Automatic Gaze-Based Detection of Mind Wandering with Metacognitive Awareness

Robert Bixler(✉) and Sidney D'Mello

Department of Computer Science, University of Notre Dame, Notre Dame, IN 46556, USA
{rbixler,sdmello}@nd.edu

Abstract. Mind wandering (MW) is a ubiquitous phenomenon where attention involuntarily shifts from task-related processing to task-unrelated thoughts. There is a need for adaptive systems that can reorient attention when MW is detected due to its detrimental effects on performance and productivity. This paper proposes an automated gaze-based detector of self-caught MW (i.e., when users become consciously aware that they are MW). Eye gaze data and self-reports of MW were collected as 178 users read four instructional texts from a computer interface. Supervised machine learning models trained on features extracted from users' gaze fixations were used to detect pages where users caught themselves MW. The best performing model achieved a user-independent kappa of .45 (accuracy of 74% compared to a chance accuracy of 52%); the first ever demonstration of a self-caught MW detector. An analysis of the features revealed that during MW, users made more regression fixations, had longer saccades that crossed lines more often, and had more uniform fixation durations, indicating a violation from normal reading patterns. Applications of the MW detector are discussed.

Keywords: Gaze tracking · Mind wandering · Affect detection · User modeling

1 Introduction

A promising strategy to improve the effectiveness of adaptive systems is to take aspects of the user's mental state into account - commonly referred to as user state estimation. Monitoring a user's mental state allows for dynamic strategies such as reorienting their attention to the interface when they become distracted [6] or detecting and responding to affective states such as confusion or boredom [3]. One user state that has received little attention until recently is mind wandering (MW). MW is a pervasive phenomenon that involves thinking about one thing while doing another. For instance, while reading a book or listening to a lecture, it is possible for an individual's thoughts to involuntarily drift toward unrelated thoughts such as unfulfilled plans and anxieties. The frequency of MW depends on the individual and environmental context, but a large-scale experience-sampling study on about 5,000 individuals estimated that MW occurs roughly 40% of the time [13].

Not only is MW frequent, it is also disruptive and detrimental to performance. This is because MW entails a shift in attention from the external environment to internal

F. Ricci et al. (Eds.): UMAP 2015, LNCS 9146, pp. 31–43, 2015.
DOI: 10.1007/978-3-319-20267-9_3

thoughts. Tasks requiring conscious focus are compromised when attention is directed toward task-unrelated thoughts. Hence, research has indicated that MW leads to performance failures on a number of tasks, such as increased error rates during signal detection tasks [20], lower recall during memory tasks [22], and poor comprehension during reading tasks [9], even when users were able to catch themselves mind wandering [18]. There are some potential benefits to MW, in that it boosts creativity and facilitates the planning of future events [14]. However, these benefits are not the norm as a recent meta-analysis on 49 independent samples found that MW was consistently negatively correlated with performance across a range of tasks [17].

The high incidence and negative influence of MW on performance suggests that there might be advantages for adaptive interfaces that reorient attention to the task at hand when MW occurs. This requires MW detection, which is a challenging proposition because MW has more covert cues than some other user states (e.g., facial expressions conveying emotions). This raises the pertinent question of how one collects labeled data (MW reports) to train supervised classifiers for MW detection. Two common methods have emerged in the literature. The first is to ask users to provide MW reports in response to thought probes (probe-caught) [22]. Users are asked to indicate if they are MW (positive instances) or not (negative instances) at the moment the probe is triggered. The second is to ask users to provide MW reports whenever they catch themselves MW (self-caught) [20]. There is a distinct difference between probe-caught and self-caught reports of MW, so it is possible that a detector built from probe-caught MW would not be useful for detecting self-caught MW. First, detection of self-caught episodes of MW does not require the use of potentially disruptive thought-probes. Second, self-caught reports rely on a user's ability to monitor their own thoughts and realize that they are MW. Therefore, they reflect a form of MW that occurs with metacognitive awareness. As discussed in the related works below, all of the previous work on MW detection has focused on probe-caught reports. In this paper we introduce the first automatic user-independent detector of MW with metacognitive awareness. As elaborated below, this raised a number of technical challenges that needed to be addressed.

Related Work. A large amount of research has been done in the field of attentional state estimation, which is a subfield of user state estimation. Attentional state estimation has been explored in a variety of domains and with a variety of end-goals. For example, attention has been used to evaluate adaptive hints in an educational game [15] and to optimize the position of news items on a screen [16]. Attentional state estimators have been developed for several tasks such as identifying object saliency during video viewing [25], and for monitoring driver fatigue and distraction [7]. Although both attentional state estimation and MW detection entail identifying aspects of a user's attention, MW detection is concerned with detecting more covert forms of involuntary attentional lapses as opposed to determining which aspects of the stimulus were being attended to.

In recent years, there have been five studies that have explicitly investigated detection of MW [1, 2, 5, 8, 10]. MW was tracked in each study with online self-reports and with behavioral measures derived from eye gaze, speech, physiology, or reading

times. Supervised machine learning was then used to predict the occurrence of each self-report from the behavioral measures.

The first study, by Drummond and Litman [8], entailed detecting MW using acoustic-prosodic information while users read a biology paragraph aloud and then provided a verbal summary. MW was tracked at set intervals where users indicated their degree of "zoning out" on a 7 point Likert scale. Their model discriminated between "high" versus "low" zone outs with an accuracy of 64%, which reflects a 22% improvement over chance. However, it is unclear if their model will generalize to new users.

The second study, by D'Mello et. al [5], used eye gaze data to detect MW during reading with reports of MW obtained in response to auditory probes that were triggered at random points during the reading session. Their best performing model yielded a detection accuracy of 60% on a down-sampled corpus containing 50% "yes" and 50% "no" responses (20% improvement over chance). This study did ensure generalizability to new users, but both the training and testing set were down-sampled prior to classification, so it is unclear if a similar level of fidelity will be observed with the authentic MW distribution.

The third study, by Franklin et. al [10], used reading times to detect MW. Users read 5000 words one at a time, using the space bar to advance to the next word. MW probes were triggered if a user spent too much or too little time on a group of ten words. They were able to classify mind wandering with an accuracy of 72% compared to an expected accuracy of 49%. However, the word-by-word reading paradigm is not necessarily representative of normal reading, and it is unclear if their method will generalize due to the method used to set parameters (parameter values were fixed rather than learned).

The fourth study, by Blanchard et. al [2], detected MW during reading using galvanic skin response and skin temperature obtained with the Affectiva Q sensor. They were able to achieve an above-chance classification accuracy of 22% in a manner that generalized to new users.

The fifth study by, Bixler et. al [1], extended the work of D'Mello et. al [5] and attempted to improve on the results of the Blanchard et. al study [2]. Using an expanded version of the data set from the Blanchard et. al study, we included additional eye-gaze features and performed a more comprehensive analysis of the models developed by D'Mello et al. [5]. Out best models attained an above chance improvement of 28% in a user-independent fashion.

Current Work. Our work entails two major contributions over previous MW detectors: (1) we use a refined and extended feature set, and (2) we focus on building models of MW with metacognitive awareness. We refined our feature set by removing features derived from the task context, as these did not contribute to the performance of previous models and are not generalizable to other tasks. We then added several new features, including those derived from blinks [11], pupil diameters [23], and saccade angles.

Our second contribution is that we focused on detecting self-caught MW (MW with metacognitive awareness). Previous work focused on probe-caught MW, which makes our work the first self-caught MW detector. The lack of a self-caught MW detector represents a gap in the MW detection literature because self-caught MW is

distinct from probe-caught MW. Probes have the potential to disrupt the MW experience, while, in reality, MW fades naturally as users realize they are MW. There is also a limit to the number of probes that can be given. Probing too often may be perceived as irritating and may be disruptive to the primary task. Self-caught reports, on the other hand, are provided whenever a user catches themselves MW, which allows reports of MW to be collected whenever it occurs, provided the user is sufficiently capable at monitoring their thoughts. Self-caught MW detectors also have unique applications as listed in the General Discussion.

In line with this, the present study focuses on the use of eye-gaze for detection of self-caught MW during reading. The dataset used is the same as in our previous study [1], but in this study we use the self-caught reports, which have not been analyzed before. As discussed above, self-caught MW is different from probe-caught MW, so it is an entirely open question as to whether it is possible to detect self-caught MW from eye gaze. Furthermore, building a model from self-caught reports has its own set of complications that need to be addressed. For one thing, there is no explicit indication of when MW does *not* occur (called negative instances). Each self-caught report is a positive instance of MW, and each instance without a self-caught report has no data on whether MW occurred or not. A user could have been MW without realizing that they were MW (i.e., MW without meta self-awareness) and thus did not report it. Deciding what constitutes a negative instance of MW is a hurdle that is not encountered when building a model based on probe-caught reports of MW, as each probe response can be explicitly labeled as either a negative or positive instance of MW. There is also the issue of how to select an appropriate window of data to consider for each MW instance. There is a simple solution when using probe-caught MW – simply consider windows that backtrack from the time of the probe. However, the same method cannot be used for negative instances of self-caught MW because they do not include an explicit probe from which to backtrack. The present paper considers multiple approaches to obtain negative instances of MW as well as multiple methods for window selection.

The present work studies MW in the context of reading - an every-day activity that is supported by a number of systems. There is also ample data to suggest that reading comprehension is impaired by MW [9, 22], so there could be considerable benefits from embedding MW detectors in systems that support large amounts of reading (e.g., educational materials, legal texts, news articles, and many others). Further, in addition to demonstrating the first detector of self-caught MW in the context of reading, the results of our systematic experimentation to address the technical challenges discussed above should be useful for researchers interested in building self-caught MW detectors for their own application domains.

2 Data Collection

Users. The users in this study were 178 undergraduate students that participated for course credit. Of these, 93 users were from a medium-sized private Midwestern university while 85 were from a large public university in the mid-South. The average

age of users was 20 years (SD = 3.6). Demographics included 62.7% female, 49% Caucasian, 34% African American, 6% Hispanic, and 4% "Other". Thus, there was considerable gender- and ethnic- diversity in our sample.

Procedure. Users read four different texts on research methods topics (i.e., experimenter bias, replication, causality, and dependent variables). Each text contained 1500 words on average (SD = 10) split into 30-36 pages with approximately 60 words per page. Texts were presented on a computer screen with 36pt Courier New font. The typeface and size were chosen to make text layout simpler and to make it easier to determine when a word was being gazed upon.

Users were given standard instructions [21] on reporting MW. MW was defined as having "no idea what you just read" and realizing that "you were thinking about something else altogether." Both self- and probe-caught MW reports were collected. Probe-caught reports were collected in response to 9 auditory probes triggered on different pages at a randomly chosen point between 4 and 12 seconds from the appearance of the page. Users were also required to supply a report if they tried to advance to the next page before the probe was triggered. Alternatively, the self-caught method entailed users reporting MW whenever they found themselves MW.

Instances of Mind Wandering. Overall, users read 33,595 pages of text. About a third of these were discarded for the present study due to failure to register eye gaze. Of the remaining pages, 30% (6,718 pages) contained a mind wandering report. 6,237 of these reports were probe-caught reports, while 481 were self-caught reports. Of the probe-caught reports, 32% were positive instances of mind wandering. Only a subset of all users (78) provided self-caught reports of mind wandering, but all 178 users were included in the analysis when selecting negative instances of MW.

3 Model Building

Both negative and positive instances of MW are tracked when using probe-caught reports, so it is readily apparent that the classification task involves distinguishing negative probe-caught reports from positive probe-caught reports. When building models using only self-caught reports, however, there are no negative reports of MW. This raises the question: what should be used as a negative instance of MW?

There were three considerations taken when selecting negative instances of MW in our study. The first consideration regarded the types of pages that should be used as negative instances of MW. Pages fell into four categories; *self-caught pages, positive probe-caught pages, negative probe-caught pages*, and *non-report pages* (pages with no report). Positive probe-caught pages were not considered further since the goal was to detect self-caught MW. Pages that included reports of both types were considered to be a self-caught page if the self-caught report occurred first, and vice versa. All self-caught pages were included in the models as positive instances of MW. Negative instances were taken to be (a) only negative probe-caught pages (PC No), (b) only non-report pages (NR), or (c) both negative probe-caught pages and non-report pages (PC No + NR).

The second consideration was how to select the number of negative instances. We chose to randomly select from all available negative instances of MW to achieve a 0.3 proportion of positive instances of MW. This proportion corresponds to the proportion of positive probe-caught reports in our data and is similar to proportions of MW reported in previous studies on MW during reading [11, 23].

The third pertinent issue was to identify a point at which to analyze gaze data for the negative instances for non-report pages. For self-caught report pages we use a window that ended two seconds prior to the self-caught report in order to avoid any confounds associated with the user pressing the report key (e.g., movement, looking at the keyboard). The length of the window was a parameter that varied as noted below. For negative probe-caught pages, the window also ended two seconds before the probe response. One of three methods was used to select comparable windows of data within non-report pages (Figure 1).

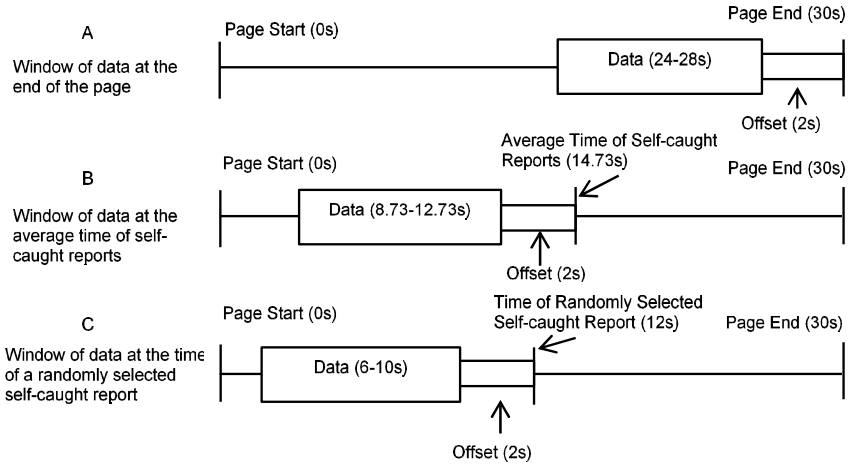

Fig. 1. Three different methods of selecting a window of data within a hypothetical 30s non-report page. The first method (A) is to use a window of data at the end of the page. The second method (B) is to use a window of data at the average time of the report for self-caught pages. The third method (C) is to use a window of data at the same time as the report within a randomly selected self-caught page. A two second offset is used to avoid confounds associated with the button press (to either advance to the next page (A) or provide a self-caught report (B, C)).

The first method was to simply use data from the *end of the page* (EoP) after including a 2-second offset to account for the key press to advance pages. For example, when selecting a four second window from a 30 second non-report page, gaze data from 24 seconds to 28 seconds within the page would be used. The second method is to select the window on the basis of the *average* time a report occurred within self-caught pages (Avg. SC). For our dataset this is 14.73 seconds. Thus, when selecting a four-second window from a 30 second non-report page, gaze data from 8.73 seconds to 12.73 seconds within the page would be used for feature calculation. The third method is to select the window based on a *randomly* selected self-caught page (Rand. SC). For example, when selecting a four second window from a 30 second non-report

page, a self-caught page would be randomly selected. If the self-caught report occurred 12 seconds into the page, then gaze data from 6 seconds to 10 seconds within the page would be used for feature calculation (last 2-seconds are considered to be the offset).

Feature Engineering. Raw data was processed into gaze fixations (points where gaze is maintained on the same location) and saccades (movements between subsequent fixations) using a fixation filter from the OpenGazeAndMouseAnalyzer (OGAMA), an open source gaze analyzer [24]. The series of gaze fixations and saccades were segmented into windows of varying length (4, 6, 8, 10, or 12 seconds), each ending at a certain point on the page as noted in the previous section. Windows that contained fewer than five fixations or windows that were shorter than four seconds were eliminated because these windows did not contain sufficient data to compute gaze features. The features used were similar to those used in our previous work [1]. Two sets of features were computed: 46 global gaze features and 20 local gaze features, yielding 66 features overall. A third set of features that relied on the context of the reading task (*context features*) were tested, but will not be discussed further because they did not improve classification accuracy and are not generalizable to different contexts.

Global gaze features were independent of the actual words being read. These included properties of eye movements such as *fixation duration* (ms)*, saccade duration* (ms)*, saccade length* (pixels), *saccade angle* (degrees between two saccade vectors), and *pupil diameter* (standardized using a within-subject z-score). For these five measurement distributions, the min, max, mean, median, standard deviation, skew, kurtosis, and range were computed, totaling 40 features. Additional features included a measure of *fixation dispersion,* the *fixation duration/saccade duration ratio*, and the *number of saccades*. Three new features were the *number of blinks*, the *proportion of time spent blinking,* and the proportion of *horizontal saccades*, which were saccades with angles less than 30 degrees above or below the horizontal axis.

In contrast to global features, *local features* were sensitive to the words being read. These included measures of the number of specific fixation types as well as the mean and standard deviations their durations. These features were calculated from *first pass fixations* (first fixation on a word during the first pass through a text)*, regression fixations* (fixations back onto words already passed)*, single fixations* (fixations on words that were only fixated on once)*, gaze fixations* (consecutive fixations on the same word)*,* and *non-word fixations*, totaling 15 features*.* Additional local features captured known relationships between fixation durations and the semantic properties of words, such as their *length, frequency* of use, number of *synonyms*, and *semantic specificity* (e.g. "blue" is more specific than "color"). The final local feature was the *ratio of reading time* to expected reading time, calculated as 200ms times the number of words read.

Tolerance analysis and feature selection were applied to each model. Tolerance analysis consisted of removing highly multicollinear features, as redundant information invites bias towards that information. Similarly, feature selection consisted of using correlation based feature selection (CFS) as implemented in Weka [12], an open source machine learning workbench, to remove features that were strongly correlated

with other features and weakly correlated with the class label. Feature selection was performed using just the training set to avoid overfitting.

Model Building. We explored five different parameters at the model building stage in order to ascertain which parameter combination resulted in the most accurate models. First, we varied which features were included in our models. Possibilities were (1) global, (2) local, and (3) global and local. This allowed us to narrow down which set of features was best able to detect MW. Second, we varied the *window sizes* used in the model. Window sizes of 4, 6, 8, 10, and 12 seconds were used in order to determine the ideal amount of data for detecting MW. Third, we used three different sampling methods for our *training set* (testing set was never sampled). We either used the entire data set, downsampled the training set, or synthetically oversampled the training set using SMOTE (Synthetic Minority Over-sampling Technique [4]). Each sampling method was applied to the same training set five separate times, and the average values of all five runs were taken. Fourth, we used three different types of outlier treatment. Outliers were either left in the dataset, trimmed, or Winsorized. Trimming consisted of removing values greater/lower than 3 standard deviations above/below the mean, while Winsorization consisted of replacing those values with the corresponding value +3 or -3 standard deviations above/below the mean.

We used 10 classifiers implemented in Weka, including Bayes net, naïve Bayes, logistic regression, SVM, and decision trees. We considered a wide array of classifiers at this early stage of the research as it is unclear which classifier is the best in this domain.

Our results were evaluated with a leave-several-user-out validation method. Data from a random 66% of the users were included in the training set, while the data from the remaining 34% were included in the training set. This process was repeated 20 times for each model and performance metrics were averaged across these iterations.

Cohen's kappa was used to evaluate model performance because it corrects for random guessing when there are uneven class distributions, as is the case in the current dataset. The kappa value is calculated using the formula K = (observed accuracy - expected accuracy) / (1 - expected accuracy), where *observed accuracy* is equivalent to recognition rate and *expected accuracy* is computed from the confusion matrix to account for the pattern of misclassifications. A kappa value of 0 indicates chance agreement while a kappa value of 1 indicates perfect agreement.

4 Results

We built seven types of models that detected self-caught (SC) MW. Models varied on the type of pages that were included for negative instances of MW (negative probe-caught, non-report pages, or both) and the method to select the window of data for non-probe report pages (end-of-page, average SC, or random SC). Results for the best model of each type are shown in Table 1. Window selection was only necessary for models with non-report pages as there was no explicit point at which MW occurred.

Table 1. Best performing models for each type of negative instance

Negative Instances	Window Selection	Instances	Classifier	Window Size (s)	Feature Types	Feature Count	Kappa
PC No	-	540	Logistic	10	G	10	.33 (.07)
NR	EoP	523	Logistic	12	G	10	.39 (.07)
NR	**Avg. SC**	**410**	**SVM**	**12**	**GL**	**18**	**.45 (.06)**
NR	Rand. SC	400	SVM	12	GL	22	.39 (.07)
PC No + NR	EoP	681	Logistic	12	G	10	.36 (.05)
PC No + NR	Avg. SC	415	SVM	12	GL	21	.43 (.07)
PC No + NR	Rand. SC	361	SVM	12	GL	21	.35 (.09)

Note. **Bolding** indicates the model with the highest kappa value. Standard deviations are in *parentheses*. PC No = Probe-caught negative instances; NR = Non-report; EoP = End of page; Avg. SC = Average self-caught; Rand SC = Random self-caught; Acc. = Accuracy; G = Global; L = Local; GL = Global + Local

The best performing model used only non-report pages as negative instances of MW, with a 12 second window located at the same point as the average time of self-caught reports. This model achieved a kappa value of 0.45, with an accuracy of 74% compared to an expected accuracy of 52%. The confusion matrix for this best model highlights the model's accuracy in classifying positive instances correctly, with a hit rate of .82 compared to a prior probability of .32. The false-alarm rate of 0.30 is not high enough to be of concern depending on the MW intervention (see Discussion).

Parameter Analysis. It is ideal to perform the least number of operations on the data in order to decrease complexity in real time systems. With that in mind, we analyzed the parameters to determine which resulted in the best models. We analyzed each of the parameters across all 108 individual models (3 sampling methods * 3 outlier treatments * 3 window sizes * 4 feature types). For each parameter, the model with the best kappa value (across the 10 classifiers) was selected for further analysis.

Window size and feature type displayed clear trends, but there was no clear trend for either sampling method or outlier treatment. In particular, kappas were higher for larger window sizes as seen in Figure 2. When analyzing feature type, it is apparent that global features were the most useful for detecting MW. The best model using both global and local features resulted in the best performance (kappa = .45), but performance of the best local feature model was quite poor (kappa = .24) compared to the best global feature model (kappa = .43). This suggests that global features were responsible for much of the performance of the combined model. This could be due to the level of precision needed for local features. When using a desk mounted eye tracker, gaze data can be affected by a user's head movements.

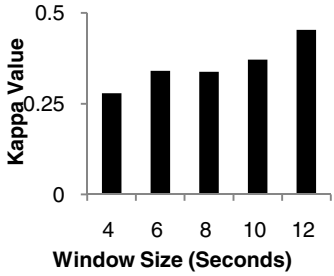

Fig. 2. The effect of window size on kappa value

Head movements have a drastic effect on local features because they require precise information on which words are being gazed upon. Global features, however, are independent of the information being displayed.

Feature Analysis. We performed a deeper analysis of how the features varied between positive and negative instances of MW in order to obtain a better understanding of a user's eye gaze during MW. We analyzed the dataset belonging to the best performing model shown in Table 1 above. Each feature was analyzed with a paired samples t-test of the difference in the mean value of the feature between positive and negative instances of MW. We analyzed data from the 31 participants for which there were both positive (self-report) and negative (non-report pages) instances of MW. There were five features that were significantly different between positive and negative instances of MW below the 0.05 level: (1) proportion of line cross saccades (Mean MW = .125 (standard deviation = .100), Mean not MW = .078 (.101)); (2) proportion of regression fixations (MW = .162 (.062), not MW = .130 (.057)); (3) minimum saccade duration (MW = 7.29 (6.74), not MW = 6.38 (5.89)); (4) fixation duration kurtosis (MW = 3.99 (4.52), not MW = 7.01 (5.75)); and (5) fixation duration skew (MW = 1.66 (.81), not MW = 2.13 (.95)). These features were indicative of a break in normal reading patterns. First of all, there was a greater proportion of regression fixations and line cross saccades during MW, and saccades had a greater duration. Furthermore, the skew and kurtosis of the fixation durations indicate that they were more uniform during MW, whereas fixation durations during normal reading would vary with word difficulty.

5 General Discussion

Our major contribution consists of building the first eye-gaze detector of self-caught MW, studying its accuracy, its parameters, and the features that were most diagnostic of MW. In the remainder of this section, we highlight our main findings, consider applications of the MW detector, and discuss limitations and avenues for future work.

Main Findings. Our results highlight a number of important findings for building detectors of self-caught MW. First, we have developed the first MW detector built using self-caught reports. The overall classification accuracy of 74% is moderate but might be sufficiently high for meaningful interventions, especially if the interventions are fail-soft in that they are not harmful if delivered when detection errors occur (more on this below). Second, we found that the best performing models used non-report pages as negative instances of MW, and had features calculated from a window of data located at the same time as the average time of reports within self-caught pages. This information might be useful to other researchers interested in building self-caught MW detectors in similar contexts. Third, similar to our previous study using probe-caught reports [1], global features were shown to be particularly useful in classifying MW for self-caught reports. This is a significant finding because global features are easier to compute, do not require very high-precision eye tracking, and are more likely to generalize to different tasks beyond reading. Fourth, we found that

larger window sizes were associated with more accurate MW detection rates. This shows that a greater amount of information at the page level is important for more accurate MW detection. Finally, an analysis of the features revealed that normal reading patterns were disrupted during MW, such that users made more regression fixations, had a larger proportion of line cross saccades, and saccades were of greater duration. Furthermore, the distribution of fixation durations was more uniform during MW, which may indicate that they were not as affected by word difficulty as they would be during normal reading.

Applications. A self-caught MW detector has a number of applications. For example, it would allow for measurement of MW without interrupting the user. It could be used to advance basic scientific research on MW itself, or to study user strategies for regaining focus. Finally, it could be embedded in any adaptive system that includes a text comprehension component. MW has been shown to negatively affect text comprehension, so any interface that includes text comprehension could be improved by dynamically responding to MW. One example of an intervention would be to recommend that a user re-read or self-explain a passage when the system detects MW. Other possibilities include presenting the information in a different format, such as showing a short animation or film; offering positive encouragement; or suggesting that the user switches topics. It is important to note that interventions should not disrupt the user if MW is detected incorrectly and should be used sparingly so users are not overwhelmed.

Limitations and Future Work. There are several limitations to the current study. First, the data was collected in a lab environment and users were limited to undergraduates from two universities located in the United States. This limits our claims of generalizability to individuals from different populations. Quantifying our method with a more diverse population and setting would boost our claims of generalizability. Second, the font size was larger than what would normally be read, an intentional decision in order to improve eye tracking precision for computing local features. Given that global features did most of the work, future studies could consider smaller font sizes. Third, an expensive, high quality eye tracker was used for data collection, which limits the scalability of using eye gaze as a modality for MW detection. However, this could eventually be addressed by the decreasing cost of consumer-grade eye tracking technology, such as Eye Tribe ($99) and Tobii EyeX ($195), or with promising alternatives that use webcams for gaze tracking [19]. Fourth, it is possible that users did not provide accurate or honest self-caught reports. However, both the probe-caught and self-caught methods have been validated in a number of studies [20, 21], and there is no clear alternative for tracking a highly internal state like MW.

Concluding Remarks. In summary, the present study demonstrated that global eye movements could be used to build a moderately accurate user-independent detector of self-caught MW, or MW with metacognitive awareness. Importantly, our approach used a relatively unobtrusive remote eye tracker that allowed for unrestricted head and body movement and involved an ecologically-valid reading activity. The next step involves integrating the detector into an adaptive system in order to trigger interventions that attempt to reorient attentional focus when MW is detected.

Acknowledgment. This research was supported by the National Science Foundation (DRL 1235958). Any opinions, findings and conclusions, or recommendations expressed are those of the authors and do not necessarily reflect the views of NSF.

References

1. Bixler, R., D'Mello, S.: Toward Fully Automated Person-Independent Detection of Mind Wandering. In: Dimitrova, V., Kuflik, T., Chin, D., Ricci, F., Dolog, P., Houben, G.-J. (eds.) UMAP 2014. LNCS, vol. 8538, pp. 37–48. Springer, Heidelberg (2014)
2. Blanchard, N., Bixler, R., Joyce, T., D'Mello, S.: Automated Physiological-Based Detection of Mind Wandering during Learning. In: Trausan-Matu, S., Boyer, K.E., Crosby, M., Panourgia, K. (eds.) ITS 2014. LNCS, vol. 8474, pp. 55–60. Springer, Heidelberg (2014)
3. Calvo, R.A., D'Mello, S.: Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. IEEE Transactions on Affective Computing **1**(1), 18–37 (2010)
4. Chawla, N.V., et al.: SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research **16**(1), 321–357 (2002)
5. D'Mello, S., et al.: Automatic Gaze-Based Detection of Mind Wandering during Reading. Educational Data Mining (2013)
6. D'Mello, S., et al.: Gaze tutor: A Gaze-Reactive Intelligent Tutoring System. International Journal of Human-Computer Studies **70**(5), 377–398 (2012)
7. Dong, Y., et al.: Driver Inattention Monitoring System for Intelligent Vehicles: A Review. IEEE Transactions on Intelligent Transportation Systems **12**(2), 596–614 (2011)
8. Drummond, J., Litman, D.: In the Zone: Towards Detecting Student Zoning Out Using Supervised Machine Learning. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010, Part II. LNCS, vol. 6095, pp. 306–308. Springer, Heidelberg (2010)
9. Feng, S., et al.: Mind Wandering While Reading Easy and Difficult Texts. Psychon Bull Rev. **20**(3), 586–592 (2013)
10. Franklin, M.S., et al.: Catching The Mind in Flight: Using Behavioral Indices to Detect Mindless Reading in Real Time. Psychonomic Bulletin & Review **18**(5), 992–997 (2011)
11. Franklin, M.S., et al.: Window to the Wandering Mind: Pupillometry of Spontaneous Thought While Reading. The Quarterly Journal of Experimental Psychology **66**(12), 2289–2294 (2013)
12. Hall, M., et al.: The WEKA Data Mining Software: an Update. ACM SIGKDD Explorations Newsletter **11**(1), 10–18 (2009)
13. Killingsworth, M.A., Gilbert, D.T.: A Wandering Mind is an Unhappy Mind. Science **330**(6006), 932–932 (2010)
14. Mooneyham, B.W., Schooler, J.W.: The Costs and Benefits of Mind-Wandering: A Review. Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale **67**(1), 11–18 (2013)
15. Muir, M., Conati, C.: An Analysis of Attention to Student – Adaptive Hints in an Educational Game. In: Cerri, S.A., Clancey, W.J., Papadourakis, G., Panourgia, K. (eds.) ITS 2012. LNCS, vol. 7315, pp. 112–122. Springer, Heidelberg (2012)
16. Navalpakkam, V., Kumar, R., Li, L., Sivakumar, D.: Attention and Selection in Online Choice Tasks. In: Masthoff, J., Mobasher, B., Desmarais, M.C., Nkambou, R. (eds.) UMAP 2012. LNCS, vol. 7379, pp. 200–211. Springer, Heidelberg (2012)
17. Randall, J.G., et al.: Mind-Wandering, Cognition, and Performance: A Theory-Driven Meta-Analysis of Attention Regulation. Psychological Bulletin (2014)

18. Schooler, J.W., et al.: Zoning Out While Reading: Evidence for Dissociations Between Experience and Metaconsciousness. In: Levin, D.T. (ed.) Thinking and Seeing: Visual Metacognition in Adults and Children, pp. 203–226. MIT Press, Cambridge (2004)
19. Sewell, W., Komogortsev, O.: Real-Time Eye Gaze Tracking with an Unmodified Commodity Webcam Employing a Neural Network. In: CHI 2010 Extended Abstracts on Human Factors in Computing Systems, pp. 3739–3744 ACM (2010)
20. Smallwood, J., et al.: Subjective Experience and the Attentional Lapse: Task Engagement and Disengagement During Sustained Attention. Consciousness and Cognition **13**(4), 657–690 (2004)
21. Smallwood, J., et al.: When Attention Matters: The Curious Incident of the Wandering Mind. Memory & Cognition **36**(6), 1144–1150 (2008)
22. Smallwood, J., Schooler, J.W.: The Restless Mind. Psychological Bulletin **132**(6), 946–958 (2006)
23. Smilek, D., et al.: Out of Mind, Out of Sight: Eye Blinking as Indicator and Embodiment of Mind Wandering. Psychological Science **21**(6), 786–789 (2010)
24. Voßkühler, A., et al.: OGAMA (Open Gaze and Mouse Analyzer): Open-Source Software Designed to Analyze Eye and Mouse Movements in Slideshow Study Designs. Behavior Research Methods **40**(4), 1150–1162 (2008)
25. Yonetani, R. et al.: Multi-Mode Saliency Dynamics Model for Analyzing Gaze and Attention. In: Proceedings of the Symposium on Eye Tracking Research and Applications, pp. 115–122. ACM, New York (2012)