Human–Computer Interaction

ISSN: 0737-0024 (Print) 1532-7051 (Online) Journal homepage: https://www.tandfonline.com/loi/hhci20

Eye-Mind reader: an intelligent reading interface that promotes long-term comprehension by detecting and responding to mind wandering

Caitlin Mills, Julie Gregg, Robert Bixler & Sidney K. D'Mello

To cite this article: Caitlin Mills, Julie Gregg, Robert Bixler & Sidney K. D'Mello (2020): Eye-Mind reader: an intelligent reading interface that promotes long-term comprehension by detecting and responding to mind wandering, Human–Computer Interaction, DOI: 10.1080/07370024.2020.1716762

To link to this article: https://doi.org/10.1080/07370024.2020.1716762

Published online: 31 Jan 2020.

HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/07370024.2020.1716762

Eye-Mind reader: an intelligent reading interface that promotes long-term comprehension by detecting and responding to mind wandering

Caitlin Mills [a], Julie Gregg[b], Robert Bixler[c], and Sidney K. D'Mello[b]

[a]Department of Psychology, University of New Hampshire, Durham, NH, USA; [b]Institute of Cognitive Science, University of Colorado Boulder, Boulder, CO, USA; [c]Department of Computer Science, University of Notre Dame, Notre Dame, IN, USA

ABSTRACT

We zone out roughly 20-40% of the time during reading – a rate that is concerning given the negative relationship between mind-wandering and comprehension. We tested if Eye-Mind Reader – an intelligent interface that targeted mind-wandering as it occurred – could mitigate its negative impact on reading comprehension. When an eye-gaze-based classifier indicated that a reader was mind-wandering, those in a MW-Intervention condition were asked to self-explain the concept they were reading about. If the self-explanation quality was deemed subpar by an automated scoring mechanism, readers were asked to re-read parts of the text in order to correct their comprehension deficits and improve their self-explanation. Each participant in the MW-Intervention condition was paired with a Yoked-Control counterpart who received the exact same interventions regardless of whether they were mind-wandering. Results indicate that re-reading improved self-explanation quality for the MW-Intervention group, but not the control group. The two conditions performed equally well on textbase (i.e. fact-based) and inference-level comprehension questions immediately after reading. However, after a week-long delay, the MW-Intervention condition significantly outperformed the yoked-control condition on both comprehension assessments (*ds* = .352 and .307). Our findings suggest that real-time interventions during critical periods of mind-wandering can promote long-term retention and comprehension.

ARTICLE HISTORY
Received 24 February 2019
Revised 12 January 2020
Accepted 12 January 2020

KEYWORDS
Intelligent UI; education; mind wandering; cognitive science

1. Introduction

Imagine reading a book. Your eyes are naturally moving across the words in the text and you may even flip to the next page. Suddenly you realize your mind is no longer focused on the reading and instead has wandered off to some unrelated topic – for example, something on your to-do list, an event on the weekend, or a fond memory from last week. The ubiquitous experience of such zone outs – referred to as 'mind-wandering' – occurs approximately 30–50% of the time during everyday activities (Killingsworth & Gilbert, 2010; Klinger & Cox, 1987; Mills, Raffaelli, Irving, Stan, & Christoff, 2018; Smith, Mills, Paxton, & Christoff, 2018) and around 20–40% during comprehension and learning tasks (D'Mello, 2018; Hutt, Mills, White, Donnelly, & D'Mello, 2016; Mills, D'Mello, & Kopp, 2015; Mills, Graesser, Risko, & D'Mello, 2017; Risko, Buchanan, Medimorec, & Kingstone, 2013; Wammes, Seli, Cheyne, Boucher, & Smilek, 2016). Although mind-wandering is associated with some benefits, such as our ability to leave the "here and now", plan future events, and creatively ideate, ample evidence suggests it may be detrimental in other situations, particularly during more

CONTACT Caitlin Mills caitlin.mills@unh.edu Department of Psychology, University of New Hampshire, 431 McConnell Hall, 15 Academic Way, Durham, NH, 03824, USA

complex tasks (D'Mello, 2018; Randall, Oswald, & Beier, 2014; Smallwood, Fishman, & Schooler, 2007). For example, a meta-analysis on 88 studies by Randall et al. (2014) found the average negative correlation between mind-wandering and performance increased from −.14 (SE = .03) to r = −.32 (SE = .04) when restricted to more complex tasks such as problem-solving and reading.

The prevalence of mind-wandering and its negative relationship with performance has led to a growing interest in understanding and mitigating its impact in a number of settings, particularly in educational contexts (Hutt et al., 2016; Kane et al., 2017; Mills et al., 2013; Risko et al., 2013; Szpunar, 2017; Szpunar, Khan, & Schacter, 2013; Wammes et al., 2016). The current research takes a step toward this goal by developing and experimentally testing the Eye-Mind Reader interface: a real-time system that uses eye gaze to detect when mind-wandering occurs, then intervenes using self-explanations and re-reading to correct short term comprehension deficits attributed to mind wandering in order to improve long-term comprehension. We were guided by two theoretical considerations. First, interventions should occur the moment readers mind wander so that potential gaps in their understanding due to mind-wandering can be corrected immediately. Second, the intervention itself should aim to promote deeper understanding through strategies that have been shown to enhance comprehension. Whereas the mind wandering detection method has been previously reported (Faber et al., 2018), we discuss here how it was embedded and tested in the Eye-Mind interface to improve comprehension outcomes.

1.1. *Theoretical background*

Converging evidence suggests that mind-wandering can have negative effects on information processing even at early perceptual levels (Kam et al., 2011; Schad, Nuthmann, & Engbert, 2012). According to Smallwood's (2011) *cascade model of inattention*, when mind-wandering interrupts basic levels of encoding during early perceptual stages, it has cascading negative consequences for later stages of processing (e.g., memory access, generation of inferences), ultimately resulting in sub-optimal representations. In the context of reading, impairments due to mind-wandering begin at the most basic (surface) level of lexical (i.e., word) encoding and cascade to deeper levels of comprehension, such as the situation model (i.e., conceptual representations of the text content – Zwaan & Radvansky, 1998) that require the generation of inferences. The situation model is critical to deep text comprehension because it guides readers' text processing by keeping track of the current state of affairs and generating predictions about what will happen next (Zwaan, Magliano, & Graesser, 1995; Zwaan & Radvansky, 1998).

In support of this account, research suggests that mind-wandering leads to a decoupling from the external stimuli, including auditory and visual information (Kam et al., 2011). This is problematic during reading since participants eye-movements (Foulsham, Farley, & Kingstone, 2013; Reichle, Reineberg, & Schooler, 2010) and reading times (Franklin, Smallwood, & Schooler, 2011; Mills et al., 2017) are no longer sensitive to the lexical properties of a text while mind-wandering – which directly impacts the ability to encode the text, even at the word level. Beyond lexical encoding, participants who reported more mind-wandering have poorer memories for specific details of the text, perhaps because they "missed" the information (Feng, D'Mello, & Graesser, 2013; Franklin, Mooneyham, Baird, & Schooler, 2013; Franklin et al., 2011; Kopp, D'Mello, & Mills, 2015; Mills et al., 2017; Unsworth & McMillan, 2012). Finally, there is also ample evidence to support the cascading effect of mind-wandering at deeper levels of processing: Smallwood, McSpadden, and Schooler (2007) reported that if readers mind-wandered while critical information was prevented in a Sherlock Holmes story (e.g., the villain was wearing a hood; Smallwood et al., 2007), they were less likely to generate key inferences later in the story (e.g., John was wearing a hood, thus John is the villain) and subsequently performed worse on a memory test. This finding is consistent with the negative relationship between mind-wandering and comprehension assessments, especially those requiring participants to make inferences across multiple parts of a text (Mills et al., 2015, 2017; Randall et al., 2014; Smallwood et al., 2007).

1.2. *Related work on mind-wandering interventions*

To date, much of the research aiming to mitigate the influence of mind-wandering has taken a proactive approach – that is, it has focused on eliminating or reducing it (Bixler, Kopp, & D'Mello, 2014; Jha, Krompinger, & Baime, 2007; Szpunar et al., 2013). For example, interpolated testing (e.g., intermittently providing quizzes throughout a learning session) has been shown to reduce overall rates of mind-wandering in video lectures from about 40% to only 20% (Szpunar et al., 2013). Additionally, a simulation study revealed the possibility of using individual differences to pre-select optimal learning conditions (e.g., easy vs. difficult texts, and high vs. low stakes value on a test after reading) that may help deter mind-wandering (Kopp, Bixler, & D'Mello, 2014), but this has yet to be empirically tested on real participants.

With these exceptions, most proactive intervention attempts have focused on mindfulness training. Mindfulness training aims to promote a non-judgmental focus on the present moment (Jha et al., 2007; Mrazek, Franklin, Phillips, Baird, & Schooler, 2013) which encourages attentional control through the practice of attention regulation. When something besides the present moment comes into focus, mindfulness training encourages individuals to reorient their attention back to the present moment immediately. Although mindfulness training has been shown to be effective in several domains (Brown & Ryan, 2003; Davidson et al., 2003; Grossman, Niemann, Schmidt, & Walach, 2004), its effectiveness for reducing mind-wandering has been less consistent (Ju & Lien, 2016; Krasich et al., 2018; Mrazek et al., 2013; Mrazek, Smallwood, & Schooler, 2012; Zeidan, Johnson, Diamond, David, & Goolkasian, 2010) – particularly for brief training exercises (less than 30 minutes).

Importantly, although proactive interventions can be effective, mind-wandering rates are never reduced to zero (Mrazek et al., 2013, 2012; Szpunar et al., 2013). This points to an important question: How should we respond to mind-wandering when it does inevitably occur? Such just-in-time *reactive* interventions have rarely been explored. Responding to mind-wandering episodes requires real-time detection and intervention.

1.2.1. *Mind-wandering detection*

The ability to intervene when someone is mind-wandering is contingent on knowing when mind-wandering occurs – i.e., having a real-time measure of mind-wandering that is not dependent on self-reports. This is difficult since mind-wandering is an elusive, internal cognitive state with few obvious behavioral correlates (unlike emotions for example). One solution to this problem is to use machine learning approaches to model the relationship between physiological and behavioral signals and self-reported instances of mind-wandering (D'Mello, Dieterle, & Duckworth, 2017), such that the learned model can then be used as an automated measure of 'mind-wandering.' This approach has been successful at detecting mind-wandering using both behavioral (Franklin et al., 2011; Mills & D'Mello, 2015; Stewart, Bosch, Chen, Donnelly, & D'Mello, 2016) and physiological signals (Bixler & D'Mello, 2014; Blanchard, Bixler, Joyce, & D'Mello, 2014; Drummond & Litman, 2010; Faber, Bixler, & D'Mello, 2017; Hutt et al., 2017, 2016) – with accuracy rates reflecting roughly 20% improvement over chance levels.

Eye-gaze data is one of the most promising signals for mind-wandering detection (Bixler & D'Mello, 2016; Faber et al., 2017; Hutt et al., 2017, 2016; Mills, Bixler, Wang, & D'Mello, 2016) based on the strong coupling between eye-movements and visual attention called the "eye-mind" link (Reichle, 2006; Reichle, Pollatsek, & Rayner, 2012; Starr & Rayner, 2001). Gaze-based mind-wandering detection models have been successfully developed using both probe-caught (i.e., self-reports in response to thought probes) as well as self-caught mind-wandering (i.e., self-reports once the participant becomes aware that they are mind-wandering) across multiple studies. The accuracy rates of eye-gaze-based mind-wandering detectors (up to 46% above chance; Hutt, Mills, et al., 2017) will likely continue to improve, but for now, we think that detection has reached a point where it can be applied to test real-time interventions as we do here. Critically, past studies have also demonstrated predictive validity in that the models' estimates of mind wandering are negatively correlated with comprehension similar to self-reports (Bixler & D'Mello, 2014; Faber et al., 2017; Hutt et al., 2019).

Finally, these detectors have been built using a leave-several-people-out cross-validation technique (or leave-one-person-out method), so there is some confidence that they will generalize to new participants. This is because data from the same person is either included in the training or testing set, but never both, thereby avoiding overfitting due to a lack of independence across the training and testing sets. Person-independent detectors are particularly useful for the purpose of real-time interventions because they need to make accurate predictions for new people without any information on them at the outset of the task.

1.2.2. *Attempts at real-time intervention*

Outside of mind-wandering, other domains have leveraged the improved accuracy, cost-effectiveness, and scalability of eye-trackers to provide real-time feedback. For example, eye gaze can be used to increase user control during video games (Smith & Graham, 2006), improve interaction with artificial agents (Vertegaal, Slagter, Van der Veer, & Nijholt, 2001), and improve driver experience and safety (Feit et al., 2017; Merat, Jamson, Lai, Daly, & Carsten, 2014). In the context of learning, one early study attempted to improve remedial reading ability by using eye gaze to detect when readers were struggling with recognition and pronunciation of words (Sibert, Gokturk, & Lavine, 2000). When the system detected a reader was having trouble, an auditory prompt would give students hints on the problematic words. Readers showed improvements in reading speed and reduced errors after using the system across multiple sessions. However, the sample size in this study was small (N = 8) and there was no control condition; thus, it is difficult to draw major conclusions from these results. More recently, D'Mello, Olney, Williams, and Hays (2012) used a commercial grade eye tracker to reengage students while they were learning from an interactive computerized tutoring system called GuruTutor. In a gaze-reactive condition, the tutor used conversational moves to direct students to reorient their attention if they were disengaged as measured by extended periods of off-screen gaze. Compared to a control group that did not receive any reactive dialog, the gaze-reactive interventions successfully promoted learning gains on deep reasoning questions.

Only one known study, conducted by our lab, has attempted to automatically respond to mind-wandering (D'Mello, Mills, Bixler, & Bosch, 2017). This study prompted students in an experimental condition to answer a textbase-level (i.e. based on factual information in the text) multiple-choice question on the content of a page they just read if an eye-gaze-based mind-wandering detector classified them as mind wandering while reading that page. If they answered incorrectly, they were given an opportunity to re-read the text, and then were either presented with the same question again or a different textbase-level question on the same content. Each intervention participant was paired to a yoked-control participant who received the same interventions on the same pages. Participants completed a posttest assessing textbase-level comprehension after reading.

This study reported no significant differences in overall posttest performance between the two conditions. However, they did identify certain cases when the experimental condition outperformed the control condition by designating each page as either high (1) or low (0) mind-wandering based on a median split of the detector's likelihood predictions. Results indicated that the experimental condition performed significantly better on textbase posttest items that corresponded to pages when the experimental condition had a low likelihood of mind-wandering but the control condition had a high likelihood of mind-wandering (d = .548); there were no differences when the reverse was true (d = .103) or for cases where both conditions were low (d = .066) or high (d = .023) mind-wandering. The authors interpret this finding as the intervention correcting comprehension deficiencies associated with mind-wandering, albeit at the textbase-level only.

Although this study demonstrates the potential feasibility of responding to mind-wandering in real-time, there is much room for improvement. In particular, the intervention targeted textbase-level comprehension which engendered shallow processing strategies and some participants even reported "skimming" the text to look for the correct answers while they were re-reading (D'Mello et al., 2017). In addition, only textbase-level comprehension was assessed immediately after reading in their study, which does not account for deeper levels of comprehension (e.g., inference-level comprehension), or the potential for a delayed-effect after a longer memory consolidation period (Ellenbogen, Hu, Payne, Titone, & Walker, 2007; Nadel, Hupbach, Gomez, & Newman-Smith, 2012; Nadel & Moscovitch, 1997).

1.3. *Current study*

Our goal is to build a reactive intervention that promotes comprehension by providing an opportunity for readers to "fill in," or strengthen information they missed as a result of mind-wandering. Interventions that simply "bring the reader back" to the text may not address the problem because they are never given an opportunity to integrate the missed material. We were focused on designing an intervention that would improve conceptual understanding rather than memory of textbase-level facts. Accordingly, we examined the reading comprehension literature to identify strategies that most effectively promote deep comprehension (McNamara, 2007). *Self-explanation* (i.e., explaining the meaning of information to oneself during a learning task) has been shown to be one of the most effective learning strategies for promoting deep comprehension (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; McNamara, 2004; VanLehn, Jones, & Chi, 1992). In particular, self-explanation leads readers to better integrate the text content with prior knowledge, generate more inferences, and more effectively monitor their comprehension.

Self-explanations may be particularly beneficial during mind-wandering, as missed or weakly-coded information may be effectively reconstructed to create an improved mental model of the text. This prediction is based on the Interactive-Constructive-Active-Passive (ICAP) Framework (Chi & Wylie, 2014), which posits that readers are more likely to benefit from an intervention that requires the readers to engage in a constructive activity in comparison to more passive interventions that do not involve any elaborative processing (e.g., simply re-directing attention). Constructive processing encourages readers to think deeply about the text while actively formulating a response that demonstrates conceptual understanding, while also evaluating their own understanding and exposing knowledge gaps. Accordingly, readers were prompted to construct self-explanations when an eye-gaze-based detector detected that they were mind wandering. Importantly, these prompts occurred before readers were able to move on in order to correct comprehension deficits thereby boosting their conceptual understanding of the text.

To the best of our knowledge, our fully-automated system is the first to track and combat mind wandering in real-time during reading. We also improved upon our previous attempt (D'Mello et al., 2017) in a number of ways. First, we developed a new self-explanation/re-reading intervention that targeted deeper levels of conceptual understanding across multiple pages of text rather than fact-based understanding on individual pages. Second, this required the reader to construct an open-ended (free text) self-explanation (vs. multiple-choice response as in previous study), which we scored using basic natural language processing techniques. Third, we improved upon the previous textbase comprehension assessment by also measuring inference-level comprehension immediately after reading. Fourth, we assessed both textbase and inference-level comprehension after a week to measure retention and to account for any potential delayed effects of the intervention after memory consolidation has occurred. Finally, we improved several usability aspects of the computerized reading interface based on feedback from the previous study.

2. Building a real-time mind-wandering detection and intervention system

Guided by the cascade model of inattention (see above), our goal was to detect and mitigate the negative effects of mind-wandering during reading. We addressed a number of considerations in the development of the intervention, which are detailed below.

2.1. *Text selection and computerized reading interface*

The intervention is intended to be text-agnostic; however, we focused on one text in this proof-of-concept implementation. The text (approximately 6,500 words) was an excerpt taken from the first 35 pages of a book entitled "Soap-bubbles and the Forces which Mould Them" (Boys, 1895). The text is about the physical behaviors of soap bubbles, how surface tension enables bubble formation, and how chemical composition affects bubble formation. We selected this text because of the mundane subject matter (i.e., the high likelihood of mind-wandering) and because we anticipated that most readers would be unfamiliar with the concepts but it is written to be understandable without prior knowledge of the topic. The original text excerpt contained multiple figures and textual references to them. We removed these to simplify the display and because the images were not necessary to understand the text. The text was otherwise consistent with the original published version.

The text was divided into 57 different screens (approximately 115 words each) that were presented as "pages" on a computer monitor in 24-point Courier New typeface, with a line height of 2.15. We chose this text presentation style to reflect relatively ecological reading with multiple sentences on screen at once. At the same time, restricting the words per screen allowed for there to be sufficient space between the lines for more accurate eye-tracking, which was important given that our remote eye tracker does not restrict head movements (see below). Readers could navigate through the text by pressing the right and left arrow keys at their own pace.

For purposes of the intervention and knowledge assessments, the 57 screens were grouped into 15 "sections" based on distinct concepts in the text. The sections were between three and five screens long (see Table 1 for an example section). The first three screens are dedicated to explaining the author's motivation for writing the book, whereas a later section describes an experiment on how a water sieve works, how evaporating alcohol contributes to the formation of "legs" on the sides of a wine glass, and why wet clothes balloon out underwater yet stick to you once you are out of the water.

2.2. *Automated mind-wandering detection*

The mind wandering detector used in this study has been described in detail elsewhere (see Faber et al., 2017). Here, we provide a brief overview of how it was developed and validated.

2.2.1. *Development and validation*

The mind-wandering detectors built by Faber et al., (2018) utilized a dataset where eye gaze data was recorded from 132 undergraduate students across two universities (see Kopp et al., 2015 for a description of the original dataset used to build the detector). In this study, students completed a computerized reading task using the same text excerpt and a similar reading procedure described above with the exception of the intervention procedure described below.

Mind-wandering was measured using a self-caught method where participants pressed a labeled key whenever they caught themselves mind-wandering. Prior to beginning the task, students received the following definition of mind wandering (we used the more colloquial term zone outs):

> *Your primary task is to read the text in order to take a short test after reading. At some points during reading, you may realize that you have no idea what you just read. Not only were you not thinking about what you are actually reading, you were thinking about something else altogether. This is called "zoning out." If you catch yourself zoning out at any time during reading, please … [REPORTING INSTRUCTIONS].*

Each page with a self-report of mind-wandering was used as a positive instance of mind-wandering, while the remaining pages were considered as negative instances of mind-wandering (i.e. the absence of mind-wandering). On average, readers reported mind-wandering 18 times across the 57 pages of text with a standard deviation of 12 and a range of 0 to 53. Five participants never reported any mind wandering, but were still included in the analyses.

Table 1. Example text section and corresponding assessments. Superscript in the text example indicates page breaks.

Text (pages 23–25):

[23] if you will only dip a paraffin candle into water. I have melted a quantity of paraffin in a dish and dipped this gauze into the melted paraffin so as to coat the wire all over with it, but I have shaken it well while hot to knock the paraffin out of the holes. You can now see on the screen that the holes, all except one or two, are open, and that a common pin can be passed through readily enough. This then is the apparatus. Now if water has an elastic skin which it requires force to stretch, it ought not to run through these holes very readily; it ought not to be able to get through at all unless forced, [24] because at each hole the skin would have to be stretched to allow the water to get to the other side. This you understand is only true if the water does not wet or really touch the wire. Now to prevent the water that I am going to pour in from striking the bottom with so much force as to drive it through, I have laid a small piece of paper in the sieve, and am pouring the water on to the paper, which breaks the fall. I have now poured in about half a tumbler of water, and I might put in more, I take away the paper but not a drop runs through. If I give the sieve a jolt then the water is driven to [25] the other side, and in a moment it has all escaped. Perhaps this will remind you of one of the exploits of our old friend Simple Simon, "who went for water in a sieve, but soon it all ran through." But you see if you only manage the sieve properly, this is not quite so absurd as people generally suppose. If now I shake the water off the sieve, I can, for the same reason, set it to float on water, because its weight is not sufficient to stretch the skin of the water through all the holes. The water, therefore, remains on the other side, and it floats even though, as I have already said, there are [26]

Self-explanation Question

What point does the author make about when water will run through a sieve?

Example Ideal Answer

Water will only run through a sieve when enough force is used to propel the water through the holes. Otherwise, the elastic skin of the water prevents it from running through the hole easily.

Keywords

sieve, forced, force, holes, run through

Textbase-level Comprehension Assessment Item

Why did the author shake the gauze coated in melted paraffin wax while hot?

(a) to knock the paraffin out of the holes (correct)
(b) to spread the paraffin so it can cover all the holes (near miss)
(c) the author did not shake the gauze while hot (thematic miss)
(d) because he burned his hand (distractor)

Inference-level Comprehension Assessment Item

Which of the following is the most similar to how water behaves if you poured it in a wax-coated thimble covered with holes?

(a) Like a colander for draining (near miss)
(b) Carrying a pail of water with a leak (thematic miss)
(c) **Like a regular cup of water (correct)**
(d) None of the above. (distractor)

A Tobii TX300 gaze tracker was used at the Midwestern university with a sampling rate of 120 Hz and a Tobii T60 was used at the Southern university with a sampling frequency of 60 Hz. Both Tobii eye-trackers were remote and therefore did not restrict head position or movement. Gaze data was collected for each eye and included the x and y position of the gaze on the screen, as well as pupil diameter. This data was converted into a series of fixations (periods where the eye remained fixated on the same point), saccades (eye movements between fixations), and blinks, from which 62 gaze features were calculated (see Faber et al., 2018 for more details on feature engineering). The specific features included: 1) eye movement descriptive features, 2) pupil diameter descriptive features, 3) blink features, and 4) miscellaneous gaze properties. The eye movement descriptive features were statistical functionals for fixation duration, saccade duration, saccade amplitude, saccade velocity, and relative and absolute saccade angle distributions. For the pupil diameter descriptive features, the eye-tracker's estimate of pupil diameter was standardized by computing the participant-level z-score. For these two groups of features, we computed the minimum, maximum, mean, median, standard deviation, skew, kurtosis, and range. The blink features were simply the number of blinks and the mean blink duration. Finally, the miscellaneous gaze properties consisted of the number of saccades, horizontal saccade proportion, fixation dispersion, and the fixation duration/saccade duration ratio.

The eye gaze features were calculated for different window sizes prior to each mind-wandering report in order to test how much data was sufficient for making accurate predictions (see Faber et al., 2018 for more details on window selection). For pages with a mind-wandering self-report, the end point of the window was positioned 3 seconds prior to the self-report in order to avoid capturing changes in the gaze data associated with reporting mind-wandering (i.e., looking down at the keyboard to press a button). For pages without a self-report, the window was 16 seconds into the page, which was based on the average time between a self-report and the beginning of the page. Pages that were shorter than the window size were discarded, as were pages with windows that contained fewer than five fixations (i.e., not enough data to compute features). This was only done during the initial detector development – no cases were discarded during deployment in the current study.

Finally, supervised classification models (e.g., support vector machines, naïve Bayes classifier, Random forest) were trained to discriminate instances of mind-wandering (pages with a self-report) from instances of normal reading (pages without a self-report). Reader-independent validation is necessary for the models to generalize to new readers. Accordingly, the models were validated using a leave-one-reader-out cross-validation method in order to ensure that data from each reader was exclusive to the training or testing set. This entailed training the model on data from $n - 1$ participants and testing the model on the remaining participant until each participant was a "test" participant.

Based on Faber et al.'s results, the best model was a sequential minimization optimization (SMO) algorithm, an implementation of a support vector machine (SVM) classifier. The model only used global features (e.g., features that are content-independent) and operated on a window size of 8-seconds. It classified a page as mind-wandering or not based on whether its predicted likelihood of mind-wandering (ranges from 0 to 1) was below or above the .5 threshold. The default .5 threshold was adopted as it led to a higher rate of true positives while maintaining a moderate rate of true negatives. The chosen model had a weighted precision of 72.2% and a weighted recall of 67.4%, which we deemed to be sufficiently accurate for intervention. Critically, the model's predictions of mind-wandering during reading were negatively correlated with textbase-level comprehension scores after reading, $r = -.374$; $p < .001$, providing evidence that the detector is sensitive to potentially harmful instances of mind-wandering.

2.2.2. *Implementing the detector and deciding when to intervene*

We embedded the classification model into the computerized reading interface for automated mind-wandering detection and intervention. Eye gaze features were computed each time a participant advanced to a new page. These values were then used as features for the detector, which resulted in a mind-wandering likelihood per page, expressed as a probability between 0 and 1. We did not include any gaze features from participants' re-reading behaviors because the model was only trained on first reads and there are known differences in processing strategies during re-reading (e.g., faster reading times, skimming, etc.; Callender & McDaniel, 2009; Phillips, Mills, D'Mello, & Risko, 2016) compared to the first-read. We also accounted for instances when eye-gaze features could not be computed (e.g., insufficient reading time, missing data for individual features). In these cases, the detector still made a probabilistic prediction based on default probabilities from the original training data used in Faber et al., (2018). For example, participants in the training data reported mind-wandering 56.3% of the time when they spent less than 4s reading a page; therefore, the detector applied the same default probability (56.3%) of predicting mind-wandering on pages where reading times were less than 4s without considering any of the gaze features.

As discussed above, the text was divided in the 15 different conceptual sections. Interventions could only occur at conceptual section breaks in the text (see Figure 1) because concepts spanned multiple pages of text. We used an aggregation and probabilistic method to determine when to intervene to mitigate the inherent error in mind wandering detection. Specifically, we first computed the average mind-wandering likelihood across all pages in a given section to increase reliability of the signal. The probabilistic decision method was implemented using a sigmoid-like function (see Figure 2). An intervention was never deployed if the section's average mind-wandering likelihood was below .3. If the likelihood fell in the middle section from (.30 – .66), the average likelihood score was used to

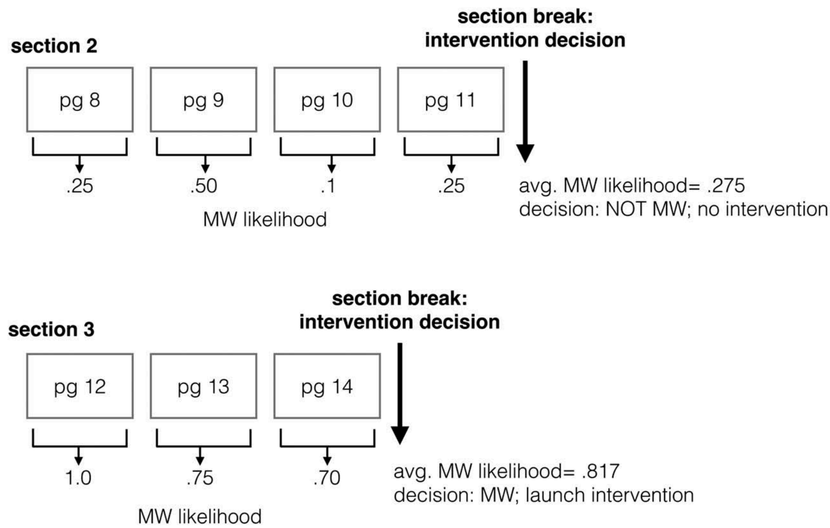

Figure 1. Example sections and decisions about mind-wandering based on average likelihood values.

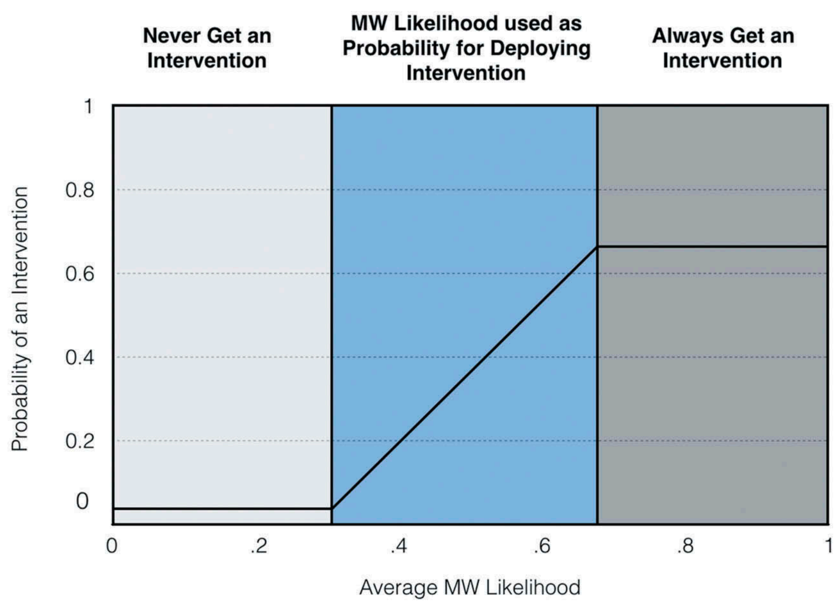

Figure 2. Sigmoid-like function used to make decisions about deploying an intervention based on the average mind-wandering likelihood in a given section. Values less than .30 will not receive intervention whereas values greater than .66 will always receive intervention. Values that fall in the middle will probabilistically receive intervention based on their average mind-wandering likelihood.

probabilistically determine if a student would receive an intervention. For example, if a student had an average mind wandering likelihood of .40, he or she had a 40% chance of getting an intervention. Finally, students always received an intervention if the average mind-wandering likelihood was above .66. These cutoff points were decided after examining histograms and scatterplots from pilot data (N = 18). At least 25% of MW likelihoods obtained in the pilot data fell within the "never receive an intervention" section of the function. Another 25% of the data fell within the "always receive an intervention" section of the function, leaving 50% left for the probabilistic decision method (see Figure 2).

2.3. Intervention development

The goal of the intervention was to ensure that the reader (i.e., the user) understood individual concepts before progressing through the text and we utilized self-explanations for this purpose.

2.3.1. Self-explanation questions

Our interventions targeted some of the mechanisms described in McNamara's (2004) definition of self-explanations, such as understanding gaps in understanding, explaining and elaborating on concepts, and tying explanations to prior knowledge. The questions required readers to elaborate on a concept described in the text in order to demonstrate understanding. For example, readers were asked to explain what the author's point was or how a concept from the text would apply to another real-world situation (e.g., "Based on an experiment described in the text, what would happen to your clothes when you jump in a pool with them on?"). The questions were previously piloted using Amazon's Mechanical Turk to ensure they were not too easy without reading the text ($N = 25$) or too difficult after exposure to the text ($N = 25$).

2.3.2. Self-explanation scoring

We needed to provide readers' feedback on their self-explanations in real-time. First, we asked research assistants to create 10 "ideal answers" per question (Guerrero & Wiley, 2019). We chose five of these answers to provide a sufficient number of comparisons and account for differences in writing styles. Next, we used a basic word overlap algorithm where readers' answers were compared to the predefined ideal answers for each question after removing stop words (e.g. "a", "an", "the", "and"). The overlap score was computed as the number of words that overlapped across the submitted answer and the ideal answers, divided by the number of unique words in the ideal answer. Readers' answers often contained synonyms that did not precisely match words in the ideal answers, but were correct in their meaning. Thus, we obtained synonyms for each of the words present in the ideal answers using WordNet (Fellbaum, 1998), a publicly available lexical database that contains groups of words that denote a specific concept (called a synset), and updated word overlap scores to include synonym overlap as well. Word overlap scores were computed for each of the five ideal answers and the highest word overlap score was taken as the reader's self-explanation score.

Finally, two human coders identified critical words and phrases (called keywords), which were essential for a correct answer (e.g., capillary action). These keywords were used to boost the self-explanation score when readers most likely generated an adequate answer, even if the reader's exact wording did not overlap with ideal answers. There were 0 to 9 keywords per question with a median of 5. Two questions did not have designated keywords because no set of words or phrases adequately captured the concept. If a readers' answer contained one of these keywords, their word overlap score was boosted by .2. This number was chosen in order to give readers a relatively small boost without guaranteeing their answer would be counted as correct.

We evaluated this scoring method by correlating automatically generated scores to human ratings (on a scale of 0 [incorrect] to 1 [correct]). We compared a total of 151 self-explanations from 18 pilot participants who interacted with the system in order to examine responses to self-explanation questions during reading. The resultant Pearson correlation of .667 (intraclass correlation = .645) was deemed acceptable based on other attempts to automatically score short-answer, open-ended responses (see Burrows, Gurevych, & Stein, 2015).

2.3.3. *Feedback and re-reading*

The self-explanation scores were used to determine the type of feedback they would receive and if readers should be given the opportunity to re-read. Similar to the mind-wandering detector, we used a sigmoid-like function with three different cutoffs. Self-explanation scores less than .30 always required readers to re-read (re-read cutoff). For scores ranging from .31 to .69, the inverse of readers' word overlap scores was used to probabilistically determine if they would be asked to re-read (probabilistic cutoff). For example, a 40% score yielded a 60% chance of re-reading. Finally, readers with scores ranging greater than .70 were always able to move forward in the text (keep-reading cutoff) but could re-read if they chose to.

One issue with the scoring of self-explanations, which is imperfect, is that giving readers incorrect feedback could be demotivating and might cause negative, reactive behaviors. The feedback mechanism was designed with this in mind. Hedged negative feedback was only given when readers' answers were below the re-read cutoff of .30 using one of the following phrases: "Not entirely! Why don't you take a second look at the last few pages before continuing?" or "Not quite! Why don't you take a second look at the last few pages before continuing?" Similarly, positive feedback was only given when readers' answers exceeded the keep-reading cutoff of .70–1.0 in the form of either, "Good job! Let's continue reading." or "Nicely done! Let's continue reading." Neutral feedback was provided when readers' answers were between .31 and .69 and varied based on whether it was followed by a re-reading intervention. When readers were able to continue reading based on the probability function, they were given the following neutral feedback: "Thanks! Let's go!" or "Thanks! Let's continue reading." However, when readers were asked to re-read based, they were given the following message, "Thanks! Now we want you to give a different answer to the question. Why don't you take a second look at the last few pages before continuing?" We did not give this any explicit feedback given their self-explanation scores were in this intermediate zone.

2.4. *Putting it all together*

The intervention mechanism (depicted in Figure 3) proceeded as follows. An intervention was triggered based on the averaged mind wandering scores computed at the section boundaries and based on the sigmoid-like function in Figure 2. A maximum of 10 (out of 15 section boundaries) interventions could be deployed per reader so as to not be overly burdensome. When an intervention was triggered, the self-explanation prompt and answer box appeared on the right side of the screen and the text they were reading shifted slightly to the left (see Figure 4). Readers were unable to navigate backwards and re-read the text but had to construct a response based on their current understanding (first attempt).

Readers submitted their self-explanation by clicking the submit button and the system automatically scored their response. If their self-explanation score was accurate based on the second sigmoid-like function shown in Figure 2, they received positive feedback and moved on. If their answer was deemed incorrect or insufficient, they received neutral or negative feedback, and were asked to re-read the last few screens (i.e. content corresponding to the self-explanation question) in order to provide an improved answer (second attempt). The answer box disappeared after receiving feedback and they were able to re-read, navigating back and forth using the arrow keys. When they were finished re-reading, they pressed a button on screen labeled "Try Again" in order to construct a new answer. Once they pressed the "Try Again" button, they were unable to navigate through the text. An empty answer box reappeared for them to type in their second self-explanation attempt. Readers received feedback on their second answer using the same feedback algorithm and then progressed to the next screen of text, regardless of the second self-explanation score.

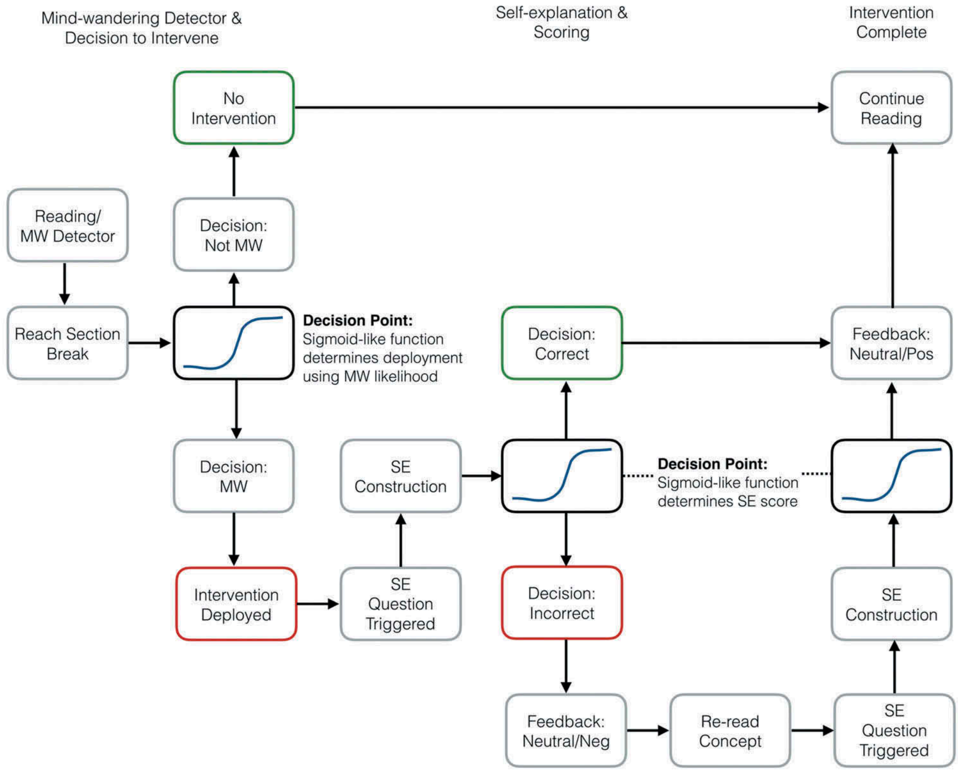

Figure 3. Intervention protocol flowchart.

Figure 4. Screenshots from interface. a) example page from the text during normal reading; b) example intervention page when readers answered self-explanation questions; c) example feedback and re-reading page.

2.5. *User study for formative testing and refinement*

The system was refined through multiple rounds of formative testing and feedback from 15 participants at the same private Midwestern university. An experimenter observed participants while they were interacting with the system and interviewed them afterward about their experience. Their interactions and responses were also analyzed. During this process, we gleaned several insights about intervention design, frequency, and the overall usability of the system. First, after learning from our previous attempt at a fact-based intervention (D'Mello et al., 2017) that students were skimming the text to look for specific answers, it was important to understand reader behavior with the new self-explanation intervention strategy. Our piloting process revealed that readers engaged in less skimming once questions were expanded to target conceptual information that spanned multiple pages of text. We did observe, however, that participants would go back and search for the conceptual answers in a section when they were allowed to navigate back and forth through the pages while answering the self-explanation. This was addressed by freezing the page once a self-explanation question was active. Second, participants requested that questions be integrated with the reading rather than a separate page, which we addressed by freezing the questions on the right-hand side of the screen during re-reading.

We also refined the self-explanation scoring algorithm and the feedback from the data collected. It became clear that the neutral feedback threshold had to be lowered from a previous value of greater than .50 since participants reported that they thought their answers were partially correct even when they received negative feedback. We therefore expanded the window where participants would get positive or neutral feedback to include scores greater than .31 in order to avoid discouraging readers when their answers were partially correct. This step also included adjusting the negative feedback to include more fail-soft language, like "Not quite!" instead of harsher phrases like, "That's incorrect."

3. Summative valdiation study

The summative evaluation was designed to answer three overarching research questions: (1) Can a mind-wandering-sensitive self-explanation intervention correct comprehension deficits in real-time *while participants are reading*? (2) Can this self-explanation intervention prevent the cascading negative effect of mind-wandering on comprehension assessed after reading and a week later?

3.1. *Method*

3.1.1. *Design*

We used a between-subjects yoked-control design with two conditions: Mind-Wandering (MW) Intervention condition (experimental condition), and a Yoked-Control condition (control condition). All participants read the same text excerpt (described above) and completed text- and inference-level comprehension assessments both immediately after reading the text as well as after a one-week delay. Scores on these tests serve as our primary dependent measure.

Participants in the MW-Intervention group received self-explanation interventions as detailed above. Each Yoked-Control participant was paired with a unique MW-Intervention participant and received the exact same self-explanation interventions at the same locations in the text as the MW-Intervention participant, irrespective of whether the Yoked-Control participant was mind-wandering. The only exception was that the Yoked-Control participants were given neutral feedback in the event that they had to re-read after receiving high word overlap scores on the first self-explanation question. This was done in order to ensure that they were still afforded the same re-exposure to the text, without giving them negative feedback when their responses were likely correct.

A yoked-control design allows us to assess whether intervening at the "right time" in the MW condition (i.e. when mind-wandering is occurring) is more effective than intervening in general (i.e.

when mind-wandering might not be occurring). The yoked conditions received the same number of interventions at the same time for purposes of comparing learning. If control participants received interventions at every conceptual break, it would not be a fair comparison since we would expect the self-explanations to have a strong positive effect that would outweigh the benefits of receiving a much smaller number of timely interventions.

A strict control condition is particularly important because the learning strategies used in our intervention make it less meaningful to compare it to a "Do-nothing" condition. First, interpolated testing (i.e. giving intermittent quiz questions during a learning session similar to the intervention) can reduce mind-wandering rates and improve performance on a comprehension test (Jing, Szpunar, & Schacter, 2016; Szpunar et al., 2013). The interventions thus give the readers an opportunity for retrieval practice and to assess their own knowledge, known as the "testing-effect" (Karpicke & Roediger, 2007; Roediger & Karpicke, 2006). Indeed, retrieval practice is one of the most effective study strategies (Eglington & Kang, 2018; Karpicke & Blunt, 2011; Rawson & Dunlosky, 2011; Roediger & Butler, 2011; Smith, Floerke, & Thomas, 2016) It was therefore important that the yoked-control condition gets the same number of self-explanation opportunities as the experimental condition as they are considered a form of interpolated testing with retrieval practice.

Second, self-explanations themselves are one of the most effective study strategies, producing superior learning gains compared to a host of other strategies, including re-exposure and directed instruction (Chi et al., 1989; Kolić-Vrhovec, Bajšanski, & Rončević Zubković, 2011; McNamara, 2004; Roy & Chi, 2005; VanLehn et al., 1992).

Third, a large body of research has shown that memory and comprehension can be improved by event boundaries (i.e. a shift from one event to another, such as spational, causal, or temporal break; Pettijohn, Thompson, Tamplin, Krawietz, & Radvansky, 2016; Radvansky & Zacks, 2017; Swallow, Zacks, & Abrams, 2009). When an event boundary is encountered, the mental model of the text is updated and information before the boundary is committed to longer-term memory storage, resulting in better retention. Our interventions are expected to provide event boundaries during reading, which is likely to improve learning.

Fourth, another benefit of the Yoked-Control condition is that it implicitly controls for time-on-task effects, an important consideration given that mind-wandering is expected to increase as a task progresses (Risko, Anderson, Sarwal, Engelhardt, & Kingstone, 2012; Thomson, Seli, Besner, & Smilek, 2014).

Finally, a Yoked-Control is an active control, which also addresses "placebo effects." One advantage of a "Do-nothing" control is that it is devoid of interruptions compared to the current conditions. Indeed, interruptions can be harmful when unrelated to the learning task (e.g., a fire alarm). However, in the context of education, uninterrupted silent reading is considered to be a passive learning strategy that is less effective compared to more constructive approaches like self-explanations (Chi & Wylie, 2014). Thus, even though the intervention interrupts the reading process, we consider this to be a beneficial interruption based on many previous studies where interruptive self-explanations led to better comprehension compared to other learning reading conditions without interruptions (e.g., re-reading; Allen, McNamara, & McCrudden, 2015; McNamara, Levinstein, & Boonthum, 2004; Ozuru, Briner, Best, & McNamara, 2010).

With this in mind, we would expect a "Do-nothing" control condition, where participants would not receive any interventions, to perform worse than both of the current conditions, which receive interventions that capitalize on a host of well-known learning strategies. Thus, a "Do-nothing" condition would yield an unfair comparison; any advantage in the MW Intervention condition could then be attributed to interpolated testing, retrieval practice, self-explanations, event-boundaries, time-on-task, and placebo effects, rather than the self-explanation interventions that specifically target mind-wandering when it occurs.

3.1.2. *Hypotheses*

Our first hypothesis was that the self-explanation and re-reading interventions would attenuate the effect of mind-wandering on comprehension *during* reading. Because participants in the MW-

Intervention condition received interventions when their comprehension was expected to be lower (i.e., because they were mind-wandering), we expected that their self-explanation quality would improve on their second attempt (i.e., after re-reading) compared to their first attempt. Participants in the Yoked-Control condition were expected to exhibit no differences in self-explanation scores across attempts since the timing of the interventions was not contingent on their mind-wandering.

Second, we hypothesized that the MW-Intervention condition would have better overall posttest comprehension measures compared to the Yoked-Control condition, both immediately after reading and after a week-long delay scores at posttest, assuming that comprehension deficits due to mind-wandering were successfully repaired during reading. Although both conditions were likely to benefit from simply receiving the learning interventions, the control condition did not receive interventions timed to their mind-wandering episodes. They were therefore expected to have lower comprehension scores overall because their mental models were not "repaired" in real-time as they progressed through the text.

Third, we hypothesized that the inference-level assessments would be more sensitive to the intervention than the textbase-level comprehension assessments because the self-explanation interventions were aimed at promoting *deep* levels of comprehension. For the same reason, we expected the delayed assessment to be more sensitive than the immediate assessment after allowing for a consolidation period which can have a considerable impact on memory (Ellenbogen et al., 2007; Kuhlmann & Wolf, 2006). Specifically, the consolidation period (i.e. neural processes contributing to the "permanent" storage of initially encoded information) is critical to the formation of long-term memories (Nadel & Moscovitch, 1997), and thus provides us with the opportunity to assess whether participants' knowledge of the text has been integrated into long-term memory.

3.1.3. *Participants*

We collected data from 113 participants enrolled in two universities. Fifty-four undergraduates from a private mid-western university participated for two research credits and were asked to return one week later to complete the delayed tests for additional .5 research credit. Accumulating a minimum number of research credits is typically a requirement for their courses. Out of the 54 participants, 36 of them (67%) opted to come back. We also collected data from 59 participants enrolled at a large public western university. Participants were given a $20 Amazon gift card for completing the first portion of the study and were offered an additional $10 gift card for completing the delayed test one week later. To improve retention rates, we administered the delayed assessment using online surveys in Qualtrics and 55 (out of 59) participants completed the delayed test (93%). The differences in compensation schemes were based on practical constraints associated with each university and are addressed as covariates in the statistical analyses. Four participants (3.6%) in the MW-Intervention condition did not receive any interventions and were excluded from the analyses (along with their yoked-counterparts) due to failed treatment in the experimental condition. In total, we analyzed data from 35 pairs of participants ($n = 70$) across both cohorts who completed both sessions and received at least one intervention.

Participants signed an informed consent prior to participating. The consent indicated they could withdraw from the research at any point without penalty. The study was approved by the Institutional Review Boards at each university, respectively. Participants were blind to condition.

Participants who responded to a demographic questionnaire ($n = 69$) were an average age of 21.09 years ($SD = 2.90$). There were 57.9% females and a majority were Caucasian (68.1%), with 23.1% Asian, 5.8% Hispanic, 1.5% African American, and 1.5% who reported as "Other."

3.1.4. *Materials*

3.1.4.1. *Eye-tracker.* A Tobii TX 300 eye tracker was used to collect eye gaze data in real-time (sampling rate was set to 120 Hz in binocular mode). Participants' eye gaze was calibrated at the beginning of the study (see Procedure) and was not monitored or updated during the task. An experimenter examined calibration quality before the reading task by assessing the alignment between gaze-points and predetermined areas of interest that mirrored text placement.

Participants were not required to have uncorrected vision as the eye-trackers could track eye movements even when participants wore glasses or contact lenses.

3.1.4.2. *Knowledge assessments.*

We did not include a pretest because participants were not expected to have much (if any) knowledge of this specific text, which we confirmed in pilot testing. Instead, participants completed one set of knowledge assessments immediately after the reading task and a second set approximately a week after. Each set included textbase-level and inference-level comprehension questions.

Textbase-level questions required factual knowledge about the text to be answered correctly (see Table 1 for example questions). There were four unique questions per conceptual section of the text, resulting in 60 questions in total (corresponding to 15 sections). Each question contained four answer options consisting of a target (the correct response to the question), a near-miss (an option that sounds correct but was not), a thematic miss (an option that follows the theme of the content but is not actually related to the question), and a distractor (an option that is not at all related). The questions were split into two subsets of 30 questions in a quasi-random fashion in order to create an even number of questions per concept. Each subset was either presented in the immediate or delayed version of the assessments (subset assignment was counterbalanced across participants).

Inference-level questions targeted conceptual knowledge of the text (see Table 1 for an example). They required participants to generate inferences or apply knowledge to answer the question correctly (Graesser, Ozuru, & Sullins, 2010; Graesser & Person, 1994). We adopted the same four-answer multiple choice structure, where correct responses required conceptual knowledge of a specific section in the text (30 total questions – 2 per section). Similar to the textbase-level questions, items were matched by section and divided into two subsets, which were then assigned for the immediate or the delayed assessment (subset assignment was counterbalanced across participants).

The questions were piloted using Amazon's Mechanical Turk to ensure they were not too easy or difficult to answer ($N = 50$). This was done by first asking participants to answer the questions without reading the text. Here, accuracy was expected to be around chance levels (25%). We modified the questions by analyzing participants' answers. For example, some multiple-choice options contained obvious incorrect cues and were therefore never selected while some incorrect options were selected too often. Questions were modified and re-tested until all questions were answered at chance levels without reading the text. In a second step, we also needed to ensure that participants would be able to answer the questions correctly after reading the text. Therefore, a different group of pilot participants answered questions after reading the corresponding conceptual section of text. Participants were able to answer all questions with above chance levels of accuracy after reading.

3.1.4.3. *Procedure.*

Participants were tested individually over a 1.5-hour session (plus a follow up session either in person or remotely) in a small testing room. Participants were seated in front of a 511 × 287 mm monitor with the eye tracker affixed at the base. They were positioned in a non-adjustable chair two feet from the monitor. We used the eye tracker's default calibration protocol, which required participants to focus on a series of five yellow dots that individually appeared in five different locations on the screen. To test calibration quality, participants were then asked to focus on a set of numbers which appeared at different locations on the screen. The experimenter checked calibration accuracy by comparing the recorded location of eye gaze relative to the location of the numbers on the screen. If the calibration was deemed accurate, the participant was allowed to move on, otherwise the calibration and checking process was repeated up to three times before allowing the participant to move on.

Next, participants received the following electronic instructions about the reading task:

> *"Your primary task is to read the text in order to take a short test after reading. The text will be displayed screen by screen and you can press the "right arrow" key to navigate through the text and the "left arrow" key to return to the previous page."*

Then, participants were given notice of the intervention questions, but they were not instructed on why, how, or when the questions would appear.

> *"While reading the text, you will occasionally be asked some questions about the pages that you just read. Depending on your answer, you will have the chance to re-read and answer again. You may have to review the last few pages that you have read. Remember, you can use the left arrow key to review what you have read."*

The reading session commenced after the instructions were delivered and the experimenter closed the door and left the testing room. The texts were displayed on a 23-inch monitor using size 36 Courier New font. There were approximately 115 words per screen (for 57 screens) with 1920 × 1080 resolution. The MW-Intervention group and the Yoked-Control group received the self-explanation interventions as described above. Participants read the text at their own pace and were instructed to notify the experimenter when they were done.

Participants completed the *immediate* knowledge assessments, consisting of the 30 textbase-level comprehension questions followed by the 15 inference-level questions, after the reading session. They then completed a brief demographics questionnaire.

Participants at the Midwestern university were given the opportunity to return for a follow-up session to complete the delayed posttest, which was scheduled for a week later using an anonymized calendar. At the western university, participants completed the delayed posttest one week after their initial session via Qualtrics. The *delayed* assessment consisted of 30 and 15 previously unseen textbase- and inference-level questions, respectively. Finally, participants were fully debriefed.

3.2. *Results*

We analyzed data from the 35 pairs (N = 70) who had complete data (i.e., scores for both the immediate and delayed tests) in order to keep the analytic approach consistent. The two conditions were compared using mixed-effects regression models in R through the lme4 package (Bates, Mächler, Bolker, & Walker, 2015). Mixed-effects models are the appropriate method of analysis for these data given the repeated (multiple subjects and interventions) and nested (participants yoked together) structure of the data (Finch, Bolin, & Kelley, 2016; West, Welch, & Galecki, 2014). A mixed-effect approach allows for the inclusion of each individual observation in the regression model without violating the independence assumption required for an analysis of variance (ANOVA).

The participants' pair number was used as a random effect in all models so that pair number was treated as a within-subject effect. This is the recommended approach according to Schweigert (1994) since the yoked design inherently relates the two participants as a result of receiving the same number of interventions at the same points during reading (i.e. they should not be considered independent). The Yoked-Control condition was set as the reference group, so positive regression coefficients indicate higher values for the MW-Intervention condition (MW > YC) and vice versa for negative values (YC > MW). Significance tests were conducted with a two-tailed α of .05 using the Type II Wald chi-square test from the 'car' package in R (Fox, Friendly, & Weisberg, 2013). At the same time, we acknowledge the limitations of p-value calculations in mixed-effects models (Luke, 2017) and encourage readers to consider the standard error (SE) of the regression coefficients, the 95% confidence intervals and effect sizes (Cohen's d) in addition to traditional significance tests. We also report Hedge's g for comparison given our small sample size (see Table 2).

Mind-wandering likelihood and self-explanation (word overlap) scores were analyzed at the concept-level using mixed effects linear regression. Text and inference-level comprehension scores were analyzed at the item level, such that binary responses (1 for correct; 0 for incorrect) were regressed on condition using mixed effects logistic regression. We include cohort (since the data was collected at two sites) as a covariate in all models but it was never significant so it was removed from the models reported here (see Table 2 for regression results).

A correlation matrix with an overview of the key variables is presented in an Appendix.

Table 2. Means, standard deviations, and effect sizes (Cohen's d and Hedge's g) assessed at the participant level. Reading Time and Reread Time were calculated based on the average time spent on each section. SE = self-explanation. 19% of data occurred before the first intervention and was removed for the "After an intervention occurred" analyses.

	MW		Yoked Control	MW-Intervention vs. Yoked-Control		
N pairs = 35	*N obs.*	*M (SD)*	*M (SD)*	*Cohen's d*	*Hedge's g*	*p-value* for regression models
All Concepts						
Reading Time (seconds)	1050	1.76 (.495)	1.71 (.458)	.103	.102	.272
Reread Time (seconds)	450	1.07 (.499)	1.13 (.478)	−.123	−.121	.314
Time Constructing First SE (ms)	352	10547 (65666)	25941 (21586)	−.315	−.311	.011
First Self-Explanation Score	352	.498 (.160)	.485 (.183)	.076	.075	.451
Second Self-Explanation Score	180	.467 (.141)	.463 (.183)	.024	.024	.675
Δ Self-Explanation Score	180	.143 (.150)	.066 (.144)	.524	.518	.002
Immediate						
Textbase Level	2100	.687 (.137)	.698 (.147)	−.077	−.077	.560
Inference Level	1050	.472 (.156)	.476 (.193)	−.023	−.023	.899
Delayed						
Textbase Level	2100	.559 (.147)	.529 (.162)	.194	.192	.152
Inference Level	1050	.436 (.165)	.394 (.154)	.263	.260	.162
After an intervention occurred						
Immediate						
Textbase Level	1698	.706 (.158)	.731 (.159)	−.158	−.156	.354
Inference Level	850	.484 (.194)	.471 (.230)	.061	.060	.944
Delayed						
Textbase Level	1698	.582 (.170)	.522 (.171)	.352	.348	.031
Inference Level	850	.449 (.200)	.389 (.191)	.307	.303	.040

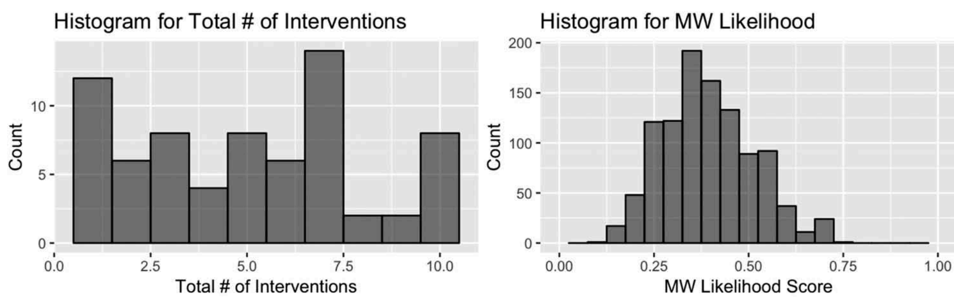

Figure 5. Histograms of total number of interventions received and mind-wandering (MW) likelihood scores across both conditions.

3.2.1. *Number of interventions*

In total, participants received 352 self-explanation interventions, with an average of 5.03 interventions each (SD = 2.96) across the 15 possible intervention concepts. This intervention rate of 33.5% is consistent with self-reported mind-wandering rates observed while reading the same text (Kopp et al., 2015). See Figure 5 for histograms of the number of interventions received and the MW Likelihood scores across all participants.

3.2.2. *Reading times*

We tested for differences in reading times in order to address a potential time-on-task confound. There were no significant differences in reading times between the two conditions during the first read (i.e. not including re-reading during self-explanations), B = 2858ms (SE = 2599 [95% CI:-2236, 7953ms]), $\chi^2(1)$ = 1.21, p = .272. The two conditions also spent similar times re-reading the text when they were asked to provide a second answer to the self-explanation question, B = −6868ms (SE = 6815 [−20218; 6591ms]), $\chi^2(1) = 1.02$, p = .314.

3.2.3. *Mind wandering likelihoods*

Participants' first attempt self-explanation scores were not associated with their mind-wandering likelihood on intervention concepts, B = −.012 (SE = .022 [−.055, .031]), $\chi^2(1) = .280$, p = .597. A separate regression revealed the same pattern for their second attempt self-explanation scores, B = −.008 (SE = .040 [95% CI: −.085, .070]), $\chi^2(1) = .037$, p = .847.

We also found that mind-wandering was not reliably related to immediate, B = −.007 (SE = .007 [−.020, .006], $\chi^2(1) = 1.17$, p = .279, or delayed textbase comprehension, B = −.009 (SE = .006 [−.020, .003]), $\chi^2(1) = 2.53$, p = .133; it was also unreliably related to immediate, B = −.012 (SE = .007 [−.025, .001], $\chi^2(1) = 3.17$, p = .075, or delayed, B = .001 (SE = .007 [−.012, .014]), $\chi^2(1) = .014$, p = .905, inference-level comprehension. These results may be attributed to the fact that, as intended, the self-explanation intervention mitigated the negative relationship between mind-wandering and comprehension by correcting comprehension deficits after re-reading (i.e., after the first self-explanation attempt).

3.2.4. *Self-explanation scores*

Participants in the Yoked-Control condition spent longer time constructing their first self-explanation compared to the MW-Intervention condition, B = −14296 ms (SE = 5608 [3285, 25305]), $\chi^2(1) = 6.50$, p = .011. Despite spending longer on their response, however, there was no effect of condition on participants' first, B = −.021 (SE = .023 [−.076, .034]), $\chi^2(1) = .569$, p = .451, or second self-explanation attempt scores, B = −.014 (SE = .033 [−.077, .050]), $\chi^2(1) = .176$, p = .675.

We also assessed if participants in the MW-Intervention condition showed greater improvement in their self-explanation scores compared to the Yoked-Control condition after re-reading. This was the case as the intervention-group's scores exhibited greater change from their first attempt to the second attempt (Second SE Score − First SE Score), indicating they may have benefited more from the opportunity to re-read, B = .106 (SE = .034 [.040; .172]), $\chi^2(1) = 9.86$, p = .002.

Taken together, the intervention appears to correct for comprehension deficits associated with mind wandering *during* reading. The pertinent question now, is whether this is reflected in comprehension assessed *after* reading and at *delay*.

3.2.5. *Comprehension assessments*

There were no significant condition differences on any of the four comprehension assessments when including data from all concepts, $ps > .15$, $ds < .2$ and 95% CI containing 0 (see Table 2). However, until the MW-Intervention condition received their first adaptive intervention, the two conditions had identical reading experiences. Therefore, we can assume that there may not be a difference between the two conditions before an intervention has even been deployed (i.e., when no manipulation has taken effect). To address this issue, we repeated the analyses after removing 19% of the items corresponding to concepts that preceded the first intervention. Items were removed for each pair separately because pairs could receive their first intervention at different places in the text. We still found no differences in textbase, B = −.101 (SE = .109 [−.314, .113]), $\chi^2(1) = .859$, p = .354, or inference-level comprehension immediately after reading, B = .010 (SE = .141 [−.267, .287]), $\chi^2(1) = .005$, p = .944. Critically, however, the MW-Intervention condition scored higher on both the textbase, B = .216 (SE = .100 [.020, .413]), $\chi^2(1) = 4.64$, p = .031 and inference assessment items, B = .293 (SE = .142 [.014, .573]), $\chi^2(1) = 4.23$, p = .040, after the week-long delay. The differences corresponded to .352 (textbase) and .307 (inference) sigma effect.

4. General discussion

The mind will inevitably wander during reading (Feng et al., 2013; Franklin et al., 2011; Phillips et al., 2016; Smallwood, 2011). Despite the ubiquity of this experience, mind-wandering can be quite problematic considering its consistent negative relationship with reading comprehension (Mills et al., 2017; Randall et al., 2014; Smallwood, 2011; Smallwood et al., 2007). With this in mind, a growing body of research has been devoted to preventing mind-wandering from occurring during learning (Kopp et al., 2014; Szpunar, 2017; Szpunar et al., 2013; Zanesco et al., 2016). However, few interventions have been designed to improve comprehension by responding to mind-wandering as it inevitably arises. Further, although it is possible to detect episodes of mind-wandering during reading with a modicum of accuracy (Bixler & D'Mello, 2016; Faber et al., 2017), with one exception of our own work (D'Mello et al., 2017), researchers have yet to leverage these detectors to create an effective learning intervention.

To address these gaps, we developed and tested the Eye-Mind Reader, a real-time gaze-based detection and intervention interface aimed to promote comprehension outcomes by combating the negative effects of mind-wandering during reading. After iterative refinement of individual system components and then the complete system itself, we tested the system in a summative yoked-randomized controlled study. An experimental condition (MW-Intervention condition) received automated self-explanation/re-reading interventions based on the classifications of an eye-gaze-based mind wandering detector. We compared this to a Yoked-Control condition that received interventions at the same points in the text regardless of their mind wandering likelihoods. Here, we discuss our main findings along with limitations and future directions.

4.1. *Main findings*

In line with our predictions, participants in the MW-Intervention condition showed greater improvement in their self-explanation scores in comparison to the Yoked-Control condition, providing evidence that the intervention was sensitive to moments when there was room to improve "online" comprehension (e.g., during the reading process). Repairing online comprehension may be the minimum viable outcome of our system.

Beyond improvements in online processing, we also tested whether a mind-wandering sensitive intervention led to differences in comprehension scores immediately after reading and after a one-week delay. We hypothesized that a reactive self-explanation intervention would improve comprehension in the MW-Intervention condition by allowing participants to fill in "gaps" in their mental model associated with mind wandering. This may help avoid a cascading negative effect of impaired comprehension for readers in the MW-Intervention condition, compared to those in the Yoked-Control condition who were not provided interventions that were sensitive to their mind-wandering episodes.

Although there were no significant differences in comprehension immediately after reading (i.e., during the same session), participants in the MW-Intervention condition outperformed the Yoked-Control condition after a week delay. The difference in conditions on the delayed tests aligns with Smallwood's (2011) prediction that mind-wandering influences comprehension across multiple levels. Here, repairing deficits in comprehension as they arise during reading helped to promote a long-term, deeper-level understanding of the text – consistent with the idea that a consolidation period often supports better memories and deeper levels of conceptual understanding (Ellenbogen et al., 2007; Nadel & Moscovitch, 1997).

4.2. *Limitations and future directions*

It is important to note the limitations of the current research program. First, participant attrition was an issue for the follow-up delayed tests. Using an in-person follow-up test may have exacerbated this issue in our Midwestern university sample as fewer participants returned for the test in person. Participant attrition does not affect the conclusions drawn here due to the yoked design, but future replications should consider recruiting a larger sample as well as alternative mechanisms or incentives for maximizing retention across sessions.

Second, this study only examined students reading one text in a single session. The chosen text was written a century ago and focused on material that was likely to be unfamiliar to any of our participants. Thus, it is possible that this particular text elicited specific reading behaviors that would not generalize to contemporary texts, or to texts that were of high interest to participants. Participants' reading behaviors may vary across multiple sessions and contexts. The effectiveness of our approach should therefore be evaluated using more varied texts and learning contexts. It is an open question as to how a similar approach might work across multiple sessions, in particular when participants were asked to read long narrative texts that require a strong situation model to maintain comprehension, for example, a historic novel? Although the detector used in this study has not been tested for generalization to different texts, there is some reason to expect it will generalize given the content-independent nature of the features used in the model. There is also growing list of mind-wandering detectors that were validated using different texts (Bixler, Blanchard, Garrison, & D'Mello, 2015; Franklin et al., 2011; Mills & D'Mello, 2015; Nguyen, Binder, Nemier, & Ardoin, 2014) and in other contexts (i.e. film viewing, online lecture viewing, and driving; Baldwin et al., 2017; Hutt et al., 2019; Mills et al., 2016; Pham & Wang, 2015; Zhao, Lofi, & Hauff, 2017); such detectors could also be adopted for testing future real-time interventions.

A third limitation relates to the scalability of this intervention outside of the laboratory as the eye-tracking technology used in this study was of research-grade and expensive. Fortunately, cost-effective eye-tracking methods are increasingly available (e.g., Tobii 4C). As reliable mind-wandering detectors are built using more scalable technologies, follow-up studies can be done to address scalability concerns. For example, a recent line of work has helped assuage concerns about scalability and ecological validity by using cost-effective eye trackers to detect mind wandering in a high school classroom (Hutt et al., 2017, 2016). The next step is to test out similar interventions with scalable eye tracking in these more ecological contexts.

Finally, the present intervention approach serves as an initial exploration on how to address mind wandering with only small to medium effect sizes (Cohen's *d*s = .352 and .307; Cohen, 1992). Promising next steps in this line of research might include explorations of different intervention strategies coupled with improved detection methods. Future work on this topic can help to further understand and refine the attentional mechanisms influencing comprehension.

4.3. *Conclusions*

The goal of this research was to test Eye-Mind Reader, an intelligent closed-loop interface designed to detect and combat the cascading negative effects of mind-wandering on reading comprehension. We accomplished two major research goals: (1) developing a self-explanation /re-reading intervention that aimed to promote deep processing of texts while addressing mind wandering, and (2) testing the intervention in a yoked-control experiment. Results indicate that the real-time intervention helped correct comprehension deficiencies *during reading* by targeting episodes of mind wandering. Furthermore, the results suggest that mind-wandering-sensitive self-explanation/re-reading interventions indeed promoted better long-term comprehension at both the text-base and inference-levels compared to a control group that received the exact same interventions but independent of mind wandering. Thus, our results highlight the advantage of closed-loop interfaces that improve learning by detecting and combating wandering minds.

Funding

This research was supported by the National Science Foundation (NSF) [ITR 0325428, HCC 0834847, DRL 1235958]. Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NSF.

Notes on contributors

Caitlin Mills (caitlin.mills@unh.edu) is a psychologist with an interest in the correlates and consequences of mind-wandering and affect during complex learning. She is an Assistant Professor in the Department of Psychology at the University of New Hampshire.

Julie Gregg (julie.gregg@colorado.edu) is a psychologist who utilizes machine learning to predict reading comprehension from eye-movement data. She is a data scientist in Brooklyn, NY.

Robert Bixler (rbixler@nd.edu) is a computer scientist who specializes in predicting mind-wandering using physiological signals such as eye-gaze during reading and film comprehension tasks. He is a Ph.D. candidate in the Computer Science department at the University of Notre Dame.

Sidney K. D'Mello (sidney.dmello@colorado.edu) conducts research in the fields of affective and attentional computing, multimodal interaction, speech and discourse processing, and intelligent learning environment. His team conducts basic research on affective and cognitive states (e.g., confusion, boredom, mind wandering) across a range of interaction contexts, develops real-time computational models of these states, and integrates the models in affect- and attention-aware intelligent technologies. He is an Associate Professor at the Institute of Cognitive Science at the University of Colorado Boulder.

ORCID

Caitlin Mills http://orcid.org/0000-0003-4498-0496

References

Allen, L. K., McNamara, D. S., & McCrudden, M. (2015). *Change your mind: investigating the effects of self-explanation in the resolution of misconceptions*. California:Proceedings of Cognitive Science Society. Pasadena.

Baldwin, C. L., Roberts, D. M., Barragan, D., Lee, J. D., Lerner, N., & Higgins, J. S. (2017). Detecting and quantifying mind wandering during simulated driving. *Frontiers in Human Neuroscience*, *11*, 406. doi:10.3389/fnhum.2017.00406

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal Of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01

Bixler, R., & D'Mello, S. (2014). *Toward fully automated person-independent detection of mind wandering*. In V. Dimitrova, T. Kuflik, D. Chin, F. Ricci, P. Dolog, & G.-J. Houben (Eds.), Proceedings of the 22nd international conference on User Modeling, Adaptation, and Personalization (pp. 37–48). Cham: Springer International Publishing.

Bixler, R., Blanchard, N., Garrison, L., & D'Mello, S. K. (2015). Automatic detection of mind wandering during reading using gaze and physiology. Proceedings of the 2015 ACM on international conference on Multimodal Interaction, 299–306. Seattle, Washington doi:10.1145/2818346

Bixler, R., & D'Mello, S. (2016). Automatic gaze-based user-independent detection of mind wandering during computerized reading. *User Modeling and User-adapted Interaction*, *26*(1), 33–68. doi:10.1007/s11257-015-9167-1

Bixler, R., & D'Mello, S. K. (2016). Automatic gaze-based user-independent detection of mind wandering during computerized reading. *User Modeling and User-adapted Interaction*, *26*, 1–36. doi:10.1007/s11257-015-9167-1

Bixler, R., Kopp, K., & D'Mello, S. K. (2014). Evaluation of a Personalized Method for Proactive Mind Wandering Reduction. 4th International Workshop on Personalization Approaches in Learning Environments, 33–41

Blanchard, N., Bixler, R., Joyce, T., & D'Mello, S. K. (2014). *Automated physiological-based detection of mind wandering during learning*. Intelligent Tutoring Systems, 55–60. Springer.

Brown, K. W., & Ryan, R. M. (2003). The benefits of being present: Mindfulness and its role in psychological well-being. *Journal of Personality and Social Psychology*, *84*(4), 822. doi:10.1037/0022-3514.84.4.822

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal Of Artificial Intelligence in Education*, *25*(1), 60–117. doi:10.1007/s40593-014-0026-8

Callender, A. A., & McDaniel, M. A. (2009). The limited benefits of rereading educational texts. *Contemporary Educational Psychology*, *34*(1), 30–41. doi:10.1016/j.cedpsych.2008.07.001

Chi, M., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219–243. doi:10.1080/00461520.2014.965823

Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*(2), 145–182. doi:10.1207/s15516709cog1302_1

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155. doi:10.1037/0033-2909.112.1.155

D'Mello, S. K., Mills, C., Bixler, R., & Bosch, N. (2017). Zone out no more: mitigating mind wandering during computerized reading. Educational Data Mining 2017. EDM, Wuhan, China

D'Mello, S. K. (2018). What do we think about when we learn? In K. Millis, J. Magliano, D. L. Long, & K. Weimer (Eds.), *Understanding deep learning, educational technologies and deep learning, and assessing deep learning*. Routledge/Taylor and Francis.

D'Mello, S. K., Dieterle, E., & Duckworth, A. (2017). Advanced, Analytic, Automated (AAA) measurement of engagement during learning. *Educational Psychologist*, *52*(2), 104–123. doi:10.1080/00461520.2017.1281747

D'Mello, S. K., Olney, A., Williams, C., & Hays, P. (2012). Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of Human-computer Studies*, *70*(5), 377–398. doi:10.1016/j.ijhcs.2012.01.004

Davidson, R. J., Kabat-Zinn, J., Schumacher, J., Rosenkranz, M., Muller, D., Santorelli, S. F., … Sheridan, J. F. (2003). Alterations in brain and immune function produced by mindfulness meditation. *Psychosomatic Medicine*, *65*(4), 564–570. doi:10.1097/01.PSY.0000077505.67574.E3

Drummond, J., & Litman, D. J. (2010). In the Zone: Towards Detecting Student Zoning Out Using Supervised Machine Learning. Intelligent Tutoring Systems: 12th International Conference. Intelligent Tutoring Systems, Pittsburgh, Pennsylvania. doi:10.1007/978-3-642-13437-1_53

Eglington, L. G., & Kang, S. H. (2018). Retrieval practice benefits deductive inference. *Educational Psychology Review*, *30*(1), 215–228. doi:10.1007/s10648-016-9386-y

Ellenbogen, J. M., Hu, P. T., Payne, J. D., Titone, D., & Walker, M. P. (2007). Human relational memory requires time and sleep. *Proceedings of the National Academy of Sciences*, *104*(18), 7723–7728. doi:10.1073/pnas.0700094104

Faber, M., Bixler, R., & D'Mello, S. K. (2017). An automated behavioral measure of mind wandering during computerized reading. *Behavior Research Methods*, 1–17. doi:10.3758/s13428-017-0857-y

Faber, M., Bixler, R., & D'Mello, S. K. (2018). An automated behavioral measure of mind wandering during computerized reading. *Behavior Research Methods*, *50*(1), 134–150. doi:10.3758/s13428-017-0857-y

Feit, S., Scheufler, N. A., Miller, R. C., Kosterich, M., Lin, J., Choi, D. J., & Grebe, C. D. (2017). Predictive human-machine interface using eye gaze technology, blind spot indicators and driver experience (United States Patent No. US10336323B2). https://patents.google.com/patent/US10336323B2/en

Fellbaum, C. (1998). A semantic network of english: the mother of all wordnets. *Computers and The Humanities*, *32*(2/3), 209–220. JSTOR.

Feng, S., D'Mello, S., & Graesser, A. C. (2013). Mind wandering while reading easy and difficult texts. *Psychonomic Bulletin & Review*, *20*(3), 1–7. doi:10.3758/s13423-012-0367-y

Finch, W. H., Bolin, J. E., & Kelley, K. (2016). *Multilevel modeling using r*. Boca Raton, FL: CRC Press.

Foulsham, T., Farley, J., & Kingstone, A. (2013). Mind wandering in sentence reading: Decoupling the link between mind and eye. *Canadian Journal of Experimental Psychology/Revue Canadienne De Psychologie Expérimentale*, *67*(1), 51–59. doi:10.1037/a0030217

Fox, J., Friendly, M., & Weisberg, S. (2013). Hypothesis tests for multivariate linear models using the car package. *The R Journal*, *5*(1), 39–52. doi:10.32614/RJ-2013-004

Franklin, M. S., Mooneyham, B. W., Baird, B., & Schooler, J. W. (2013). Thinking one thing, saying another: The behavioral correlates of mind-wandering while reading aloud. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-013-0468-2

Franklin, M. S., Smallwood, J., & Schooler, J. W. (2011). Catching the mind in flight: Using behavioral indices to detect mindless reading in real time. *Psychonomic Bulletin & Review*, *18*(5), 992–997. doi:10.3758/s13423-011-0109-6

Graesser, A., Ozuru, Y., & Sullins, J. (2010). What is a good question? In Bringing reading research to life (pp. 112–141). New York, NY: Guilford Press

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring: american educational research journal, 31(1), 104–137. *American Educational Research Journal*, *10 10 31*, 3102/00028312031001104. doi:10.3102/00028312031001104

Grossman, P., Niemann, L., Schmidt, S., & Walach, H. (2004). Mindfulness-based stress reduction and health benefits: A meta-analysis. *Journal of Psychosomatic Research*, *57*(1), 35–43. doi:10.1016/S0022-3999(03)00573-7

Guerrero, T. A., & Wiley, J. (2019). *Using "idealized peers" for automated evaluation of student understanding in an introductory psychology course*. International conference On Artificial Intelligence In Education, 133–143. Springer. doi:10.3917/jibes.301.0133

Hutt, S., Hardey, J., Bixler, R., Stewart, A., Risko, E., & D'Mello, S. K. (2017). Gaze-based detection of mind wandering during lecture viewing. 10th international conference on Educational Data Mining, 226–231. Wuhan, China

Hutt, S., Krasich, K., Mills, C., Bosch, N., White, S., & Brockmole, J. R., & D'Mello, S. K. (2019). Automated gaze-based mind wandering detection during computerized learning in classrooms. User Modeling and User-Adapted Interaction, *29*(4), 821–867

Hutt, S., Krasich, K., Mills, C., Bosch, P. N., White, S., & Brockmole, J. R., & D'Mello, S. K. (2019). Automated gaze-based mind wandering detection during computerized learning in classrooms. User Modeling and User-Adapted Interaction, *29*(4), 821–867. doi:10.1007/s11257-019-09228–5

Hutt, S., Mills, C., Bosch, N., Krasich, K., Brockmole, J., & D'Mello, S. (2017). "out of the fr-eye-ing pan": towards gaze-based models of attention during learning with technology in the classroom. Proceedings of the 25th conference on User Modeling, Adaptation and Personalization, 94–103. Bratislava, Slovakia. doi:10.1145/3079628

Hutt, S., Mills, C., White, S., & Donnelly, P. J., & D'Mello, S. K. (2016). The eyes have it: gaze-based detection of mind wandering during learning with an intelligent tutoring system. Proceedings of the 9th international conference on Educational Data Mining, International Educational Data Mining Society, 86–93. Raleigh, North Carolina. EDM. doi:10.1177/1753193415617756

Jha, A. P., Krompinger, J., & Baime, M. J. (2007). Mindfulness training modifies subsystems of attention. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(2), 109–119. doi:10.3758/CABN.7.2.109

Jing, H. G., Szpunar, K. K., & Schacter, D. L. (2016). Interpolated testing influences focused attention and improves integration of information during a video-recorded lecture. *Journal of Experimental Psychology: Applied*, *22*(3), 305–318. doi:10.1037/xap0000087

Ju, Y.-J., & Lien, Y.-W. (2016). Better control with less effort: The advantage of using focused-breathing strategy over focused-distraction strategy on thought suppression. *Consciousness and Cognition*, *40*, 9–16. doi:10.1016/j.concog.2015.12.002

Kam, J. W., Dao, E., Farley, J., Fitzpatrick, K., Smallwood, J., Schooler, J. W., & Handy, T. C. (2011). Slow fluctuations in attentional control of sensory cortex. *Journal of Cognitive Neuroscience*, *23*(2), 460–470. doi:10.1162/jocn.2010.21443

Kane, M. J., Smeekens, B. A., von Bastian, C. C., Lurquin, J. H., Carruth, N. P., & Miyake, A. (2017). A combined experimental and individual-differences investigation into mind wandering during a video lecture. *Journal of Experimental Psychology: General*, *146*(11), 1649–1674. doi:10.1037/xge0000362

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*(6018), 772–775. doi:10.1126/science.1199327

Karpicke, J. D., & Roediger, H. L., III. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*(2), 151–162. doi:10.1016/j.jml.2006.09.004

Killingsworth, M. A., & Gilbert, D. T. (2010). A wandering mind is an unhappy mind. *Science*, *330*(6006), 932. doi:10.1126/science.1192439

Klinger, E., & Cox, W. M. (1987). Dimensions of thought flow in everyday life. *Imagination, Cognition and Personality*, *7*(2), 105–128. doi:10.2190/7K24-G343-MTQW-115V

Kolić-Vrhovec, S., Bajšanski, I., & Rončević Zubković, B. (2011). The role of reading strategies in scientific text comprehension and academic achievement of university students. *Review of Psychology*, *18*(2), 81–90.

Kopp K., Bixler R., D'Mello S. (2014) Identifying Learning Conditions that Minimize Mind Wandering by Modeling Individual Attributes. In: Trausan-Matu S., Boyer K.E., Crosby M., Panourgia K. (eds) Intelligent Tutoring Systems. ITS 2014. Lecture Notes in Computer Science, vol 8474. Springer, Cham.

Kopp, K., D'Mello, S., & Mills, C. (2015). Influencing the occurrence of mind wandering while reading. *Consciousness and Cognition: an International Journal*, *34*, 52–62. doi:10.1016/j.concog.2015.03.003

Krasich, K., Hutt, S., Mills, C., Spann, C. A., Brockmole, J. R., & D'Mello, S. K. (2018). *"Mind" ts: testing a brief mindfulness intervention with an intelligent tutoring system*. International conference on Artificial Intelligence in Education, 176–181. Springer. doi:10.1142/S2424835518500170

Kuhlmann, S., & Wolf, O. T. (2006). Arousal and cortisol interact in modulating memory consolidation in healthy young men. *Behavioral Neuroscience*, *120*(1), 217. doi:10.1037/0735-7044.120.1.217

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502. doi:10.3758/s13428-016-0809-y

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes*, *38*(1), 1–30. doi:10.1207/s15326950dp3801_1

McNamara, D. S. (2007). *Reading comprehension strategies: theories, interventions, and technologies. mahwah*. New Jersey: Lawrence Erlbaum Associates Publishers.

McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy training for active reading and thinking. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 222–233. doi:10.3758/BF03195567

Merat, N., Jamson, A. H., Lai, F. C., Daly, M., & Carsten, O. M. (2014). Transition to manual: Driver behaviour when resuming control from a highly automated vehicle. *Transportation Research Part F: Traffic Psychology and Behaviour*, *27*, 274–282. doi:10.1016/j.trf.2014.09.005

Mills, C., Bixler, R., Wang, X., & D'Mello, S. K. (2016). Automatic gaze-based detection of mind wandering during film viewing. Proceedings of the international conference on Educational Data Mining, 30–37. International Educational Data Mining Society. Raleigh, North Carolina

Mills, C., D'Mello, S., Lehman, B., Bosch, N., Strain, A., & Graesser, A. (2013). What makes learning fun? Exploring the influence of choice and difficulty on mind wandering and engagement during learning. Artificial Intelligence in Education - 16th International Conference, Proceedings, 71–80. https://doi.org/10.1007/978–3–642–39112–5–8

Mills, C., & D'Mello, S. K. (2015). Toward a real-time (day) dreamcatcher: detecting mind wandering episodes during online reading. Proceedings of the 8th international conference on Educational Data Mining, 69–76. Madrid, Spain

Mills, C., D'Mello, S. K., & Kopp, K. (2015). The influence of consequence value and text difficulty on affect, attention, and learning while reading instructional texts. *Learning and Instruction*, *40*, 9–20. doi:10.1016/j.learninstruc.2015.07.003

Mills, C., Graesser, A., Risko, E. F., & D'Mello, S. K. (2017). Cognitive coupling during reading. *Journal of Experimental Psychology: General*, *146*, 872–883. doi:10.1037/xge0000309

Mills, C., Raffaelli, Q., Irving, Z. C., Stan, D., & Christoff, K. (2018). Is an off-task mind a freely-moving mind? Examining the relationship between different dimensions of thought. *Consciousness and Cognition*, *58*, 20–33. doi:10.1016/j.concog.2017.10.003

Mrazek, M. D., Franklin, M. S., Phillips, D. T., Baird, B., & Schooler, J. W. (2013). Mindfulness training improves working memory capacity and gre performance while reducing mind wandering. *Psychological Science, Vol*, *20 20 24*, 0956797612459659. doi:10.1177/0956797612459659

Mrazek, M. D., Smallwood, J., & Schooler, J. W. (2012). Mindfulness and mind-wandering: Finding convergence through opposing constructs. *Emotion*, *12*(3), 442. doi:10.1037/a0026678

Nadel, L., Hupbach, A., Gomez, R., & Newman-Smith, K. (2012). Memory formation, consolidation and transformation. *Neuroscience & Biobehavioral Reviews*, *36*(7), 1640–1645. doi:10.1016/j.neubiorev.2012.03.001

Nadel, L., & Moscovitch, M. (1997). Memory consolidation, retrograde amnesia and the hippocampal complex. *Current Opinion in Neurobiology*, *7*(2), 217–227. doi:10.1016/S0959-4388(97)80010-4

Nguyen, K.-V., Binder, K. S., Nemier, C., & Ardoin, S. P. (2014). Gotcha! Catching kids during mindless reading. *Scientific Studies of Reading*, (ahead-of-print), 18: 1–17.

Ozuru, Y., Briner, S., Best, R., & McNamara, D. S. (2010). Contributions of self-explanation to comprehension of high- and low-cohesion texts. *Discourse Processes*, *47*(8), 641–667. doi:10.1080/01638531003628809

Pettijohn, K. A., Thompson, A. N., Tamplin, A. K., Krawietz, S. A., & Radvansky, G. A. (2016). Event boundaries and memory improvement. *Cognition*, *148*, 136–144. doi:10.1016/j.cognition.2015.12.013

Pham, P., & Wang, J. (2015). *Attentive learner: improving mobile mooc learning via Implicit Heart Rate Tracking*. International conference on Artificial Intelligence in Education, 367–376. Springer.

Phillips, N. E., Mills, C., D'Mello, S., & Risko, E. F. (2016). On the influence of re-reading on mind wandering. *The Quarterly Journal of Experimental Psychology*, *69*(12), 2338–2357. doi:10.1080/17470218.2015.1107109

Radvansky, G. A., & Zacks, J. M. (2017). Event boundaries in memory and cognition. *Current Opinion in Behavioral Sciences*, *17*, 133–140. doi:10.1016/j.cobeha.2017.08.006

Randall, J. G., Oswald, F. L., & Beier, M. E. (2014). Mind-wandering, cognition, and performance: A theory-driven meta-analysis of attention regulation. *Psychological Bulletin*, *140*(6), 1411–1431. doi:10.1037/a0037428

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General*, *140*(3), 283. doi:10.1037/a0023956

Reichle, E. D. (2006). Computational models of eye-movement control during reading: Theories of the eye-mind link. *Cognitive Systems Research*, *7*(1), 2–3. doi:10.1016/j.cogsys.2005.07.001

Reichle, E. D., Pollatsek, A., & Rayner, K. (2012). Using EZ reader to simulate eye movements in nonreading tasks: A unified framework for understanding the eye–mind link. *Psychological Review*, *119*(1), 155. doi:10.1037/a0026473

Reichle, E. D., Reineberg, A. E., & Schooler, J. W. (2010). Eye movements during mindless reading. *Psychological Science*, *21*(9), 1300–1310. doi:10.1177/0956797610378686

Risko, E. F., Anderson, N., Sarwal, A., Engelhardt, M., & Kingstone, A. (2012). Everyday attention: Variation in mind wandering and memory in a lecture. *Applied Cognitive Psychology*, *26*(2), 234–242. doi:10.1002/acp.v26.2

Risko, E. F., Buchanan, D., Medimorec, S., & Kingstone, A. (2013). Everyday attention: Mind wandering and computer use during lectures. *Computers & Education*, *68*, 275–283. doi:10.1016/j.compedu.2013.05.001

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. doi:10.1016/j.tics.2010.09.003

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. doi:10.1111/j.1467-9280.2006.01693.x

Roy, M., & Chi, M. T. (2005). The self-explanation principle in multimedia learning. *The Cambridge Handbook of Multimedia Learning*, 271–286.

Schad, D. J., Nuthmann, A., & Engbert, R. (2012). Your mind wanders weakly, your mind wanders deeply: Objective measures reveal mindless reading at different levels. *Cognition*, *125*(2), 179–194. doi:10.1016/j.cognition.2012.07.004

Schweigert, W. A. (1994). *Research methods and statistics for psychology*. CA:Thomson Brooks/Cole, Belmont.

Sibert, J. L., Gokturk, M., & Lavine, R. A. (2000). The reading assistant: Eye gaze triggered auditory prompting for reading remediation. Proceedings of the 13th annual ACM symposium on User Interface Software and Technology, 101–107. San Diego, CA doi:10.1145/354401

Smallwood, J. (2011). Mind-wandering while reading: Attentional decoupling, mindless reading and the cascade model of inattention. *Language and Linguistics Compass*, *5*(2), 63–77. doi:10.1111/lnco.2011.5.issue-2

Smallwood, J., Fishman, D. J., & Schooler, J. W. (2007). Counting the cost of an absent mind: Mind wandering as an underrecognized influence on educational performance. *Psychonomic Bulletin & Review*, *14*(2), 230–236. doi:10.3758/BF03194057

Smallwood, J., McSpadden, M., & Schooler, J. W. (2007). The lights are on but no one's home: Meta-awareness and the decoupling of attention when the mind wanders. *Psychonomic Bulletin & Review*, *14*(3), 527–533. doi:10.3758/BF03194102

Smith, A. M., Floerke, V. A., & Thomas, A. K. (2016). Retrieval practice protects memory against acute stress. *Science*, *354*(6315), 1046–1048. doi:10.1126/science.aah5067

Smith, G. K., Mills, C., Paxton, A., & Christoff, K. (2018). Mind-wandering rates fluctuate across the day: Evidence from an experience-sampling study. *Cognitive Research: Principles and Implications*, *3*(1), 54. doi:10.1186/s41235-018-0141-4

Smith, J. D., & Graham, T. C. (2006). Use of eye movements for video game control. Proceedings of the 2006 ACM SIGCHI international conference on Advances in Computer Entertainment Technology, 20. Hollywood, CA doi:10.1145/1178823

Starr, M. S., & Rayner, K. (2001). Eye movements during reading: Some current controversies. *Trends in Cognitive Sciences*, *5*(4), 156–163. doi:10.1016/S1364-6613(00)01619-3

Stewart, A., Bosch, N., Chen, H., & Donnelly, P. J., & D'Mello, S. K. (2016). Where's your mind at?: Video-based mind wandering detection during film viewing. Proceedings of the 2016 conference on User Modeling Adaptation and Personalization, 295–296. ACM. Halifax, Canada

Swallow, K. M., Zacks, J. M., & Abrams, R. A. (2009). Event boundaries in perception affect memory encoding and updating. *Journal of Experimental Psychology: General*, *138*(2), 236–257. doi:10.1037/a0015631

Szpunar, K. K. (2017). Directing the wandering mind. *Current Directions in Psychological Science*, *26*(1), 40–44. doi:10.1177/0963721416670320

Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences*, *110*(16), 6313–6317. doi:10.1073/pnas.1221764110

Thomson, D. R., Seli, P., Besner, D., & Smilek, D. (2014). On the link between mind wandering and task performance over time. *Consciousness and Cognition*, *27*, 14–26. doi:10.1016/j.concog.2014.04.001

Unsworth, N., & McMillan, B. D. (2012). Mind wandering and reading comprehension: Examining the roles of working memory capacity, interest, motivation, and topic experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. doi:10.1037/a0029669

VanLehn, K., Jones, R. M., & Chi, M. T. (1992). A model of the self-explanation effect. *The Journal of the Learning Sciences*, *2*(1), 1–59. doi:10.1207/s15327809jls0201_1

Vertegaal, R., Slagter, R., & Van der Veer, G., & Nijholt, A. (2001). Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. Proceedings of the SIGCHI conference on Human Factors in Computing Systems, 301–308. Seattle, Washington

Wammes, J. D., Seli, P., Cheyne, J. A., Boucher, P. O., & Smilek, D. (2016). Mind wandering during lectures II: Relation to academic performance. *Scholarship of Teaching and Learning in Psychology*, *2*(1), 33–48. doi:10.1037/stl0000055

West, B. T., Welch, K. B., & Galecki, A. T. (2014). *Linear mixed models: a practical guide using statistical software*. Boca Raton, FL: Chapman and Hall/CRC. doi:10.1201/b17198

Zanesco, A. P., King, B. G., MacLean, K. A., Jacobs, T. L., Aichele, S. R., Wallace, B. A., … Saron, C. D. (2016). Meditation training influences mind wandering and mindless reading. *Psychology of Consciousness: Theory, Research, and Practice*, *3*(1), 12–33. doi:10.1037/cns0000082

Zeidan, F., Johnson, S. K., Diamond, B. J., David, Z., & Goolkasian, P. (2010). Mindfulness meditation improves cognition: Evidence of brief mental training. *Consciousness and Cognition*, *19*(2), 597–605. doi:10.1016/j.concog.2010.03.014

Zhao, Y., Lofi, C., & Hauff, C. (2017). *Scalable mind-wandering detection for MOOCs: a webcam-based approach*. European conference on Technology Enhanced Learning, 330–344. Springer.

Zwaan, R. A., Magliano, J. P., & Graesser, A. C. (1995). Dimensions of situation model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(2), 386.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162. doi:10.1037/0033-2909.123.2.162

Appendix

Table A1. Pearson *Correlations (r) between key variables.*

	Total # Intervention	Reading Time	MW Likelihood	SE Score	Textbase Imm	Textbase Delayed	Inference Imm
Reading Time Per Page	.205						
MW Likelihood	.356**	.393***					
SE Score 1	.297*	.242*	−.111				
Textbase Imm.	−.100*	.200	−.102	.107			
Textbase Delayed	−.002	.072	−.287*	.110	.546***		
Inference Imm.	−.031	.242*	−.042	.061	.511***	.376**	
Inference Delayed	−.112	.066	−.120	.000	.332**	.490***	.000[1]

Notes. * indicates $p < .05$, ** indicates $p < .01$, *** indicates $p < .001$. Correlations calculated at the participant level; SE Score 1 = first self-explanation attempt; Imm = Immediate comprehension test. [1]Due to the zero correlation found here, we recomputed this correlation separately by Condition: r = −.175 in the MW Intervention Condition (p = .314); r = .156 in the Yoked-Control Condition (p = .372).