User Model User-Adap Inter
DOI 10.1007/s11257-015-9167-1

Automatic gaze-based user-independent detection of mind wandering during computerized reading

Robert Bixler[1] · **Sidney D'Mello**[1]

Received: 21 January 2015 / Accepted in revised form: 31 August 2015

Abstract Mind wandering is a ubiquitous phenomenon where attention involuntarily shifts from task-related thoughts to internal task-unrelated thoughts. Mind wandering can have negative effects on performance; hence, intelligent interfaces that detect mind wandering can improve performance by intervening and restoring attention to the current task. We investigated the use of eye gaze and contextual cues to automatically detect mind wandering during reading with a computer interface. Participants were pseudorandomly probed to report mind wandering while an eye tracker recorded their gaze during the reading task. Supervised machine learning techniques detected positive responses to mind wandering probes from eye gaze and context features in a user-independent fashion. Mind wandering was detected with an accuracy of 72 % (expected accuracy by chance was 60 %) when probed at the end of a page and an accuracy of 67 % (chance was 59 %) when probed in the midst of reading a page. Global gaze features (gaze patterns independent of content, such as fixation durations) were more effective than content-specific local gaze features. An analysis of the features revealed diagnostic patterns of eye gaze behavior during mind wandering: (1) certain types of fixations were longer; (2) reading times were longer than expected; (3) more words were skipped; and (4) there was a larger variability in pupil diameter. Finally, the automatically detected mind wandering rate correlated negatively with measures of learning and transfer even after controlling for prior knowledge, thereby providing evidence of predictive validity. Possible improvements to the detector and applications that utilize the detector are discussed.

✉ Robert Bixler
Rbixler@nd.edu

Sidney D'Mello
sdmello@nd.edu

[1] Department of Computer Science, University of Notre Dame, Notre Dame, IN 46556, USA

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

Keywords Gaze tracking · Mind wandering · User modeling

1 Introduction

Most of us have had the experience of reading, listening to a lecture, or engaging in a personally-relevant task only to realize that our attention has gradually drifted away from the task at hand to off-task thoughts, such as dinner, childcare, or everyday worries and anxieties. It has been estimated that these unintentional attentional shifts toward internal task-unrelated thoughts, referred to as *mind wandering*, or zoning out, occur between 20–50 % of the time depending on the task and the environmental context (Kane et al. 2007; Killingsworth and Gilbert 2010; Schooler et al. 2004; Smilek et al. 2010). For example, one recent large-scale study tracked mind wandering in 5000 people with pseudorandom prompts from an iPhone app and discovered that people reported mind wandering for 46.9 % of the prompts (Killingsworth and Gilbert 2010). Mind wandering is not merely incidental but consequential as well. It negatively effects performance on tasks requiring conscious control because a person cannot simultaneously focus on both the task at hand and task-unrelated thoughts. It has been shown that mind wandering results in increased error rates during signal detection tasks (Robertson et al. 1997; Smallwood et al. 2004), lower recall during memory tasks (Seibert and Ellis 1991; Smallwood and Schooler 2006), and poor comprehension during reading tasks (Feng et al. 2013; Smallwood et al. 2007). In addition, a recent meta-analysis of 49 independent samples found that mind wandering was consistently negatively correlated with performance across a range of tasks (Randall et al. 2014). Based on this research, it is evident that performance on tasks that require attentional focus can be hindered by off task thoughts (i.e., mind wandering). This suggests that there is an opportunity to improve task performance by attempting to reduce the negative effects of mind wandering.

There are two fundamental strategies that can be taken to reduce the negative effects of mind wandering: it can either be prevented before it occurs (i.e., proactive) or it can be addressed once it occurs (i.e., reactive). Mindfulness training through meditation is perhaps the most common strategy to proactively counteract mind wandering (Davidson 2010; Jha et al. 2007; Lutz et al. 2009; Mrazek et al. 2013). Mindfulness is an opposing construct to mind wandering that entails an element of sustained attentiveness to one's own thoughts (Mrazek et al. 2012). By increasing awareness of one's own thoughts, the possibility of catching oneself mind wandering is also increased, thereby reducing overall levels of mind wandering. Although mindfulness training regimens may reduce the overall propensity to mind wander, they might lack scalability in terms of time and effort. For example, a typical mindfulness training regimen can last as much as eight weeks (Baer 2003), which requires time, commitment, and sometimes a meditation expert to lead the training. Short mindfulness training exercises that range from a few minutes to a few hours have recently been explored (Mrazek et al. 2012; Zeidan et al. 2010), but their effectiveness is yet to be fully understood.

Another method of mind wandering prevention entails tailoring the environment to the user in order to reduce their propensity to mind wander. Individual-difference user attributes such as distractibility (Forster and Lavie 2014) and working memory

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

capacity (Kane et al. 2007) have been found to be associated with mind wandering. Thus, tailoring the environment to accommodate these individual-difference user attributes may help prevent mind wandering. For example, we recently attempted to use machine learning techniques to identify a learning configuration (i.e., difficulty of materials, incentives, etc.) that was most conducive to a lowered rate of mind wandering for individual users in a reading task (Kopp et al. 2014). We did this by taking into consideration measures of user attributes such as reading comprehension, scholastic aptitude, and boredom proneness, amongst others. We were able to predict whether easy or difficult texts would result in a lower mind wandering rate for a participant with an accuracy of 64 %. In a follow up study (Bixler et al. 2014), we assessed the effectiveness of this method by comparing it to two alternative methods. The first alternative method was to select the learning configuration that was most conducive to a lowered rate of mind wandering for all participants on average. The second alternative method was to select a learning configuration at random. We found that our method resulted in a lower mind wandering rate than randomly selecting a learning condition, but was not significantly different from the mind wandering rate obtained when selecting the learning condition that was best on average. In other words, our method showed some promise, but more research is needed to determine the effectiveness of this individual-difference based proactive approach to reducing mind wandering.

One limitation of proactive strategies in general is that they might reduce mind wandering, but are unlikely to eliminate it entirely. They also have no mechanism to address mind wandering when it inevitably occurs. Thus, in contrast to exploring ways to prevent mind wandering, the current study focuses on reactive systems that intervene when mind wandering occurs. Intervention requires detection, but a reliable, scalable, and unobtrusive method to detect mind wandering has yet to be developed. Fortunately, recent research on mind wandering lends insight into one promising approach towards development of a mind wandering detector. In particular, decades of scientific evidence in support of an *eye-mind* link posits that there could be a link between external information and eye movements during reading (Just and Carpenter 1980; Rayner 1998; Reichle et al. 1998). This has motivated a recent exploration of the link between mind wandering and eye behaviors such as eye blinks, pupil diameters, and fixations (periods where the eye is gazing at the same location). For instance, a number of studies have shown that blinks occur more frequently during mind wandering (Frank et al. 2015; Grandchamp et al. 2014; Smilek et al. 2010; Uzzaman and Steve 2011). Similarly, pupil diameter has been shown to increase during mind wandering (Franklin et al. 2013; Smallwood et al. 2011; Uzzaman and Steve 2011). Research into gaze fixations during mind wandering has found that fixation durations are longer (Frank et al. 2015; Foulsham et al. 2013; Reichle et al. 2010) and subsequent fixations on the same word are more frequent (Uzzaman and Steve 2011). To preview one recent study, Schad et al. (2012) investigated the relationship between gaze duration and word frequency in both normal and mindless reading. They found that fixation gaze durations on infrequently occurring long words were significantly greater than those for frequently occurring long words during normal reading, but not when mind wandering.

This body of research suggests that mind wandering is associated with eye gaze behaviors that are distinctly different from normal reading. Can we leverage these differences to automatically detect mind wandering? This paper makes a first step

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

towards this goal by using eye gaze to automatically detect mind wandering in near real-time in a manner that generalizes to new users. The remainder of the paper is organized as follows. Section 2 details work related to our study, including work on attentional state estimation and mind wandering detection, and concludes with a brief overview of the current study. Sections 3 and 4 cover the data collection procedure and our supervised machine learning method, respectively. Section 5 includes an exposition of our results. We conclude with a discussion of the main findings, potential applications, limitations, and future work.

2 Related work

Our work on mind wandering detection is closely related to the field of user mental state estimation. We begin by discussing relevant work in the field of attentional state estimation (Roda and Thomas 2006), a subfield of user mental state estimation that is particularly related to mind wandering detection. We then cover all related work on mind wandering detection, and conclude by detailing the novelty of the current study.

2.1 Attentional state estimation

Attentional state estimation entails analyzing a user's behavior to determine the degree and locus of attentional focus. It is driven in part by the desire to increase user productivity, safety, and satisfaction by modeling attention. Knowledge of a user's attentional state can be used for a number of purposes, but common applications aim at improving productivity in real world environments such as an office space (Ba and Odobez 2006; Börner et al. 2014; Dong et al. 2010; Horvitz et al. 1999, 2003; Matsumoto et al. 2000; Selker 2004; Stiefelhagen et al. 2001; Stiefelhagen 2002; Stiefelhagen and Zhu 2002; Vertegaal et al. 2006; Voit and Stiefelhagen 2008), monitoring driver inattention to increase safety (Bergasa et al. 2006; Dong et al. 2011; D'Orazio et al. 2007; Fletcher and Zelinsky 2009; Knipling et al. 1994; Su et al. 2006; Tawari et al. 2014; Torkkola et al. 2004; Yeo et al. 2009), and improving interaction in virtual environments (Barbuceanu et al. 2011; Horvitz et al. 2003; Muir and Conati 2012; Navalpakkam et al. 2012; Roda and Nabeth 2007; Toet 2006; Vertegaal et al. 2006; Ugurlu 2014; Yonetani et al. 2012).

There have been many attempts to improve productivity in the workplace through attention-aware systems. In real world environments such as a work meeting, user attention can be estimated via gaze location, head position, and posture (Ba and Odobez 2006; Börner et al. 2014; Dong et al. 2010; Matsumoto et al. 2000; Selker 2004; Stiefelhagen et al. 2001; Stiefelhagen 2002; Stiefelhagen and Zhu 2002; Voit and Stiefelhagen 2008). In these studies, a user's focus of attention is usually attributed to the object or individual that the user is looking at. Additional information such as acoustic information (Stiefelhagen et al. 2001; Stiefelhagen 2002), actions in the environment (Ba and Odobez 2006), and contextual information, such as the time of day or the composition of the user's daily schedule (Horvitz et al. 2003, 1999), have been used as well.

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

Safety critical activities such as driving a car can be heavily impacted by user attention (Sussman et al. 1985). As such, there have been a number of studies on driver inattention, specifically distractibility and fatigue (Dong et al. 2011). In addition to gaze data (Bergasa et al. 2006; Fletcher and Zelinsky 2009; Su et al. 2006; Tawari et al. 2014), a wide variety of other methods have been used to estimate the attentional state of the driver. Some examples include neurophysiological measures such as EEG (Yeo et al. 2009), driving context measures such as seat pressure and steering angle (Furugori et al. 2005; Torkkola et al. 2004), and analyses of the position and closures of the eyes (D'Orazio et al. 2007).

There are also opportunities to improve a user's satisfaction with a virtual environment by considering the user's attention. For example, Barbuceanu et al. (2011) estimated intended actions of a user in a virtual reality environment. The order in which objects in the environment were gazed at was used as input to hidden Markov models that predicted which action was likely, resulting in an accuracy of 88 %. Further examples involving attention estimation include educational systems (Muir and Conati 2012; Roda and Thomas 2006; Ugurlu 2014), dynamic scene viewing (Yonetani et al. 2012), and information dissemination services (Börner et al. 2014; Navalpakkam et al. 2012).

Although attentional state estimation is related to mind wandering detection, it is important to distinguish the two. Both entail identifying the focus of a user's attention, but mind wandering detection is different in that it is concerned with detecting more covert forms of involuntary attentional lapses. Many of the previous studies assumed that the user was paying attention to an aspect of their environment. For example, in the Navalpakkam et al. study (2012), attention was estimated based on where the user was looking. The assumption is that a user's gaze location indicates their locus of attention. However, this does not account for situations when a user is looking at an aspect of the environment but thinking about something else entirely. This is the case during mind wandering. Mind wandering is characterized by a disconnect between the external environment (the items being displayed) and internal thought, commonly referred to as 'perceptual decoupling' (Schooler et al. 2011; Smallwood et al. 2008). Although the location of a user's gaze can be used as an estimation of attention, it alone cannot be used to detect mind wandering because mind wandering is akin to "looking but not seeing." A different approach is needed for mind wandering detection as reviewed below.

2.2 Mind wandering detection

A few recent studies have investigated the behavioral correlates of mind wandering (Feng et al. 2013; Schooler et al. 2011; Smallwood et al. 2008). Two methods of tracking mind wandering have been used in the literature. The first is to have participants provide a mind wandering report in response to thought probes placed throughout the task (*probe-caught*). The second is to allow participants to provide a mind wandering report whenever they catch themselves mind wandering (*self-caught*). Researchers have linked these mind wandering reports to neurological signals (Christoff et al. 2009), acoustic and prosodic information (Drummond and Litman 2010), physiolog-

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

ical signals (Blanchard et al. 2014; Pham and Wang 2015; Smallwood et al. 2004), behavioral measures (Franklin et al. 2011; Mills & D'Mello in press), and eye behaviors (Foulsham et al. 2013; Frank et al. 2015; Franklin et al. 2013; Grandchamp et al. 2014; Reichle et al. 2010; Smallwood et al. 2011; Smilek et al. 2010; Uzzaman and Steve 2011) as discussed below.

2.2.1 Non eye-gaze based research

One area of mind wandering research is concerned with discovering the neurological correlates of mind wandering (Gruberger et al. 2011). It is hypothesized that a collection of brain regions, known as the default mode network, play a role in mind wandering (Christoff et al. 2009; Mason et al. 2007). One study by Christoff et al. (2009) used functional magnetic resonance imaging (fMRI) to investigate the brain activations of participants prior to probe-caught reports of mind wandering during a Sustained Attention to Response Task (SART— a lab-based attentional task) (Robertson et al. 1997). They found that regions associated with both the default mode network and executive control network were activated prior to mind wandering reports. Despite this positive result, the prohibitive cost and obtrusiveness of fMRI, limits their use for real-world mind wandering detection. Wearable EEG, which is cheaper and less obtrusive than the equipment used in fMRI, might have some promise. However, to the best of our knowledge, EEG has yet to be used for mind wandering detection.

Drummond and Litman (2010) used acoustic-prosodic (e.g., pitch) features in what was perhaps the first attempt to automatically detect mind wandering (operationalized as "zoning out"). Their model attempted to discriminate "high" versus "low" instances of zoning out while participants read a biology text aloud. Mind wandering was measured using a 7 point Likert scale which was subsequently discretized into "high zoning out" (consisting of responses 1–3) and "low zoning out" categories (consisting of responses 4–7). Their accuracy of 64 % reflects an important first step in mind wandering detection. However, it is unclear if their validation approach used independent training and testing sets, so generalization to new users is unknown.

Physiological signals such as galvanic skin response (GSR) and skin temperature are part of the sympathetic nervous system. The relationship between the sympathetic nervous system and attentional states indicates that GSR and skin temperature may be useful measures of mind wandering. Indeed, Smallwood et al. (2004) found that GSR was lower in the 30 second period prior to reports of mind wandering during a SART. Along these lines, Blanchard et al. (2014) recently used supervised machine learning methods to detect mind wandering using GSR and skin temperature. They used an Affectiva Q device to record both GSR and skin temperature signals while participants provided mind wandering reports in response to auditory thought probes during a computerized reading task. The best model achieved a kappa value of .22 in a user-independent fashion, which is a promising result. In another example, Pham and Wang (2015) detected mind wandering using heart rate and lecture content features while participants viewed MOOC-style lectures on a mobile phone. Participants were required to hold their finger over the back camera of the phone and heart rate was detected using Photoplethysmography sensing methods. Mind wandering was tracked using auditory probes that were triggered every three minutes on average. They were able to achieve a

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

kappa value of .22 when evaluating their model in a user-independent fashion, similar to the study by Blanchard et al. (2014). While this method is currently only applicable in a mobile phone context, the increasing prevalence of mobile phones and the lack of a need for hardware modifications make this approach an exciting prospect.

Franklin et al. (2011) detected mind wandering using behavioral measures. In this study, participants read approximately 5000 words of a novel using a self-paced word-by-word reading paradigm. Participants were assigned to a control condition or a thought probe condition. In the latter condition, thought probes were provided whenever an algorithm estimated that the participant was either mind wandering or reading normally. The algorithm only considered periods where the previous ten words were deemed difficult. If the participant was reading fast during this period they were considered to be mind wandering. They were considered to be reading normally if they were reading slowly. The assessment of whether a participant was reading fast or slow was determined based on thresholds derived from pilot data. They were able to correctly classify thought probes 72 % of the time compared to an expected accuracy of 49 %. Although this study yielded a moderate classification rate for mind wandering detection, it had two limitations. First, word-by-word reading is not naturalistic reading behavior, which raises questions of generalizability in real-world contexts. Second, the thresholds were selected based on a limited pilot sample of 29 participants. Few additional details of the pilot study were provided which makes it difficult to reproduce the methodology.

2.2.2 Eye-gaze based research

Mind wandering has also been linked to various eye behaviors: blinks (Grandchamp et al. 2014; Smilek et al. 2010), pupil diameters (Franklin et al. 2013; Grandchamp et al. 2014; Smallwood et al. 2011), and eye movements (see Fig. 1) such as fixations (periods where the eye is gazing at the same location) and saccades (movements between fixations) (Foulsham et al. 2013; Frank et al. 2015; Reichle et al. 2010; Uzzaman and Steve 2011). For instance, Smilek et al. (2010) investigated mind wandering while participants responded to ten auditory thought probes while reading two passages for 15 minutes each. Eleven out of the 12 participants blinked more frequently in the five second intervals preceding a mind wandering report. Reichle et al. (2010) investigated mind wandering while four participants read an entire novel in 12–15 h-long sessions during which they provided self-caught and probe-caught mind wandering reports every 2–4 min. The results indicated that fixations tended to be more erratic during mind wandering, with fewer fixations on words as well as from one word to another. Pupil diameter has also been shown to be indicative of mind wandering. One study by Franklin et al. (2013) compared pupil diameters during mindless and normal reading. Participants read the text word-by-word and were presented with pseudo-random thought probes while reading. Pupil diameter was found to be significantly larger during periods of mind wandering. Another study, by Smallwood et al. (2011) investigated pupil diameters during a working memory task and a choice reaction time task. They analyzed the difference in pupil diameters between correct responses and incorrect responses on the tasks; incorrect responses were taken as a measure of inattention. They found no difference in pupil diameter for the choice reaction time task,

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

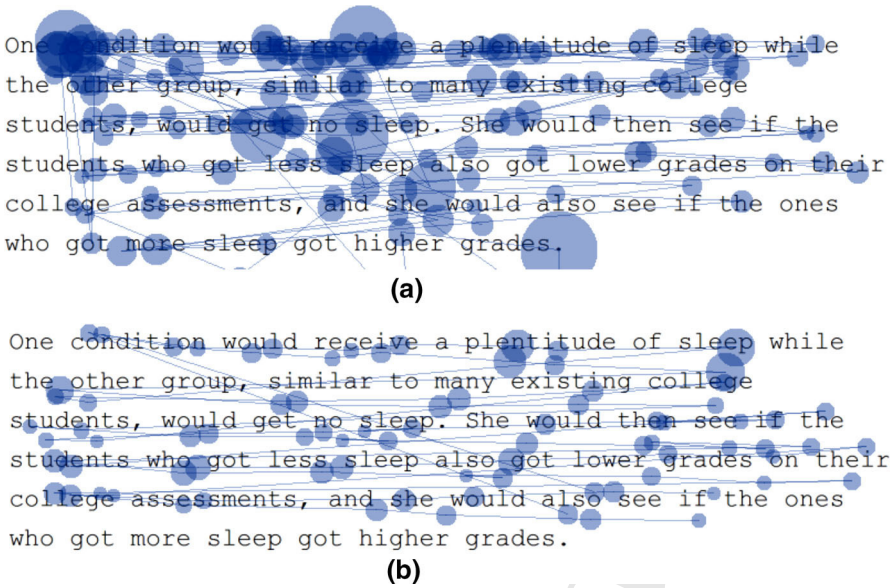


Fig. 1 Gaze fixations (*circles*) and saccades (*lines*) of the same page read by two different participants. The first image (**a**) depicts mind wandering and the second image (**b**) depicts normal reading. *Circle* diameter is proportional to fixation duration

but found that incorrect responses on the working memory task were associated with a larger mean pupil diameter and a larger standard error of the mean.

These studies suggest that there might be distinct eye behavior patterns during mind wandering. It is important to note, however, that eye movement patterns were not entirely consistent across studies. For example, Grandchamp et al. (2014) found that pupil diameter was smaller during mind wandering, which contradicts findings from other studies (Franklin et al. 2013; Smallwood et al. 2011). However, as Grandchamp et al. (2014) mention, this could be due to the differences in the tasks and methodologies. Grandchamp et al. (2014) used a single fixation cross for the duration of the experiment compared to the reading task and working memory task used by Franklin et al. (2013) and Smallwood et al. (2011), respectively, so the introduction of visual stimuli such as numbers and words could account for the differences in pupil diameter. There was also some inconsistency in the duration and number of fixations during mind wandering. Several studies (Foulsham et al. 2013; Frank et al. 2015; Reichle et al. 2010) found that fixation durations were greater during mind wandering, while others either found no effect (Smilek et al. 2010) or the opposite effect (Uzzaman and Steve 2011). Similarly, while one study found that there were more fixations during mind wandering (Foulsham et al. 2013), there were several that found the opposite effect (Frank et al. 2015; Smilek et al. 2010; Uzzaman and Steve 2011).

Despite these inconsistencies, these studies indicate that mind wandering is in fact reflected in eye gaze, which suggests that eye gaze might be a suitable channel for mind wandering detection. Furthermore, eye movements are ubiquitous in most interfaces for users without any visual impediments. We recently made an initial attempt to use

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

eye gaze data to detect mind wandering during reading (D'Mello et al. 2013). In this study, 84 participants were asked to read four texts on research methods while a Tobii T60 eye tracker recorded their gaze. Thought probes were triggered when participants fixated upon certain words in the text or when they pressed the spacebar to advance to the next page. The feature set consisted of 12 local features that were dependent upon the words being read and 17 global features that were not (these features are discussed in Sect. 4.1). The best performing user-independent mind wandering detector yielded an accuracy of 60 % after downsampling the data to remove class imbalance. Downsampling was performed by randomly undersampling the majority class (not mind wandering) so as to yield a 50 % chance of mind wandering. These results are promising, but they are limited by the fact that classification accuracy was not very impressive and the data was downsampled prior to classification. The goal of this paper is to expand upon this initial attempt at gaze-based mind wandering detection.

2.3 Current study

This study reports the development and validation of one of the first (aside from exceptions described above) fully automated user-independent detectors of mind wandering during computerized reading. We focus on a computerized reading task since reading is a critical component of many real-world tasks and reading comprehension is negatively associated with mind wandering (Feng et al. 2013; Schooler et al. 2004; Smallwood et al. 2008). Furthermore, it is our hope that by focusing on a general activity (reading) as opposed to a more specific interaction context, our eye gaze-based mind wandering detector can be more broadly applied to other interfaces. Our approach to mind wandering detection entails collecting eye gaze data and self-reports of mind wandering using the probe-caught method (discussed above) while users read texts on a computer screen. We then extract features from the eye gaze signal and contextual cues (discussed below) and use supervised classification techniques to build models that discriminate instances of mind wandering from normal reading. The models are constructed and validated in a user-independent fashion, so we have some confidence that they generalize to new users.

The present research is novel in a number of respects. First, previous work on attentional state estimation has not considered mind wandering, and other than the Drummond & Litman study (2010), the Franklin et al. study (2011), and our preliminary attempts (Blanchard et al. 2014; D'Mello et al. 2013), this work represents the first large-scale attempt at fully automated user-independent detection of mind wandering. Second, it significantly expands upon our preliminary work with a much larger and more diverse data set. Data was collected from two universities in the current study, compared to only one in our previous attempt at building a mind wandering detector from gaze data. Third, we considered an enhanced set of gaze features in order to improve classification accuracy when compared to our preliminary attempt using an impoverished feature set. Our previous work used 27 features, while the current study uses 80 features. Fourth, eye gaze features were complemented with contextual cues, such as high-level text characteristics (e.g., difficulty) and reading behaviors (e.g., reading rate), as context might help disambiguate noisy gaze signals.

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

3 Data collection

3.1 Participants

Participants were 178 undergraduate students from two U.S. universities that participated for course credit. Of the 178 students, 93 were from a medium-sized private Midwestern university and 85 were from a large public university in the mid-South. The average age of participants was 20 years ($SD = 3.6$). Demographics included 62.7 % female, 49 % Caucasian, 34 % African American, 7 % Asian, 6 % Hispanic, and 4 % "Other."

3.2 Texts and experimental manipulations

Participants read four different texts on research methods topics (experimenter bias, replication, causality, and dependent variables) adapted from a set of texts used in the educational game *Operation ARA!* (Halpern et al. 2012). We chose these topics because our student sample was primarily composed of psychology undergraduate students and these topics align with what they would encounter in their classes. On average, the texts contained approximately 1500 words ($SD = 10$) and were split into 30–36 pages with approximately 60 words per page. Texts were presented on a computer screen with size 36 Courier New typeface.

There were two experimental manipulations: difficulty and value. The difficulty manipulation consisted of presenting either an easy or a difficult version of each text. A text was made more difficult by replacing words and sentences with more complex alternatives while retaining semantics, length, and content. Value referred to the extrinsic value of the text. Participants were presented with either a *high-value* text or a *low-value* text. They were informed that a posttest followed the reading and questions from the *high-value* texts would be worth three times as much on the posttest as questions from *low-value* texts. Participants were instructed that an additional reading would follow the posttest if their posttest score was not high enough. This presumably incentivized them to do well on the posttest (and thus on the *high-value* texts) to avoid an unappealing additional task. The difficulty and value manipulations were part of a larger research study and are only used here as context features when building our mind wandering detector.

3.3 Mind wandering probes

Auditory thought probes were used to measure mind wandering. Thought probes are a standard and validated method for collecting online mind wandering reports (Mooneyham and Schooler 2013; Smallwood et al. 2008).

Nine pseudorandom pages in each text were identified as *probe pages*. An auditory probe (i.e., a beep) was triggered on probe pages at a randomly chosen time interval 4–12 seconds after the page appeared. These probes were considered to be *within-page probes*. An *end-of-page probe* was triggered if the page was a probe page and the participant tried to advance to the next page before the within-page probe was

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

Table 1 Incidence of mind wandering

Response type	Yes	No	Total	Percent (yes) (%)
End-of-page	209	651	860	24.3
Within-page	1278	2839	4117	31.0
Total	1487	3490	**4977**	**29.9**

triggered. Participants were instructed to indicate if they were mind wandering or not by pressing keys marked "yes" or "no," respectively. Participants could also report mind wandering at any time by pressing the "yes" key. Pages that included these types of reports were considered *self-caught* pages. These self-caught reports ($N = 481$) were collected for use in other analyses and are not considered further in this study because the self-caught report may interfere with the mind wandering process prior to the probe. The instructions defined mind wandering as having "no idea what you just read" and realizing that "you were thinking about something else altogether." Although misreports are possible, there is no clear alternative for tracking such a highly internal phenomena.

Table 1 provides a summary of analyzed mind wandering reports. This table only includes reports that were used to build models and thus excludes those that occurred within 4 s from the beginning of the page or that were preceded by fewer than 5 fixations (see details in Sect. 4.2). The mind wandering rates of 24 % for end-of-page and 31 % for within-page probes are similar to previous studies involving reading (Schooler et al. 2004; Smallwood et al. 2008).

3.4 Procedure

All procedures were approved by the ethics board of both Universities prior to any data collection. After signing an informed consent, participants were seated in front of either a Tobii TX 300 or Tobii T60 eye tracker depending on the university (both were in binocular mode). The Tobii eye trackers are remote eye trackers, so participants could read freely without any restrictions on head position or movement. Next, they completed one of two multiple choice pretests. Each pretest was comprised of 24 deep-reasoning questions from a total of 48 questions (see Appendix 1 for example items). The remaining 24 questions were held aside for a posttest. The questions in the pretest and posttest were counterbalanced across participants.

Participants completed a brief 60-s standard calibration procedure. Participants were then instructed how to respond to the mind wandering probes based on instructions from previous studies (Feng et al. 2013). Next, they then read four texts for an average of 32.4 minutes ($SD = 9.09$) on a page-by-page basis, using the space bar to advance. The value of each text was displayed to the participants before reading. The order of the four texts, experimental conditions, and assignment of condition to text were counterbalanced across participants using a Graeco-Latin Square.

Once participants completed reading all four texts, they completed a posttest comprised of the remaining half of the 48 questions. At this point they also completed a transfer test that evaluated their ability to use previously learned knowledge in new

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

situations (see Appendix 1 for example items and performance). Participants were then fully debriefed.

4 Supervised classification

The goal was to build supervised machine learning models from short windows of data (4–10 s) prior to each mind wandering report. This consisted of three steps. First we detected eye movements from the raw gaze signal and computed features based on these eye movements within each window. Second, we performed a variety of different operations on the datasets used for our models, such as resampling the training set and removing outliers. Finally, we evaluated our models with 20 iterations of a participant-independent leave-several-participants-out cross-validation method as detailed below.

4.1 Feature engineering

The first step was to convert the raw gaze data into eye movements. Fixations (i.e., points where gaze was maintained on the same location) and saccades (i.e., eye movements between fixations) were estimated from raw gaze data using a dispersion based fixation filter from the open gaze and mouse analyzer (OGAMA), an open source gaze analyzer (Voßkühler et al. 2008). We then calculated features using data from a specific period of time (*window)* within any given page. The time series of gaze fixations was segmented into windows of varying length (4, 6, 8, and 10 s), each ending with a mind wandering probe. The windows ended immediately before the auditory probe was triggered in order to avoid confounds associated with motor activities in preparation for the key press in response to the probe. Windows that contained less 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. Three sets of features were computed: 46 global gaze features, 23 local gaze features, and 11 context features, yielding 80 features overall.

4.1.1 Global gaze features

Global gaze features (listed in Table 2) were independent of the words. There were two categories of gaze features: eye behavior descriptives and miscellaneous gaze properties. Eye behavior descriptives consisted of descriptive statistics of five eye behaviors, including (1) *fixation duration,* (2) *saccade duration,* (3) *saccade distance,* (4) *saccade angle,* and (5) *pupil diameter*. *Fixation duration* was the duration of each fixation in milliseconds. *Saccade duration* was measured as the number of seconds between two subsequent fixations. Similarly, *saccade distance* was measured as the number of pixels between two subsequent fixations. *Saccade angle* was calculated using the line segment between two subsequent fixations. The angle between this line segment and the x axis was used as the saccade angle. Pupil diameter was the pupil diameter measured by the eye tracker that was standardized by calculating the participant-level z-score for each pupil diameter in the window.For each of these

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

Table 2 Global features

Feature	Description
Fixation duration	Duration in milliseconds of a fixation
Saccade duration	Duration in milliseconds between two subsequent fixations
Saccade distance	Distance in pixels between two subsequent fixations
Saccade angle	Angle in degrees between the x-axis and the saccade
Pupil diameter	Diameter of pupil (standardized within-participant)
Number of saccades	Total number of saccades within window
Horizontal saccade proportion	Proportion of saccades with angles no more than 30 degrees above or below the horizontal axis
Fixation dispersion	Root mean square of the distances from each fixation to the average fixation position in the window
Fixation saccade ratio	Ratio of fixation duration to saccade duration
Blink count	Total number of blinks within window
Blink duration	Proportion of time spent blinking

Bold indicates that the mean, median, min, max, std. dev., range, kurtosis, and skew of the distribution of each measurement were used as features

five behavior measurements, we computed the min, max, mean, median, standard deviation, skew, kurtosis, and range, thereby yielding 40 features.

The remaining six global features consisted of miscellaneous gaze properties. The first was the *number of saccades*. The second was the *horizontal saccade proportion*, which was the proportion of saccades that had an angle no more than 30 degrees above or below the x axis. The third was*fixation dispersion,* which was calculated as the root mean square of the distance of each fixation to the average fixation position in the window. The fourth was the*fixation duration/saccade duration ratio*, which was the ratio of the sum of all the fixation durations to the sum of all the saccade durations. The final two global features were calculated from blinks, which were detected as periods where the eye tracker suddenly and briefly lost track of both eyes (Holmqvist et al. 2011). These included *blink count* and *blink duration*. *Blink count* was the number of blinks within the window. *Blink duration* was the sum of the durations of the blinks within the window.

4.1.2 Local gaze features

Unlike global features, *local gaze features* (listed in Table 3) were sensitive to the words being read. There were three categories of local features: fixation types, word characteristics, and eye movement metrics. Fixation types included (1) *first pass fixations,* (2) *regression fixations,* (3) *gaze fixations,* (4) *single fixations,* and (5) *non-word fixations* (Rayner 1998). To help illustrate these further, an example window with a set of fixations and saccades is shown in Fig. 2. The circles are fixations and the lines between circles are saccades. The number above each word is that word's index with respect to the text on the screen. The darker fixations are the first and last fixations; the first fixation is on the word "she", while the last fixation is on the word "sugar." A

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

Table 3 Local features

Feature	Description
First pass fixations	First fixation on each word during the first pass through text
Regression fixations	Fixations on words that were already passed
Gaze fixations	Consecutive fixations on the same word
Single fixations	Fixations on words that were only fixated on once
Non-word fixations	Fixations not on a word
End-of-clause Fixations	Number of fixations on the last word of a sentence
Word length	Number of characters within a word
Hypernym depth	Semantic specificity of a word (i.e., "crimson", is more specific than "red", which is more specific than "color")
Global frequency	Overall frequency of a word in English as measured by the CELEX data (Baayen et al. 1995)
Synset size	Number of synonyms of a word
Line cross saccades	Proportion of saccades with a vertical distance greater than the height of a line of text
Words skipped	Proportion of words that were skipped between fixations on subsequent words
Reading time ratio	Ratio of actual to expected reading time (200 ms times the # of words read)

For the bolded fixation measurements, the mean and standard deviation of the durations along with the proportion of fixations of that type were taken as features. For the italicized word characteristic measurements, the correlation of each measurement with the mean fixation duration was taken as a feature

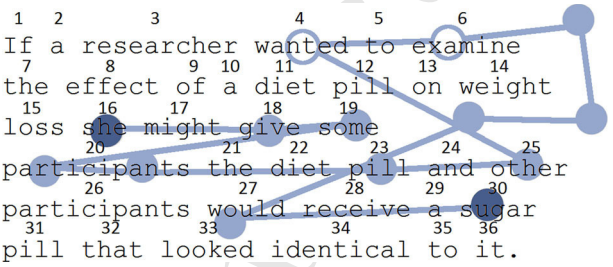


Fig. 2 An example of fixations and saccades within a window of time on a page. Fixations are represented as *circles*, while saccades are represented as the lines between each fixation. The index of each word is shown directly above the word. The beginning and ending fixation are displayed using a darker shade, and regression fixations are displayed as unfilled *circles*

first pass fixation is the first fixation on a word during the first pass through the text, such as the fixations on the words "she", "other", and "would." *Regression fixations* are fixations on earlier text. The fixations on the words "wanted" and "examine" are regression fixations because they occur after fixations on words with a higher index, such as "some" and "other." A *gaze fixation* is any fixation on the same word after accounting for the first pass fixation (e.g., the second fixation on the word "participants"). Except for the gaze fixations on "participants", the fixations on the other words in the example are *single fixations*. The three fixations that occur after the fix-

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

ation on the word "examine" are *non-word fixations*, as they do not fall on a word. Specific local features extracted from these different types of fixations included the proportion of each fixation type (compared to the total number of fixations), and the mean and standard deviation of the duration of each fixation type. This resulted in 15 local features. In addition, the number of *end-of-clause* fixations was used as a feature, based on the well documented sentence wrap-up effect (Warren et al. 2009), which posits that fixations at the ends of clauses are longer because these clauses take longer to process.

Word characteristics covered the extent to which well-known relationships between characteristics of words (in each window) and gaze fixations were observed. These included 4 features, which were correlations between fixation durations and: (1) *word length*, (2) *hypernym depth*, (3) *global frequency* of a word (Baayen et al. 1995), and (4) *synset size* of a word. These features exploit known relationships during normal reading, such as a positive correlation between word length and fixation duration. These relationships should break down during mind wandering due to the perceptual decoupling that occurs. The Pearson correlation between each of these word characteristics and fixation duration was computed and used as a feature, resulting in 4 features. Eye movement metrics captured the relationship between the movement of the eye across the screen and the position of the words in the text. These included 3 features: (1) *line cross saccades,* (2) *words skipped,* and (3) *reading time ratio*. *Line cross saccades* was the proportion of saccades that passed from one line of text to another. Each line was 75 pixels in height, so saccades between two fixations separated by a distance greater than 75 pixels on the y axis were considered line cross saccades. *Words skipped* was the proportion of words that were skipped between each pair of adjacent fixations divided by the total number of words covered. For example, in Fig. 2 one word was skipped between "she" and "give," two words were skipped between "participants" and "pill," and so on. The total number of words covered in this example is 26, as the lowest index is 4 and the highest index is 30.

The *reading time ratio* was a measure of the amount of time spent to read a portion of text in relation to how much time it should take to read that text. This was computed by taking the ratio of the actual reading time to the expected reading time. The actual reading time was calculated as the amount of time between the first and last fixations on words. The expected reading time was calculated by multiplying the number of words covered by 200ms, which is the average length of fixations during normal reading (Holmqvist et al. 2011; Rayner 1998). Calculating the number of words covered was slightly different from the calculation for number of the words skipped feature because regressions were not included. For example, in Fig. 2, the starting index is 16, as the word "she" is the first word fixated upon. The ending index for this example is 30, so the number of words covered is 14. As each word should take 200 ms to read on average, the expected reading time is 2.8 s (14×200), which makes the reading time ratio 4/2.8 or 1.43. A reading time ratio below 1 indicates that less time was taken to cover the portion of text than expected, while a value greater than 1 indicates that more time was taken than expected. Therefore, the reading time ratio of 1.43 indicates that it took longer than expected to read the words in the example.

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

4.1.3 Context features

Context features were an amalgamation of reading times and situational factors. They included timing features and the difficulty and value of the text (see Sect. 3.2). *Session time*, *text time*, and *page time* were the elapsed time between the mind wandering probe and the beginning of the session, text, and page, respectively. *Session page number* and *text page number* were the number of pages read from the beginning of the session and text, respectively. *Average page time* was the average amount of time spent reading all previous pages. *Previous page time* was the time spent reading the previous page. *Previous page time ratio* was the ratio of the previous page time to the average page time. *Current difficulty* and *current value* were the difficulty and value of the current text, respectively. *Previous difficulty* and *previous value* were the difficulty and value of the previous text, respectively. In all, there were 11 context features.

4.2 Model building

The models were built using the aforementioned sets of features. Features were calculated from the window of data preceding each probe. Each instance was then labeled according to what the participant reported in response to the thought probe ("yes" for mind wandering; "no" for not mind wandering). Different datasets were constructed for window sizes of 4, 6, 8, and 10 s in length. Twenty supervised machine learning algorithms from Weka (Hall et al. 2009) were used to build models that discriminated mind wandering from normal reading (responding "yes" vs. responding "no" to a mind wandering probe). We consider a large set of classifiers because we have no a priori prediction about the type of model that is best suited for this classification task. The following Weka implementations (with default hyperparameters) were used: bagging (with REPTree as a base learner); Bayes net; naïve Bayes; logistic regression; simple logistic regression; SMO (SVM); SPegasos (SVM); voted perceptron; k-nearest neighbors; conjunctive rule; decision table; JRip; ridor; decision stump; AdaBoost; C4.5 decision tree; grafted C4.5 decision tree; REPTree; random forest; and random tree. Multiple models were built by varying a number of external parameters. This was done in order to identify the most accurate models as well as to explore how different factors affect classification accuracy. In addition to varying classifiers, we varied six additional parameters: (1) type of *mind wandering report*; (2) *window size;* (3) *minimum number of fixations*; (4) *feature types*; (5) *feature selection*; and (6) *outlier treatment/sampling*.

First, data sets included either *end-of-page or within-page* mind wandering reports as defined above (see Sect. 3.3). These report types were analyzed separately because they occur at different moments during reading and might potentially be associated with different gaze characteristics.

Second, we calculated features using four different window lengths (4, 6, 8, and 10 s) to ascertain the amount of gaze data needed to predict mind wandering.

Third, we varied the *minimum number of fixations* that were required in each window before it was included in the data set. A lack of fixations could indicate gaze tracking problems, prolonged eye closure, off-screen gaze, etc. To account for this, each window

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

was first required to have at least 5 fixations. There were three other model types that included an additional requirement as well, resulting in four model types overall. The other three model types were required to have either 1, 2, or 3 fixations per second of window size. For instance, if the requirement was 2 fixations per second, then a window of 6 s would need a minimum of 12 fixations to be included in the data set.

Fourth, we varied the *feature types* that were used in each model in order to study the utility of each individually and in concert. Models were built with just global features, just local features, both global and local features, or global, local, and context features.

Fifth, *feature selection* was applied to the *training set only* (more details below) in order to narrow down the number of features in each model. This was done in order to remove the negative influence of features that convey the same information (e.g., number of fixations and number of saccades) and to identify the most diagnostic features. Features that were strongly correlated with other features but weakly correlated with mind wandering reports were ranked using correlation-based feature selection (CFS) (Hall 1999). The top 30, 40, 50, or 60 % features ranked by CFS were kept—this percentage was another parameter that was varied.

Finally, the *training* data was modified in seven ways, each encompassing various combinations of *outlier treatment* and*sampling*. The data was: (1) unmodified (raw); (2) trimmed; (3) Winsorized; (4) trimmed and downsampled; (5) Winsorized and downsampled; (6) trimmed and oversampled; or (7) Winsorized and oversampled. Outlier treatment was performed because outliers can cause model instability—especially for parametric models. 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. Sampling methods were also varied in data sets that were trimmed or Winsorized as there was an uneven class distribution (i.e., "no" mind wandering responses accounted for 70 % of all responses), which can have adverse effects on classification accuracy. Downsampling consisted of removing instances from the majority class (i.e., "no" responses) at random until the classes were balanced. Oversampling consisted of using the SMOTE (Synthetic Minority Over-sampling Technique) algorithm (Chawla et al. 2011) as implemented in Weka to oversample the minority class. SMOTE creates synthetic data points from a randomly selected pair of similar instances of the minority class. For each feature, the difference between the two instances is computed and multiplied by a random number between 0 and 1, which results in a new data point of the minority class. Importantly, the sampling methods were only applied to the training data.

It should be noted that varying a large number of parameters could potentially increase Type 1 error akin to performing multiple comparisons when performing inferential statistics. However, this problem does not apply here because we are not doing any significance testing in order to identify the best model. We simply determine which model out of the various configurations tested resulted in the highest mind wandering detection accuracy akin to a parameter search.

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

```
for each preprocessing parameter combination (report type, window size,
minimum number of fixations, feature type, outlier treatment):
  preprocess data based on parameters

  do 20 times:
    training set = data from random 66% of participants
    testing set = data from remaining 34% of participants
    store dataset pair
  end

  for each post processing parameter combination (feature selection size,
  sampling method, classifier):

    for each dataset pair:
      do 5 times:
        feature selection set = data from random 66% of training set
        perform feature selection and store feature rankings
      end
      selected features = top % of ranked features

      do 5 times:
        resample training set (either oversampling or downsampling)
        build model with training set and classify testing set
        compute and store evaluation metrics
      end
      average evaluation metrics for each resampling iteration and store
    end
    average evaluation metrics across all dataset pairs and store
  end
end
```

Fig. 3 Pseudocode of model building and validation

4.3 Model validation

A leave-several-participants-out nested cross validation method was used to ensure that data from each participant was exclusive to either the training or testing set. Figure 3 contains pseudocode for this process. The outermost loop contained one iteration for each possible combination of report type, window size, minimum number of fixations, feature type, and outlier treatment. The data was then split into 20 data set pairs consisting of a training set and a testing set. Data from a random 66 % of the participants were placed in the training set, while data from the remaining 34 % were placed in the testing set. At this point an inner loop varied the feature selection percentage, sampling method, and classifier for each of the data set pairs so that the same data set pairs were used across these parameters. A nested cross validation was then performed on the training set in order to select the best performing features. Data from a random 66 % of the participants in the training set were used to perform feature selection in order to avoid overfitting. Feature selection was repeated five times for each data set pair in order to minimize the variance caused by a random selection of participants, with a new random selection of participants for each data set pair. The feature rankings were averaged over these five iterations, and a certain percentage (top 30, 40, 50, or 60 % based on feature selection threshold parameter) of the highest ranked features were chosen to be used in the model. Sampling was performed after

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

feature selection, and was also repeated five times for each data set pair. Each of the five resampled training sets was used to build a model and classify the testing set, with accuracy metrics averaged across the five resampled training sets. These accuracy metrics were then further averaged across the 20 data set pairs.

The kappa value (Cohen 1960) was used to evaluate model performance as it corrects for chance. The kappa value is calculated as kappa = (observed accuracy—expected accuracy)/(1—expected accuracy), where observed accuracy is equivalent to recognition rate and expected accuracy is computed from the marginal probabilities in the confusion matrix. Kappa values of 0, 1, >0, and <0 indicate chance, perfect, above chance, and below chance agreement, respectively. Precision and recall were also considered as additional performance metrics. Precision is measured as the ratio of the number of true positives for a given class to the total number of instances classified as that class. Recall is measured as the ratio of the number of true positives of a given class to the total number of instances that actually belong to that class. The precision and recall were calculated as a weighted average based on the number of instances of each class (positive mind wandering and negative mind wandering).

5 Results

We analyzed our models in six ways. First, we analyzed the best performing model (i.e., the one with the highest kappa across all parameter combinations) for end-of-page and within-page mind wandering reports. Second, we investigated the distribution of kappa values for all of the models in order to determine how many of our models performed similarly to the best models. Third, we analyzed which feature set performed the best by comparing the best models obtained for each type of feature. Our fourth step was to study the influence of each of the external parameters on classification accuracy. Fifth, we investigated differences in the gaze features for positive vs. negative instances of mind wandering in order to study how mind wandering was manifested in our set of gaze features. Finally, we studied the predictive validity of the mind wandering detector by correlating the detector's estimates of mind wandering with important learning outcomes (i.e., prior knowledge, learning, and transfer).

5.1 Best performing models

We begin by analyzing the best (highest kappa) models for end-of-page and within-page mind wandering reports (see Tables 4, 5). A Bayes net yielded the best model for end-of-page mind wandering reports. This model had a kappa of .31 (i.e., roughly 31 % above chance) and the precision was slightly higher than the recall. The confusion matrix for this model (see Table 6) shows a 57 % chance of accurately classifying mind wandering responses (hits) versus incorrectly classifying mind wandering as normal reading (misses).

A naïve bayes classifier yielded the highest kappa of .18 (accuracy of 18 % above chance) and equivalent precision and recall for the within-page reports. The miss rate for this model selected on the basis of kappa alone was deemed to be too high (.63). Therefore, we identified an additional within-page model with a lower kappa value but

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

Table 4 Parameter values for best models

Type of probe	Classifier	Outlier treatment	Sampling	Window size (s)	Fixations/ second	Feature type	Feature count
End-of-page	Bayes net	Winsorized	SMOTE	6	1	Global+local+ context	31
Within-page	Naïve bayes	Trimmed	Downsampled	4	2	Global	16
Alternative within-page	Naïve bayes	Trimmed	SMOTE	4	2	Global	16

Table 5 Results for best models

Type of probe	Kappa	Accuracy (%)	Expected accuracy (%)	Precision	Recall
End-of-page	.31 (.06)	72 (4)	60 (5)	.76 (.05)	.72 (.04)
Within-page	.18 (.06)	67 (3)	59 (3)	.66 (.03)	.67 (.03)
Alternative within-page	.15 (.07)	60 (4)	52 (2)	.65(.03)	.60 (.04)

Standard deviations in parenthesis

Table 6 Confusion matrices for best models

	Actual	Classified		Prior
		Yes	No	
End-of-page	Yes	.57 (hit)	.43 (miss)	.23
	No	.23 (FA)	.77 (CR)	.77
Within-page	Yes	.37 (hit)	.63 (miss)	.31
	No	.20 (FA)	.80 (CR)	.69
Alternative within-page	Yes	.57 (hit)	.43 (miss)	.31
	No	.39 (FA)	.61 (CR)	.69

Values are proportionalized and averaged over iterations
FA false alarm, *CR* correct rejection

with a lower miss rate (.43) as well. This model did result in an increase in false alarms, so the choice of models should likely be based on the needs of the target application (i.e., whether misses are considered to be more detrimental than false-alarms and vice versa. The subsequent analyses focus on the "best" within-page model rather than this alternative model in order to maintain a consistent selection criterion (best kappa).

An analysis of the best models indicated that several aspects of the models were similar. First, both models were Bayesian which indicates that Bayesian models might be suitable for this type of data and classification task. Second, outlier treatment was applied to both models. Although the type of treatment differed across models (the end-of-page model benefitted from Winsorization while the within-page model was trimmed), the fact that outlier treatment was used in both of the best models might suggest that outliers are detrimental when detecting mind wandering. Third, both models included global features, suggesting that these features are particularly useful

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

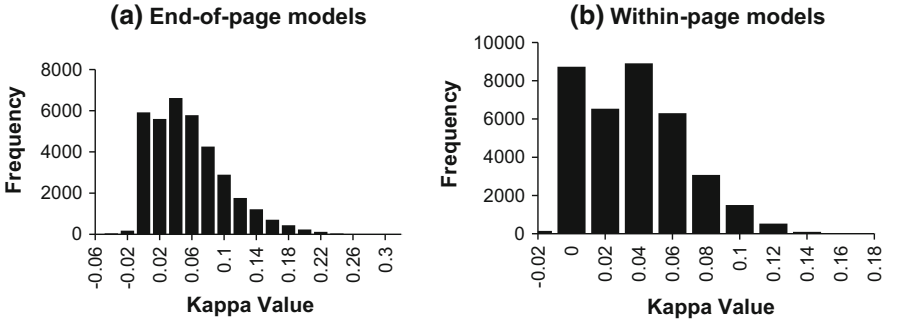


Fig. 4 Histograms of the kappa values for end-of-page (**a**) and within-page (**b**) models

for detecting mind wandering. Fourth, the training sets for each model were either downsampled or oversampled using SMOTE. This indicates that a balanced training set leads to more accurate classification of the testing set, which was not sampled in any way. The only difference between the alternative within-page model and the best within-page model was the type of sampling method used. The alternative model used SMOTE, while the best model was downsampled. This could indicate that the synthetic oversampling of the minority class (mind wandering) via SMOTE resulted in more balanced classifications. Other minor differences between the best models pertained to the window size (6 s vs. 4 s), minimum number of fixations/sec required in the window (1 vs. 2) and number of features in the model (31 vs. 16).

5.2 Comparisons of best models with other models

We also analyzed the best models in comparison with the rest of the models. Figure 4 shows histograms of the kappa values for all the end-of-page and within-page models. There was not a large difference in kappa value between the best model and the second best model for either type of probe; the difference was .01 for end-of-page models and .03 for within-page models. However, it can be seen that as the kappa value increased, the number of models with that kappa value decreased. Although there were models that obtained similar kappa values to the best models chosen above, there were very few of them in comparison to the total number of models.

5.3 Feature type comparison

It is important to determine which features result in the most accurate model because the different types of features vary in both their generalizability to different contexts and in their ease of computation. Our overall best models all included global features which indicates that these might be more useful for detecting mind wandering than local or context features. In order to investigate this further, we compared the highest kappa value obtained for each feature set across all models (Fig. 5). This analysis clearly shows that global features result in higher kappa values than local features for

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

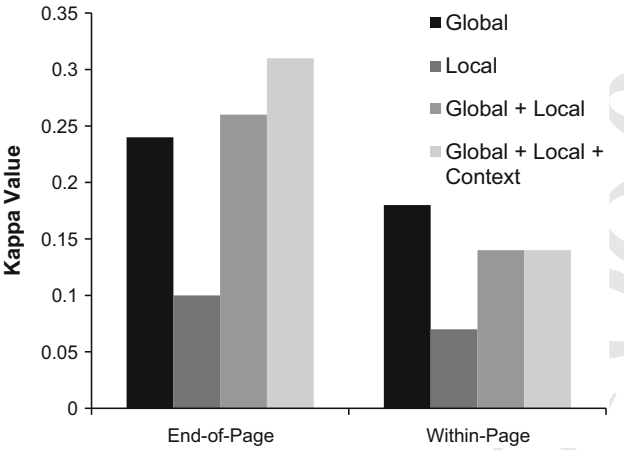


Fig. 5 Effect of *feature type* on kappa value

both end-of-page and within-page models. A combination of global, local, and context features resulted in a small improvement in kappa value over the global features alone for the end-of-page models, while the global features outperformed the combined feature set for within-page models, thereby highlighting the importance of global features for mind wandering detection.

Global features likely performed better than local and context features for two reasons. First, they are less susceptible to noise than local features. Many local features rely on an accurate alignment between fixations and specific words on the display. If the fixations are aligned with the wrong word then many of the local features will be incorrectly computed. In comparison, as long as the length and relative positions of the fixations are accurate, global features will still be correct even if the fixation positions are all shifted from their actual locations. Second, global features contain more information on localized behavior than the context features. Context features are mostly based on previous page reading times and have little to do with the actual behavior of the participant on the page. Therefore, global features seem to be more diagnostic of mind wandering than either local or context features.

5.4 Parameter comparison

We also investigated how each of the external parameters performed across all models. This is an important analysis because operations such as outlier treatment or down-sampling increase complexity in real-time systems. It is beneficial to determine if certain operations add little to classification accuracy and can therefore be omitted or reduced in scale. The analysis proceeded as follows. The kappa value associated with the best performing classifier for each of the parameter configurations was retained. Then, for each parameter value, the kappa value of all the models built with that parameter value were averaged to obtain the average kappa value for that parameter value. For example, the kappa value of each model that was built with a window size of 4

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

Fig. 6 Outlier treatment effect

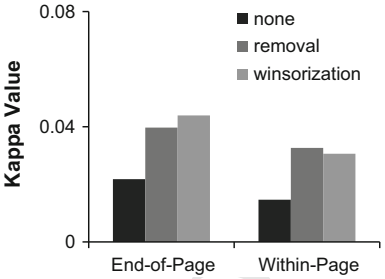


Fig. 7 Sampling method effect

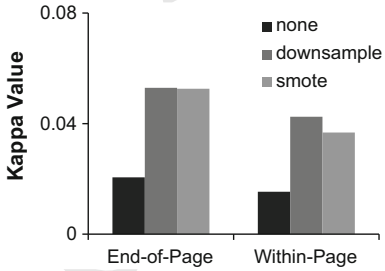
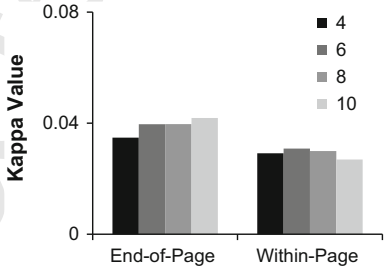


Fig. 8 Window size effect



s was averaged to obtain the average kappa value for a window size of 4 s. End-of-page models and within-page models were analyzed separately. Parameters that were analyzed in this way included outlier treatment (Fig. 6), sampling method (Fig. 7), window size (Fig. 8), minimum number of fixations (Fig. 9), and feature selection threshold (Fig. 10).

Clear trends were observed for outlier treatment and sampling method, but not for window size, minimum number of fixations, and feature selection threshold. In particular, either trimming outliers or Winsorizing outliers resulted in better performance than when not using outlier treatment (Fig. 6). With respect to sampling, the kappa values were lowest when no sampling method was used, while both downsampling and oversampling resulted in similar performance improvements (Fig. 7). This indicates that sampling was essential in this data set because it prevents the models from being skewed towards the majority class due to an uneven class distribution.

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

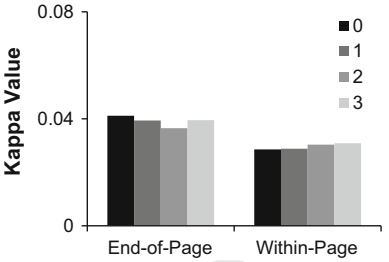


Fig. 9 Minimum number of fixations effect

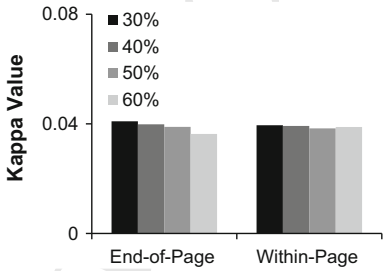


Fig. 10 Effect of *feature selection threshold* on kappa value

5.5 Feature analysis

Another goal of this study was to understand eye-gaze behavior during mind wandering. Therefore, we analyzed how the features differed between positive and negative instances of mind wandering. Data from end-of-page models and within-page models were combined for this analysis because using data from just end-of-page models resulted in too few data points for a meaningful analysis (the end-of-page models with a window size of 6 s had less than half as many data points as the within-page models with a window size of 4 s). Features were analyzed using a two-tailed paired samples *t*-test of the difference between the mean value of the feature for positive vs. negative instances of mind wandering. A non-parametric Wilcoxon signed-rank test, which does not assume normality, was also performed as a non-parametric complement to the*t*-test. Features that were significant ($p < .05$) with both tests are listed in Table 7 along with the effect sizes (Cohen's d).

There are four main conclusions regarding the differences in participants' eye gaze during mind wandering and normal reading. First, the duration and variation of specific types of fixations were greater during mind wandering, which is partly consistent with a greater fixation duration associated with mind wandering as reported in a previous study (Reichle et al. 2010). Specifically, the mean and standard deviation for gaze fixations (consecutive fixations on the same word), first pass fixations (fixations on a word during the first pass through the text), and single fixations (fixations on words that are only fixated upon once) were greater during mind wandering. The maximum fixation duration was also greater during mind wandering. Second, observed reading time exceeded expected reading time during mind wandering as indicated by a higher reading time ratio. Third, more words were skipped between fixations when the participant was mind wandering. Fourth, there was a larger variance in pupil diameter during

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

Table 7 Differences in features between positive and negative instances of mind wandering

Feature	Mean feature value		*df*	*t*	Wilcox-*Z*	Cohen's *d*
	Positive instances	Negative instances				
Fixation duration max (ms)	584 (203)	545 (136)	130	−1.98	−2.10	−.17
Gaze duration mean (ms)	240 (57)	229 (38)	124	−2.30	−2.05	−.21
Gaze duration *SD* (ms)	109 (73)	92 (34)	124	−2.62	−2.15	−.24
First pass duration mean (ms)	248 (51)	237 (36)	130	−2.42	−3.38	−.21
First pass duration *SD* (ms)	119 (67)	105 (38)	130	−2.27	−2.78	−.20
Single duration mean (ms)	249 (60)	237 (39)	130	−2.25	−2.84	−.20
Single duration *SD* (ms)	115 (72)	99 (35)	130	−2.48	−2.79	−.22
Reading time ratio*	1.58 (.063)	1.46 (.037)	130	−2.48	−1.79	−.22
Words skipped	.80 (.45)	.71 (.26)	130	−2.28	−2.46	−.20
Pupil diameter Z-Score *SD*	.59 (.15)	.56 (.08)	130	−2.73	−2.093	−.24

SD standard deviation; *df* degrees of freedom. Standard deviations are in parentheses
* The reading time ratio was only marginally significant for the non-parametric test, with a *p*-value of .07

mind wandering. Taken together, these results lend a clearer picture of participants' eye gaze associated with mind wandering. Participants spend a greater amount of time reading than can be expected during mind wandering, which can be attributed to the longer fixation duration. In addition, a greater proportion of the words are skipped, ostensibly indicating a shallower depth of processing.

5.6 Predictive validity

Our final analysis was to investigate if our models were making valid predictions in that the predicted mind wandering rate should negatively correlate with learning. The learning measures were the proportion of correct answers on the pretest, on the posttest, and on the transfer test. We first performed correlations between the three learning measures and the *actual mind wandering rate* of each participant. The actual mind wandering rate was calculated as the proportion of positive mind wandering reports to the total number of mind wandering reports for each participant. The next step was to calculate a *predicted mind wandering rate* for each participant.We built an end-of-page model and within-page model for each participant by training the models on data from all the other participants akin to a leave-on-participant-out validation method. Each model was built using the parameters of the best models (Table 4). The predicted mind wandering rate of each participant was calculated as the proportion of instances classified as mind wandering to the total number of instances classified. The mind wandering rate distributions were not normal, so we performed non-parametric correlations (i.e., Spearman's rho) as reported in Table 8.

The results indicate that the actual mind wandering rate was negatively correlated with all three learning measures. The correlations were negative for both models, but were only significant for the end-of-page models. The predicted mind wandering rate was also negatively correlated with the three learning measures, and was significant for

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

Table 8 Correlations between mind wandering rate (MWR) and learning measures

	End-of-page		Within-page	
	Actual MWR	Predicted MWR	Actual MWR	Predicted MWR
Pretest proportion correct	**−.272**	**−.470**	−.088	**−.430**
Posttest proportion correct	**−.248**	**−.556**	−.095	**−.496**
Transfer proportion correct	**−.266**	**−.415**	−.090	**−.431**

Bolding indicates significance at .05

Table 9 Partial correlations between mind wandering rate (MWR) and learning measures when controlling for pretest

	End-of-page		Within-page	
	Actual MWR	Predicted MWR	Actual MWR	Predicted MWR
Posttest prop correct	−.084	**−.361**	−.046	**−.299**
Transfer prop correct	−.132	**−.187**	−.099	**−.235**

Bolding indicates significance at .05

both types of models. These negative correlations are consistent with studies that show lowered reading comprehension during mind wandering (Feng et al. 2013; Randall et al. 2014; Smallwood et al. 2007), thereby giving us confidence that our models were accurately approximating mind wandering.

We further investigated if mind wandering rates were predictive of learning after controlling for prior-knowledge (pretest scores). The partial correlations shown in Table 9 indicate that the actual mind wandering rates were not significantly correlated with learning though they were all in the expected negative direction. However, the predicted mind wandering rates were significantly negatively correlated with learning, thereby providing evidence for the incremental predictive validity of the automated mind wandering detectors.

6 General discussion

Mind wandering is a frequent phenomenon that has a significant negative impact on performance. This suggests that intelligent systems that support critical tasks such as learning, vigilance, and decision making could benefit from some mechanism to detect and address mind wandering. As an initial step in this direction, the purpose of this paper was to build a system capable of automatically detecting mind wandering using eye gaze and contextual features in a manner that generalizes to new users. In the remainder of this section, we discuss our major findings, applications, limitations, and future work.

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

6.1 Main findings

Six main conclusions can be drawn from our results. First we have shown that it is possible to detect mind wandering during reading by analyzing eye gaze features and aspects of the reading context. We were able to unobtrusively detect mind wandering while reading both at the end of a page and within the page with accuracies of 72 and 67 % respectively. These roughly correspond to above-chance improvements of 31 and 18 %. Our work expanded on previous research by analyzing a richer eye gaze feature set that was complemented with an entire new set of context features. This resulted in an improvement in mind wandering classification accuracy over previous studies. Furthermore, our validation method ensured participant independence across training and testing sets, thereby providing evidence that our models should generalize to new users in similar contexts. Additional confidence in the generalizability of our results comes from the diversity of the training data as participants were recruited from two universities with different demographic characteristics.

Second, our results indicated that classification rates were much higher for end-of-page mind wandering reports (kappa = .31) compared to within-page mind wandering reports (kappa = .18). One possible explanation for this difference is that gaze patterns occurring at the end of a page are distinct from those within a page. Alternatively, mind wandering itself could adopt different behavioral patterns when occurring at the end of vs. within a page. It could be that text at the end of the page cues a participant to reflect on what was just read, similar to the wrap-up effect mentioned earlier (Warren et al. 2009). This could result in a form of mind wandering characterized by awareness, which is fundamentally different from the mind wandering without awareness that might occur within a page. Analyzing the differences between eye gaze patterns at the end of a page and within a page could yield a deeper understanding of mind wandering itself, which could then be leveraged to improve mind wandering classification accuracies.

Third, we have shown that feature type had a profound effect on classification accuracy. The overall best models all contained global features, and a comparison of the best models for each feature set showed that global features resulted in higher kappa values than local features. This suggests that much of the performance of the models combining global, local, and context features was due to the inclusion of global features. Hence, it might be more important to track overall gaze patterns (global features) rather than focusing on the specific words being read (local features). This is a significant finding because the global features are easier to compute and are more likely to generalize to different tasks beyond reading.

Fourth, we have shown that certain external parameters resulted in increased classification accuracies, while others were shown to have little or no effect. Specifically, both outlier treatment and sampling method had an effect on classification accuracy, while window size, minimum number of fixations, and feature selection threshold did not. Models that were either trimmed or Winsorized obtained higher kappa values than models without any outlier treatment, suggesting that outliers have an adverse effect on classification accuracy. Similarly, models that were either downsampled or oversampled (training data only) resulted in higher kappa values than models without any type of sampling method applied, indicating that an even class distribution in the

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

training set contributes to better classification accuracy in this context. Thus, both outlier treatment and sampling should be included in future work.

Fifth, we found that certain features were more predictive of mind wandering than others. These features in turn shed light on how eye movements differ between mind wandering and normal reading. We compared the difference in the mean value of each feature between positive and negative instances of mind wandering for a combined end-of-page and within-page data set. There were 10 significantly different features that provided evidence that during mind wandering, specific types of fixations were longer in duration, participants skipped more words, and ultimately took longer to read than expected. They also revealed that pupil diameter varied more during mind wandering. It is not entirely clear if these patterns are consistent with previous studies because of the inconsistency across previous studies. For example, although a longer fixation duration associated with mind wandering was found in a number of studies (Foulsham et al. 2013; Frank et al. 2015; Reichle et al. 2010), several others did not find the same effect (Smilek et al. 2010; Uzzaman and Steve 2011). It may be difficult to reach a conclusive resolution to the inconsistencies in these studies without a comprehensive analysis of the methodological differences between each study and attempts to replicate their findings. However, we hope that our results can help add to the dialogue and strengthen one set of findings over another. That being said, based on our results and those of the majority of previous studies, we would expect that future studies would also find greater fixation durations during mind wandering as we have found here.

Sixth, both theory and research would suggest that an increase in mind wandering should result in a decrease in learning (Smallwood et al. 2011). We thus verified the predictive validity of our mind wandering detectors by correlating them with measures of learning. We compared the correlations between measures of learning and the actual mind wandering rate versus the predicted mind wandering rate. In all cases, the predicted mind wandering rate was significantly negatively correlated with the learning measures, even after controlling for prior knowledge. The predicted mind wandering rates also correlated more strongly with the learning measures than the actual mind wandering rate, which might be due to limitations on participants' ability to self-report mind wandering.

6.2 Applications

Mind wandering inhibits learning from text, which is the standard way to learn. Thus, user interfaces involving reading comprehension could be improved by detecting and responding to mind wandering. In addition to reading, it is possible that gaze-based mind wandering detection during a wider array of tasks and contexts could be attempted. Attentional state estimation has already been studied in a variety of areas (see Sect. 2.2.1), and any interface that would benefit from modeling attentional states would likely also benefit from modeling mind wandering.

There are three possible ways that mind wandering detection could be used to improve systems. First, it could be used in conjunction with interventions to counteract the negative effects of mind wandering on task performance. For instance, in a system that includes several passages of critical information, a mind wandering detector could be used to detect when a user's attention is diverted from the passages and

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

the information is not likely to be retained. Then, an intervention could be deployed, such as asking the user a question about the information and prompting the user to re-read the text if the question is answered incorrectly. Second, a mind wandering detector could be used to make post task decisions. In a learning context this could mean adding items to a posttest or selecting follow up material from information on sections where mind wandering was detected, which would provide another opportunity for the material to be learned. Third, the detector could be used to diagnose and evaluate a system. For example, a mind wandering detector could detect at which points a typical user mind wanders when interacting with the system. If mind wandering occurs consistently at certain points for many users, it could indicate that parts of the interface corresponding to those points should be modified with different content or a different presentation method. We are currently testing these possibilities.

There is also the possibility that aspects of our model might be generalizable to different interaction contexts, such as watching a film or viewing an online lecture. The best performing models used global features such as fixation duration and saccade length, which can both be measured regardless of the information being viewed. These features are much more likely to generalize to different contexts compared to our local features which are only applicable if text is being displayed. Therefore, it is possible that our model could be applied to a much broader range of systems than those that display text, such as online lectures or information visualization systems. This is of course, a speculative claim that needs to be empirically verified.

6.3 Limitations

There were a number of limitations to our study that stemmed from choices in hardware, methodology, and the strength of our results. First, our choice of hardware limits the scalability of our study. The eye tracker used in our study, a Tobii TX300 eye tracker, is too expensive to be deployed in a large capacity. In general, the cost of high quality eye trackers limits the scalability of eye gaze as a mind wandering detection modality. It is possible that this will be resolved in the short term due to the steadily decreasing cost of consumer-grade eye tracking technology such as Eye Tribe ($99) and Tobii EyeX ($195) eye trackers, or with webcam based eye tracking (Sewell and Komogortsev 2010).

There were also methodological choices that restrict the generalizability of our results. Although the participants in this study were more diverse than in the previous attempts at mind wandering detection, the sample was still restricted to data from undergraduate students collected in a lab setting. Hence, quantifying performance on a more diverse population and in more diverse settings would boost some of our claims of generalizability. In addition, the data was collected in a lab environment, which can further limit the generalizability of our models. Furthermore, mind wandering was tracked using self-reports and there is the possibility for participants to self-report inaccurately or dishonestly. That being said, self-reports of mind wandering have been validated in a number of studies (Smallwood et al. 2004, 2008), and there is currently no clear alternative for tracking mind wandering. A further methodological limitation was the choice of text size. We chose a text size that was larger than what would normally be read in order to improve eye tracking precision when computing

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

local features. Given that local features did most of the work, future studies could consider smaller text sizes.

There were some limitations with our results. It was surprising that predicted mind wandering rates were more strongly correlated with learning than actual (or self-reported) mind wandering rates. It could be the case that the model tends to classify reports from the same participant in an all or nothing fashion. This would lead to distributions of mind wandering rates that are inflated with 0s or 1s, which may have increased the strength of the correlations. Alternatively, the stronger correlations might be attributed to more precise mind wandering estimates due to well-known limitations of self-reports. Future research is needed to investigate among these and other possibilities.

Another limitation with the results pertains to the highest classification accuracy of 72 %, hit rate of .57, and correct rejection rate of .77, which are moderate at best. This mind wandering detection accuracy is comparable, if not higher, than existing mind wandering detection systems (see Sect. 2.2.2). The classification task itself is difficult as mind wandering is an internal, evasive, and noisy phenomenon and generalizability to new users was emphasized by using a stringent validation method which guarantees independent training and testing sets. Accuracy would ostensibly be higher if models were optimized for individual users. That being said, an improvement is needed before the technology be deployed to drive real-world interventions. In particular, applications that involve real-time interventions during the task would require a much lower false alarm rate. If the intervention is triggered too often, users would likely begin to ignore the intervention might even become annoyed if it is too disruptive. For these applications, further work is needed to develop a mind wandering detector with a much lower false alarm rate. Alternatively, a higher hit rate at the expense of a low false alarm rate could be acceptable in applications where the intervention occur offline (e.g., tagging content for restudy). Further research would be required to determine an acceptable threshold of hits vs. false alarms across various applications, which might necessitate the use of cost-sensitive classification methods.

6.4 Future work

Future work could be focused in several areas. These include considering multiple modalities, using different methods to track mind wandering, considering different tasks, and investigating interventions that alleviate mind wandering when it is detected.

First, multimodal mind wandering detection might yield enhanced classification accuracy compared to unimodal detection due to the increased bandwidth of available information. In particular, it is possible that easily collected attributes of a user such as a user's predisposition for mind wandering or baseline measures of reading behavior could be used to improve detection rates when combined with eye gaze features. There is also the potential for detection of mind wandering from physiological signals (Blanchard et al. 2014; Pham and Wang 2015) and facial expressions in conjunction with eye gaze, individual difference attributes, and contextual cues.

Second, the generalizability of our features to other task contexts should be explored. Future research is needed to understand how a global mind wandering model

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

built in one task context (e.g., reading) can generalize to closely related (e.g., text-diagram integration) or unrelated (e.g., watching a film) contexts.

Third, our predicted mind wandering rates were more strongly correlated with learning than actual mind wandering rates, which may warrant further investigation. Poor learners might have distinct eye movement patterns, especially when they are trying to comprehend difficult material. The models could be predicting mind wandering when learners are having difficulty comprehending the material even if they are not mind wandering. This would cause an increase in the predicted mind wandering rate for participants that performed worse on the posttest regardless of their actual mind wandering rate.

Finally, possible interventions to restore attention when mind wandering is detected could be explored. In educational applications this might include flagging content ostensibly missed due to mind wandering, displaying this content in an alternate format, quizzing the student on missed content, or asking them to self-explain.

6.5 Concluding remarks

Mind wandering is a frequent phenomenon that negatively influences performance across a range of tasks. This suggests that intelligent systems that support attentional-demanding tasks could benefit from mechanisms that detect and respond to mind wandering. As an initial step in this direction, the purpose of this paper was to build a system capable of automatically detecting mind wandering using eye gaze in a manner that generalizes to new users. We demonstrated that gaze data coupled with contextual cues can be effective in automatically detecting mind wandering during reading. Our approach has several advantages. We used remote gaze trackers which are relatively unobtrusive as they use no-contact sensors. We did not require the use of a headrest and this allowed for unrestricted head and body movement during an ecologically-valid reading activity. Further, our machine learning method explicitly focused on generalizability to new users. The next challenge is to expand this program of research to different interaction contexts, to move out of the lab and into the wild by using scalable low cost eye-tracking, and to design and test automated interventions that direct attention to the task at hand when wandering is detected.

Acknowledgments We would first like to thank our collaborators at the University of Memphis for assistance with data collection. We also thank Kris Kopp, Caitlin Mills, Nigel Bosch, Jennifer Neale, Jacqueline Kory, Jonathan Cobian, and Matthew Hunter for help with data collection and analysis. The authors would also like to thank the individuals who reviewed the initial draft of this paper prior to publication. This research was supported by the National Science Foundation (NSF) (ITR 0325428, HCC 0834847, DRL 1235958).

Appendix 1

Sample items from learning assessments

Sample item from posttest

Which scenario would be the best to use a double blind study?

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

Table 10 Performance on learning assessments

Learning assessment	Proportion correct
Pretest	.515 (.198)
Posttest	.622 (.214)
Transfer test	.451 (.235)

Standard deviations in parentheses

(a) Researchers are testing the relationship between test anxiety and GPA (thematic miss)
(b) Researchers are testing the effects of lotion on making people look younger (correct answer)
(c) Researchers are testing the effect of a new soap on bacteria cells (near miss)
(d) None of the above (distractor)

Sample item from transfer test

A sports psychologist tested whether visualizations alone have an impact on muscle tone. Visualization involves picturing an activity (e.g., shooting a basketball) in your mind, but not physically doing the activity. Earlier studies compared physical exercise only to a group that did physical and visualization exercise, but the sports psychologist added two groups to her study: (1) visualization exercise only and (2) a control group that did neither. All participants were first given a standardized test of muscle tone (biceps) by a trained professional. Participants were randomly assigned to one of the four conditions. For four weeks, participants (except the control group) spent 20 minutes every morning performing exercises (physical, visualized, or physical and visualized) designed to strengthen biceps. After four weeks, bicep muscle tone was measured by the same expert, unaware of participants' condition. Both physical exercise and visualization exercise had significant effects on muscle tone, confirming earlier results that visualization exercise effectively increases muscle tone. The researcher repeated the study and found the same results.

The conclusions being drawn by the researchers are…

(a) Causal because the researcher measured muscle tone before and after the treatment conditions.
(b) Correlational because the researcher measured muscle tone before and after the treatment conditions.
(c) Causal because the participants were randomly assigned to conditions (correct answer)
(d) Correlational because the participants were randomly assigned to conditions Table 10.

References

Baayen, R., Harald, P., Richard, G.: The CELEX Lexical Database (Release 2) (1995)
Baer, R.A.: Mindfulness training as a clinical intervention: a conceptual and empirical review. Clin. Psychol. **10**(2), 125–143 (2003)

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

Barbuceanu, F., Antonya, C., Duguleana, M., et al.: Attentive User Interface for Interaction Within Virtual Reality Environments Based on Gaze Analysis. Human-Computer Interaction, pp. 204–213. Springer, Interaction Techniques and Environments, New York (2011)

Ba, Sileye O., Odobez, J-M.: A Study on Visual Focus of Attention Recognition from Head Pose in a Meeting Room, pp. 75–87. Springer, In Machine Learning for Multimodal Interaction, New York (2006)

Bergasa, L.M., Nuevo, J., Sotelo, M.A., et al.: Real-time system for monitoring driver vigilance. IEEE Trans. Intell. Transp. Sys. **7**(1), 63–77 (2006)

Bixler, R., Kopp, K., D'Mello, S.: Evaluation of a Personalized Method for Proactive Mind Wandering Reduction. In: Proceedings of the 4th Workshop on Personalization Approaches for Learning Environments (PALE 2014), pp. 33, (2014)

Bixler, R., D'Mello, S.: Toward fully automated person-independent detection of mind wandering. In: User Modeling, Adaptation, and Personalization, pp. 37–48. Springer, New York (2014)

Blanchard, N., Bixler, R., Joyce, T., et al.: Automated physiological-based detection of mind wandering during learning. In: Intelligent Tutoring Systems, pp. 55–60. Springer, New York (2014)

Börner, D., Kalz, M., Specht, M.: Lead me gently: facilitating knowledge gain through attention-aware ambient learning displays. Comput. & Educ. **78**, 10–19 (2014)

Chawla, N.V., Bowyer, K. W., Hall, L.O., et al.: SMOTE: Synthetic Minority Over-Sampling Technique.' arXiv preprint arXiv:1106.1813. Accessed Aug 25, 2014 (2011)

Christoff, K., Gordon, A.M., Smallwood, J., et al.: experience Sampling during fMRI reveals default network and executive system contributions to mind wandering. Proc. Natl. Acad. Sci. **106**(21), 8719–8724 (2009)

Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Meas. **20**(1), 37–46 (1960)

Davidson, R.J.: Empirical explorations of mindfulness: conceptual and methodological conundrums. Emotion **10**(1), 8–11 (2010)

Dong, L., Di, H., Tao, L., et al.: Visual Focus of Attention Recognition in the Ambient Kitchen. In: Computer Vision–ACCV 2009, pp. 548–559. Springer, New York (2010)

Dong, Y., Hu, Z., Uchimura, K., et al.: Driver inattention monitoring system for intelligent vehicles: a review. IEEE Trans. Intell. Transp. Syst. **12**(2), 596–614 (2011)

D'Orazio, T., Leo, M., Guaragnella, C., et al.: A visual approach for driver inattention detection. Pattern Recognit. **40**(8), 2341–2355 (2007)

Drummond, J., Litman, D.: In the Zone: towards Detecting Student Zoning Out Using Supervised Machine Learning, pp. 306–308. Springer, New York, In Intelligent Tutoring Systems (2010)

Feng, S., D'Mello, S., Graesser, A.C.: Mind wandering while reading easy and difficult texts. Psychon. Bull. & Rev. **20**(3), 586–592 (2013)

Fletcher, L., Zelinsky, A.: Driver inattention detection based on eye gaze-road event correlation. Int. J. Robot. Res. **28**(6), 774–801 (2009)

Forster, S., Lavie, N.: Distracted by your mind? Individual differences in distractibility predict mind wandering. J. Exp. Psychol. **40**(1), 251–260 (2014)

Foulsham, T., Farley, J., Kingstone, A.: Mind wandering in sentence reading: decoupling the link between mind and eye. Can. J. Exp. Psychol./Rev. Can de Psychol Exp **67**(1), 51–59 (2013)

Frank, D.J., Nara, B., Zavagnin, M., et al.: Validating older adults' reports of less mind-wandering: an examination of eye movements and dispositional influences. Psychol. Aging **30**(2), 266–278 (2015)

Franklin, M.S., Broadway, J.M., Mrazek, M.D., et al.: Window to the wandering mind: pupillometry of spontaneous thought while reading. Q. J. Exp. Psychol. **66**(12), 2289–2294 (2013)

Franklin, M.S., Smallwood, J., Schooler, J.W.: Catching the mind in flight: using behavioral indices to detect mindless reading in real time. Psychon. Bull. & Rev. **18**(5), 992–997 (2011)

Furugori, S., Yoshizawa, N., Iname, C., et al.: Estimation of driver fatigue by pressure distribution on seat in long term driving. Rev. Automot. Eng. **26**(1), 053–058 (2005)

Grandchamp, R., Braboszcz, C., Delorme, A.: Oculometric variations during mind wandering. Front. Psychol. **5**, 1000–1078 (2014)

Gruberger, M., Ben-Simon, E., Levkovitz, Y., et al.: Towards a neuroscience of mind-wandering. Front. Hum. Neurosci. **5**, 56–60 (2011)

Hall, M.A.: Correlation-based feature selection for machine learning, (1999)

Hall, M., Frank, E., Holmes, G., et al.: The WEKA data mining software: an update. ACM SIGKDD Explor. Newsl. **11**(1), 10–18 (2009)

Journal: 11257-USER Article No.: 9167 TYPESET DISK LE CP Disp.:2015/9/9 Pages: 36 Layout: Small-X

Halpern, D.F., Millis, K., Graesser, A.C., et al.: Operation ARA: a computerized learning game that teaches critical thinking and scientific reasoning. Think. Skills Creat. **7**(2), 93–100 (2012)

Holmqvist, K., Nyström, M., Andersson, R., et al.: Eye Tracking: a Comprehensive Guide to Methods and Measures. Oxford University Press, Oxford (2011)

Horvitz, E., Jacobs, A., Hovel D.: Attention-sensitive alerting'. In: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, pp. 305–313. Morgan Kaufmann Publishers Inc. (1999)

Horvitz, E., Kadie, C., Paek, T., et al.: Models of attention in computing and communication: from principles to applications. Commun. ACM **46**(3), 52–59 (2003)

Jha, A.P., Krompinger, J., Baime, M.J.: Mindfulness training modifies subsystems of attention. Cogn. Affect. & Behav. Neurosci. **7**(2), 109–119 (2007)

Just, M.A., Carpenter, P.A.: A theory of reading: from eye fixations to comprehension. Psychol. Rev. **87**(4), 329 (1980)

Kane, M.J., Brown, L.H., McVay, J.C., et al.: For whom the mind wanders, and when an experience-sampling study of working memory and executive control in daily life. Psychol. Sci. **18**(7), 614–621 (2007)

Killingsworth, M.A., Gilbert, D.T.: A wandering mind is an unhappy mind. Science **330**(6006), 932–932 (2010)

Knipling, R.F., Wierwille, W.W.: America, IVHS: Vehicle-Based Drowsy Driver Detection: Current Status and Future Prospects. National Highway Traffic Safety Administration,Office of Crash Avoidance Research (1994)

Kopp, K., Bixler, R., D'Mello, S.: Identifying Learning Conditions that Minimize Mind Wandering by Modeling Individual Attributes, pp. 94–103. Springer, New York, In Intelligent Tutoring Systems (2014)

Lutz, A., Slagter, H.A., Rawlings, N.B., et al.: Mental training enhances attentional stability: neural and behavioral evidence. J. Neurosci. **29**(42), 13418–13427 (2009)

Mason, M.F., Norton, M.I., Van Horn, J.D., et al.: Wandering minds: the default network and stimulus-independent thought. Science **315**(5810), 393–395 (2007)

Matsumoto, Y., Ogasawara, T., Zelinsky, A.: Behavior recognition based on head pose and gaze direction measurement. In: Intelligent Robots and Systems, 2000. (IROS 2000). Proceedings. 2000 IEEE/RSJ International Conference on, IEEE pp. 2127–2132 (2000)

D'Mello, S., Cobian, J., Hunter, M.: Automatic gaze-based detection of mind wandering during reading. Educ. Data Min. (2013)

Mills, C., D'Mello, S.: In Press. Toward a real-time (Day) dreamcatcher: detecting mind wandering episodes during online reading. In: *Proceedings of the 8th International Conference on Educational Data Mining*, International Educational Data Mining Society pp. XX–XX

Mooneyham, B.W., Schooler, J.W.: The costs and benefits of mind-wandering: a review. Can. J. Exp. Psychol./Rev. Can. de Psychol. Exp. **67**(1), 11–18 (2013)

Mrazek, M.D., Franklin, M.S., Phillips, D.T., et al.: Mindfulness training improves working memory capacity and gre performance while reducing mind wandering. Psychol. Sci. **24**(5), 776–781 (2013)

Mrazek, M.D., Smallwood, J., Schooler, J.W.: Mindfulness and mind-wandering: finding convergence through opposing constructs. Emotion **12**(3), 442–448 (2012)

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., et al. (eds.) Intelligent Tutoring Systems, Lecture Notes in Computer Science, pp. 112–122. Springer, Heidelberg (2012)

Navalpakkam, V., Kumar, R., Li, L., et al.: Attention and selection in online choice tasks. In: Masthoff, J., Mobasher, B., Desmarais, M.C., et al. (eds.) User Modeling, Adaptation, and Personalization, Lecture Notes in Computer Science, pp. 200–211. Springer, Heidelberg (2012)

Pham, P., Wang, J.: Attentive learner: improving mobile MOOC learning via implicit heart rate tracking. In: Conati, C., Heffernan, N., Mitrovic, A., et al. (eds.) Artificial Intelligence in Education, pp. 367–376. Springer International Publishing, New York (2015)

Randall, J.G., Oswald, F.L. Beier, M.E.: Mind-wandering, cognition, and performance: a theory-driven meta-analysis of attention regulation. Psychol. Bull. (2014)

Rayner, K.: Eye movements in reading and information processing: 20 years of research. Psychol. Bull. **124**(3), 372 (1998)

Reichle, E.D., Pollatsek, A., Fisher, D.L., et al.: Toward a model of eye movement control in reading. Psychol. Rev. **105**(1), 125 (1998)

Reichle, E.D., Reineberg, A.E., Schooler, J.W.: Eye movements during mindless reading. Psychol. Sci. **21**(9), 1300–1310 (2010)

Journal: **11257-USER** Article No.: **9167** TYPESET DISK LE CP Disp.:**2015/9/9** Pages: **36** Layout: **Small-X**

Robertson, L.H., Manly, T., Andrade, J., et al.: Oops!: performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. Neuropsychologia **35**(6), 747–758 (1997)

Roda, C., Nabeth, T.: Supporting attention in learning environments: attention support services, and information management. In: Creating New Learning Experiences on a Global Scale, pp. 277–291. Springer, New York (2007)

Roda, C., Thomas, J.: Attention aware systems. Encycl. Hum. Comput. Interact. **58**, 38 (2006)

Schooler, J.W., Reichle, E.D., Halpern D.V.: Zoning out while reading: evidence for dissociations between experience and metaconsciousness. In: Daniel T.L. (ed.) Thinking and Seeing: Visual Metacognition in Adults and Children, pp. 203–226. Cambridge, Mass.: MIT Press (2004)

Schooler, J.W., Smallwood, J., Christoff, K. et al.: Meta-awareness, perceptual decoupling and the wandering mind. Trends Cognit. Sci. (2011)

Seibert, P.S., Ellis, H.C.: Irrelevant thoughts, emotional mood states, and cognitive task performance. Mem. & Cognit. **19**(5), 507–513 (1991)

Selker, T.: Visual attentive interfaces. BT Technol. J. **22**(4), 146–150 (2004)

Sewell, W., Komogortsev, O.: Real-time eye gaze tracking with an unmodified commodity webcam employing a neural network. In: CHI'10 Extended Abstracts on Human Factors in Computing Systems, ACM pp. 3739–3744, (2010)

Smallwood, J.: Mind-wandering while reading: attentional decoupling, mindless reading and the cascade model of inattention. Lang. Linguist. Compass **5**(2), 63–77 (2011)

Smallwood, J., Beach, E., Schooler, J.W., et al.: Going AWOL in the brain: mind wandering reduces cortical analysis of external events. J. Cognit. Neurosci. **20**(3), 458–469 (2008)

Smallwood, J., Brown, K.S., Tipper, C., et al.: Pupillometric evidence for the decoupling of attention from perceptual input during offline thought 'ed. Sam Gilbert. PLoS ONE **6**(3), e18298 (2011)

Smallwood, J., Davies, J.B., Heim, D., et al.: Subjective experience and the attentional lapse: task engagement and disengagement during sustained attention. Consci. Cognit. **13**(4), 657–690 (2004)

Smallwood, J., Fishman, D.J., Schooler, J.W.: Counting the cost of an absent mind: mind wandering as an underrecognized influence on educational performance. Psychon. Bull. & Rev. **14**(2), 230–236 (2007)

Smallwood, J., McSpadden, M., Schooler, J.W.: When attention matters: the curious incident of the wandering mind. Mem. & Cognit. **36**(6), 1144–1150 (2008)

Smallwood, J., Schooler, J.W.: The restless mind. Psychol. Bull. **132**(6), 946–958 (2006)

Smilek, D., Carriere, J.S.A., Cheyne, J.A.: Out of mind, out of sight: eye blinking as indicator and embodiment of mind wandering. Psychol. Sci. **21**(6), 786–789 (2010)

Stiefelhagen, R.: Tracking focus of attention in meetings. In: Proceedings of the 4th IEEE International Conference on Multimodal Interfaces, IEEE Computer Society pp. 273, (2002)

Stiefelhagen, R., Yang, J., Waibel, A.: Estimating focus of attention based on gaze and sound. In: Proceedings of the 2001 workshop on Perceptive user interfaces, ACM pp. 1–9 (2001)

Stiefelhagen, R., Zhu, J.: Head orientation and gaze direction in meetings. In: CHI'02 Extended Abstracts on Human Factors in Computing Systems, ACM pp. 858–859 (2002)

Su, Mu-Chun, Hsiung, Chao-Yueh, Huang De-Yuan (2006) 'A Simple Approach to Implementing a System for Monitoring Driver Inattention'. In Systems, Man and Cybernetics, 2006. SMC'06. IEEE International Conference on, IEEE pp. 429–433

Sussman, E., Bishop, H., Madnick, B., et al.: Driver inattention and highway safety. Transp. Res. Rec. **1047**, 40–48 (1985)

Tawari, A., Sivaraman, S., Trivedi, M.M. et al.: Looking-in and looking-out vision for urban intelligent assistance: estimation of driver attentive state and dynamic surround for safe merging and braking. In: Intelligent Vehicles Symposium Proceedings, 2014 IEEE, IEEE pp. 115–120 (2014)

Toet, A.: Gaze directed displays as an enabling technology for attention aware systems. Comput. Hum. Behav. **22**(4), 615–647 (2006)

Torkkola, K., Massey, N., Wood. C.: Driver inattention detection through intelligent analysis of readily available sensors. In: Intelligent Transportation Systems, 2004. Proceedings. The 7th International IEEE Conference on, IEEE pp. 326–331 (2004)

Ugurlu, Y.: User attention analysis for e-learning systems using gaze and speech information. In: Information Science, Electronics and Electrical Engineering (ISEEE), 2014 International Conference on, IEEE pp. 1–5 (2014)

Uzzaman, S., Steve, J.: The eyes know what you are thinking: eye movements as an objective measure of mind wandering. Consci. Cognit. **20**(4), 1882–1886 (2011)

Journal: **11257-USER** Article No.: **9167** TYPESET DISK LE CP Disp.:**2015/9/9** Pages: **36** Layout: **Small-X**

Vertegaal, R., Shell, J.S., Chen, D., et al.: Designing for augmented attention: towards a framework for attentive user interfaces. Comput. Hum. Behav. **22**(4), 771–789 (2006)

Voit, M., Stiefelhagen, R.: Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios. In: Proceedings of the 10th International Conference on Multimodal Interfaces, ACM pp. 173–180 (2008)

Voßkühler, A., Nordmeier, V., Kuchinke, L., et al.: open gaze and mouse analyzer): open-source software designed to analyze eye and mouse movements in slideshow study design. Behav. Res. Methods **40**(4), 1150–1162 (2008)

Warren, T., White, S.J., Reichle, E.D.: Investigating the causes of wrap-up effects: evidence from eye movements and e-z reader. Cognition **111**(1), 132–137 (2009)

Yeo, M.V.M., Li, X., Shen, K., et al.: 'Can SVM be used for automatic EEG detection of drowsiness during car driving? Saf. Sci. **47**(1), 115–124 (2009)

Yonetani, R., Kawashima, H., Matsuyama, T.: Multi-mode saliency dynamics model for analyzing gaze and attention. Proceedings of the Symposium on Eye Tracking Research and Applications, ETRA '12, New York, NY, pp. 115–122. ACM, New York (2012)

Zeidan, F., Johnson, S.K., Diamond, B.J., et al.: Mindfulness meditation improves cognition: evidence of brief mental training. Consci. Cognit. **19**(2), 597–605 (2010)

Robert Bixler is a Ph.D. candidate in Computer Science at the University of Notre Dame. He received his B.A. in Computer Science and English from the Alma College in 2012. His primary interests lie in the areas of affective computing and machine learning. More specific interest focuses on building models of user states such as boredom and mind wandering based on keystrokes and eye movements respectively. His paper covers his most recent work on using eye movements for mind wandering detection.

Sidney D'Mello is an Assistant Professor in the departments of Psychology and Computer Science at the University of Notre Dame. D'Mello received his B.S. in Electrical Engineering from Christian Brothers University, his M.S. in Mathematical Science, and his Ph.D. in Computer Science from the University of Memphis. His interests include affective computing, attentional computing, intelligent learning environments, speech and language processing, human-computer interaction, and computational models of cognition. He has co-edited five books and has authored over 180 journal papers, book chapters, and conference proceedings in these areas.

Journal: **11257-USER** Article No.: **9167** ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:**2015/9/9** Pages: **36** Layout: **Small-X**