Toward a Real-time (Day) Dreamcatcher: Sensor-Free Detection of Mind Wandering During Online Reading

Caitlin Mills
University of Notre Dame
Department of Psychology
Notre Dame, IN 46556
cmills4@nd.edu

Sidney D'Mello
University of Notre Dame
Department of Psychology
Department of Computer Science
Notre Dame, IN 46556
sdmello@nd.edu

ABSTRACT
This paper reports the results from a sensor-free detector of mind wandering during an online reading task. Features consisted of reading behaviors (e.g., reading time) and textual features (e.g., level of difficulty) extracted from self-paced reading log files. Supervised machine learning was applied to two datasets in order to predict if participants were mind wandering as they navigated from one screen of text to the next. Mind wandering was detected with an accuracy of 20% above chance (Cohen's kappa = .207; AUC = .609), which was obtained via leave-one-participant-out cross-validation. Similar to actual rates of mind wandering, predicted rates of mind wandering were negatively related to posttest performance, thus providing some evidence for the predictive validity of the detector. Applications of the detector to attention-aware educational interfaces are discussed.

Keywords
Mind wandering, attention, machine learning, reading

1. INTRODUCTION
It is not uncommon to experience looking up from a book only to realize you have no idea what you just read. In fact, it has been documented that people can read up to 17 words of gibberish before even realizing that they have zoned out [32]. Since students often have trouble realizing when they have zoned out themselves, it can be especially difficult to determine when someone is not paying attention through observation. For example, a student who is deeply engaged in learning can often look quite similar to another student who is thinking about something else completely.

This phenomenon, known as *mind wandering*, is an involuntary shift in attention away from the external task towards task-unrelated thoughts [36]. Mind wandering is detrimental during learning, as learning requires consolidating external information into mental structures. During episodes of mind wandering, however, students are unable to integrate external information with their existing internal representations. Thus, missed information is not processed and mental models are not updated, limiting overall understanding. Given the negative impact of mind wandering on learning [14, 30, 32, 33], it is important to develop systems that can reorient attention when students mind wander in order to facilitate engagement and learning. Building detectors of mind wandering is an essential first step towards this goal and is the focus of the present paper.

1.1 Related Work
One of the first known studies related to mind wandering detection was conducted by Drummond and Litman [13]. In their study, students read a paragraph about biology aloud then performed a learning task (i.e., paraphrase or self-explanation). Students periodically self-reported how frequently they were thinking about off-task thoughts on a scale from 1 (all the time) to 7 (not at all). Supervised machine learning trained on acoustic-prosodic features was used to classify whether students were "high" in zoning out (1-3 on the scale) versus "low" in zoning out (5-7 on the scale). Results indicated an accuracy of 64% in discriminating "low" versus "high" zone outs. This pivotal study on mind wandering was innovative with respect to automatically detecting zone outs during a learning task. However, they used a leave-one-instance-out cross-validation method (rather than a leave-one-participant-out cross-validation method), so generalizability of the detector to new students is unclear.

Recent research has also attempted to detect mind wandering during online reading using both gaze [5] and peripheral physiology [6]. In both of these studies, mind wandering was collected via thought probes that occurred on pseudo-random pages (i.e., computer screens) during reading. Students responded either "yes" or "no" about whether they were mind wandering at the time of the probe. In the first study, a detector of mind wandering achieved an accuracy of 72% (Cohen's kappa = .28) using features extracted from gaze data collected with a Tobii eye tracker [5]. In the second study, a detector of mind wandering built using physiological features (i.e., skin conductance and temperature) achieved an accuracy of 74% (Cohen's kappa = .22). Both of these detectors used a leave-several-subjects-out validation method to ensure generalizability to new students.

These detectors display impressive results given the elusive nature of mind wandering. However, the equipment and sensors required for eye-gaze and physiology tracking might impair scalability. In particular, one issue faced by online learning environments is that sensors are not readily available. For example, students using an ITS deployed online from their home computer would not have access to an eye tracker or a way to measure skin conductance at their convenience. A key question then is how to detect mind wandering based on information that is readily available, for example, in interaction log files. Along these lines, the aim of the current study is to identify a set of features that 1) are theoretically

related to mind wandering, and 2) can be extracted from log files during online learning.

Interaction-based detectors trained from interaction log files have been used to successfully build detectors of other "off-task" states, such as gaming the system and off-task conversation [4, 7–9]. While mind wandering is related to other forms of "off-task" states, such as boredom, behavioral disengagement, and distractions [1, 3, 4, 8, 9, 26, 42], it is inherently distinct because it is involuntary and involves internal thoughts rather than overt expressive behaviors. The involuntary, unconscious nature of mind wandering makes detection particularly difficult. First, whereas other off-task states often involve some behavioral markers to denote disengagement, mind wandering is a completely internal state that can look similar to on-task states. Second, the onset and duration of mind wandering episodes cannot be precisely measured because people are often unaware their attention has been directed away from the external task. Thus, finding features that will pick up on subtle differences in attention is extremely difficult.

To date, one study has attempted sensor-free mind wandering detection (see Table 1 for a summary of mind wandering detectors). Franklin et al. [15] attempted to classify if readers were "mindlessly reading" using two criterion: (1) difficulty and (2) reading time. For the first criterion, readers could only be classified as "mind wandering" while reading difficult text. To establish the level of difficulty, each word was assigned a difficulty score based on the average of three binary ratings: (1) length (at least four letters = 1, less than four letters = 0), (2) syllables (at least two syllables = 1, under two syllables = 0), and (3) familiarity (based on a psycholinguistic database where above average = 1, below average = 0). Then, the average difficulty across a running window of 10 words had to be above a threshold set at .45 for a reader to be classified as "mindless reading." The second criterion was based on reading time. Participants read one word on a screen at a time. Using a running window of 10 words, a specific threshold (based on pilot data) was applied to determine when readers were reading either too fast or too slow.

Table 1. Overview of Previous Mind Wandering Detectors

	Key Features	Classification Accuracy	Validation Method
Bixler et al. (2014)	Eye Gaze	72% correct	leave-several-subjects-out
Blanchard et al. (2014)	Physiology	74% correct	leave-several-subjects-out
Drummond et al. (2010)	Prosodic/ Lexical	64% correct	leave-one-instance-out
Franklin et al. (2011)	Difficulty/ Reading Time	72% correct	thresholds derived from pilot data

This study provided some evidence that reading time, combined with textual features such as difficulty, might be indicative of mind wandering (accuracy = 72%). However, since reading times were collected by presenting one word on the screen at a time, their methods and predetermined thresholds for fast and slow reading may not be generalizable to other, more natural, reading contexts. Additionally, mind wandering was never predicted to occur during "easy" portions of the text, which may not accurately reflect the real-life occurrence of this phenomenon. For example, mind wandering still occurs around 20% during easy texts [27], even though it is more frequent during difficult texts. Furthermore, their method relied on a number of pre-set thresholds with little information on how these thresholds were established, thereby complicating attempts to replicate their results.

1.2 Current Study

This paper reports a person-independent detector of mind wandering during a more natural, computerized self-paced reading task using basic information that can be extracted from reading logs. In an attempt to provide a foundation for an easily-scalable way to capture when mind wandering occurs, the detector is completely sensor-free.

The mind wandering detector was trained on two unpublished datasets in which participants attempted to learn about scientific research methods by reading texts presented online. Participants completed a posttest after reading in order to assess learning. Importantly, these datasets include diversity with respect to population, methods, and level of text difficulty. For example, dataset 1 was collected via Mechanical Turk, a validated online data collection platform [23], and had an average age of 35 years. Dataset 2 was collected from a Midwestern university subject pool and had an average age of 19 years. Therefore, building a detector of mind wandering using more than one dataset with varying conditions will increase our confidence in its relative generalizability.

2. DATASETS

The datasets were originally collected to investigate mind wandering under various conditions, such as varying levels of difficulty and text presentations. In addition, a posttest was completed after reading in order to assess how mind wandering relates to learning. In both datasets, participants were instructed to read the text carefully and notified that they would be asked to answer questions about content from the text after reading. Dataset 1 (N = 177) was collected on Amazon's Mechanical Turk, an online data collection platform that has been validated for high quality data [23, 28]. Participants were compensated $2.50 after completing the experiment. Dataset 2 (N = 141) was collected via an online subject pool at a Midwestern university in the United States. Participants received class credit after completing the study.

Table 2 provides an overview of the experimental designs and manipulations used in each dataset. The Text Difficulty manipulation involved participants reading texts that were experimentally manipulated to be either "easy" or "difficult" (see section 2.1 for manipulation details). The Text Presentation manipulation involved participants reading either one sentence or one paragraph at a time on the screen.

2.1 Reading Materials

The two texts used in the existing datasets were adapted from texts used in the serious game, *Operation ARA!* [25]. Each text focused on a concept related to research methods: (1) the dependent variable and (2) making causal claims, both of which are key concepts relevant to understanding the scientific method. In the existing datasets, easy and difficult versions of each text

were used in order to investigate the effect of text difficulty on mind wandering.

Easy versions of the text were more narrative in nature, and consisted of shorter sentences and fewer low frequency words (average Flesh-Kincaid Grade Level = 9). Difficult versions of the text consisted of longer, more complex sentences with more low frequency words (average Flesh-Kincaid Grade Level = 13). Both versions had the same conceptual content and were approximately 1500 words in length. An example of an easy sentence is, "People who know about the scientific method do not fall for unsupported claims like this one." The difficult version of the same sentence was, "So many citizens fall for these dubious claims, but people who comprehend the scientific method are not victimized by these unsupported claims."

2.2 Procedure

Participants first completed an electronic consent form. They were then given instructions for the self-paced learning task. Participants pressed the space bar to move through each screen of the text. Texts were presented on screen either one sentence at a time or one paragraph at a time based on experimental manipulation (see Table 2).

Mind wandering was tracked via auditory thought probes in both datasets. A standard description of mind wandering [36] was employed: "At some point during reading the texts, you may realize that you have no idea what you just read. Not only were you not thinking about the text, you were thinking about something else altogether." The probe consisted of an auditory beep that occurred on pseudo-random screens throughout each text. Probes were triggered when participants pressed the space bar to advance to the next portion of the text. Participants were instructed to press the "Y" key if they were mind wandering or the "N" key if they were not. Participants were not able to advance to the next screen until they had responded to the mind wandering probe. A total of six auditory mind wandering probes were inserted in each text. Probes were placed in an identical location with respect to content within each text. That is, regardless of whether the text was presented one sentence or paragraph at a time, the probe would occur after reading identical content.

Table 2. Overview of Two Datasets

	Dataset 1	**Dataset 2**
Sample	Mechanical Turk	University subject pool
# Texts	1	2
# Participants	177	141
Manipulations:		
Text Difficulty	Easy/Difficult	Difficult only
Text Presentation	Par/Sen	Par/Sen

Notes. Par = Paragraph-by-paragraph; Sen = sentence-by-sentence

Participants completed a posttest after reading each topic. Posttests consisted of four-alternative multiple-choice questions that tapped two levels of comprehension: (1) surface level, and (2) inference level. Surface level questions were based on factual or text level characteristics of the text. Inference questions were designed to elicit patterns of reasoning and require participants to use inference or apply a learned concept to a novel example in order to answer the question correctly [19]. For dataset 2, participants answered an 18-item posttest that covered both topics, which included six inference and 12 surface level multiple-choice questions. Since only one text was read during dataset 1, the posttest was limited to the 9 corresponding questions (3 inference and 6 surface level questions).

2.3 Mind Wandering Reports

Every screen of text where a probe was triggered was classified as either "Mind Wandering" or "Not Mind Wandering" based on participants' response to the probe. The two datasets were pooled in order to maximize training and validation data. In total, there were 2754 probe screens that were used to build the models. Participants indicated they were mind wandering in response to 31.3% of all the probes. Thus, our data set contained 861 instances of Mind Wandering and 1893 instances of Not Mind Wandering.

3. MODEL BUILDING

3.1 Feature Engineering

A considerable amount of empirical research has been dedicated to understanding mind wandering through experimental manipulations, such as comparing mind wandering across various conditions. Other studies have focused on explaining the behavioral correlates and temporal patterns of mind wandering [14, 16, 16, 27, 34, 38, 40]. The features in the current research were informed by the following discoveries about mind wandering: First, mind wandering is affected by the difficulty of a task [14, 27]. Second, mind wandering is related to response times and lexical features [15, 29]. Third, mind wandering rates vary as a function of time on task [30, 40]. In line with these findings, a total of 13 features were computed based on information that can found in log files. The 13 features can be subdivided into three categories: (1) Reading Behavior Features (2 features), (2) Textual Features (8 features), and (3) Context Features (3 features).

Reading Time Features. Participants' reading time (i.e. how long they spent on each screen) was collected during the reading task. Importantly, the thought-probe was triggered as participants attempted to move on to the next screen. Therefore, we can use reading behaviors from the current screen of text (screen K) to detect whether they are mind wandering or not before they moved on to the next screen (K+1).

The first reading behavior feature was *Reading Time*, which was simply the amount of time spent reading a given paragraph before pressing the space bar to advance onto the next screen. Reading Time was computed at the paragraph level in order to account for differences in reading times across the Text Presentation manipulation. When texts were presented one paragraph at a time, *Reading Time* was simply how long they spent on the screen leading up to the thought-probe. When texts were presented one sentence at a time, sentences were aligned with the content from the paragraph presentation condition. Thus, *Reading Time* was calculated as the amount of time spent reading identical content before the thought-probe regardless of presentation style.

The second reading behavior feature was called *Decoupling* [41]. *Decoupling* is a theoretically-driven metric based on the idea that reading times should increase with more complex text characteristics, such as sentence length and other discourse features [18]. If participants are not appropriately allocating resources (i.e., increasing reading times when text complexity increases) to meet the current task demands, then we might expect deviation from this linear relationship thus indicating decoupling

from the reading task. *Decoupling* was computed on the alignment (or misalignment) of reading times and text complexity. Text complexity was assessed using Flesh-Kincaid Grade Level (FKGL; [22]). The formula used to calculate decoupling was: |*z*-score standardized reading times – *z*-score standardized FKGL|. It is important to point out that decoupling was computed using the absolute value of the difference between reading time and text complexity, such that higher values would occur both when reading times were both over and under appropriated relative to text complexity. Thus, we are primarily interested in how well the overall magnitude of deviation in the relationship between reading time and text complexity can predict mind wandering.

Textual features. Eight textual features were computed in total. The first feature was simply the *Number of Characters* in the current paragraph. The second feature was the *Number of Words* in the current paragraph. Both features were used because they may differ notably between easy and difficult conditions, as easy texts were specifically manipulated to contain shorter words. Regardless of whether the screen was being presented one paragraph at a time or one sentence at a time, these features were used to represent the length of the current unit of text being processed. Longer paragraphs may require increased cognitive resources (related to mind wandering [24]) when a single idea must be kept in working memory across larger amounts of text. The third feature was *FKGL* [22], an indicator of reading level that is derived from the number of syllables and word length in a sentence. The current FKGL was also computed based on the current paragraph being read, as this metric is not reliable for extremely small portions of text, such as a single sentence.

The remaining five textual features were computed using Coh-Metrix, a program that analyzes texts across multiple levels of cognition and comprehension [17, 18]. We used five different features from Coh-Metrix: (1) Narrativity, (2) Deep Cohesion, (3) Referential Cohesion, (4) Syntactic Simplicity, and (5) Word Concreteness. *Narrativity* is computed based on how well the text aligns with the narrative genre, by conveying a story, procedure, or sequence of actions. *Deep Cohesion* is computed based on how well different ideas in the text are cohesively tied together in order to signify causality or intentionality. *Referential Cohesion* is based on how words and ideas are connected to each other across the span of the story or text. *Syntactic Simplicity* is computed based on the simplicity of the syntactic structures in the text. Lastly, *Word Concreteness* is based on the degree to which context words evoke concrete mental images, rather than abstract or conceptual representations.

Context features. Three context features were also computed based on the context of the reading task. *Current Paragraph Number* is the number of paragraphs read from the beginning of the text. *Current Difficulty* is whether the text was experimentally manipulated as easy or difficult. *Current Presentation* is whether the text was being presented one sentence at a time or one paragraph at a time.

3.2 Supervised Classification and Validation

We used supervised machine learning to build detectors of mind wandering for each screen that included a thought-probe. The goal of the paper was to create a detector that would accurately predict whether participants responded "yes" or "no" to the mind wandering probes. RapidMiner, a popular machine learning tool, was used to train binary classifiers to make this distinction. In total, four binary classifiers provided in RapidMiner were used, including Naïve Bayes, Bayes Net, RIPPER (JRip implementation), and C4.5 (J48 implementation). Down-sampling was used to create equal classes for the training data only. This was achieved by randomly selecting 45.4% of the Not Mind Wandering instances and 100% percent of the Mind Wandering instances for training. The original distributions were not changed in the testing data to preserve the validity of the results.

Manual feature selection was applied by removing one feature at a time and assessing performance on held-out testing data (see below). If model performance decreased after a feature was removed, it was preserved for the final model[1].

All models were evaluated using leave-one-participant-out cross-validation, in which *k*-1 participants are used in the training data set. The model was then tested on the participant who was not used in the training data. This process was repeated *k* times until every participant was used as the testing set once. Cross-validating at the participant level increases confidence that models will be more generalizable when applied to new participants because the testing and training sets are independent.

Classification accuracy was evaluated using two metrics: (1) Area Under the ROC Curve (AUC), and (2) Cohen's kappa. AUC is statistically similar to A′ [21] and ranges from 0 to 1, where 0.5 is chance level of accuracy and 1 is perfect accuracy. Cohen's kappa [10] indicates the degree to which the model is better than chance (kappa of 0) at correctly predicting Mind Wandering or Not Mind Wandering. A kappa of 1 indicates the detector performs perfectly. We also report percent correctly classified (accuracy), but note that this should be interpreted cautiously since class imbalance tends to inflate accuracy.

4. RESULTS

4.1 Classification Accuracy

Four classification algorithms (J48, JRIP, Naïve Bayes, and Bayes Net) were applied to the two combined datasets. The final models reported in this section were selected based on the highest AUC achieved after testing all four classification algorithms. A final combined feature model (combined model) was achieved with the J48 decision tree classifier using six features from the feature subtypes: *Reading Time*, *Decoupling*, *Number of Characters*, *Number of Words*, *FKGL*, and *Referential Cohesion.* Importantly, the combined model performed at rates above chance (AUC = .609; kappa = .207; accuracy = 63%). Despite using information solely obtained from log files and text characteristics, these accuracy rates are only slightly lower than the sensor-based detectors of mind wandering reported in Table 1.

We also examined the confusion matrix for the final combined model (see Table 3). The model had a relatively high rate of misses (.427), where actual instances of Mind Wandering were predicted as Not Mind Wandering. However, the model also displayed more correct rejections (.653), such that Not Mind Wandering instances were accurately classified as Not Mind Wandering. This was complemented by a low rate of false alarms as well (.347).

We were also interested in exploring how each of the three feature subtypes (i.e., reading behaviors, textual, and context features) were able to predict mind wandering independently. Each group of feature subtypes was therefore tested independently using the same four classification algorithms (J48, JRIP, Naïve Bayes, and Bayes Net). A summary of the classification accuracies for the

1 We also tested models using all 13 features, which exhibited lower performance (assessed via AUC) than the combined model using feature selection.

best performing models (selected based on highest AUC) can be found in Table 4.

Table 3. Confusion Matrices of Combined Model

	Pred. MW	**Pred. Not MW**	**Priors**
Actual MW	.573 (hit)	.427 (miss)	.313
Actual Not MW	.347 (false alarm)	.653 (correct rejection)	.687

Note. Pred. = Predicted; MW = Mind Wandering

All three models built from the feature subtypes performed above chance levels (AUC > .5). However, none of these models performed as well as the combined model. For example, the Textual Features Only model did not perform as well in the absence of reading time behaviors and vice versa. This suggests that using a range of feature types might help with classification accuracies rather than a subset of features.

Based on the confusion matrices, it appears that the three feature subtype models exhibited different patterns of classification (see Table 5). Although the Reading Behaviors Only model (*Reading Time* and *Decoupling*) displayed the lowest hit rates (.439), this model also had the highest rate of correct rejections. Conversely, the Textual Features Only (five Coh-Metrix dimensions, *Number of Characters*, and *Number of Words*) and the Context Features Only (*Current Presentation*, *Current Difficulty*, and *Current Paragraph Number*) models had similar higher hit rates, but fewer correct rejections compared to the Reading Behaviors Only Model.

Table 4. Performance Metrics

Features in model	**AUC**	**Kappa**	**Classifier**
Combined Model	.609	.207	J48
Reading Behaviors Only	.560	.122	J48
Textual Features Only	.591	.115	Bayes Net
Context Features Only	.542	.104	JRIP

It is important to point out that the combined model's confusion matrix also shared some similarities with the feature subtype models. The Reading Behavior Only model had the highest correct rejections (.687), which were on par with the combined model (.653). Similarly, the Textual Features Only and Context Features Only models had the best hit rates (.554 and .557), which were also on par with the hit rates in the combined model (.573). Thus, the combined model appears to strike a balance between hits and correct rejection, which is why it yields the highest AUC compared to the individual models.

4.2 Feature Analysis

Since our features were modeled after empirically-supported relationships of mind wandering (see Section 3.1), we explored how our features related to the model's predictions of mind wandering. For each participant, we computed the mean of each feature as well as the proportion of predicted mind wandering (based on the combined model's predictions). As an additional step, the averages were *z*-score standardized across the two datasets to account for the differences in methods. Predicted mind wandering was then regressed on each of the six features included in the combined model, $F(6,317) = 35.5$, $p < .001$, $R^2_{adjusted} = .395$. The regression allowed us to examine the relationship between each of the features and predicted mind wandering while controlling for the other features in the model. Table 6 presents a summary of the features used the combined model, as well as the standardized regression coefficient (β) for each feature.

Table 5. Confusion Matrices for Each Feature Set Separately

Reading Behavior	*Pred. MW*	*Pred. Not MW*
Actual MW	.439 (hit)	.561 (miss)
Actual Not MW	.313 (false alarm)	.687 (correct rejection)
Textual Features	*Pred. MW*	*Pred. Not MW*
Actual MW	.554 (hit)	.446 (miss)
Actual Not MW	.424 (false alarm)	.576 (correct rejection)
Context Features	*Pred. MW*	*Pred. Not MW*
Actual MW	.557 (hit)	.443 (miss)
Actual Not MW	.432 (false alarm)	.568 (correct rejection)

Note. Pred. = Predicted; MW = Mind Wandering

Reading Time was negatively related to predicted mind wandering, indicating that mind wandering predictions were associated with faster reading times. The second reading behavior feature, *Decoupling*, was positively related to predicted mind wandering. Mind wandering was more likely to be predicted when decoupling scores were higher, since higher decoupling scores indicate a misalignment between reading times compared to text complexity.

Number of Characters and *Number of Words* were both positively related to predicted mind wandering, suggesting that more content in general is associated with greater predictions of mind wandering. This is also related to the idea that longer paragraphs may have demanded increased cognitive resources, which is theoretically related to episodes of mind wandering [24].

Table 6. Standardized coefficients for regressing predicted mind wandering on features in the combined model (*β*)

Features Included in Combined Model	**Standardized Coefficient (β)**
Reading Behavior Features	
Reading Time	**-.750**
Decoupling	**.493**
Textual Features	
Number of Characters	**.139**
Number Words	.099
Referential Cohesion	**-.139**
FKGL	**.239**

Notes. Bold = significant at $p < .05$; FKGL = Flesch Kincaid Grade Level.

Referential Cohesion was also negatively related to predicted mind wandering. This relationship is theoretically plausible, as

breakdowns in *Referential Cohesion* are indicative of increased difficulty [20]. Indeed, difficulty has been found to be related to mind wandering during reading [14, 27].

None of the Context features were included in the combined model. This was an unexpected result, since time on task has previous been correlated to mind wandering [40] and the previous detectors of mind wandering have utilized context features [5, 6]. It is possible that one of the Context Features, *Current Difficulty*, may not have been useful in the combined model due, in part, to the fact that the textual features were essentially more sensitive measures of difficulty. For example, FKGL and Referential Cohesion may be more sensitive measures of *Current Difficulty*.

4.3 Predictive Validity

In order to establish predictive validity for the detector, we ascertained if *predicted* mind wandering relates to learning similar to actual (self-reported) mind wandering rates? Based on previous research, we expect a negative relationship between actual mind wandering and learning [11, 32, 39]. To address this question, posttest performance was first correlated with *actual* rates of mind wandering (i.e., responses to the thought probes). Participants' posttest performance was calculated as the proportion of correct responses for the surface- and inference-level questions separately. The variables were standardized across the two datasets to account for any differences in populations. Indeed, *actual* mind wandering was negatively related to both surface (Spearman's *rho* = -.338, $p < .001$) and inference level (*rho* = -.288, $p < .001$) comprehension on the posttest.

To establish the predictive validity of the detector, we ascertained if *predicted* mind wandering was related to posttest performance similar to actual mind wandering. *Predicted* mind wandering rates (from the combined detector) was negatively correlated with surface level (*rho* = -.294, $p < .001$) as well as inference level performance on the posttest *(rho* = -.193, $p = .008$). The negative correlations with both types of posttest performance gives us some confidence in our model's predictive validity, since predicted mind wandering shows similar relationships with learning as actual self-reported mind wandering. This finding is notable since the model predicted mind wandering correctly around 20% above chance (kappa = .207), yet *predicted* mind wandering related almost as well to posttest scores as *actual* rates of mind wandering.

5. GENERAL DISCUSSION

Mind wandering is a ubiquitous phenomenon that is negatively related to learning [11, 32, 39]. Mind wandering can have a detrimental impact on comprehension when pieces of information are not accurately integrated into a learner's mental model of the instructional texts. Over time, information missed during episodes of mind wandering can accumulate, leaving deficits in the learner's overall understanding of a text. The development of attention-aware systems may provide opportunities to restore learners' attention in real-time to facilitate learning. However, we must first be able to detect mind wandering in order to respond to its occurrence.

We attempted to address this issue by developing a participant-independent detector of mind wandering through analyzing log files and textual characteristics collected during an online reading task. Two diverse datasets were used to ensure further generalizability. The detector was able to accurately classify mind wandering 20% above chance (kappa = .207; *AUC* = .609). Given that mind wandering is an elusive internal state of attention and we used completely sensor-free data, modest classification accuracies are to be expected. Additionally, the classification accuracy found in this study (63%) is only slightly lower than those reported for previous detectors built using sensor-based approaches including eye gaze and physiology (See Table 1; [5, 6].

Three types of features were used to build the mind wandering detector: (1) reading behaviors, (2) textual features, and (3) context features. An independent model was built for each subtype of features, which allowed us to better understand how the subtypes of feature perform independently. Each set of features was able to correctly classify mind wandering independently at levels above chance, though performance varied across models. None of these models outperformed the combined model, so we conclude that combining different types of features was optimal in the current detector. Thus, future research may consider using one or more of these subtypes of features, as they are relatively easy to extract from log files.

Many of the features were included based on previous psychological and educational research on mind wandering. The relationships between the features and predicted rates of mind wandering were revealing in a number of ways. For example, a negative relationship between Referential Cohesion and predicted mind wandering directly supports the situation model view of text comprehension [14, 35]. This view posits that reading involves the construction of a *situation model*, which is a constantly-updated mental representation of a text's meaning [18, 43]. Situation models are harder to construct during difficult texts due to inconsistencies or lack of cohesion. Poorly constructed situation models consume fewer attentional resources, leaving extra resources available for off-task thoughts. Therefore, this theory would predict a negative relationship between mind wandering Referential Cohesion, which is what we find.

Response times as well as reading time information have been utilized in previous detectors of off-task states like disengagement [4, 7, 8]. Thus, it is not surprising that both reading time behavior features were related to predicted mind wandering. A negative relationship with Reading Time indicates that shorter reading times were indicative of increased mind wandering predictions. It is also worth noting that Decoupling, which is derived from a theoretically-supported relationship between reading time and text complexity, was positively related to predicting mind wandering. Indeed, these relationships suggest features based on reading times may be used a behavioral indices of attention during reading.

Our detector also showed some evidence for predictive validity. Predicted mind wandering was negatively related to posttest performance, similar to actual mind wandering. Future work should explore other avenues of establishing validity using other online measures of engagement and comprehension. Similar to [15], another method of validation would be to trigger thought probes on the pages where mind wandering is predicted in real-time. We could then evaluate responses to the predicted episodes of mind wandering in order to determine how accurate the model performs in a real-time detection setting.

It is important to note that these models are not without limitations. First, these models were built in the context of an instructional reading task, which may not generalize to other learning environments. Second, although two independent datasets were used, our results cannot currently be generalized beyond the current sample. Third, although self-reports of mind wandering using a thought-probe method have been validated in previous studies [35, 36], they depend on participants accurate

and honest responses. Additionally, given the internal nature of mind wandering, external coders are not a viable option. Therefore, future work may consider using a different method of probing, where participants might self-monitor and report instances of mind wandering at any point during reading [31] (as opposed to only at times when thought-probes occur). Finally, there is no known research establishing a way to determine the onset of mind wandering in real-time [37]. Thus, while detectors to date are able to predict instances of self-reported mind wandering (which is inherently realized), no method has been established to indicate how long the episode lasts or when it began.

Future work may include attempts to improve these models using additional features. For example, additional sensor-free features, such as trait-based features like prior knowledge and interest might further improve prediction rates. In addition, combining features developed here with previous detectors of mind wandering may also improve prediction rates (e.g., eye gaze). It is possible that combining multiple channels of data may have an additive effect to improve prediction rates.

In summary, this paper provides some initial evidence for a sensor-free detector of mind wandering during online instructional reading. A sensor-free detector of mind wandering may open up new avenues for interventions and instructional designs in order to facilitate attention. Previous detectors for disengagement behaviors, such as gaming the system and Gaze Tutor, have been used in the design of interventions, such as reintroducing the content that is missed due to gaming [2] and providing engaging dialogue to redirect students' attention [12]. The detector presented in this paper is an initial step for interventions that can occur when the mind wanders away from the current task. We believe further development of these types of models is promising for creating an attention-aware system that can respond in real-time.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation (NSF; DRL 1235958). 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] Arroyo, I. et al. 2007. Repairing disengagement with non-invasive interventions. *AIED* (2007), 195–202.

[2] Baker, R.S.J. et al. 2006. Adapting to when students game an intelligent tutoring system. *Intelligent Tutoring Systems* (2006), 392–401.

[3] Baker, R.S.J. 2007. Modeling and understanding students' off-task behavior in intelligent tutoring systems. *Proceedings of the SIGCHI conference on Human factors in computing systems* (2007), 1059–1068.

[4] Beck, J.E. 2004. Using response times to model student disengagement. *Proceedings of the ITS2004 Workshop on Social and Emotional Intelligence in Learning Environments* (2004), 13–20.

[5] Bixler, R. and D'Mello, S. 2014. Toward Fully Automated Person-Independent Detection of Mind Wandering. *User Modeling, Adaptation, and Personalization*. Springer. 37–48.

[6] Blanchard, N. et al. 2014. Automated Physiological-Based Detection of Mind Wandering during Learning. *Intelligent Tutoring Systems* (2014), 55–60.

[7] Cocea, M. and Weibelzahl, S. 2006. Can Log Files Analysis Estimate LearnersLevel of Motivation?. *LWA* (2006), 32–35.

[8] Cocea, M. and Weibelzahl, S. 2007. Cross-system validation of engagement prediction from log files. *Creating New Learning Experiences on a Global Scale*. Springer. 14–25.

[9] Cocea, M. and Weibelzahl, S. 2011. Disengagement Detection in Online Learning: Validation Studies and Perspectives. *IEEE Transactions on Learning Technologies*. 4, 2 (Apr. 2011), 114–124.

[10] Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*. 20, 1 (1960), 37–46.

[11] Dixon, P. and Bortolussi, M. 2013. Construction, integration, and mind wandering in reading. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*. 67, 1 (2013), 1.

[12] D'Mello, S. et al. 2012. Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of Human-Computer Studies*. 70, 5 (2012), 377–398.

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

[14] Feng, S. et al. 2013. Mind wandering while reading easy and difficult texts. *Psychonomic bulletin & review*. (2013), 1–7.

[15] Franklin, M.S. et al. 2011. Catching the mind in flight: Using behavioral indices to detect mindless reading in real time. *Psychonomic Bulletin & Review*. 18, 5 (Oct. 2011), 992–997.

[16] Franklin, M.S. et al. 2013. Thinking one thing, saying another: The behavioral correlates of mind-wandering while reading aloud. *Psychonomic Bulletin & Review*. (Jun. 2013).

[17] Graesser, A.C. et al. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*. 36, 2 (2004), 193–202.

[18] Graesser, A.C. et al. 2011. Coh-Metrix Providing Multilevel Analyses of Text Characteristics. *Educational Researcher*. 40, 5 (2011), 223–234.

[19] Graesser, A.C. et al. 2010. What is a good question? *Bringing reading research to life*. Guilford Press. 112–141.

[20] Graesser, A.C. and McNamara, D.S. 2011. Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*. 3, 2 (2011), 371–398.

[21] Hanley, J.A. and McNeil, B.J. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 143, 1 (1982), 29–36.

[22] Klare, G.R. 1974. Assessing Readability. *Reading Research Quarterly*. 10, 1 (Jan. 1974), 62–102.

[23] Mason, W. and Suri, S. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*. 44, 1 (2012), 1–23.

[24] McVay, J.C. and Kane, M.J. 2010. Does mind wandering reflect executive function or executive failure? Comment on Smallwood and Schooler (2006) and Watkins (2008). (2010).

[25] Millis, K. et al. 2011. Operation ARIES!: A serious game for teaching scientific inquiry. *Serious games and edutainment applications*. (2011), 169–195.

[26] Mills, C. et al. 2014. To Quit or Not to Quit: Predicting Future Behavioral Disengagement from Reading Patterns. *Intelligent Tutoring Systems* (2014), 19–28.

[27] Mills, C. et al. 2013. What Makes Learning Fun? Exploring the Influence of Choice and Difficulty on Mind Wandering and Engagement during Learning. *Artificial Intelligence in Education* (2013), 71–80.

[28] Rand, D.G. 2012. The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of theoretical biology*. 299, (2012), 172–179.

[29] Reichle, E.D. et al. 2010. Eye Movements During Mindless Reading. *Psychological Science*. 21, 9 (Sep. 2010), 1300–1310.

[30] Risko, E.F. et al. 2012. Everyday attention: variation in mind wandering and memory in a lecture. *Applied Cognitive Psychology*. 26, 2 (2012), 234–242.

[31] Schooler, J.W. et al. 2011. Meta-awareness, perceptual decoupling and the wandering mind. *Trends in cognitive sciences*. 15, 7 (2011), 319–326.

[32] Smallwood, J. et al. 2007. Counting the cost of an absent mind: Mind wandering as an underrecognized influence on educational performance. *Psychonomic Bulletin & Review*. 14, 2 (2007), 230–236.

[33] Smallwood, J. 2011. Mind-wandering while reading: Attentional decoupling, mindless reading and the cascade model of inattention. *Language and Linguistics Compass*. 5, 2 (2011), 63–77.

[34] Smallwood, J. et al. 2009. Shifting moods, wandering minds: negative moods lead the mind to wander. *Emotion*. 9, 2 (2009), 271.

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

[36] Smallwood, J. and Schooler, J.W. 2006. The restless mind. *Psychological bulletin*. 132, 6 (2006), 946.

[37] Smallwood, J. and Schooler, J.W. 2015. The science of mind wandering: empirically navigating the stream of consciousness. *Annual review of psychology*. 66, (2015), 487–518.

[38] Szpunar, K.K. et al. 2013. Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences*. 110, 16 (2013), 6313–6317.

[39] Szpunar, K.K. et al. 2013. Mind wandering and education: from the classroom to online learning. *Frontiers in psychology*. 4, (2013).

[40] Thomson, D.R. et al. 2014. On the link between mind wandering and task performance over time. *Consciousness and cognition*. 27, (2014), 14–26.

[41] Vega, B. et al. 2013. Reading into the Text: Investigating the Influence of Text Complexity on Cognitive Engagement. *Proceedings of the 6th international conference on educational data mining* (2013), 296–299.

[42] Wixon, M. et al. 2012. WTF? detecting students who are conducting inquiry without thinking fastidiously. *User Modeling, Adaptation, and Personalization*. Springer. 286–296.

[43] Zwaan, R.A. and Radvansky, G.A. 1998. Situation models in language comprehension and memory. *Psychological bulletin*. 123, 2 (1998), 162.