus those used by the computer.

1.3 Current Study

We use data from a previous study that recorded videos of computer users reading an e-text for approximately 30-mins [54]. Users self-reported mind wandering as they read (see Section 3.1 for details on the validity of self-reports to measure mind wandering), which we used as ground-truth labels to train machine learning models that automatically classify 10-second video clips prior to the self-reports as positive (mind wandering) or negative (not mind wandering) [11]. We collected third-party mind wandering ratings of the same video clips from nine human judges (henceforth referred to as observers or humans) recruited from Amazon Mechanical Turk (AMT). We compared the observers' accuracy to machine accuracy by comparing estimates of each to self-reported mind wandering of the original users. Similarly, we compared a computer-human fusion, which incorporates predictions from both as well as examine the minimum number of human observers needed for the task. We also asked the observers to provide brief justifications for their classification of each clip and used natural language processing techniques to uncover systematicities in their explanations to potentially identify sources of discrepancies and/or commonalities between computers and observers. These explanations provide insight into how observers and the computer make their decisions,

how they operationalize mind wandering in terms of visual cues, and ways in which human and computer decisions are complementary.

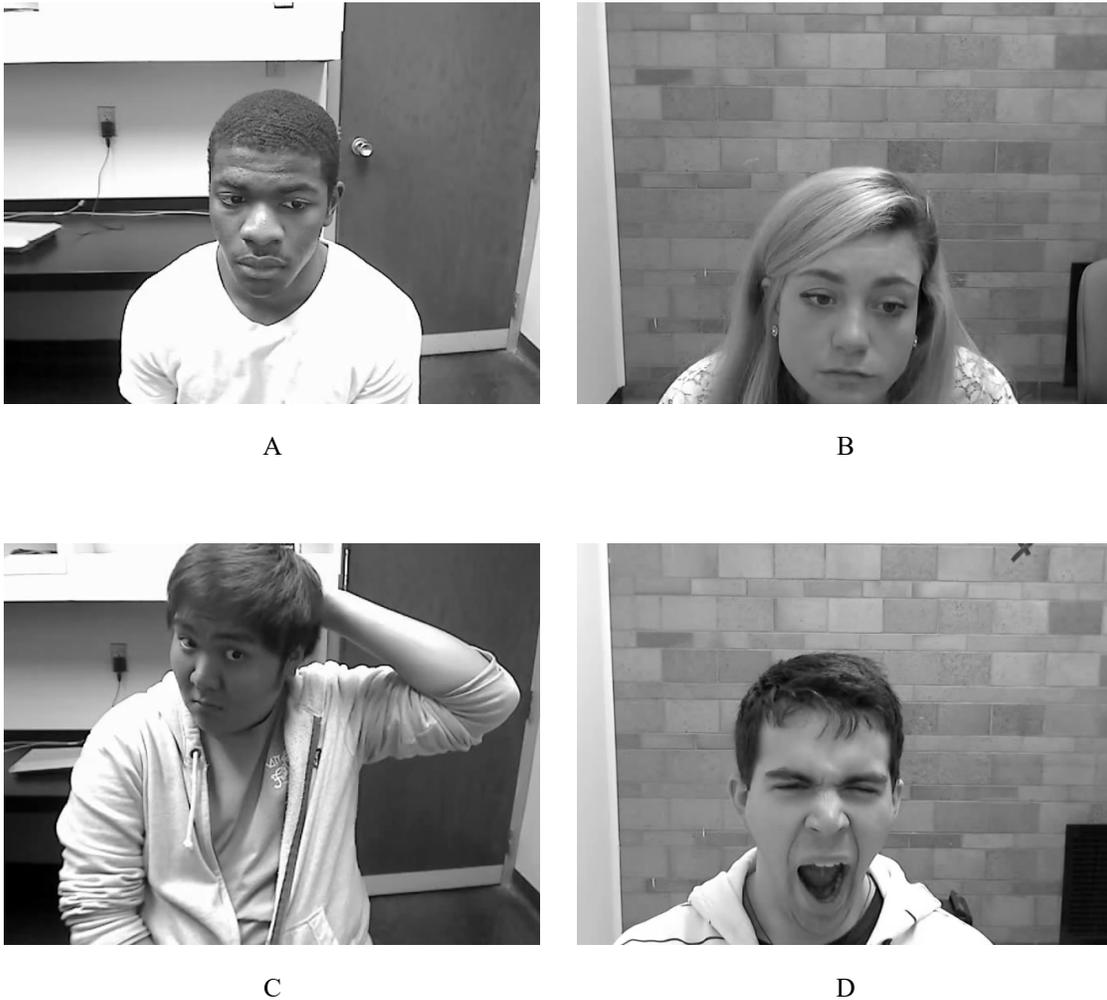


Figure 1: Users who reported (A) paying attention, (B) mind wandering, and – less obviously – (C & D) paying attention.

2 Related Work

To date, no study has compared how accurately human observers can detect mind wandering, let alone compared observers to automated methods or human–computer decision fusion. However, comparative detection accuracy of observers vs. computers has been assessed with respect to other psychological constructs, which might provide some guidance.

2.1 Accuracy of Human Detection of Related Psychological Constructs

Mind wandering is closely related to other psychological constructs that are relevant in the context of HCI applications, including engagement – a broader construct consisting of affective, cognitive, and behavioral components [25,42]. Mind wandering falls primarily under the cognitive component of engagement, and does not necessarily have the related behavioral or affective characteristics (though research suggests mind wandering correlates negatively with happiness [53]). Research on assessing engagement based on markers

like looking at the screen, which can easily be assessed by human annotators [96], captures engagement defined in terms of overt behaviors. A computer user who is mind wandering, on the other hand, may be gazing at the screen yet not be cognitively engaged. There may indeed be behavioral markers of mind wandering (see RQ3), but they are likely more difficult to observe since they are correlates and byproducts of cognitive (not behavioral) disengagement. Mind wandering also relates to research on human observations of whether a user's thoughts are directed inward or outward. Research has shown that observers were able to judge the direction of computer users' thoughts while performing internally- or externally-directed tasks, and that observers relied especially on users' eyes to make judgments [7]. However, mind wandering, though internal [84], consists of off-task thought, rather than internally-focused on-task thought.

Though mind wandering is primarily cognitive, it may have emotional correlates that have been investigated more thoroughly. For example, we expect mind wandering to negatively associated with affective states such as confusion, frustration, and interest [2,22,27], and positively associated with boredom and sadness [2]. Previous work on emotion recognition suggests that observers outperform automatic (computer) methods in some contexts. Note that context is critical, since facial expressions vary depending on context [21]. An early study by Schindler, Van Gool, & de Gelder [77] found that observers were 6% more accurate than a computerized method for classifying stereotypical, posed expressions of emotion in images. In a more recent study, Holkamp & Schavemaker [48] extracted video clips of participants who watched affectively-charged videos and then self-reported their valence levels. They recruited 15 third-party human observers, who judged valence levels from facial expressions using a custom web-based annotation tool. Researchers compared observer ratings to computer vision methods, finding that observers outperformed the computer (75% versus 66% accuracy in terms of alignment with self-reports). Furthermore, Yitzhak et al. [99] found that observers (undergraduate students) outperformed computers when it came to detecting subtle, non-stereotypical facial expressions. They noted that emotion detection research has traditionally focused on acted, exaggerated expressions of emotion, which can be easily recognized with computer vision methods. For acted emotions, they found that computers exhibited 89% accuracy in a 7-way classification task, while observers' annotation accuracy was similarly high with 88% accuracy. However, observers rated naturalistic expressions about 61% as intense as acted expressions. The computer vision system had a much lower (21% accuracy) compared to observers (79% accurate) on a naturalistic emotion dataset, where emotions were rated as less intense but more natural in appearance than acted expressions. Recently, researchers also compared human judges to eight different commercially-available facial expression recognition tools, and found that the observers significantly outperformed all of them for both acted and naturalistic emotion expressions, with much larger differences for naturalistic expressions [29].

There are also cases where computers have outperformed observers when detecting emotions from facial expressions [52], pain [5], and deception [100]. Janssen et al. [52] adopted a multimodal approach combining facial expressions, speech, and physiological data to automatically detect five *experimentally-elicited* emotions. The computerized approach was notably more accurate than observers on the same dataset (82% versus 63% accuracy). However, the comparison should be taken with a modicum of caution because the protocols may have been biased in favor of the computer as strict person-level independence between training and testing sets was not enforced. In more of an apples-to-apples comparison, Bartlett et al. [5] compared observers to computer vision methods for distinguishing between facial expressions reflecting genuine (experimentally-elicited) vs. fake pain. They found that observers' accuracy was at chance-levels (50%) without training and only increased to 55% with training. The computer easily outperformed the observers with an accuracy of 85%, which was attributed to its ability to model the subtle dynamics of mouth movements associated with genuine pain.

There is reason to expect that the aforementioned findings from facial expression recognition and experimentally induced emotions might not transfer to mind wandering. This is because some emotions have relatively robust visual indicators (e.g., eyebrow lowering during confusion; [59]) in service of their communication and social signaling functions [56]. Mind wandering, on the other hand, is an internal cognitive state not easily defined by observable behaviors [84]. Similar to deception, it is also likely socially advantageous to disguise displays of mind wandering, compared to, for example, displays of happiness (in

some circumstances). As such, it is unclear whether there are visual cues that indicate mind wandering. If so, what might those cues be?

2.2 Automatic Mind Wandering Detection

Recent research has revealed some insights into possible cues. One of the first studies that tracked mind wandering while participants read an e-text found that they blinked significantly more often when mind wandering compared to normal reading [86], thereby demonstrating one possible visible indicator. The same study found that participants were less likely to fixate their eyes on the same location frequently when mind wandering. Other studies have also revealed additional indicators of mind wandering such as fewer, longer gaze fixations and more irregular saccades (eye movements; [8,40,72]); however, gaze patterns are task-dependent in that reading will incur very different patterns than scene viewing [38]. It is unclear from previous research whether observers are able to visually identify these eye-gaze-related reading behaviors from videos. It is also unclear if these links between mind wandering and reading behaviors are context-free or might be moderated by differences in difficulty, length, presentation, and other aspects of the tasks, which range from reading entertaining fiction one word at a time to reading scientific text one page at a time [36,41]. It also remains to be seen if human observers can infer mind wandering from these indicators.

In lieu of eye gaze features, several of which might not be perceptible to human observers, researchers have leveraged facial features to automatically detect mind wandering. In one study, Stewart et al. [87] recorded videos of participants' faces as they watched a 35-min narrative film and self-reported when they caught themselves mind wandering. They utilized computer vision software [58] to extract facial action units (AUs), which reflect facial muscle activations including eyebrow lowering, cheek raising, lip corner lowering, and others. They also extracted body movement features from the videos and trained user-generalizable machine learning classifiers to detect mind wandering from body movement and facial AUs. Their best model achieved an F_1 score of .390, which was 13% above a random-chance baseline, demonstrating the possibility of face-based mind wandering detection, but also the difficulty of the task. Furthermore, the researchers found that head nodding and lack of facial muscle movement (i.e., neutral expression) were the clearest facial indicators of mind wandering.

Stewart et al. [88] expanded on their analyses by investigating how their mind wandering detection model trained on a narrative film viewing task generalized to a different task: reading an e-text. They found two AUs that were predictive across tasks (AU23: lip tightener, and AU26: jaw drop). However, a model with a large combination of features generalized more effectively than any individual features, suggesting that a complex combination of facial cues are likely required for mind wandering detection. Overall, this study reported mind wandering accuracy scores reflecting 25% improvements over random chance (within-task) and 21-22% above chance when generalizing detectors across interaction tasks.

2.3 Human–Machine Fusion

Research on face-based mind wandering detection (or related tasks) has not explored the possibility of human–machine decision fusion, though research in other domains suggests it may offer advantages. For example, in a person recognition task, researchers found that fusing algorithmic and human judgments enabled faster and more accurate identification of individuals in photographs from the American Civil War [62]. In another case, humans combined their own judgment with predictions from an algorithmic tool to improve their decisions about whether or not to recommend reports of child maltreatment for investigation [23]. A form of human–machine fusion called computer-assisted diagnosis (CAD) is prevalent in medical applications, where physicians combine their diagnoses with a computer's diagnosis to improve accuracy or reduce labor [28]. CAD has been successfully applied across a range of medical imaging domains, such as chest X-rays [93], magnetic resonance images of the brain [33], and mammograms [51], with some promising results. For example, accuracy of a computer prediction fused with a physician's prediction yielded similar accuracy to that of two physicians [44]. Researchers have also explored tradeoffs between accuracy and the time required of people to make annotations – for example, by dividing labor in different ways between humans and a machine system for detecting underwater mines in images [97], a topic we explore in this

paper. Specifically, this paper extends previous work on human–machine decision fusion by exploring a new problem domain (detecting mind wandering users in videos), and varying the amount of human involvement in decision making to quantify the effort vs. accuracy tradeoff in this domain.

2.4 Interim Summary

In summary, prior research has shown that facial features offer a promising means for automatic mind wandering detection including potential generalizability across tasks. Previous work also showed that in related tasks such as naturalistic emotion recognition, human observers may be more accurate than algorithmic computer methods; however, we do not know how observers compare to computers in a mind wandering detection task. Furthermore, though previous research identified some aspects of facial expressions that relate to mind wandering (e.g., blinking, specific AUs), little is known about the cues human observers rely on to detect mind wandering. The current study addresses these issues by comparing and combining machine learning methods with human observations of mind wandering across almost three thousand video clips as well as analyzing the justifications provided by the human observers.

3 Data Sources

We used data from a previous study involving participants from two universities in the United States. Key data details are described here, but see primary analysis publication for additional details [54]. We obtained university ethics board approval before collecting any data analyzed in this paper. All participants consented prior to the study; we analyzed data only from participants who further agreed to have their videos recorded and analyzed for research purposes.

3.1 Participants & Procedure

Data were collected from 152 participants who read the beginning 6,501 words of a text entitled “*Soap-bubbles, and the forces which mould them*” by C.V. Boys [12] on a computer screen while their faces and upper bodies were recorded at 12.5 frames per second and at a 640x480 pixel resolution with a webcam placed at the top of the computer monitor. Figure 1 shows example frames from these video recordings. While reading, participants pressed marked keyboard keys to report episodes where they caught themselves mind wandering. Mind wandering (colloquially referred to as *zoning out*) was defined to participants before they started reading as follows:

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”. If you catch yourself zoning out at any time during reading, please indicate what you are thinking about at that moment during reading.

When zoning out:

If you are thinking about the task itself (e.g., how many pages are there left to read, this text is very interesting) or how the task is making you feel (e.g., curious, annoyed) but not the actual content of the text, please press the key that is labeled "task".

OR

If you are thinking about anything else besides the task (e.g., what you ate for dinner last night, what you will be doing this weekend) please press the key that is labeled "other". Please familiarize yourself with where these two keys on the keyboard now so that you will know their location when you begin reading.

Please be as honest as possible about reporting zoning out. It is perfectly natural to zone out while reading. Responding that you were zoning out will in no way affect your scores on the test or your progress in this study, so please be completely honest with your reports. If you have any questions

about what you are supposed to do, please ask the experimenter now. Please press the right arrow key to begin.

We relied on self-reports of mind wandering because mind wandering is an inherently internal phenomenon [85], which may be difficult to discern from external indicators (an issue we explore in this paper). Additionally, self-reports have been previously validated as a measure of mind wandering in multiple studies [16,40,80,83,92]. Furthermore, self-reports of mind wandering correlate with objective outcome measures, thereby demonstrating predictive validity [41,60,71,78]. In these data, self-reported mind wandering correlated $r = -.229$ ($p = .012$) with a reading comprehension posttest, which is what was expected.

3.2 Extracting Video Clips

Of the 152 participants, data from 10 participants were removed due to video recording errors and data from 3 were removed because they declined to sign a data release agreement, resulting in a dataset with 139 participants.

Participants provided 2,577 self-caught mind wandering reports (positive mind wandering instances only) across 7,923 pages of text reading, or approximately 1 report every 3 pages. Participants self-reported mind wandering for an average of 18.5 pages ($SD = 13.5$). The average mind wandering report occurred 16s after the beginning of a page. In selecting video clip lengths, there is a tradeoff between clip length and the number of usable clips, given that reports occurred at different times within pages. We selected 10s as the clip length to compromise between obtaining many (but perhaps uninformative) short clips versus long (but fewer) clips. We also chose to remove clips that spanned multiple pages, since page turning movements (participant keypresses) interfered with the intended goal of analyzing facial features related only to mind wandering. Ideally, predictions could be made for all possible clips; however, including page turn clips reduced mind wandering detection accuracy to near chance levels (perhaps because of the added noise from page turning movements), so we proceeded with removing these. Since there are several seconds of gaps between consecutive page turn events (when participants are reading), removing these clips still allows frequent – if not continuous – predictions to be made.

An additional 207 clips were removed because the face could not be detected for at least 1s within the 10s clip, a requirement for the automatic approach as elaborated below. Face detection failures occurred when participants were out of the camera's view, a common occurrence (e.g., Figure 1b). Thus, we obtained 1,031 valid clips of mind wandering.

Participants reported when they were mind wandering (positive mind wandering cases), but not when they were paying attention (negative mind wandering cases), though both positive and negative cases are required to train a supervised classifier. Thus, we extracted negative cases from periods of time with no mind wandering reports (Figure 2). We did so by dividing the reading session into 10s clips (possible negative cases), removing any that were within a 30s window before a mind wandering report (because indicators of mind wandering may emerge 20-25s before a self-report; [55]), and further removing clips that overlapped with page turn events (to eliminate spurious facial movements attributable to page turn keypresses). From these negative cases, we randomly selected 2,406 instances to yield a mind wandering rate of 30% (1,031 positive out of 3,437 total clips), which is consistent with the current study and previous findings on the incidence of mind wandering during reading [24].

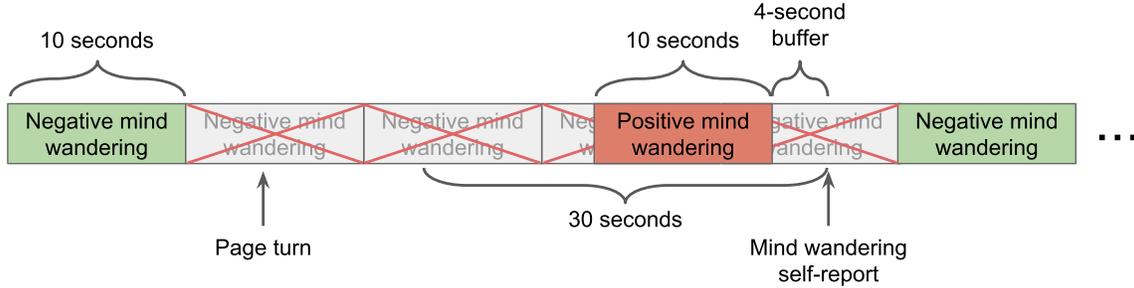


Figure 2: Illustration of positive and negative mind wandering instances extracted from self-reports and periods of time between self-reports. Instances that included page turn events were removed to avoid key-pressing actions being captured in video clips.

4 Automated Mind Wandering Detector

We examined predictions from a previously-developed face-based mind wandering detector [11]. In this section we only discuss the details needed for the main study, which compares the accuracy of the automated detector to human observers on the task of classifying video clips as mind wandering or not.

4.1 Facial Feature Extraction

We focused on five different types of facial features to serve as inputs for the supervised classification methods. Features were extracted on a frame by frame basis and then aggregated across individual frames in the 10s clips. First, we estimated *gross upper-body movement* from video clips using a validated background subtraction technique [95]. Specifically, we subtracted each video frame from a continuously-updated estimate of the video background and measured the proportion of changed pixels as an estimate of motion. Second, *local binary pattern (LBP) features* represent textures, which could be discriminative of mind wandering by capturing, for example, the appearance of teeth (indicating mouth opening), or other facial expressions associated with skin texture changes. We used OpenFace to automatically obtain the position of the eyes and center of the mouth [3], and then extracted LBP patches from those positions. OpenFace is an open-source software package for face detection and facial feature extraction, utilizing deep neural networks and support vector machines trained on several datasets to provide state-of-the-art accuracy. Third, *facial action unit (AU) features* describe activations of facial muscles (e.g., AU1 is the *inner brow raiser* muscle; [32]). We automatically extracted 19 AUs that span the eyes, cheeks, nose, and mouth regions of the face with EmotientSDK, which is a commercial version of the Computer Expression Recognition Toolbox (CERT). CERT is a validated computer vision tool for automatic extraction of AUs from face videos [58]. Note that EmotientSDK is no longer commercially available; however, OpenFace also provides 18 AU estimates [3], and could perhaps serve as a replacement in future implementations. Fourth, we computed *AU co-occurrence features* to capture co-occurring (in the 10-sec window) AUs, for example, using muscle activations near the eyes to distinguish between smiles of enjoyment and other smiles – even if the mouth muscle activations appear similar [31]. Specifically, we computed the Jensen-Shannon divergence (JSD) distribution similarity between all pairs of AUs [57]. Finally, we used *dynamic AU features* to model temporal dynamics in AUs over time. We utilized an approach similar to Bartlett et al. [5] in which we convolved 1-dimensional Gabor filters across AU time series with wavelengths ranging from 1 to 12 seconds, thereby distinguishing between fast and slow facial expression changes.

In previous work we also analyzed *head pose* features as a proxy for gaze direction. However, these features were minimally effective for detecting mind wandering [11], so we focused on the first five facial feature types in this study.

4.2 Supervised Machine Learning

We trained radial basis function support vector machine (SVM) classifiers to automatically detect mind wandering from the facial features described above. We trained one detector for each of the five sets of features, with person-independent nested cross-validation for selecting hyperparameters and evaluating classifier accuracy. In particular, we selected features and tuned the SVM hyperparameters C (from 10^{-2} to 10^2) and $gamma$ (from 10^{-5} to 10^2) via nested cross-validation within training data only. We selected, trained, and evaluated all models via the *scikit-learn* Python library [67]; see primary publication for additional details [11]. Finally, we fused the predictions of the five detectors (one for each feature set) via majority vote to generate the final machine-based prediction of mind wandering for each video clip (described in detail below). This model represents the state of the art for face-based mind wandering detection utilizing both AU features reported in previous work [87,88] and new features described above, the inclusion of which improved upon the previous approach. Thus, this model serves as a point of comparison to human observations on the same video clips, which is the main focus of this paper.

5 Collecting Human Observations of Mind Wandering

We recruited human observers (Turkers) from Amazon’s Mechanical Turk² web-based crowdsourcing platform to provide judgments about whether the person in each video clip was mind wandering or not. Video clips were the same as those classified with computer vision methods, thus affording a direct comparison between computer and observer accuracy.

5.1 Reliability of Mechanical Turk Studies

Previous research has shown that Turkers are representative of typical Internet user demographics [89], and more representative of the U.S. population than typical university students [10]. Furthermore, recent research has reported that studies conducted on Mechanical Turk (MTurk) consistently replicate findings from laboratory studies across a variety of domains, including HCI [18,45,47,49,65,82,90]. Thus, we expected MTurk would serve as a valid source of mind wandering ratings from typical humans.

5.2 Recruitment of Turkers

We recruited Turkers by submitting web-based human intelligence tasks (HITs; described below) to the platform, which Turkers viewed in a list of other HITs and selected if they wished to participate. We compensated Turkers with above minimum pay based on average time taken to complete the tasks (\$15.85 per hour). In order to obtain multiple independent judgments for each clip, we did not allow Turkers to complete the same HIT more than once. Additionally, we required Turkers to be at least 18 years old and located in the United States. In the end, 898 unique Turkers qualified and participated in the study.

5.3 Design of Rating Task

Each HIT contained 10 randomly-selected video clips, presented one at a time. Turkers first electronically consented to participate. Next, we displayed the following instructions:

As you may know, sometimes it is difficult to stay focused. It is often times the case that your attention starts to drift away and you “zone out.” We call this mind wandering.

For this task, you will be viewing 10 video clips of people as they read from a computer screen. In some videos the person is mind wandering, while in others they are not. There will not necessarily be an equal number of videos of each type. We would like you to tell us if you think that the person in the video is mind wandering or not, and why you think that.

² <https://www.mturk.com>

Although Turkers received instructions on the annotation task, we did not train them via examples or feedback. Whereas training may improve accuracy, it is not representative of the typical (untrained) way in which users assess whether others are paying attention.

The interface consisted of an HTML form with the instructions, a video clip, two radio buttons labeled *Mind wandering* and *Not mind wandering*, and a free-response text box where we asked Turkers to justify their mind wandering judgments (screenshot in Figure 3). Specifically, we asked them “What gestures or parts of the face did you use to make your decision for this video?” Finally, we asked Turkers to enter a two-digit number which was shown for two seconds after the end of each clip. We included this step to verify that Turkers had indeed watched the clip and followed the HIT instructions. They submitted the correct verification number for almost all clips (98.9% correct). To ensure validity, we re-submitted HITs for cases where verification numbers were incorrect. Turkers were also able to report a confidence level for their judgments (1 to 6), and could report situations where the face was not fully visible. Finally, Turkers pressed a *Next* button to rate the next clip, and could assess their progress through the HIT via a progress bar indicator.



Please indicate if the person in the video was mind wandering or not

Mind wandering
 Not mind wandering

How confident are you in your decision for this video?

1 (Least confident)
 2
 3
 4
 5
 6 (Most confident)

What gestures or parts of the face did you use to make your decision for this video?

Only check the check box below if the person's face was NOT fully visible

The person's face was not fully visible

Please input the number shown at the end of the video

Next

Figure 3: Screenshot of the mind wandering annotation interface shown to Amazon Mechanical Turk workers.

5.4 Data Collection 1: Inter-rater Reliability Estimation

Individual observers (raters) may come to different conclusions about the same classification task for various reasons, such as personal biases (e.g., one observer may tend to predict mind wandering more often than another observer), careless errors (e.g., not watching the video clips), or ambiguities in the classification task. Although we attempted to minimize careless errors by including a verification number for each observation, some such errors may have occurred. Moreover, the classification task is expected to be difficult (as can be seen in Figure 1), and reliability between two independent sets of observations could be low. However, if

there is at least some degree of consensus (i.e., the average pairwise correlation between sets of observations is $r > 0$), however small, then averaging predictions from additional observers will approach consensus. Individual judgment errors tend to “cancel out” due to the principle of aggregation, resulting in a higher *effective reliability* than the reliability of individual pairs of observers [76].

For tasks with high pairwise reliability (strong correlations between sets of observations), few sets of observations will be required to achieve high effective reliability – perhaps only two sets of observations to verify that reliability is high. For more difficult tasks, pairwise correlations between observations might be much lower, and thus more observations will be required to approach consensus. Previous research has not studied the difficulty of third-party human observation of mind wandering; thus, we first collected two independent observations for each video clip to estimate the reliability between human observers. We measured agreement between two sets of ratings via Pearson’s r (or, equivalently, ϕ or Spearman’s ρ , since observations are binary judgments of positive or negative mind wandering cases). We chose r over tetrachoric correlation because mind wandering was measured on a binary scale for both the recorded users and the observers. Reliability for two independent sets of observations on all 3,437 clips was $r = .152$, indicating low – but above zero – reliability. More observations were thus needed to achieve medium ($r = .3$), large ($r = .5$), or better effective reliability effect sizes [19].

5.5 Data Collection 2: Improving Reliability

We estimated effective reliability for additional rounds of observation using the Spearman–Brown prediction formula, which yields the effective reliability (r^*) achieved by two or more sets of observations given a mean pairwise reliability value:

$$r^* = \frac{\text{number of observations per clip} \times \text{average pairwise correlation}}{1 + \text{average pairwise correlation} \times (\text{number of observations per clip} - 1)}$$

This enabled forecasting the impact of adding additional observations on effective reliability in advance of data collection 2. We estimated effective reliability for additional raters (up to 20 total raters) assuming mean pairwise reliability of $r = .152$, as was obtained in data collection 1. Figure 4 shows estimated reliability after collecting additional observations, illustrating the diminishing returns of additional observations as majority voting between observers approaches consensus. We focused on odd numbers of observations to avoid ties in majority voting, and found that we would need 5 observations to achieve effective reliability (i.e., $r^* \geq 0.4$), 7 observations to reach $r^* \geq 0.5$, 9 to reach $r^* \geq 0.6$, and 15 to reach $r^* \geq 0.7$ (Figure 4). We chose nine raters as a compromise between study cost and effective reliability, and thus collected seven additional observations for each clip in data collection 2.

Figure 4 also shows the effective reliability that we obtained as we collected additional observations. Mean pairwise correlation across all nine rounds of observation was $r = .214$, which was higher than the initial estimate from data collection 1 of $r = .152$. Thus, effective reliability was higher than expected as well: $r^* = 9 \text{ [raters]} \times .214 \text{ [average pairwise correlation]} / (1 + .214 \times [9 - 1]) = .710$, which we deemed sufficient for present purposes. Final pair-wise inter-rater reliability as measured via Krippendorff’s alpha was .210.

Turkers generally reported high confidence in their judgments with little variance ($M = 4.97$ on a 1–6 scale, $SD = 0.419$). Preliminary analyses indicated no accuracy advantage in weighing Turker’s decisions by their reported confidence scores, so we did not utilize these scores further. Additionally, there was only one clip for which they agreed that the face was not fully visible, though for this clip mean confidence was 4.13 (out of 6) and 8 of 9 Turkers agreed that the person in the clip was mind wandering; thus, we did not remove the clip from analyses.

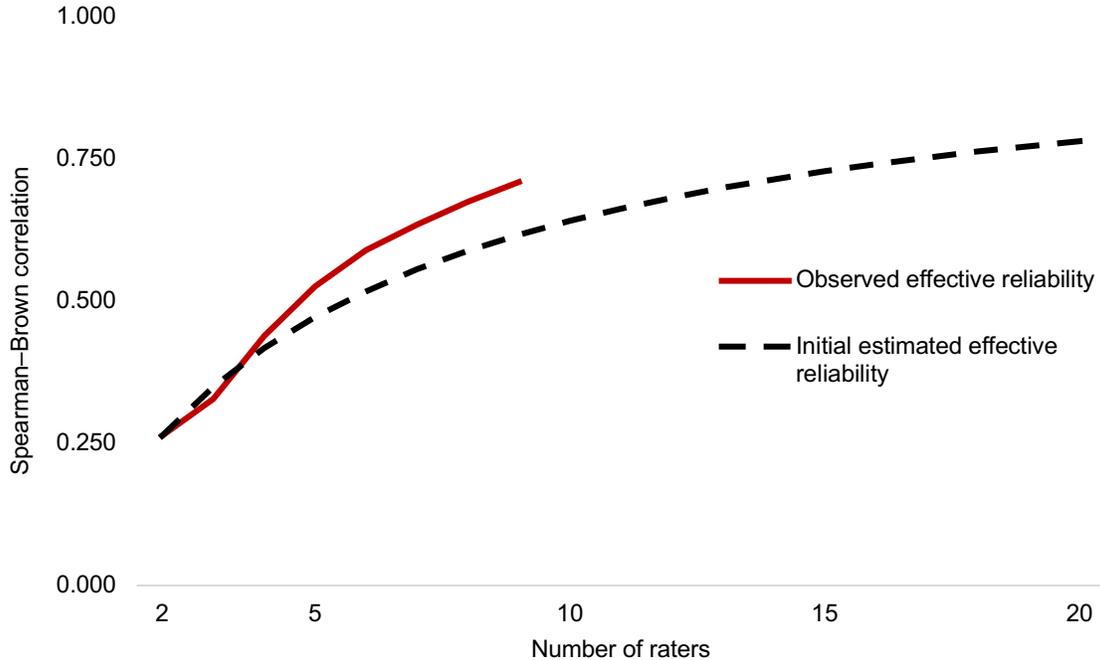


Figure 4: Estimated effective reliability for multiple rounds of data collection on Mechanical Turk (computed before performing data collection 2) and observed effective reliability after data collection was complete.

5.6 Analyzing Observations with Open-Vocabulary Models

We sought to understand the visual cues that observers used to make their judgments as part of RQ3. To do so, we trained open-vocabulary natural language processing models to find associations between observers’ justifications (text responses) and the judgments that they made.

The model building process proceed as follows. First, we concatenated the nine observers’ justifications for each clip to form one justification per clip. Next, we performed stemming to enable matching of different forms of the same word (e.g., “looking” and “looked” became “look”; [69]). We then computed counts of n-grams including unigrams (single words), bigrams (pairs of consecutive words), and trigrams (three consecutive words) for each clip. We filtered the set of n-grams by calculating the pointwise mutual information (PMI) of each bigram and trigram within the training data [79]. PMI measures the probability of an n-gram occurring relative to the probability of its individual constituent words, thereby favoring n-grams with words that primarily occur together (e.g., “glance down”) over those that occur independently as well (e.g., “might”, “be”). Note that bigrams and trigrams could occur across sentences or concatenated justifications, but would have low PMI unless the constituent words were statistically related, and thus they would not be included.

We utilized 10-fold clip-level cross validation for these models, selecting hyperparameters by grid-searching with nested 5-fold cross-validation in the training data only. Hyperparameters for selecting n-grams included minimum PMI (2, 3, or 4), minimum proportion of documents (justifications for a clip) an n-gram was represented in (0, .01, ..., .06), and n-gram size ($n \leq 1, 2, \text{ or } 3$). We also tuned a Laplace smoothing hyperparameter for the multinomial naïve Bayes model ($\alpha = .10, .25, .50, .75$) to improve accuracy.

Model AUC was .727; additionally, accuracy for positive mind wandering instances was $F_1 = .724$ (versus .336 random-chance baseline³), and for negative instances $F_1 = .877$ (versus .664 chance level). Thus,

³ Random-chance baseline was computed by making random mind wandering predictions aligned with the base rate in the data (i.e., the human judgments of mind wandering), then computing F_1 . We constructed the negative mind wandering baseline in the same manner.

the model predicted observers' labels at levels well above random chance, indicating that n-grams were closely related to observers' judgments.

Finally, in order to identify which words were most effective at discriminating positive from negative mind wandering responses, we computed the correlation (Pearson's r) between each n-gram and the self-reported mind wandering labels (1 or 0) in each cross-validation fold, then calculated the mean across all folds. Each of the 10 cross-validation folds could have potentially utilized a different set of n-grams since n-grams were filtered based on the hyperparameters described above. If an n-gram did not appear in a particular fold, we set the correlation for that fold to 0 so that it would penalize the overall mean. This analysis revealed which words and phrases observers used to indicate mind wandering along with the strength of these relationships.

6 RQ1: Comparing Observer vs. Computer Assessments of Mind Wandering

Research question 1 (RQ1) asks *how accurately can human observers detect mind wandering from videos of users engaged in computer-mediated tasks and how does their accuracy compare to automatic methods?* We computed accuracy of observers to answer the first part of RQ1, and compared that to the computer to answer the second part of RQ1. We measured accuracy for both computers and observers by comparing their judgments of mind wandering to *in situ* mind wandering self-reports from the participants who read the e-text. For observers, we calculated the classification probability for each clip as the proportion of positive (mind wandering) votes among the nine Turkers for that particular clip. For example, if 6 of the 9 votes were positive, the final probability would be $6 / 9 = .667$.

Similarly, the computer classification probability was based on majority voting from the five individual machine learning models trained on each of the five sets of features as discussed above. For example, if 2 of the 5 models made positive (mind wandering) predictions for a particular instance (video clip), we would calculate the predicted probability as $2 / 5 = .400$ for that instance. Another possibility would be to average the confidences of the individual models with the distance of instances from the SVM hyperplane serving as a measure of confidence. However, hyperplane distances are not comparable across models, and thus we focused on majority vote (as with human observations). Further, as also noted in our previous work [11], slightly higher AUC can be obtained via stacking (training an additional classifier on the outputs of base classifiers); however, this resulted in predictions that were biased in favor of positive mind wandering reports, yielding machine predictions distribution that were dissimilar to the distribution of observers' predictions. Since this violated the distribution exchangeability assumption of our AUC comparisons [94], we focused on the majority vote computer model.

6.1 Overall Results

Table 1 reports precision, recall, and F_1 scores of positive (mind wandering) and negative (not mind wandering) video clips based on majority votes; Table 2 reports the confusion matrices. For positive cases, both the computer and observers had equivalent precision, but the computer had higher recall, leading to higher F_1 scores and more true-positive predictions. In contrast, the computer had higher precision, but observers had higher recall, for negative cases. Whereas these results reflect accuracy comparisons at one particular decision threshold (.500), we can compare the precision-recall tradeoff across all possible threshold with the receiver operating characteristic (ROC) curves of computers and observers. In particular, we compared area under the ROC curve (AUC) using the bootstrap method [15], and we compared the curves themselves with Venkatraman's test [94].

We first statistically compared overall AUC between the computer (AUC = .598) and all nine observers voting together (AUC = .589) with the bootstrap test (using 10,000 bootstrap samples) for the difference in area between two correlated ROC curves, implemented in the *pROC* package in *R* [70,75]. The difference was not significant ($p = .572$), indicating that AUCs were indeed similar. However, both AUCs exceeded chance (AUC of 0.5) suggesting that both the computer and observers were detecting a signal, albeit somewhat faintly.

We also observed that the shape of the ROC curves appeared notably different for observers and the computer (Figure 5), despite the AUCs being similar. Thus, we compared the ROC curves with Venkatraman’s test, also with 10,000 sampling iterations. Venkatraman’s test measures whether the shapes of the ROC curves themselves are different, even if the area under those curves may be similar. The test was significant ($p < .005$), supporting the hypothesis that the shape of the ROC curves differed.

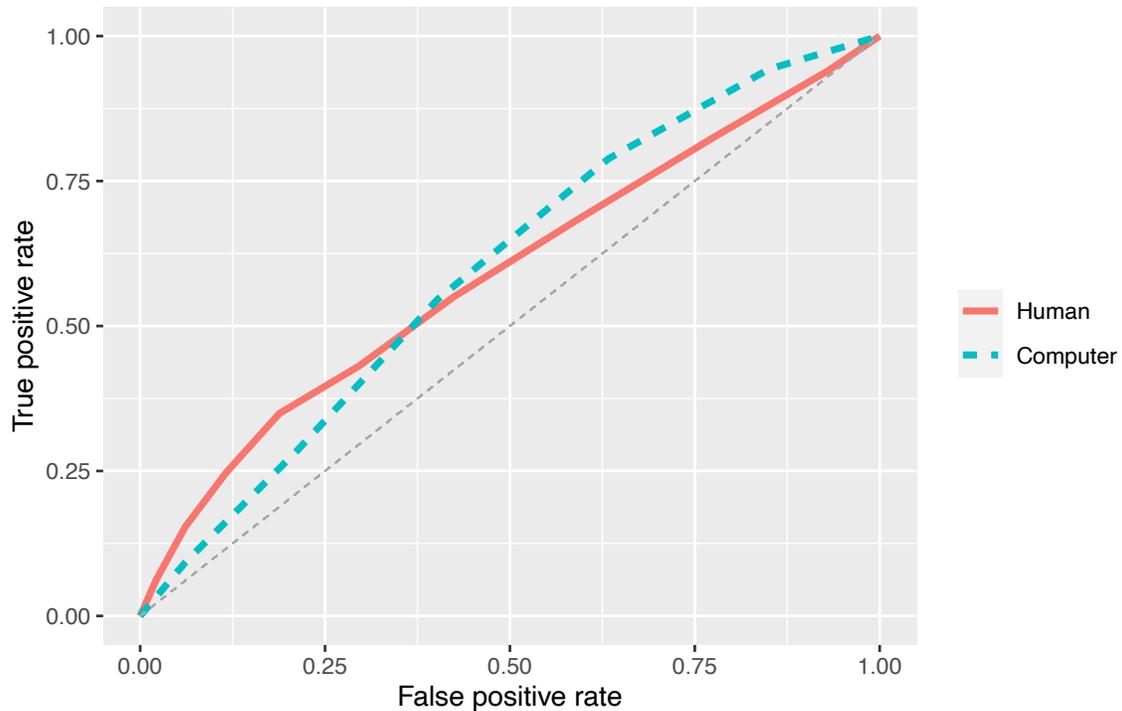


Figure 5: Receiver operating characteristic (ROC) curves for human (nine observers) and computer classifications of mind wandering.

Given that the ROC curves differed, we computed partial AUC (pAUC) values for the high precision (low false-positive rate) and high recall (high true-positive rate) portions of the ROC curve separately, splitting the ROC curve at the point where the computer and observer curves intersected (at true-positive rate = .496, false-positive rate = .366; see Figure 5). After splitting the ROC curve into two parts, we rescaled each pAUC to match the typical range of AUC scores – i.e., [0, 1] with .5 as the random chance level. For the low false-positive rate ROC portion, the computer’s pAUC was .634, while observers’ pAUC was .717. We compared these two partial AUCs with the bootstrap method, which showed that observers were indeed significantly more accurate in the low false-positive ROC portion ($p < .005$). Similarly, we compared the high true-positive rate portion for both curves, where the computer’s pAUC was .586 and observers’ pAUC was .547. This difference was also significant ($p < .005$), confirming that the computer was more accurate in the high true-positive portion of the AUC curve. Additionally, the computer appeared to be more consistent in terms of accuracy than the observers with the computer absolute pAUC difference = $|.586 - .634| = .048$, versus observers’ difference = $|.716 - .546| = .170$.

Table 1: Mind wandering detection accuracy for automatic computer vision-based method (computer) versus human observers.

	Computer performance	Observers performance	Base rate	Predicted rate (computer)	Predicted rate (observers)
<i>Positive mind wandering</i>					
F ₁	.439	.406	.300	.445	.336
Precision	.368	.384	.300	.445	.336
Recall	.545	.431	.300	.445	.336
<i>Negative mind wandering</i>					
F ₁	.667	.723	.700	.555	.664
Precision	.754	.743	.700	.555	.664
Recall	.599	.704	.700	.555	.664
AUC	.598	.589			

Table 2: Confusion matrices for computer and human observers' classification of mind wandering, assuming a 0.5 (i.e., majority vote) decision threshold.

		Computer's predicted mind wandering	
		Positive	Negative
Self-reported mind wandering	Positive	562 (.545)	469 (.455)
	Negative	966 (.401)	1440 (.599)
		Observers' predicted mind wandering	
		Positive	Negative
Self-reported mind wandering	Positive	444 (.431)	587 (.569)
	Negative	711 (.296)	1695 (.704)

6.2 Varying the Number of Observers

With nine observers, observer and computer accuracies differed in different regions of the ROC curve, but overall accuracies were not statistically different. However, in practical HCI applications nine observers is likely an unrealistically large number. We thus varied the number of observers by randomly selecting one or more observers per video clip, then averaging votes as before (Figure 6). With one randomly-chosen observer per clip, observer AUC was just .548 – significantly lower than the computer's .598 AUC ($p < .001$). With two observers, AUC was .563 (also significantly lower than the computer; $p = .014$). However, with three observers, AUC was .574, and did not significantly differ from the computer ($p = .096$). Thus, three or more observers were needed to match the accuracy of the computer.

To explore the trend more generally, we quantified the improvement in AUC due to each additional observer by sampling $n=1..9$ observers repeatedly (1000 times per n) and averaging AUCs obtained. We then

calculated the improvement in AUC for n observers versus $n-1$ observers as the percentage increase (past the accuracy of $n-1$ observers) in AUC obtained for n observers after mean-centering at chance level (i.e., $[AUC_n - .5] / [AUC_{n-1} - .5] - 100\%$). Results in Figure 7 show that the improvement per added observer decreased quickly, but that small improvements were obtained even with the ninth observer. In fact, adding a second observer produced over 35% improvement in AUC for the human predictions. Conversely, adding a fifth observer produced less than 5% improvement for both humans, and the eighth and ninth observers improved AUC by less than 2.5%.

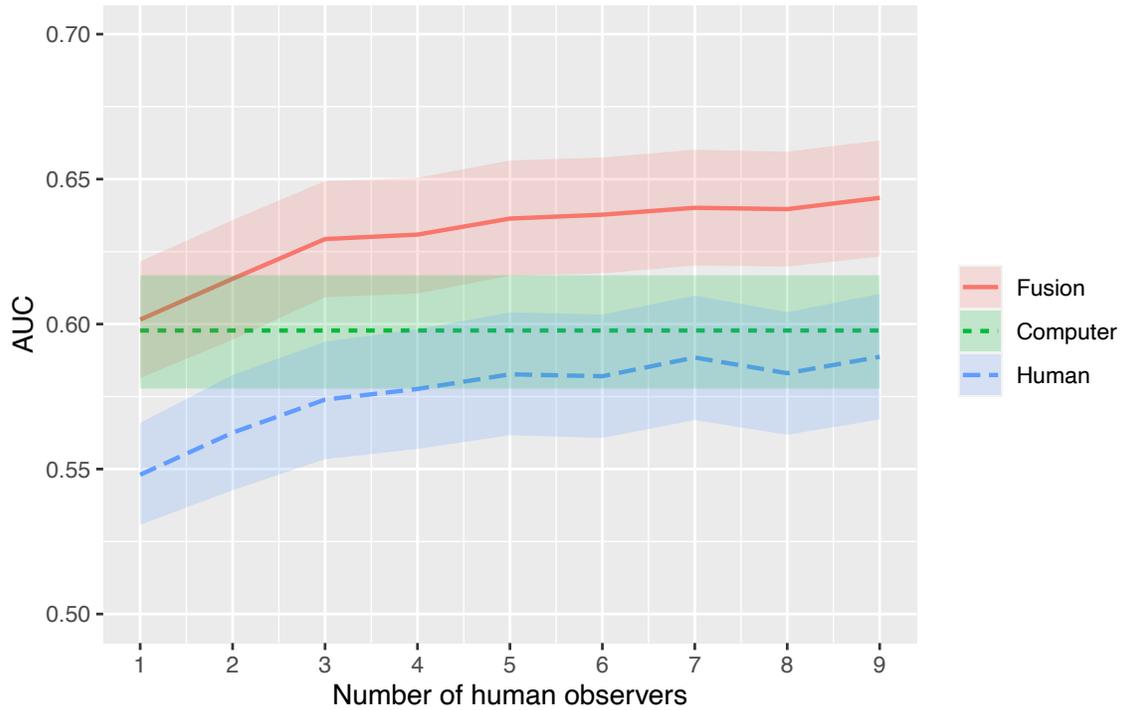


Figure 6. AUC as a function of the number of observers included in the human predictions (RQ1) and the fusion predictions (RQ2). Bootstrap 95% confidence intervals are shown with shading.

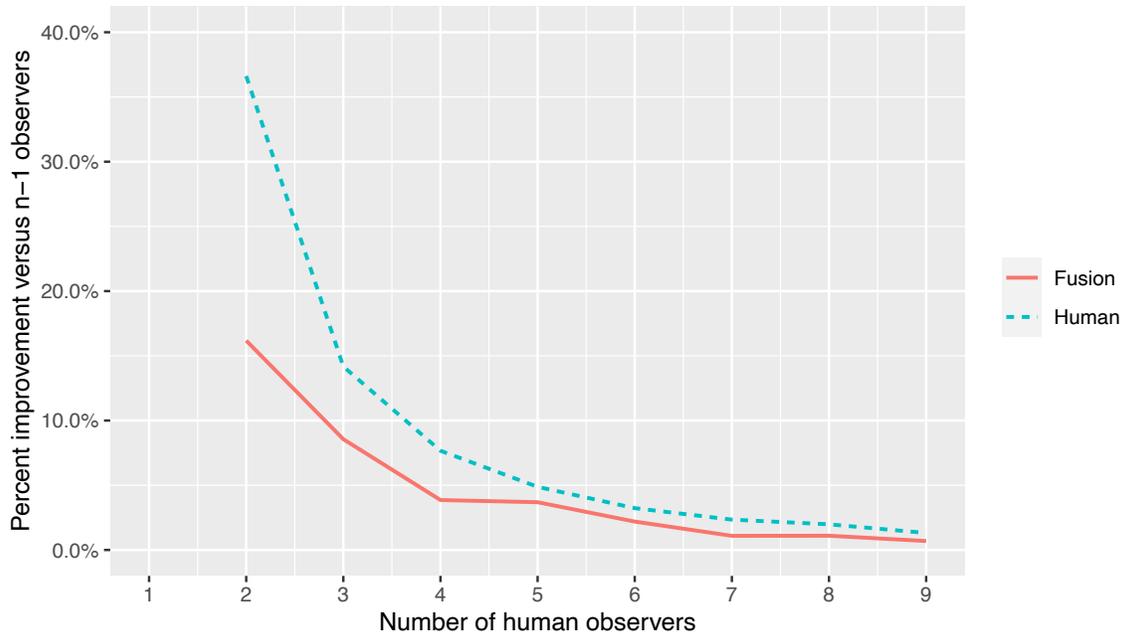


Figure 7. Additive value of each additional observer as a percentage of improvement in AUC.

7 RQ2: Fusing Observer and Computer Assessments of Mind Wandering

RQ2 builds on RQ1 by examining cooperation between observers and the computer. Specifically, RQ2 asks *does a fusion of human- and machine- predictions outperform either one independently?* We fused observers' and the computer's predictions by averaging the prediction from each. We then compared the fused predictions to the computer and the observers alone. Similar to RQ1, we also varied the number of observers included in the fusion to determine whether adding a limited number of observers would yield improvements over a computer-only approach. Additionally, we examined conditional fusion in which observers only provide judgments for a specific case (i.e., when the computer predicts positive mind wandering), thus reducing the labor required from observers.

7.1 Overall Results

Table 3 compares F_1 , precision, recall, and AUC for human-computer decision fusion models alongside the individual observer and computer accuracies. Comparisons showed that the human-computer fusion AUC (.644) was significantly better than observers (AUC = .589; $p < .001$) and the computer (AUC = .598; $p < .001$). Results for positive and negative mind wandering cases separately showed that the fusion model combined the strengths of the computer and observers. For positive mind wandering, the fusion had a similar F_1 score (.444) compared to the computer (.439), and slightly higher than observers (.406). For negative mind wandering, the fusion model had an F_1 score similar to the observers (.726 versus .723), and higher F_1 than the computer (.667). A similar pattern is apparent for precision, where the fusion model matched or exceeded the observers (.407 versus .384) and the computer (.368) for positive mind wandering, and matched observers and the computer for negative mind wandering (.760 versus .743 and .754, respectively). In terms of recall, however, fusion was between the computer and observers for positive mind wandering (.488 versus .545 and .431), but surpassed the computer for negative mind wandering (.695 versus .599) and matched the observers (.704). Taken together, these results indicate that the fusion matched the better of the computer's and observers' predictions, and thus was more accurate overall; however, this fusion required all nine observers to contribute to the decision.

Table 3. Classification metrics for fusion models, with individual computer and observer values reproduced from Table 1 for comparison purposes

	Computer performance	Observers performance	Observers + computer fusion	Positive-only fusion	Base rate
<i>Positive mind wandering</i>					
F ₁	.439	.406	.444	.389	.300
Precision	.368	.384	.407	.397	.300
Recall	.545	.431	.488	.381	.300
<i>Negative mind wandering</i>					
F ₁	.667	.723	.726	.746	.700
Precision	.754	.743	.760	.739	.700
Recall	.599	.704	.695	.752	.700
AUC	.598	.589	.644	.612	

7.2 Varying Observer Involvement in Fusion

As in section 6.1, we varied the number of observers included in the observers' classification decision. In this case, we added one randomly-chosen observer at a time to the fusion model to determine how many observers were needed for the fusion to outperform the computer-only decisions. With one randomly-chosen observer, the fusion's AUC of .602 was not significantly ($p = .790$) better than the computer (AUC = .598). With two observers in the fusion model, AUC was .616 ($p = .217$), and with three observers AUC was .629 ($p = .026$). Thus, at least three observers were needed in the fusion model to significantly outperform the computer-only model. However, including all nine observers further outperformed fusion with three ($p = .009$; see Figure 6).

We measured the improvement per observer for the fusion method, like in section 6.1, and found a similar trend such that the first few observers had notable positive impacts on AUC while the later observers yielded marginal improvement (Figure 7). In particular, adding a second observer produced over 15% improvement versus one observer, while the sixth through ninth observers provided less than 2.5% improvement. Additional observers were less impactful, as a percentage, for fusion compared to observer-only predictions since the computer predictions constituted only a portion of the overall accuracy for the fusion.

We also explored reducing human labor by having observers provide predictions only where the computer predicted positive mind wandering. The resulting fusion had an AUC of .612, which was significantly higher than the computer ($p = .003$) but not better than the observers alone ($p = .090$). However, this method required judgments for only 44.5% of instances, rather than all instances as in the observers-only method. Table 3 shows that the positive-only fusion primarily improved on the computer predictions in terms of precision for positive mind wandering cases (.398 versus .368) and recall for negative mind wandering (.752 versus .599). The confusion matrix (Table 4) also shows that the positive-only fusion model had many more true negatives than the computer model (1824 versus 1440). This was expected since, in this fusion, only the computer's positive mind wandering predictions were fused with observer predictions (i.e., potentially turning a borderline positive prediction into a negative prediction). The result is a model that is more selective about predicting positive mind wandering cases than the computer-only approach, which is more accurate in the high false-positive region of the ROC curve (see RQ1).

The improved accuracy of fusion models compared to individual computer or human predictions indicates that these predictions are complementary. In fact, the two sets of predictions were negatively correlated ($r = -.124, p < .001$), despite both being positively related to the same outcome (mind wandering). This suggests that observers and the computer differed substantially in how they made their predictions, which we explore in depth in RQ3.

Table 4: Confusion matrices for the human–machine decision fusion model with observer input on only positive computer predictions; computer’s matrix reproduced from Table 2 for comparison purposes

		Computer’s predicted mind wandering	
		Positive	Negative
Self-reported mind wandering	Positive	562 (.545)	469 (.455)
	Negative	966 (.401)	1440 (.599)
		Positive-only fusion predicted mind wandering	
		Positive	Negative
Self-reported mind wandering	Positive	393 (.381)	638 (.619)
	Negative	596 (.248)	1810 (.752)

8 RQ3: Visual Cues of Mind Wandering

Research question 3 asks *what visual cues do observers utilize to make their decisions, and how are those cues complementary to those used by the machine?* We examined the justifications observers provided for their mind wandering labels to identify cues that the observers used to inform their decision. On average, there were 5.09 words per justification (SD = 5.35). There were 8.66 justifications per clip (out of a maximum possible 9), indicating that nearly all Turkers provided justifications. For this analysis we adapted an open-vocabulary analysis similar to Park et al. [66] with the goal of using natural language processing and machine learning to identify words and phrases (n-grams) that are diagnostic of observers’ mind wandering ratings in a generalizable manner. We inspected machine learning model features to discover which n-grams related to positive or negative mind wandering judgments, and thus whether observers were assessing mind wandering via previously-observed indicators of mind wandering – such as gaze, blinks, and fidgeting – or via other behaviors. Accordingly, we trained multinomial naïve Bayes models to predict the observers’ labels from the n-gram counts (features) using nested-cross validation (see section 5.6 for methodological details) to ensure generalizability across video clips.

8.1 N-gram Word Clouds from Open-Vocabulary Models

We generated word clouds where the size of each word (n-gram) was proportional to the correlation (Pearson’s r) between that n-gram and both the observers’ majority vote label and the computer’s predicted label (top row of Figure 8). Several broad patterns are evident. Observers provided specific justifications about positive instances of mind wandering, such as *close* (referring to eyes), *asleep*, *yawn*, *blink*, and other words that refer to physical motions apparent in the videos. However, observers also made more subjective judgments about the appearance of mind wandering, such as participants who *are drift[ing] away* or *are zon[ing out]*. Similarly, observer justifications for negative mind wandering labels include specific actions related to effortful reading such as *follow*, *side to [side]*, and *move*, as well as subjective descriptions such as

focus and not drift [away]. Results differed for the automated computer method. Correlations between observer n-grams and computer predictions (bottom row of Figure 8) showed little overlap with the observer word-clouds, which is unsurprising since the computer models were trained on self- and not observer- reports. Nevertheless, it is interesting that some n-grams appeared in both computer and observer word clouds but in the opposite direction (e.g., *are drift[ing] away*, *yawn*, *bod[ily/y]*, and *follow*). The n-grams correlated with computer predictions were predominately related to body movement, mouth movement, and blinking, whereas observers' predictions were correlated with higher-level behaviors inferred from these activities, such as reading, drifting away, zoning out, and sleeping. Both relied on eye-related cues, though the specific cues differed (e.g., eyes closing versus blinking).

Previous work shows that among the individual types of features used by the computer for automatic mind wandering detection (see section 2.2), LBP features provided the most accurate results [11]. These features represent relatively low-level facial textures only a few pixels wide, in contrast to the higher-level features like AUs (e.g., jaw drop) that are more similar to justifications given by observers. Hence, the computer appeared to attend to some visual characteristics that observers did not, though it is difficult to interpret these low-level features to discover what they represent and whether observers could potentially be trained to recognize the same features.

To characterize observers' judgments further, we selected the n-grams with the largest correlations to observers' mind wandering predictions and matched them to the full justification text (reproduced without editing spelling; Table 5). This further illustrates many of the specific actions that observers observed when categorizing mind wandering. For example, yawning, scratching, and eyes closing were all correlated with positive mind wandering labels. Conversely, participants' eyes following the text repeatedly occurred in justifications for negative mind wandering observations – e.g., “eyes focused on text”, “eyes were moving side to side.” Observers were not entirely consistent, however; head scratching was noted as a sign of mind wandering by one person, but as alertness by another.

Table 5: Example observer justifications for the 10 n-grams that were most positively or negatively correlated with observer mind wandering labels.

n-gram	<i>r</i>	Example observer justification
<i>Positive mind wandering judgments</i>		
close	.347	He had his head laying on his hands and his eyes were closing as if he was falling asleep.
fall	.303	Looked like she was falling asleep.
out while	.283	Eyes are zoning out while reading.
are zone	.283	Eyes are zone out while reading
are drift away	.256	Eyes are drift away while reading.
asleep	.250	Falling Asleep
almost	.208	Her eyes's movement. She is almost slept
yawn	.199	She let out a BIG yawn, showing she was tired, and she was scratching her head (boredom). She simply looked fatigue.
fell	.175	He fell asleep while reading.
sleepi	.145	Seems disinterested & sleepy - no focus!
<i>Negative mind wandering judgments</i>		
not drift	-.295	Eyes are not drift away while reading.
read	-.218	Eyes moving as if reading entire time. Mouth moving as if speaking the words aloud.
follow	-.198	Granted he was yawning, but his eyes followed the text.
move	-.198	Her eyes were moving side to side and she appeared to be reading aloud to herself
away while	-.144	Eyes are not drift away while reading. He is concentrating.
focus	-.128	eyes focused on text and moving with it
text	-.116	She was not only following text with her eyes, but she appeared to be reading the text aloud as well.
alert	-.111	I think she stayed with it, eyes kept their alertness and attention. She scratched her head and shifted a bit to stay alert I think.
what	-.101	His eyes were following along with what he was reading
look care	-.090	The person is looking carefully at the screen

8.2 Visual Inspection of Example Observations

We extracted frames from video clips of positive and negative mind wandering instances to ground the provided justifications more clearly in visual examples. In particular, we examined video clips where observer justifications included the three justifications with the highest most positive or negative correlation to observer mind wandering labels (Figure 9).

For the positive mind wandering clips (rows A, B, and C in Figure 9), it appears that all three participants were disengaging from the task, either by drifting toward sleep or simply ignoring the task. The participant who observers labeled as “zoning out” (row C) has closed eyes in some frames as well. This participant did, however, self-report mind wandering for this clip, indicating that they self-caught mind wandering.

We note three different behaviors for the negative mind wandering clips. In Figure 9 row D is a participant whose eyes are scanning the text; in row E, the participant is reading aloud, and row F features a participant who yawns but apparently is reading the e-text.

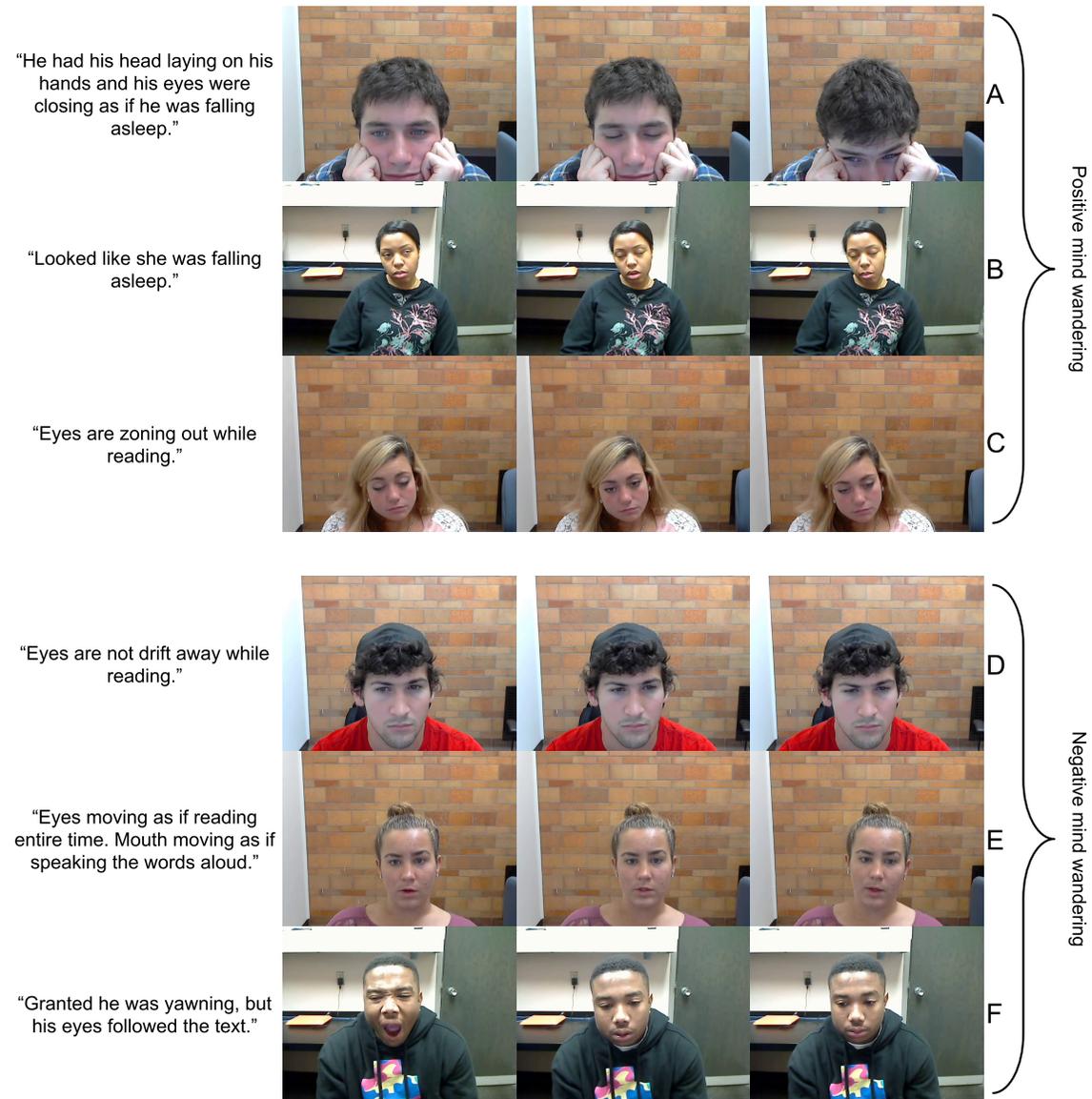


Figure 9: Examples of correctly-identified positive and negative mind wandering video clips, with sample observer justifications for each clip.

9 Discussion and Conclusions

The ability to identify mind wandering from video has implications for a variety of HCI contexts including education, medicine, safety, and so on. Automatic computerized methods for assessing mind wandering may be able to help users with observation, either by improving accuracy or reducing the labor required. We addressed two research questions centered on these issues, as well as a third question exploring the differences between observer and computer perceptions of mind wandering. In this section we discuss the main findings from our three research questions and their implications for HCI.

9.1 Main Findings

It was possible that mind wandering would be only subtly apparent from facial expressions during the present computerized reading context. Indeed, the subtlety of expressions was apparent in our results as evidenced by the modest detection accuracies achieved by both observers and computers (RQ1). Although we expected observers might outperform computers, given related work judging subtle emotional expressions [99], we found that overall accuracies were similar. Importantly, when examining detection accuracy across different decision thresholds in ROC curves, we found that the automated method outperformed observers in the high false-positive, low false-negative region of the ROC curve, while observers outperformed the automated method for the high false-negative, low false-positive region. These results indicated that observers and computers made complementary, but non-redundant predictions, which suggested that a fusion of the two might yield improved accuracy by combining the strengths of each.

Our experiments with human-computer decision fusion indeed showed that human and computer predictions were complementary (negatively correlated with each other), and that accuracy improved significantly when observer and computer predictions were combined (RQ2). We also found that only three observers were sufficient to significantly outperform the computer-only approach, though including all nine observers improved accuracy further. The fusion approach in which observers only provided their judgments for cases where the computer predicted positive mind wandering also improved accuracy over the computer-only approach, and resulted in more selective mind wandering predictions (higher precision, but lower recall).

Analyses of observers' open-ended justifications for their predictions (RQ3) showed that observers made their predictions based on high-level user behavioral states inferred from videos, including reading, sleepiness, focusing, and zoning out. They also made observations about lower-level, specific user behaviors, such as blinking, yawning, and hand movements, but these were largely uncorrelated with their predictions of positive or negative mind wandering (with the exception of eyes closing). Conversely, the computer's predictions were more correlated with low-level behaviors noted in observers' justifications. These results highlight the fact that observers and the computer were relying on different cues to make their judgments, which explains why fusions of observer and computer predictions produced additive results (improved accuracy) rather than being redundant. Differences between observer and computer predictions also highlight aspects of the videos that observers deemed important but were overlooked by the computer, and could be explored to improve the computer's predictions in future work.

To this point, one aspect that observers frequently noted were reading-related eye movements when they classified a clip as negative mind wandering (see Table 5 and Figure 9). For example, justifications such as "eyes were moving side to side", "following text with her eyes", and "eyes moving as if reading" indicate that observers were paying close attention to eye movements. On the other hand, observers also mentioned "sleep", fidgeting ("body", "hand", "neck" movements), "rub[bing]" (Figure 8), and similar appearances as justifications for rating a clip as a positive mind wandering episode. However, these may be more directly related to boredom and only tangentially related to mind wandering, which the computer (with no preconceptions of mind wandering) would be better able to avoid. Users are perhaps more likely to be familiar with related constructs like behavioral engagement (e.g., looking at the screen), boredom, confusion, and frustration, rather than mind wandering. In other words, observers' top-down biases appear to be influencing

their judgments, which was an effective strategy for negative cases of mind wandering but not for positive mind wandering instances. The computer, with no preconceived biases, is largely driven by the stimulus itself.

9.2 Implications of Results for HCI Research and Application

Our findings have important implications for 1) research scenarios where accuracy is essential, and 2) software applications where the accuracy of the observers, computer, and human-machine decision fusion provides guidance on how to measure mind wandering while minimizing user input. We discuss these both of these aspects here with respect to our three research questions.

RQ1 Implications. Accuracy with a single human observer (AUC = .548) was only slightly better than chance level, indicating that a lone observer is not likely to be able to effectively judge mind wandering in other users in a computer-mediated interaction. A team of nine observers was significantly better, and comparable to the automatic computerized method. However, additional observers beyond the first few yielded rapidly diminishing improvements in accuracy, suggesting only 3-4 observers are needed to approach consensus. In some applications this level of cooperation between observers may be feasible; in retrospective analysis of video clips for research purposes, it certainly is. However, given the comparable accuracy of the computer method, an automatic approach is a potentially viable replacement for tedious human observation, most definitely for real-time applications and when multiple users are involved (e.g., simultaneously monitoring the attention of 10 participants in an online study over Zoom).

RQ2 Implications. Fusion of observer and computer predictions of mind wandering did indeed improve accuracy, suggesting that it is a viable approach. When accuracy is paramount, computer predictions fused with a team of observers' predictions produces the highest accuracy by a significant margin. For less-sensitive applications, such as online learning, a fusion including a team of just three observers (e.g., teaching assistants) significantly outperformed the computer alone. Our findings also inform the number of observers needed based on how much each additional observer improves prediction accuracy in the fusion model; specifically, additional observers beyond 3-4 offer little accuracy improvement for fusion, as in the observers-only model. Additionally, we found that observer effort could be reduced by more than half by having observers only provide judgments in the cases where the computer predicts positive mind wandering; this significantly improved accuracy over the computer alone. Reducing observer effort is essential for applications where it is impractical to have observers remain constantly vigilant for an extended period of time.

RQ3 Implications. We found that the computer's predictions aligned with observers' observations of low-level behaviors (e.g., blinks) rather than higher-level inferences like reading or zoning out. This finding offers some insight into how the computer makes its predictions, and how that compared to human observations. This is an especially important for HCI applications given previous research that shows users will form their own "folk theories" (which may be quite inaccurate) of how algorithms work if they are not given explanations [34], and that misunderstandings can lead to negative experiences like attributing incorrect algorithmic predictions to the behaviors of other users [35]. Further, findings in RQ3 imply that there are opportunities to improve automatic mind wandering assessment methods by developing features for higher-level behaviors like reading. Automatic, video-based approaches to estimate gaze are still in the early stages of research, but are becoming increasingly accurate (e.g., [3]). These approaches might eventually improve to the point that gaze-related reading features – such as fixations and saccades – can be accurately extracted from video and used to improve the accuracy of mind wandering detectors.

Observers' justifications also covered many different behaviors and facial expressions, indicating that there is not a one-to-one mapping of mind wandering to a facial expression, even in this specific context. Rather, there are many expressions (which could also vary across contexts), implying that it may be difficult to train observers to detect mind wandering based on a canonical "mind wandering expression"; see [4] for a parallel argument for emotion detection. Finally, RQ3 findings suggest that users may sometimes equate mind wandering with related, but theoretically distinct constructs, such as boredom (as mentioned explicitly by an observer in Table 5). Future work training users to recognize mind wandering in video-based interactions might include distinctions between related constructs, and compare users' ability to distinguish between these constructs.

9.3 Limitations and Future Work

Our study has some limitations that should be addressed. First, we only evaluated mind wandering detection in the context of task (e-text reading on a computer). Given the task-specific justifications observers provided for their negative mind wandering reports (e.g., eyes moving with text), accuracy might be notably different for different tasks where eye movements might be less predictable (e.g., scene viewing [55]). Thus, replication in situations where users were performing different tasks, such as viewing a film, is warranted.

Second, computer-mediated interactions are often social, and observers may detect mind wandering more (or less [39]) accurately in social contexts in which they are active participants and have a stake in the outcomes. Moreover, facial expressions themselves may differ in some social situations, given that facial expressions are often only present as a social communication tool [20]. In contrast, the present stimuli, which feature non-interactive videos of a person reading an e-text, were devoid of social cues that might have aided in their judgments. As such, replicating the study by using videos collected in more engaging social contexts might be an important next step. However, the current context is representative of several computer-mediated interactions, such as online video-based lectures with limited interactivity.

Third, the laboratory context in which users read the e-text allowed precise control of potential distractions, lighting, camera placement, seating, thereby mitigating environmental factors that could introduce variance into video recordings of users. It remains an open question whether computers and observers can recognize mind wandering “in the wild” when confounding factors may make the task more difficult. Conversely, as mentioned above, these factors may add valuable context to the recognition task.

Finally, our results showed that human observers, while capable of recognizing mind wandering at above-chance levels, were far from perfect. Many of the false-positive and false-negative classifications made by observers could be due to the subtle, internal nature of mind wandering, but it is also possible that training observers to detect mind wandering may improve their accuracy for a specific context in which the space of mind wandering facial expressions may be limited (though perhaps not limited enough, as RQ3 findings suggest). Whereas our goal was to investigate the accuracy of *untrained* observers, future work could include strategies such as providing observers with labeled examples before annotation, or providing feedback after each annotation they make to promote learning. Previous research on detection of pain from facial expressions found that training observers resulted in minimal improvements [5], so we do not anticipate training would make a substantial difference here, though this has yet to be empirically tested.

10 Concluding Remarks

This study is part of a broader field of HCI research seeking to understand the characteristics of mind wandering during computer-based tasks and how mind wandering can be measured, and thus accounted for, in these tasks. We were particularly interested in how third-party human observers perceive mind wandering, and whether observers are more accurate at detecting it than a computer vision-based machine learning approach. We confirmed that mind wandering is very difficult to visually detect for both observers and computers. However, both outperformed chance baselines, suggesting that there is a signal amidst the noise; moreover, combining the computer’s predictions to those of the observers was effective for improving accuracy and reducing the need for human labor.

Importantly, the computers and observers have complementary strengths and weaknesses, with the computer outperforming the observers for high true-positive decision thresholds and observers outperforming the computer for low false-positive thresholds. We also found that the visual cues that observers use in making their decisions were different from the computer’s cues, which can be incorporated to improve the accuracy of the computer models. Further research will lead to a better understanding of mind wandering and methods to detect it, thereby opening the door for automated measurement of a ubiquitous component of every human-computer interaction and to attention-aware systems that aim to make interactions with computers more engaging, enjoyable, and effective by attending to attention.

ACKNOWLEDGMENTS

The authors would like to thank Cathlyn Stone for implementing natural language processing code used in this work. This research was supported by the National Science Foundation (NSF) (DRL 1235958, IIS 1523091, DRL 1920510). 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] Benjamin Baird, Jonathan Smallwood, Michael D. Mrazek, Julia W. Y. Kam, Michael S. Franklin, and Jonathan W. Schooler. 2012. Inspired by distraction: Mind wandering facilitates creative incubation. *Psychological Science* 23, 10 (October 2012), 1117–1122. DOI:<https://doi.org/10.1177/0956797612446024>
- [2] Ryan S. Baker, Sidney K. D’Mello, Ma. Mercedes T. Rodrigo, and Art Graesser. 2010. Better to be frustrated than bored: The incidence, persistence, and impact of learners’ cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies* 68, 4 (April 2010), 223–241. DOI:<https://doi.org/10.1016/j.ijhcs.2009.12.003>
- [3] Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, IEEE, Piscataway, NJ, 59–66. DOI:<https://doi.org/10.1109/FG.2018.00019>
- [4] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. 2019. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychol Sci Public Interest* 20, 1 (July 2019), 1–68. DOI:<https://doi.org/10.1177/1529100619832930>
- [5] Marian Stewart Bartlett, Gwen C. Littlewort, Mark G. Frank, and Kang Lee. 2014. Automatic decoding of facial movements reveals deceptive pain expressions. *Current Biology* 24, 7 (March 2014), 738–743.
- [6] Bogdan Batrinca and Philip C. Treleaven. 2015. Social media analytics: a survey of techniques, tools and platforms. *AI & Society* 30, 1 (February 2015), 89–116. DOI:<https://doi.org/10.1007/s00146-014-0549-4>
- [7] Mathias Benedek, David Daxberger, Sonja Annerer-Walcher, and Jonathan Smallwood. 2018. Are you with me? Probing the human capacity to recognize external/internal attention in others’ faces. *Visual Cognition* 26, 7 (August 2018), 511–517. DOI:<https://doi.org/10.1080/13506285.2018.1504845>
- [8] Robert Bixler, Nathaniel Blanchard, Luke Garrison, and Sidney K. D’Mello. 2015. Automatic detection of mind wandering during reading using gaze and physiology. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI ’15)*, ACM, New York, NY, 299–306. DOI:<https://doi.org/10.1145/2818346.2820742>
- [9] Nathaniel Blanchard, Robert Bixler, Tera Joyce, and Sidney K. D’Mello. 2014. Automated physiological-based detection of mind wandering during learning. In *Proceedings of the 12th International Conference on Intelligent Tutoring Systems (ITS 2014)* (Lecture Notes in Computer Science), Springer, Cham, CH, 55–60. DOI:https://doi.org/10.1007/978-3-319-07221-0_7
- [10] John Bohannon. 2011. Social science for pennies. *Science* 334, 6054 (October 2011), 307–307. DOI:<https://doi.org/10.1126/science.334.6054.307>
- [11] Nigel Bosch and Sidney K. D’Mello. in press. Automatic detection of mind wandering from video in the lab and in the classroom. *IEEE Transactions on Affective Computing* (in press). DOI:<https://doi.org/10.1109/TAFFC.2019.2908837>
- [12] Charles Vernon Boys. 1890. *Soap-bubbles, and the forces which mould them*. London, England: Society for Promoting Christian Knowledge.
- [13] Taylor A. Burke, Brooke A. Ammerman, and Ross Jacobucci. 2019. The use of machine learning in the study of suicidal and non-suicidal self-injurious thoughts and behaviors: A systematic review. *Journal of Affective Disorders* 245, (February 2019), 869–884. DOI:<https://doi.org/10.1016/j.jad.2018.11.073>
- [14] Rafael A. Calvo and Sidney K. D’Mello. 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing* 1, 1 (January 2010), 18–37. DOI:<https://doi.org/10.1109/T-AFFC.2010.1>

- [15] James Carpenter and John Bithell. 2000. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine* 19, 9 (2000), 1141–1164. DOI:[https://doi.org/10.1002/\(SICI\)1097-0258\(20000515\)19:9<1141::AID-SIM479>3.0.CO;2-F](https://doi.org/10.1002/(SICI)1097-0258(20000515)19:9<1141::AID-SIM479>3.0.CO;2-F)
- [16] Kalina Christoff, Alan M. Gordon, Jonathan Smallwood, Rachelle Smith, and Jonathan W. Schooler. 2009. Experience sampling during fMRI reveals default network and executive system contributions to mind wandering. *Proceedings of the National Academy of Sciences* 106, 21 (May 2009), 8719–8724. DOI:<https://doi.org/10.1073/pnas.0900234106>
- [17] Kalina Christoff, Zachary C. Irving, Kieran C. R. Fox, R. Nathan Spreng, and Jessica R. Andrews-Hanna. 2016. Mind-wandering as spontaneous thought: a dynamic framework. *Nature Reviews Neuroscience* 17, 11 (November 2016), 718–731. DOI:<https://doi.org/10.1038/nrn.2016.113>
- [18] Scott Clifford, Ryan M Jewell, and Philip D Waggoner. 2015. Are samples drawn from Mechanical Turk valid for research on political ideology? *Research & Politics* 2, 4 (October 2015), 1–9. DOI:<https://doi.org/10.1177/2053168015622072>
- [19] Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed. ed.). Lawrence Erlbaum, Hillsdale, NJ.
- [20] Carlos Crivelli and Alan J. Fridlund. 2018. Facial displays are tools for social influence. *Trends in Cognitive Sciences* 22, 5 (May 2018), 388–399. DOI:<https://doi.org/10.1016/j.tics.2018.02.006>
- [21] Carlos Crivelli and Alan J. Fridlund. 2019. Inside-out: From basic emotions theory to the behavioral ecology view. *Journal of Nonverbal Behavior* 43, 2 (June 2019), 161–194. DOI:<https://doi.org/10.1007/s10919-019-00294-2>
- [22] M. Csikszentmihalyi. 1990. *Flow: The psychology of optimal experience*. New York: Harper and Row.
- [23] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (CHI '20), ACM, New York, NY, 1–12. DOI:<https://doi.org/10.1145/3313831.3376638>
- [24] Sidney K. D’Mello. 2019. What do we think about when we learn? In *Deep Comprehension: Multi-Disciplinary Approaches to Understanding, Enhancing, and Measuring Comprehension*, Keith Millis, J. Magliano, D. Long and K. Wiemer (eds.). Routledge, New York, NY, 52–67.
- [25] Sidney K. D’Mello, Ed Dieterle, and Angela L. Duckworth. 2017. Advanced, analytic, automated (AAA) measurement of engagement during learning. *Educational Psychologist* 52, 2 (April 2017), 104–123.
- [26] Sidney K. D’Mello and Jacqueline Kory. 2015. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys* 47, 3 (February 2015), 43:1-43:36.
- [27] Sidney K. D’Mello, Blair Lehman, Reinhard Pekrun, and Art Graesser. 2014. Confusion can be beneficial for learning. *Learning and Instruction* 29, 1 (2014), 153–170.
- [28] Kunio Doi. 2007. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and Graphics* 31, 4 (June 2007), 198–211. DOI:<https://doi.org/10.1016/j.compmedimag.2007.02.002>
- [29] Damien Dupré, Eva G. Krumhuber, Dennis Küster, and Gary J. McKeown. 2020. A performance comparison of eight commercially available automatic classifiers for facial affect recognition. *PLOS ONE* 15, 4 (April 2020), 0231968:1–17. DOI:<https://doi.org/10.1371/journal.pone.0231968>
- [30] Dominic B. Dwyer, Peter Falkai, and Nikolaos Koutsouleris. 2018. Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology* 14, 1 (2018), 91–118. DOI:<https://doi.org/10.1146/annurev-clinpsy-032816-045037>
- [31] Paul Ekman, Richard J. Davidson, and Wallace V. Friesen. 1990. The Duchenne smile: Emotional expression and brain physiology: II. *Journal of Personality and Social Psychology* 58, 2 (1990), 342–353.
- [32] Paul Ekman and Wallace V. Friesen. 1978. *Facial action coding system: A technique for the measurement of facial movement*. Palo Alto, CA: Consulting Psychologists Press.
- [33] El-Sayed A. El-Dahshan, Heba M. Mohsen, Kenneth Revett, and Abdel-Badeeh M. Salem. 2014. Computer-aided diagnosis of human brain tumor through MRI: A survey and a new algorithm. *Expert Systems with Applications* 41, 11 (September 2014), 5526–5545. DOI:<https://doi.org/10.1016/j.eswa.2014.01.021>
- [34] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I “like” it, then I hide it: Folk theories of social feeds. In

- Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), ACM, New York, NY, 2371–2382. DOI:<https://doi.org/10.1145/2858036.2858494>
- [35] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. “I always assumed that I wasn’t really that close to [her]”: Reasoning about invisible algorithms in news feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), ACM, New York, NY, 153–162. DOI:<https://doi.org/10.1145/2702123.2702556>
- [36] Myrthe Faber, Robert Bixler, and Sidney K. D’Mello. 2018. An automated behavioral measure of mind wandering during computerized reading. *Behavior Research Methods* 50, 1 (2018), 134–150. DOI:<https://doi.org/10.3758/s13428-017-0857-y>
- [37] Myrthe Faber and Sidney K. D’Mello. 2018. How the stimulus influences mind wandering in semantically rich task contexts. *Cognitive Research: Principles and Implications* 3, 35 (2018), 1–14. DOI:<https://doi.org/10.1186/s41235-018-0129-0>
- [38] Myrthe Faber, Kristina Krasich, Robert E. Bixler, James R. Brockmole, and Sidney K. D’Mello. in press. The eye–mind wandering link: Identifying gaze indices of mind wandering across tasks. *Journal of Experimental Psychology: Human Perception and Performance* (in press). DOI:<https://doi.org/10.1037/xhp0000743>
- [39] Myrthe Faber, McKenzie Rees, and Sidney K. D’Mello. 2018. Mind wandering during conversations affects subjective but not objective outcomes. In *CogSci 2018*.
- [40] David J. Frank, Brent Nara, Michela Zavagnin, Dayna R. Touron, and Michael J. Kane. 2015. Validating older adults’ reports of less mind-wandering: An examination of eye movements and dispositional influences. *Psychology and Aging* 30, 2 (2015), 266. DOI:<http://dx.doi.org/10.1037/pag0000031>
- [41] Michael S. Franklin, James M. Broadway, Michael D. Mrazek, Jonathan Smallwood, and Jonathan W. Schooler. 2013. Window to the wandering mind: Pupillometry of spontaneous thought while reading. *The Quarterly Journal of Experimental Psychology* 66, 12 (December 2013), 2289–2294. DOI:<https://doi.org/10.1080/17470218.2013.858170>
- [42] Jennifer A. Fredricks, Phyllis C. Blumenfeld, and Alison H. Paris. 2004. School engagement: Potential of the concept, state of the evidence. *Review of Educational Research* 74, 1 (2004), 59–109. DOI:<https://doi.org/10.3102/00346543074001059>
- [43] Cédric Galéra, Ludivine Orriols, Katia M’Bailara, Magali Laborey, Benjamin Contrand, Régis Ribéreau-Gayon, Françoise Masson, Sarah Bakiri, Catherine Gabaude, Alexandra Fort, Bertrand Maury, Céline Lemercier, Maurice Cours, Manuel-Pierre Bouvard, and Emmanuel Lagarde. 2012. Mind wandering and driving: Responsibility case-control study. *BMJ* 345, (December 2012), e8105. DOI:<https://doi.org/10.1136/bmj.e8105>
- [44] Fiona J. Gilbert, Susan M. Astley, Maureen G.C. Gillan, Olorunsola F. Agbaje, Matthew G. Wallis, Jonathan James, Caroline R.M. Boggis, and Stephen W. Duffy. 2008. Single reading with computer-aided detection for screening mammography. *New England Journal of Medicine* 359, 16 (October 2008), 1675–1684. DOI:<https://doi.org/10.1056/NEJMoa0803545>
- [45] Joseph K. Goodman, Cynthia E. Cryder, and Amar Cheema. 2013. Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making* 26, 3 (July 2013), 213–224. DOI:<https://doi.org/10.1002/bdm.1753>
- [46] Romain Grandchamp, Claire Braboszcz, and Arnaud Delorme. 2014. Oculometric variations during mind wandering. *Frontiers in Psychology* 5, (February 2014). DOI:<https://doi.org/10.3389/fpsyg.2014.00031>
- [47] Jeffrey Heer and Michael Bostock. 2010. Crowdsourcing graphical perception: Using Mechanical Turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '10), ACM, New York, NY, USA, 203–212. DOI:<https://doi.org/10.1145/1753326.1753357>
- [48] Y. H. Holkamp and J. Schavemaker. 2014. A comparison of human and machine learning-based accuracy for valence classification of subjects in video fragments. In *Proceedings of Measuring Behavior 2014*, Noldus Information Technology, Wageningen, NL, 73–76.
- [49] John J. Horton, David G. Rand, and Richard J. Zeckhauser. 2011. The online laboratory: Conducting experiments in a real labor market. *Exp Econ* 14, 3 (September 2011), 399–425. DOI:<https://doi.org/10.1007/s10683-011-9273-9>

- [50] Stephen Hutt, Kristina Krasich, Caitlin Mills, Nigel Bosch, Shelby White, James R. Brockmole, and Sidney K. D’Mello. 2019. Automated gaze-based mind wandering detection during computerized learning in classrooms. *User Modeling and User-Adapted Interaction* 29, 4 (2019), 821–867. DOI:<https://doi.org/10.1007/s11257-019-09228-5>
- [51] Afsaneh Jalalian, Syamsiah B. T. Mashohor, Hajjah Rozi Mahmud, M. Iqbal B. Saripan, Abdul Rahman B. Ramli, and Babak Karasfi. 2013. Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: A review. *Clinical Imaging* 37, 3 (May 2013), 420–426. DOI:<https://doi.org/10.1016/j.clinimag.2012.09.024>
- [52] Joris H. Janssen, Paul Tacken, J. J. G. Gert-Jan de Vries, Egon L. van den Broek, Joyce H. D. M. Westerink, Pim Haselager, and Wijnand A. IJsselsteijn. 2013. Machines outperform laypersons in recognizing emotions elicited by autobiographical recollection. *Human-Computer Interaction* 28, 6 (November 2013), 479–517. DOI:<https://doi.org/10.1080/07370024.2012.755421>
- [53] Matthew A. Killingsworth and Daniel T. Gilbert. 2010. A wandering mind Is an unhappy mind. *Science* 330, 6006 (November 2010), 932–932. DOI:<https://doi.org/10.1126/science.1192439>
- [54] Kristopher Kopp, Sidney K. D’Mello, and Caitlin Mills. 2015. Influencing the occurrence of mind wandering while reading. *Consciousness and Cognition* 34, (July 2015), 52–62.
- [55] Kristina Krasich, Robert McManus, Stephen Hutt, Myrthe Faber, Sidney K. D’Mello, and James R. Brockmole. 2018. Gaze-based signatures of mind wandering during real-world scene processing. *Journal of Experimental Psychology: General* 147, 8 (2018), 1111–1124. DOI:<https://doi.org/10.1037/xge0000411>
- [56] Jukka M. Leppänen and Charles A. Nelson. 2009. Tuning the developing brain to social signals of emotions. *Nature Reviews Neuroscience* 10, 1 (January 2009), 37–47. DOI:<https://doi.org/10.1038/nrn2554>
- [57] J. Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* 37, 1 (January 1991), 145–151.
- [58] Gwen Littlewort, J. Whitehill, Tingfan Wu, Ian Fasel, M. Frank, J. Movellan, and M. Bartlett. 2011. The computer expression recognition toolbox (CERT). In *2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011)*, 298–305.
- [59] Bethany T. McDaniel, Sidney K. D’Mello, Brandon G. King, Patrick Chipman, Kristy Tapp, and Art Graesser. 2007. Facial features for affective state detection in learning environments. In *Proceedings of the 29th Annual Cognitive Science Society*, 467–472.
- [60] Jennifer C. McVay and Michael J. Kane. 2009. Conducting the train of thought: Working memory capacity, goal neglect, and mind wandering in an executive-control task. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 35, 1 (2009), 196–204. DOI:<http://psycnet.apa.org/doi/10.1037/a0014104>
- [61] Jennifer C. McVay, Michael J. Kane, and Thomas R. Kwapil. 2009. Tracking the train of thought from the laboratory into everyday life: An experience-sampling study of mind wandering across controlled and ecological contexts. *Psychonomic Bulletin & Review* 16, 5 (October 2009), 857–863.
- [62] Vikram Mohanty, David Thames, Sneha Mehta, and Kurt Luther. 2019. Photo Sleuth: Combining human expertise and face recognition to identify historical portraits. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI ’19)*, ACM, New York, NY, 547–557. DOI:<https://doi.org/10.1145/3301275.3302301>
- [63] Benjamin W. Mooneyham and Jonathan W. Schooler. 2013. The costs and benefits of mind-wandering: A review. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 67, 1 (2013), 11–18.
- [64] Jaclyn Ocumpaugh, Ryan S. Baker, and Ma. Mercedes T. Rodrigo. 2015. *Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 technical and training manual*. Technical Report. New York, NY: Teachers College, Columbia University. Manila, Philippines: Ateneo Laboratory for the Learning Sciences.
- [65] Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5, 5 (June 2010), 411–419.
- [66] Gregory Park, H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Michal Kosinski, David J. Stillwell, Lyle H. Ungar, and Martin E. P. Seligman. 2015. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology* 108, 6 (2015), 934–952. DOI:<https://doi.org/10.1037/pspp0000020>

- [67] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, (November 2011), 2825–2830.
- [68] Phuong Pham and Jingtao Wang. 2015. AttentiveLearner: Improving mobile MOOC learning via implicit heart rate tracking. In *Artificial Intelligence in Education* (Lecture Notes in Computer Science), Springer, Cham, CH, 367–376. DOI:https://doi.org/10.1007/978-3-319-19773-9_37
- [69] Martin F. Porter. 2001. Snowball: A language for stemming algorithms. Retrieved from <http://snowball.tartarus.org/texts/introduction.html>
- [70] R Core Team. 2013. R: A language and environment for statistical computing. (2013).
- [71] Jason G. Randall, Frederick L. Oswald, and Margaret E. Beier. 2014. Mind-wandering, cognition, and performance: A theory-driven meta-analysis of attention regulation. *Psychological bulletin* 140, 6 (2014), 1411–1431.
- [72] Erik D. Reichle, Andrew E. Reineberg, and Jonathan W. Schooler. 2010. Eye movements during mindless reading. *Psychological Science* 21, 9 (September 2010), 1300–1310.
- [73] Evan F. Risko, N. Anderson, A. Sarwal, M. Engelhardt, and Alan Kingstone. 2012. Everyday attention: Variation in mind wandering and memory in a lecture. *Applied Cognitive Psychology* 26, 2 (2012), 234–242.
- [74] Evan F. Risko, Dawn Buchanan, Srđan Medimorec, and Alan Kingstone. 2013. Everyday attention: Mind wandering and computer use during lectures. *Computers & Education* 68, (October 2013), 275–283. DOI:<https://doi.org/10.1016/j.compedu.2013.05.001>
- [75] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12, 1 (March 2011), 77. DOI:<https://doi.org/10.1186/1471-2105-12-77>
- [76] Robert Rosenthal. 2005. 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.). Oxford University Press, New York, NY, 199–234.
- [77] Konrad Schindler, Luc Van Gool, and Beatrice de Gelder. 2008. Recognizing emotions expressed by body pose: A biologically inspired neural model. *Neural Networks* 21, 9 (November 2008), 1238–1246. DOI:<https://doi.org/10.1016/j.neunet.2008.05.003>
- [78] Jonathan W. Schooler, Erik D. Reichle, and David V. Halpern. 2005. Zoning out while reading: Evidence for dissociations between experience and metacognition. In *Thinking and seeing: Visual metacognition in adults and children*, Daniel T. Levin (ed.). Cambridge, MA: MIT Press, 204–226.
- [79] H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS ONE* 8, 9 (September 2013). DOI:<https://doi.org/10.1371/journal.pone.0073791>
- [80] Paul Seli, Jonathan S. A. Carriere, D. R. Thomson, J. Allan Cheyne, K. A. E. Martens, and Daniel Smilek. 2014. Restless mind, restless body. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 40, 3 (2014), 660–668. DOI:<http://dx.doi.org/10.1037/a0035260>
- [81] Paul Seli, Michael J. Kane, Jonathan Smallwood, Daniel L. Schacter, David Maillet, Jonathan W. Schooler, and Daniel Smilek. 2018. Mind-wandering as a natural kind: A family-resemblances view. *Trends in Cognitive Sciences* 22, 6 (June 2018), 479–490. DOI:<https://doi.org/10.1016/j.tics.2018.03.010>
- [82] Daniel J. Simons and Christopher F. Chabris. 2012. Common (mis)beliefs about memory: A replication and comparison of Telephone and Mechanical Turk survey methods. *PLoS ONE* 7, 12 (December 2012), e51876. DOI:<https://doi.org/10.1371/journal.pone.0051876>
- [83] Jonathan Smallwood, Emily Beach, Jonathan W. Schooler, and Todd C. Handy. 2008. Going AWOL in the brain: Mind wandering reduces cortical analysis of external events. *Journal of Cognitive Neuroscience* 20, 3 (2008), 458–469.
- [84] Jonathan Smallwood and Jonathan W. Schooler. 2006. The restless mind. *Psychological Bulletin* 132, 6 (2006), 946–958.

- [85] Jonathan Smallwood and Jonathan W. Schooler. 2015. The science of mind wandering: Empirically navigating the stream of consciousness. *Annual Review of Psychology* 66, 1 (2015), 487–518. DOI:<https://doi.org/10.1146/annurev-psych-010814-015331>
- [86] Daniel Smilek, Jonathan S. A. Carriere, and J. Allan Cheyne. 2010. Out of mind, out of sight: Eye blinking as indicator and embodiment of mind wandering. *Psychological Science* 21, 6 (2010), 786–789.
- [87] Angela Stewart, Nigel Bosch, Huili Chen, Patrick J. Donnelly, and Sidney K. D’Mello. 2016. Where’s your mind at? Video-based mind wandering detection during film viewing. In *Proceedings of the 2016 Conference on User Modeling, Adaptation, and Personalization (UMAP 2016)*, ACM, New York, NY, 295–296. DOI:<https://doi.org/10.1145/2930238.2930266>
- [88] Angela Stewart, Nigel Bosch, and Sidney K. D’Mello. 2017. Generalizability of face-based mind wandering detection across task contexts. In *Proceedings of the 10th International Conference on Educational Data Mining (EDM 2017)*, International Educational Data Mining Society, 88–95.
- [89] Neil Stewart, Christoph Ungemach, Adam J. L. Harris, Daniel M. Bartels, Ben R. Newell, Gabriele Paolacci, and Jesse Chandler. 2015. The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision Making* 10, 5 (2015), 479–491.
- [90] Siddharth Suri and Duncan J. Watts. 2011. Cooperation and contagion in web-based, networked public goods experiments. *PLoS ONE* 6, 3 (March 2011), e16836. DOI:<https://doi.org/10.1371/journal.pone.0016836>
- [91] Karl K. Szpunar, Novall Y. Khan, and Daniel L. Schacter. 2013. Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences* 110, 16 (April 2013), 6313–6317. DOI:<https://doi.org/10.1073/pnas.1221764110>
- [92] Sarah Uzzaman and Steve Joordens. 2011. The eyes know what you are thinking: Eye movements as an objective measure of mind wandering. *Consciousness and Cognition* 20, 4 (December 2011), 1882–1886. DOI:<https://doi.org/10.1016/j.concog.2011.09.010>
- [93] B. Van Ginneken, B.M. Ter Haar Romeny, and M.A. Viergever. 2001. Computer-aided diagnosis in chest radiography: A survey. *IEEE Transactions on Medical Imaging* 20, 12 (December 2001), 1228–1241. DOI:<https://doi.org/10.1109/42.974918>
- [94] E. S. Venkatraman and Colin B. Begg. 1996. A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika* 83, 4 (December 1996), 835–848. DOI:<https://doi.org/10.1093/biomet/83.4.835>
- [95] Jacqueline Kory Westlund, Sidney K. D’Mello, and Andrew M. Olney. 2015. Motion Tracker: Camera-based monitoring of bodily movements using motion silhouettes. *PLoS ONE* 10, 6 (June 2015).
- [96] J. Whitehill, Z. Serpell, Yi-Ching Lin, A Foster, and J.R. Movellan. 2014. The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing* 5, 1 (January 2014), 86–98.
- [97] David P. Williams, Michel Couillard, and Samantha Dugelay. 2014. On human perception and automatic target recognition: Strategies for human-computer cooperation. In *2014 22nd International Conference on Pattern Recognition, IEEE, Piscataway, NJ*, 4690–4695. DOI:<https://doi.org/10.1109/ICPR.2014.802>
- [98] Matthew R. Yanko and Thomas M. Spalek. 2014. Driving with the wandering mind: The effect that mind-wandering has on driving performance. *Human Factors* 56, 2 (March 2014), 260–269. DOI:<https://doi.org/10.1177/0018720813495280>
- [99] Neta Yitzhak, Nir Hiladi, Tanya Gurevich, Daniel S. Messinger, Emily B. Prince, Katherine Martin, and Hillel Aviezer. 2017. Gently does it: Humans outperform a software classifier in recognizing subtle, nonstereotypical facial expressions. *Emotion* 17, 8 (2017), 1187–1198. DOI:<http://dx.doi.org/10.1037/emo0000287>
- [100] X. Yu, S. Zhang, Z. Yan, F. Yang, J. Huang, N. E. Dunbar, M. L. Jensen, J. K. Burgoon, and D. N. Metaxas. 2015. Is interactional dissynchrony a clue to deception? Insights from automated analysis of nonverbal visual cues. *IEEE Transactions on Cybernetics* 45, 3 (March 2015), 492–506. DOI:<https://doi.org/10.1109/TCYB.2014.2329673>

Author Statement

This paper is most closely related to our previous work on automatic face-based detection of mind wandering [11]. We utilized the predictions made by our previously-developed automatic method for the current paper, as noted in the paper. However, the research questions we propose and address in this paper are new.