Gaze-based predictive models of deep reading comprehension

Authors:

*Rosy Southwell[a], Caitlin Mills[b], Megan Caruso[a], Sidney K. D'Mello[a]

[a]*Institute of Cognitive Science, University of Colorado, Boulder, USA*
[b]*Department of Psychology, University of New Hampshire, Durham, New Hampshire, USA*

roso8920@colorado.edu *corresponding author
caitlin.mills@unh.edu
megan.caruso@colorado.edu
sidney.dmello@colorado.edu

Declarations

This paper or a similar version is not currently under review by a journal or conference. This paper is void of plagiarism or self-plagiarism as defined by the Committee on Publication Ethics and Springer Guidelines.

We report novel analyses using data from published work [1] – this is described as "Study 2" in the attached manuscript, and provides a replication and validation of the model trained on the main dataset (Study 1) which has never been published. An early version of the analysis of Study 2 was presented in a 5-minute talk at the Society for Text and Discourse annual conference in July 2020. In sum, the majority of the attached manuscript consists of unpublished work. This manuscript is not under consideration for publication elsewhere.

1. Mills, C., Gregg, J., Bixler, R., & D'Mello, S. K. (2021). Eye-Mind reader: an intelligent reading interface that promotes long-term comprehension by detecting and responding to mind-wandering. Human–Computer Interaction, 1–27. https://doi.org/10.1080/07370024.2020.1716762

Funding

This research was supported by the National Science Foundation (NSF) (DRL 1235958, DRL 1920510). Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NSF.

Conflicts of interest/Competing interests

None reported

Author contributions:

RS: methodology, formal analysis, writing - original draft preparation, reviewing and editing.
CM: methodology, data collection, writing – reviewing and editing, investigation.
MC: writing - original draft preparation, reviewing and editing.
SD: conception, supervision, project administration, funding acquisition, writing – reviewing and editing

Acknowledgements

The authors thank Robert Bixler for contributions to data analysis and methodology, and Candace Peacock for providing feedback on the manuscript.

Abstract

Eye gaze patterns can reveal user attention, reading fluency, corrective responding, and other reading processes, suggesting they can be used to develop automated, real-time assessments of comprehension. However, past work has focused on modeling factual comprehension, whereas we ask whether gaze patterns reflect deeper levels of comprehension where inferencing and elaboration are key. We trained linear regression and random forest models to predict the quality of users' open-ended self-explanations (SEs) collected both during and after reading and scored on a continuous scale by human raters. Our models use theoretically-grounded eye tracking features (number and duration of fixations, saccade distance, proportion of regressive and horizontal saccades, spatial dispersion of fixations, and reading time) captured from a remote, head-free eye tracker (Tobii TX300) as adult users read a long expository text (6500 words) in two studies (N=106 and 131; 247 total). Our models: (1) demonstrated convergence with human-scored SEs (r = .322 and .354), by capturing both within-user and between-user differences in comprehension; (2) were distinct from alternate models of mind-wandering and shallow comprehension; (3) predicted multiple-choice posttests of inference-level comprehension (r = .288, .354) measured immediately after reading and after a week-long delay beyond the comparison models; and (4) generalized across new users and datasets. Such models could be embedded in digital reading interfaces to improve comprehension outcomes by delivering interventions based on users' level of comprehension.

Keywords: self-explanation, gaze tracking, reading, comprehension, machine learning, automated assessment

1 Introduction

Comprehension of text is critical for thriving life in modern society: while reading print media for pleasure may be in decline, computerized reading is integral to functioning at work (with the average professional receiving 120 emails per day; Chui et al., 2012). Reading is ubiquitous in a host of user interfaces for education (e.g., digital textbooks: Tyner, 2014), entertainment (e-readers for fiction), information (e.g., news articles), law and policy (e.g., legal statutes), and so on. Understanding and *learning* from complex texts is thus an invaluable skill for an informed citizenry (Alexander, 2012; Britt et al., 2014). Yet, the ubiquity of reading belies its complexity: a host of cognitive processes must work in unison with the visual system as we move our eyes across the text and process what we are reading (Just & Carpenter, 1976). Reading comprehension is a complex, multi-dimensional process that involves parsing sentences, identifying and decoding the meaning of each word, integrating incoming information with what has been previously read, and situating the new information in existing knowledge – all in service of constructing a rich mental representation of the text. Failures at any stage of this process can negatively impact comprehension, and often the user is insufficiently aware of the gaps in their own comprehension, particularly if they are less-fluent readers (Kinnunen & Vauras, 1995). Further, fluency does not guarantee comprehension, as users may mind wander (Smallwood, 2011), lack sufficient background knowledge (Kendeou & Broek, 2010), and not be sufficiently motivated or capable of processing the text deeply, instead relying on skimming and other shallow forms of comprehension (Duggan & Payne, 2009). Simply put, reading is not the same as reading for understanding (Britt et al., 2014; Graesser et al., 1994; McNamara et al., 2007), which is difficult to achieve.

Thus, comprehension monitoring could be integrated into computerized reading interfaces, providing intervention or guidance when warranted by the comprehension model. Increasingly, learning environments are digital experiences (Boulanger & Kumar, 2019), and technological advancements allow personalization of content and presentation to assist learners. Predictive models can be used to trigger adaptation of a learning environment in real-time (Mousavinasab et al., 2018). Some examples of effective interventions already exist in the field of attention-aware learning environments (D'Mello, 2016; D'Mello et al., 2017; Hutt et al., 2016; Mills et al., 2020), where attentional focus is modeled during a task, and informs modifications to the learning environment such as *what* or *how* information is displayed. One such example is an attention-aware reading interface which uses eye gaze to estimate when users are mind-wandering, upon which it prompts users to explain their understanding and re-read if necessary (Mills et al., 2021). Feedback can also be provided to the users on their level of comprehension to support their metacognitive awareness (for example: Bondareva et al., 2013; Roll et al., 2011) which, in turn, is beneficial for learning (Amini et al., 2020). Beyond education, implicit knowledge of a user's comprehension status could inform user interfaces – whether as a "comprehension-aware" responsive interface (analogous to the related field of attention-aware user interfaces; Roda & Thomas, 2006) or simply to inform design choices for digital and print-based textual materials (Everdell, 2014). Accordingly, our aim is to develop a non-intrusive, automated, real-time measure of deeper levels of user comprehension (i.e., focusing on understanding not rote memorization).

There are many design decisions in developing such a model. For one, we need a ground-truth measure of comprehension. This entails deciding the type, format, and timing of the measure. Shallow reading comprehension can be assessed with multiple-choice questions targeting rote recall of verbatim information from the text (Graesser et al., 2009). Multiple-choice questions can also be designed to target

deeper, integrative levels of processing; for instance, to probe the quality of inference formation (e.g., Glass, 2009). However, multiple-choice questions give us a unidimensional, binary (success/failure) view, and potentially are biased toward measuring prior knowledge rather than comprehension *per se* (Ozuru et al., 2013). The timing of the assessment is another important factor. Comprehension assessments are either interspersed during or immediately after reading. Interruption of reading with unrelated tasks can disrupt comprehension (Foroughi et al., 2015), but on the other hand interpolated testing on the text content can improve subsequent comprehension (e.g. Roediger and Karpicke 2006) by supporting the user's metacognitive awareness of gaps in their understanding (Agrawal et al., 2012). Comprehension assessments that occur after reading are subject to interference and decay effects, limiting their utility as measures of unfolding comprehension processes.

Given these considerations, we specifically focused on predicting reading comprehension at a *deep* level using *self-explanations,* which are freeform responses to carefully constructed, open-ended prompts (McNamara, 2004; McNamara et al., 2007). Self-explanations often contain indicators of deep comprehension such as inferences and elaboration (McNamara, 2004) and recognition of connections between the reader's prior knowledge and different parts of the text. Self-explanation thus renders the ongoing comprehension process more visible than, for instance, multiple-choice questions. In addition, self-explanation is itself a reading *strategy* which readers can initiate themselves in order to improve their own comprehension (Bielaczyc et al., 1995; McNamara, 2004), and indeed self-explanation training has proven beneficial for struggling readers (Chi et al., 1994). To address the issue of timing and interruptions of the assessments, we collected written self-explanations both during and after reading.

The second critical issue pertains to the objective measurement features used to automate the comprehension assessment. Our modeling approach relies on the "eye-mind link": the coupling between eye movement and cognition during reading (Foulsham et al., 2013; Nilsson, 2012; Rayner, 1998; Reichle, 2006) which has long flagged eye movement data as a useful modality to inform adaptive learning systems (e.g., Shute & Zapata-Rivera, 2012). As reviewed in more detail below, the pattern and timing of certain eye movement features during reading are sensitive to linguistic properties of words and sentences, reader skill and comprehension difficulty. Although earlier studies primarily focused on tightly controlled experiments with short, disconnected sentences or paragraphs designed to isolate specific linguistic factors, many findings are corroborated by studies using longer texts and materials presented in a more naturalistic manner, aided by advancements in eye-tracking technology.

2 Background

2.1 Reading comprehension

Reading comprehension theories vary in their exact description of core cognitive processes and their interactions, but all agree on the following outline. Word forms are recognized (Coltheart et al., 2001; Grainger & Holcomb, 2009), their semantic associations are retrieved (Pattamadilok et al., 2017; Price et al., 1997), and integrated into an unfolding mental construction of the text (Zwaan & Radvansky, 1998). Eye movements are driven by both top-down information from the user's current mental model (Booth & Weger, 2013; Just & Carpenter, 1980), and bottom-up information from visual cues such as those available in parafoveal vision (Reichle & Reingold, 2013; Reingold et al., 2012; Schotter et al., 2012).

A common distinction is drawn between mental representations of a passage at the *surface code, textbase* and *situation model* levels (Dijk & Kintsch, 1983). Surface code representation refers to a literal encoding of the words in a sentence. The textbase refers to the meaning of the sentence, independent of the exact wording. The situation model incorporates *integration* across multiple sentences or even paragraphs (Zwaan & Radvansky, 1998). The situation model represents the combination of the reader's own prior knowledge with the interpreted meaning of the text itself (Graesser et al., 1994).

Cognitive theories such as Kintsch's Construction-Integration model (Kintsch, 1988, 1998) explain how inferences and generalizations can be made from the initial, more literal meaning of the text. Accounts differ in whether building the situation model is an automatic, bottom-up process (Gerrig & O'Brien, 2005) suggesting no route for inferential processes to strongly affect eye movements, or whether reading – like virtually everything the brain does – is a predictive process in which readers actively form predictions about and inferences from the text (Graesser et al., 1994). Both mechanisms – automated, feedforward and reflective, top-down processing – are likely involved in reading comprehension (Rapp & Broek, 2005). Certain elaborative processes are thought not to occur during reading unless prompted, such as projecting consequences into the future (Graesser & Bertus, 1998). Regardless of this debate, if initial encoding at the textbase level is unsuccessful, this could theoretically have downstream effects on developing the situation model (Smallwood et al., 2007). For instance, if a user's mind wanders during a specific section of text where key information is presented, the formation of the textbase representation is disrupted, hindering their ability to later form inferences pertaining to that information (Smallwood et al., 2008).

2.2 Eye movements in reading

Eye movements are the most prevalent human behavior (Carpenter, 2000) and are a rich source of information on reading strategies (Just & Carpenter, 1980; Raney et al., 2014) and the cognitive state of the user. Fixations, saccades, and regressions are the main components of eye movements. Fixations are pauses (around 225 ms) when the eyes are fixed on a location to encode information. Saccades are the movements between these pauses (during reading, saccades advance around eight characters). Regressions are saccades back to text that has already been read, occurring 10-25% of the time (Rayner et al., 2012). Only two thirds of words are fixated during typical reading: for instance, function words are often skipped, yet multiple fixations can be made on others (Rayner et al., 2006).

Eye movement features are highly variable depending on characteristics of the text, indicating that comprehension processes and eye movements are linked. For instance, lexical (word-level) variables, such as word frequency in the lexicon, affect fixation times (Rayner & Duffy, 1986). Words which are predictable given the preceding context have shorter fixations (McDonald & Shillcock, 2003), and conversely, sentence-final words which are surprising in context receive more, longer fixations, and increased regressions to earlier in the sentence (Ehrlich & Rayner, 1981; Rayner & Well, 1996a). In longer texts, increased difficulty in a passage leads to increased number of fixations and longer fixation durations on average (Cook & Wei, 2019; Rayner et al., 2006).

What do such findings on the gaze correlates of reading comprehension tell us about the mechanics of the eye-mind link? Some models of the reading process explicitly incorporate the control of eye movements and cognitive variables during reading. In the E-Z reader model (Reichle et al., 2009), readers can only process one word at a time due to attention limits, starting with word recognition and proceeding to

higher-level integrative processing. The speed at which meanings can be retrieved for the attended word influences the programming of saccades and progression of attention to the next word. If integration of the word into context fails, a regressive saccade back to the word in question may be triggered (Reichle et al., 2009). This model replicates some key empirical findings concerning semantic and lexical effects on fixation durations: for example, the parafoveal preview effect whereby fixation durations can be decreased when properties of upcoming words can be perceived already in parafoveal vision (Reichle et al., 2003, Schotter et al., 2012), and "spillover" effects where fixations are longer on words following a difficult-to-process word (Rayner & Duffy, 1986; Reichle et al., 2003). Similarly, according to the SWIFT model (Engbert et al., 2005) processes of word recognition and saccade generation are interdependent, and attention is not constrained to only process words in serial.

2.3 The eye-mind link as a window onto ongoing comprehension?

The critical question for the present study is whether measuring a user's eye movements can tell us not just about the properties of the text, which are known variables, but about the internal cognitive workings of the user which determine comprehension outcomes. Regressive eye movements are a critical component of skilled reading and are triggered when the user recognizes gaps or weaknesses in their understanding (Booth & Weger, 2013; Metzner et al., 2017). Recently Moort et. al. (2020) found that text-based and knowledge-based incongruences resulted in different regression patterns. However not all comprehension repair is reflected in eye movements; for instance, resolution of comprehension incongruencies can occur without regressive eye movements (Chi, 2000; Meseguer et al., 2002).

Skim reading can be helpful to extract key information from a text (Duggan & Payne, 2009; Masson, 1983; Taylor, 1965) but is often associated with impaired comprehension (Rayner et al., 2012; Strukelj & Niehorster, 2018) and is reflected in gaze as fewer and shorter fixations (Masson, 1983), longer saccade distances and more skipping of words (Strukelj & Niehorster, 2018). Fewer regressive saccades occur during speed reading (Miyata et al., 2012), which when combined with more skipping leads to larger areas of unread text, accompanied by lower comprehension of those areas (Just et al., 1982).

Eye movements are also affected by the attentional state of the user. Mind-wandering (defined here as task-unrelated thought) is consistently linked to decreased comprehension (e.g., Faber et al., 2018; Reichle et al., 2010; D'Mello & Mills, 2021) and has a somewhat consistent gaze signature during reading. When users report mind-wandering during reading, eye tracking data reveals fewer, longer, and more dispersed fixations (Bixler & D'Mello, 2015; Faber et al., 2018; Reichle et al., 2010). However, some studies find no change or even an increase in fixation count (Smilek et al., 2010; Steindorf & Rummel, 2019). Eye movements during mind-wandering become less sensitive to lexical and linguistic changes in the text (Reichle et al., 2010; Steindorf & Rummel, 2019, Franklin et al., 2011). Mind-wandering is associated with reduced corrective regressions (Reichle et al., 2010; Foulsham et al., 2013), which can lead to impaired comprehension. Overall alignment of gaze with the text has been linked to attentive reading (Biedert et al., 2012; Buscher et al., 2008), captured with spatial saccade patterns such as horizontal saccade proportion and fixation dispersion (Biedert et al., 2012; Southwell et al., 2020).

One challenge to modeling deep comprehension from eye movements is the sheer number of processing stages, spread over time and neural pathways, that are involved in comprehension. For instance, as reviewed above, comprehension can be constructed in a "covert" manner by cognitive (re)processing without affecting eye movement patterns. This places an upper limit on how faithfully comprehension can

be inferred from gaze properties. However, although eye movements are somewhat automatic and decided by low-level lexical factors (Yang & McConkie, 2001), there is at least some effect of comprehension difficulties on eye movements. Therefore, even if deep comprehension processes do not themselves affect eye movements measurably, once can potentially infer the success of discourse-level comprehension from eye movements which impact encoding at the surface or textbase levels. Thus, we hope to demonstrate that, in principle, eye movements contain signatures attributable to higher-level processing during reading.

2.4 Related work

Machine-learned models have proven fruitful for predicting reading comprehension and related processes. To keep scope manageable, we focus on studies that use eye gaze during reading, but also briefly discuss a few other relevant studies. Such models are ideally designed with constraints from findings and theories in the psychological and cognitive sciences (D'Mello et al., 2020). Approaches differ in their success at prediction, generalizability across users and reading scenarios, and interpretability.

2.4.1 Automated models of reading comprehension

Reading comprehension has previously been modeled from gaze (D'Mello et al., 2020; Southwell et al., 2020; Ahn et al., 2020; Copeland et al., 2014, Copeland & Gedeon, 2013; Makowski et al., 2019; Martínez-Gómez & Aizawa, 2014; Underwood et al., 1990; Wallot et al., 2015, Rajendran et al., 2018), physiology (Daley et al., 2014), facial cues (Lai et al., 2019) and neural signals (Broadway et al., 2015; Yuan et al., 2014) as we review below.

Summary metrics of gaze computed at the level of pages or entire passages of text have proven somewhat promising for predicting user understanding, as assessed both during (D'Mello et al., 2020) and immediately after reading (Southwell et al., 2020). Copeland and colleagues (Copeland et al., 2014, Copeland & Gedeon, 2013) trained artificial neural networks to predict comprehension scores at the level of individual questions (both objective quiz scores, and the reader's self-assessed comprehension) on a computerized reading assignment from gaze features including proportion of gaze regressions and mean fixation duration, but they did not test generalizability of their models to new (unseen) users. Summary metrics of fixations, saccades and pupil diameter, as well as textual characteristics spanning lexical, syntactic, semantic and discourse levels were used in models to predict user-level comprehension and skill (Martínez-Gómez et al. 2014), however performance at predicting comprehension scores was only significant when scores (originally continuous) were binarized and the dataset was restricted to the top and bottom quartile of performers, essentially removing the most difficult cases. Likewise, Ahn et al. (2020) found that their models only perform well for binarized comprehension scores using a convolutional neural network and their model did not generalize to unseen participants. Wallot et al. (2015) found a linear regression model could significantly predict comprehension scores from power-law scaling factors computed from gaze position and fixation duration timeseries. However, the model was not cross-validated so it may not generalize across users. Second, the reading materials were chosen to be easy or difficult, so it is unclear to what extent their model was predicting text characteristics rather than user characteristics. A prevalent issue is that some models are not assessed on unseen users (Underwood et al., 1990, Copeland & Gedeon, 2014, Rajendran et al., 2018) or do not generalize well across users when explicitly tested (Ahn et al., 2020, Copeland & Gedeon, 2013, Martínez-Gómez & Aizawa, 2014).

In addition, there is a lack of studies which demonstrate a generalizable model of deeper, discourse-level understanding.

Beyond eye movements, other sensors can provide informative data on reading comprehension. Daley et al. (2014) exploited a known physiological response associated with emotion, the respiratory sinus arrhythmia, to predict reading comprehension scores; however, this study used reading aloud in children so it is unclear whether this would generalize to silent reading that is more prevalent outside of a primary educational setting. Multimodal configurations incorporating eye-tracking with EEG and facial expression have also proven successful: Lai et al. (2019) used coarse metrics from these sensors to predict user-level pass/fail on a brief science reading and assessment with precision and recall of 0.8. EEG alone can also be a powerful predictor of comprehension, as evidenced by its ability to reveal mind-wandering which in turn is related to comprehension (Broadway et al., 2015), however this study used a constrained, word-by-word presentation paradigm rather than free reading. Still, very little EEG data – as short as 4-second samples – may be needed to predict comprehension in a manner generalizable across users (Yuan et al., 2014). Yet, the use of an EEG headset can be cumbersome in a real-world HCI setting, at least with the technologies readily available. Therefore, we propose eye tracking over other physiological measures as particularly suitable for tracking reading comprehension noninvasively.

2.4.2 Automated models of related cognitive processes

Gaze data has also been used to measure users' cognitive states within intelligent user interfaces (e.g., Mills et al., 2021; Scheiter et al., 2019; see Conati et al., 2013 for a review). For instance, an early gaze-based system for reading support (Sibert et al., 2000) used the dwell time on an individual word to infer word-level processing difficulty: if the user paused on a word for long enough, an automated Reading Assistant provided an audio pronunciation of the word. Mind-wandering can be also successfully estimated from gaze summary metrics (Faber et al., 2018; Mills et al., 2021, Hutt et al., 2019, 2017) collected during reading. Skimming – which is often associated with lower comprehension outcomes – is discriminable from eye movement patterns (Biedert et al., 2012; Krejtz et al., 2019). Interestingly, a reader's subjective assessment of their own understanding can be predicted from gaze features (Copeland & Gedeon, 2013), even if this does not match objective measures of comprehension (Sanches et al., 2018). Loboda et al., (2011) found that relevance of specific words in an information-seeking task can be inferred from reading time, fixation duration and number of fixations.

Finally, the general level of reading or language skill can be modeled from user-level gaze features (Lou et al., 2017; Underwood et al., 1990). For instance, eye movements and pupil diameter have been used to predict general reading ability (Zhan et al., 2016) as measured by standardized tests at the level of an individual reader, rather than comprehension on a particular text, which could vary within readers. Ultimately, the ideal comprehension model would be able to predict not only between-individual differences in reading comprehension, but also within-individual differences perhaps at the level of individual pages within a text.

3 Current study: aims and novelty

In this study, we build predictive user-independent machine-learned computational models (D'Mello et al., 2020) of reading comprehension during computerized reading, assessed via self-explanations during (Study 2) or after (Study 1) reading, from eye movements measured with an eye tracker. We focused on a

small number of gaze features based on theoretical and experimental research on reading comprehension or factors (such as processing difficulty, attentional state and reading skill) which are correlated with comprehension outcomes. This was done to ensure that the features were interpretable in terms of existing theoretical frameworks from the cognitive, psychological and education fields (as discussed in Paquette et al., 2019; D'Mello et al., 2020). We examined the convergent, discriminant, and predictive validity of our models along with their generalizability across users and studies.

Our motivation is to model ongoing comprehension of long, connected texts at the deep, discourse level – the level of integration of the text into a situation model over the course of multiple paragraphs – using eye movements. Our approach is grounded in theoretical and empirical research on reading comprehension and eye movements during reading. Despite the vast literature, there is no established unified account of reading that connects low-level eye movements with high-level discourse processes (Rayner & Reichle, 2010; Reichle et al., 2009). Similarly, despite recent work on modeling reading comprehension from eye gaze, this work has focused on short texts (e.g. Martínez-Gómez et al. 2014) with limited evidence of generalizability to new users (Copeland & Gedeon, 2013, 2014; Ahn et al., 2020). The two studies that developed user-independent models of comprehension of long, connected texts (D'Mello et al., 2020; Southwell et al., 2020) have emphasized shallow rote comprehension. Thus, the present study reflects the first attempt to develop and validate a user-independent, gaze-based model of deep comprehension of long, connected texts.

We focused on self-explanations as our ground-truth measure of comprehension as noted above. Our goal is to model comprehension in real time, however collecting self-explanations during reading can cause interruptions and may itself trigger critical deep comprehension processes such as inferencing. Therefore, in Study 1, self-explanations were collected immediately after uninterrupted reading whereas they were collected during reading in Study 2. In both studies, self-explanation prompts were structured such that a given prompt could be linked to a specific section of the text, allowing us to align each segment of gaze data with corresponding comprehension scores. Accordingly, we used supervised-machine learning techniques (linear regression and random forest classifiers) to predict self-explanation scores from gaze features during the relevant sections of text. The models were trained in a user-independent fashion, where all data from the same user was in the training *or* testing set, but never in both.

In addition to self-explanations, we included two posttest (i.e., post-reading) multiple choice measures of learning and comprehension. Textbase-level comprehension was measured using rote items referring to specific details in the text whereas inference-level comprehension items targeted conceptual knowledge that was implicit in the text and required inference on the part of the reader to answer correctly. Knowledge, such as that acquired during reading, can be substantially reprocessed over time, particularly following consolidation during sleep (Nadel et al., 2012). Thus, we also administered parallel versions of these assessments after a week-long delay to investigate retention.

We address the following four research questions (RQs):

RQ1. How accurately can self-explanations be modeled from eye gaze? We examined alignment of the models' predictions to ground-truth self-explanations to measure accuracy. We also assessed to what degree the model captures differences in comprehension between readers (i.e., variability across readers) versus within readers (i.e., variability within a text), and identified features that were most predictive of self-explanation scores.

RQ2. How does a self-explanation model compare with prior gaze-based models of comprehension? Gaze-based predictive models of shallow reading comprehension have already been developed (as reviewed above) and we generated predictions of shallow, rote-level comprehension from one such model (D'Mello et al., 2020) and compared it to our self-explanation model. Given mind-wandering is negatively correlated with comprehension, eye movements characteristic of mind-wandering can be indirectly informative of a reader's comprehension outcomes. Accordingly, we also compared our model to a gaze-based model of mind-wandering (Faber et al., 2018). Finally, because we use both eye gaze and reading time, we compared our model to one trained exclusively on reading-time to quantify the value added of eye gaze.

RQ3: To what extent does the model predict posttest measures of learning and comprehension? We examined whether model-predicted self-explanation scores correlated with textbase and inference-level comprehension measures after reading and after a one-week delay, allowing us to quantify the persistence of the relationship between gaze and comprehension.

RQ4: Does the model generalize to new users and different reading contexts? We then used the self-explanation models from each study to generate predictions on the other, thereby examining how the models generalize across datasets in predicting both self-explanations scores and the posttest comprehension measures.

4 Method

Methods were similar across both studies, so we jointly present both with key differences noted. The data from Study 1 has not been previously published. The data from Study 2 was previously published (Mills et al., 2021), but the previous focus was on learning outcomes whereas the present analysis, which focuses on eye gaze, is new.

4.1 Participants

Participants in Study 1 were university undergraduates (N=136) from two universities in the USA; N=51 from a public Eastern university and N=85 from a public Western university. The mean age of participants was 20.7 (range 18-44), with 69% female, 30% male, and 1.4% other gender. Ethnicities of participants were 1.4% African American, 8.6% Asian, 79% Caucasian, 5.0% Hispanic, and 5.8% Other. Participants were compensated with a $30 Amazon gift certificate.

Participants in Study 2 (N=113) were undergraduates; n = 54 from a private Midwestern university who participated for course credits and *n* = 59 from a public Western university who were compensated with a $30 Amazon gift certificate. Demographic details are available for 69 of the participants who completed an optional demographics questionnaire. The mean age of respondents was 21.1 (range 18-28), with 58% female and 42% male. Ethnicities of respondents were 1.5% African American, 23% Asian, 68% Caucasian, 5.8% Hispanic, and 1.5% Other.

All participants provided informed consent before participating and were informed they were able to withdraw from the study without penalty at any time. The studies and consent procedure were approved by the Institutional Review Board at the respective universities. Participants were required to be students at least 18 years old, English speakers and self-identified as not having significant uncorrected visual impairments. Contact lenses or eyeglasses were permitted as the eyetracker has been demonstrated to work with such vision correction.

4.2 Materials & measures

Text. Participants read a long, connected text while their eye movements were recorded. The text was non-fiction: a 6500-word excerpt from the start of a book on surface tension in liquids (Boys, 1890). This century-old science text was selected to minimize the likelihood that participants would have previously read it. While the original book included diagrams, the text was read by the researchers and deemed sufficient for comprehension without the diagrams. It had a Flesh-Kincaid grade score of 11.8, indicating moderate difficulty for the age group. The reading materials were split into 15 sections, each representing one focal concept with 1-7 pages per concept.

Self-explanations. Self-explanation prompts pertaining to specific sections (one per concept for a total of 15) were structured to encourage readers to elaborate on and generalize concepts presented in the text. See Table 1 for an example question and responses.

Knowledge assessments. Participants also completed four-alternative multiple-choice assessments to assess both "shallow" textbase-level and "deep" inference-level comprehension. There were four shallow and two inference items for each of the 15 concepts.

Textbase-level questions addressed factual knowledge presented in the text. For example:

> Does the sieve the author used in an experiment float on water?
>
> a) yes, the weight of the sieve is not sufficient to stretch the skin of the water through the holes
> b) no, the water passes through the holes
> c) no, the sieve is too heavy to float
> d) none of the above

The answer (a) to this rote item can be obtained directly from the following sentence in the text:

> *"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."*

Inference-level questions targeted deeper comprehension which required the formation of inferences and integration with existing knowledge (Graesser et al., 2009). For example, the following requires an inference on the part of the participant that the thimble with holes in it will behave like the sieve example presented in the text, i.e. option (c) below:

> 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
> b) carrying a pail of water with a leak

c) like a regular cup of water
d) none of the above

Eye Tracking. Throughout reading for both studies, the eye gaze position of both eyes was tracked using an infrared eye tracking system sampling at 120Hz that allowed for free movement of the head (Tobii TX300, Tobii Pro AB, VA, USA). Stimuli were displayed on a 23" monitor integrated with the eyetracker and had a resolution of 1920x1080. Participants were seated such that eyes were approximately 65cm from the screen.

4.3 Procedure

Participants were individually tested in a quiet research lab at the respective universities. Upon providing consent, participants were seated in front of the eye tracker monitor. Upon completion of preliminary activities that varied by study (see below), the main reading task commenced.

Participants were given the following reading goal: *"Your primary task is to read the text in order to take a short test after reading."*. The text was split across 57 screens with an average of 115 words on each (SD = 8, range 90-129), and was self-paced with the participant pressing the right arrow key to progress to the next page. For Study 2 only, participants also had the option to press the left arrow key to go back to the previous page.

The shallow and inference assessments occurred at two points: first, following reading the text *(Shallow-Immediate* and *Inference-Immediate)* and second after a delay of approximately one week (*Shallow-Delayed* and *Inference-Delayed*). Half of the items for each concept were used at each assessment point (main session or delayed follow-up), such that questions in the delay session were previously unseen, counterbalanced across participants. Participants were not able to refer to the text when answering these assessments, and each question appeared on its own screen.

The timing of the self-explanation prompts and experimental conditions varied by study as noted below.

Study 1. Self-explanation prompts occurred immediately *after* reading the entire text, interspersed with the multiple-choice questions. Thus, reading was uninterrupted. Prior to reading, participants watched brief, 3-minute movie clips chosen to influence their affective states (happy – N=44; sad – N=44; neutral – N=43). This manipulation was collected as an exploratory investigation into mood and text processing. It is not relevant to this study beyond the influence it may have as a confounding variable between gaze variables and comprehension assessment outcomes. To address this, we report model performance by experimental condition.

Study 2. Unlike in Study 1, self-explanation prompts were interspersed during reading, so reading was interrupted. Interventions, where participants were given self-explanation prompts, were probabilistically triggered using a gaze-based mind-wandering detector (Faber et al., 2018; Mills et al., 2021) for the half of participants assigned to the "Mind-Wandering Intervention" group. For this group, self-explanations were scored using an automated scoring algorithm based on word overlap between the response and a set of ideal answers generated by the researchers and the inclusion of critical keywords and phrases. While writing their initial self-explanation response, participants were unable to refer to the text. If the automated scoring deemed the self-explanation unsatisfactory, the participant was able to re-read the pages in the preceding section and given another opportunity to revise their initial self-explanation (see

Mills et al., 2021 for details). The other half of the participants ("yoked-control") were each matched with a participant in the Intervention group, and they simply received interventions with matched timings to their experimental counterpart irrespective of their mind-wandering. Although the automated mind-wandering detector is incidental to the present analysis, it would be a confound because self-explanations were prompted when mind-wandering was deemed likely in the intervention group. To address this, we report model performance split by experimental condition.

4.4 Data treatment

4.4.1 Scoring self-explanations

Self-explanations were scored after the study by expert raters on a continuum between 0 and 1 (SE score). Six researchers independently scored archetypal answers to each question. Criteria for scoring included similarity between each response and the archetypal answers, and inclusion of particular keywords deemed critical by the researchers. To assess reliability, two raters scored a subset of responses. The correlation between raters on this subset was sufficiently high (Study 1: r = .89, n=30; Study 2: r = .70, n=151), so one rater scored the remaining self-explanations. See Table 1 for representative examples of responses alongside the given scores.

4.4.2 Computing gaze features

Gaze position was averaged across both eyes, then fixations and saccades were extracted using the Open Gaze and Mouse Analyzer software (Voßkühler et al., 2008), upon which page-level (a page corresponds to a screen of text) gaze summary metrics were computed. Areas-of-interest on the page were defined around each word and this was used to identify fixations as regressions; any fixations falling on an area-of-interest corresponding to any word earlier than the previous fixation was counted as a regression fixation. We derived 6 key gaze features, the first four of which were chosen based on strong support from the experimental literature for their connection to reading comprehension (see Introduction) and all six are the same as previously used in gaze-based models of comprehension (see D'Mello et al., 2020; Southwell et al., 2020 for detailed justification). The features were *fixation count* and *mean fixation duration*, proportion of *regression fixations* (as a fraction of fixation count), mean *saccade distance* (also known as saccade amplitude), proportion of *horizontal saccades*, and *fixation dispersion*. In contrast to prior studies, fixation count was divided by the number of words on the page to derive *fixations per word*, therefore disentangling the effect of word count on fixation count. Horizontal saccade proportion was the proportion of saccades on a page falling within 30° of the horizontal, either above or below. Fixation dispersion takes the mean (x,y) coordinates over all fixations on a page and computes the average Euclidean distance of all fixations to this mean point. The latter two are non-traditional, but have been used in previous gaze-based modeling studies of reading (D'Mello et al., 2020; Southwell et al., 2020), and were chosen as a proxy for the alignment of eye movement patterns to the physical layout of the text.

Reading time was used as an additional feature, computed as the time from page onset on-screen to the participant's key press to progress to the next page, and divided by word count to give *reading time per word*. Though it is not strictly an eyetracker-derived feature, it is an important predictor of reading comprehension so we included it here (Mills, Graesser, Risko, & D'Mello, 2017); this is consistent with previous studies cited above.

For Study 2, where re-reading was possible following an intervention, we only included eye gaze from the first read of a page and their first submitted self-explanation for a given concept. For both studies, *concept-level* gaze features were computed by averaging page-level features over all pages within a given conceptual section of the text (regardless of whether the concept had a corresponding human-scored self-explanation). These concept-level gaze features were used to predict comprehension performance on the items pertaining to each concept (see Modeling section).

4.4.3 Data exclusion and statistics

For Study 1, self-explanations were collected for 131 participants, covering all except one concept due to an experimental error, resulting in a total of 1834 scores. As the missing concept was the same across all participants, gaze data and posttest scores from this concept was excluded from further analysis. After removing instances without sufficient gaze data during reading of the corresponding concept (see below), 1612 self-explanation scores remained (mean score = .41; SD = .39. Self-explanations per participant: mean 12.3, range 1-14).

For Study 2, only 106 participants had both gaze data and self-explanation scores; the remaining 7 participants were excluded from further analysis. As noted above, only the first self-explanation response for a given concept was analyzed here; further self-explanations completed after the intervention are disregarded. There were 515 self-explanation instances (mean 4.9 scores per participant; range 1-10) with a mean score of .62 (standard deviation= .36).

The raw gaze data comprised 7524 pages from 133 participants for Study 1 and 6042 pages from 106 participants for Study 2. Pages with fewer than 3 fixations or a reading time under 2 seconds per page were excluded as unread. The cutoff of 3 fixations was due to 3 being the minimum number of instances required to compute some of the summary metrics. For Study 1, 1328 pages (17.7%) were dropped due to meeting the unread criterion, leaving 6196 pages. A further 370 pages corresponding to concept 14 were dropped due to having no self-explanation score. This remaining data was from 131 participants with valid gaze data for an average of 82% of pages. For Study 2, 642 pages were unread (10.6%) resulting in gaze data from 5397 pages remaining, with each participant having valid gaze data for on average 89% of pages. Averaging over pages to yield concept-level features resulted in 1612 observations of each of the 7 concept-level gaze features for Study 1, and 1531 for Study 2.

Summary statistics of the gaze features are shown in Table 2, and correlations between features (and with SE Score) are shown in Table 3. Generally, the magnitudes of the correlations are small. Features were assessed for multicollinearity by computing the variance inflation factor (VIF) and all VIFs were below 2.5 for both datasets which indicates low multicollinearity (Dormann et al., 2013). Some of the correlations between features, and with SE score, are in opposite directions between the two studies. However, in these cases the absolute magnitude of the correlation is very small, ultimately suggesting that they might not be very reliable correlations when considered individually. Further, the studies are not identical replications (e.g., self-explanations in Study 2 were triggered by a mind-wandering detector and were interspersed with reading whereas reading was uninterrupted in Study 1). Thus, it is plausible that the zero-order correlations would vary by study.

4.5 Machine learning

4.5.1 Main self-explanation (SE) models.

We fit supervised machine learning models to predict the expert-scored SE scores from the concept-level gaze features. Models were fit using the *caret* package (version 6.0-86, Kuhn, 2008) in *R* (version 3.6.3, 2014). All models were fit with 4-fold cross validation, meaning that for each training iteration data from 75% of the participants were used to train the model, but performance was assessed on the held-out portion. Assignment to folds was constrained such that all observations from a given participant were in the same fold. This participant-level cross validation ensures the model generalizes to unseen participants. For each model this process was repeated over 100 runs, each with a different partitioning of participants into folds.

Models were ordinary least-squares linear regression, or random forest regression. Random forests use an ensemble of decision trees, each modeling random subsets of the data (both in terms of features and samples), the predictions of which are averaged over all the trees in the 'forest' (Breiman, 2001). Unlike the linear regression model, random forests are capable of modeling nonlinearities and interactivity between features. We used forests with 100 trees, and on each of the 100 runs described above, hyperparameter tuning over all possible values of *mtry* (the number of features selected at each branching point, with possible values from 1 to 7). Linear regression models are incapable of modeling such interactions between features, and the standard linear model we used has no hyperparameters.

Models were fit and evaluated on data from 14 concepts (out of the 15 – see above) for Study 1. By design of the intervention, some of the concepts for Study 2 had gaze data but no corresponding SE (9.6 concepts per participant on average). For Study 2, the model was used to generate estimates of SE scores on all concepts (i.e., even for pages without an actual SE). Specifically, the model parameters of the median-performing model from the fold where each participant was held out were used to generate predictions for the remaining concepts with no ground-truth SE score for a given participant. This preserves participant-level independence between training and testing data.

4.5.2 Comparison models

Several supplementary models were also fit to aid in interpreting the main model.

Shuffled model: Variance in the SE scores across the dataset could arise from two main sources: difference in comprehension between concepts, and between participants. To assess the degree to which the model captured within-participant variability in performance, we also fit models to a shuffled surrogate dataset where the SE scores were shuffled with respect to the concept-level gaze features within each participant. This maintained the participant-wise mean and variance in scores while breaking the link between scores on specific pages and gaze features. Note that the effectiveness of this shuffling procedure depends on the within-subject variability in SE scores (i.e., shuffling will have no effect if there is little to no variability). To better determine the effect of shuffling labels on the resulting model, for each run we computed the mean participant-level absolute difference in SE score before and after shuffling. The dataset was then median-split on the participant-level mean absolute error (MAE) and the model correlation was separately evaluated for each split, for both the main and shuffled models. We would expect that if the model captured within-participant variability, the shuffled model would have a lower correlation for the *high MAE* split (where shuffling was more successful because there was a large difference between pre- and post- shuffled scores) than for the *low MAE* split. Alternatively, if the

shuffled model did not have a lower correlation, this would indicate that the model depends predominantly on between-participant differences in gaze and comprehension.

Reading time model: We also fit models on just the reading time feature for the purposes of measuring the importance of gaze features in contributing to comprehension estimates.

Shallow-online model: A previously published model of shallow comprehension using data from an earlier study (D'Mello et al., 2020) was designed to predict page-level accuracy (correct or incorrect) of responses to shallow multiple-choice questions presented during reading based on the gaze data on the corresponding page. This model was a random forest on the same 7 gaze features and reading time used in the present paper. Concept-level gaze features from Study 1 and Study 2 were input to this model to generate shallow, online comprehension predictions on the present data.

Mind-wandering model: We also use a mind-wandering detector (Faber et al., 2018) consisting of a support vector machine classifier, trained on instances of self-caught mind-wandering during computerized reading from another earlier study. This model was designed to compute predictions of mind-wandering probability at the page-level using 62 summary statistics derived from 8-second windows of gaze data obtained in the window from 11 to 3 seconds before the mind-wandering self-report. This model provided page-level estimates of mind-wandering, which we subsequently aggregated to the concept-level.

Whereas the shuffled and reading time models were trained on the present data, the shallow online and mind-wandering models were trained on data from a different set of participants but on the same text.

4.5.3 Accuracy metrics

For all models, (cross-validated) predicted and actual SE scores for each concept were averaged to the participant-level (to improve reliability) and their correlation was computed as the measure of model accuracy. This was done for all 100 runs and the model with a median correlation in each study was taken as the representative model. We used Meng's method for comparing correlation coefficients which gives a z-statistic and associated p-value (Meng et al., 1992). In addition, to ascertain the degree of divergence of the SE model from the comparison models from other studies, we also computed participant-averaged model probabilities derived from the shallow-online model, and the mind-wandering model.

5 Results

We fit both random forest (RF) and linear regression (LR) models to predict SE scores. The LR model was overall superior when considering our research questions, so we focus solely on the LR results below (A comparison of the two model types is presented in Section 5.5 and detailed RF results are in Supplemental Materials).

5.1 RQ1: How accurately can self-explanations be modeled from eye gaze?

5.1.1 How do the main models compare to reading-time and shuffled-surrogates?

SE model results are summarized in Table 4 alongside the reading-time-only models. At the participant level, SE estimates from the main model were moderately correlated with human-scored SEs (r = .322, r

= .354 for Studies 1 and 2[1] respectively). Figure 1 shows the distributions of predicted and actual scores averaged at the user level, where we note that the predicted scores were more "peaky" and less variable than the actual scores. Indeed, the standard deviation of predicted SE score was .06 for both studies, but .20 and .24 for actual scores.

The reading-time-only model did not predict SE scores in Study 1, but was a significant predictor for Study 2 with a marginally lower correlation than the main model: r = .254 vs. .354, z =1.91, p = .056).

The model fit to features shuffled within participants was used to determine the degree of within-participant variability captured by the model. To better determine the effect of shuffling labels on the resulting model, participants were split by median change in SE score upon shuffling: this corresponds to low and high effectiveness of shuffling[2]. For participants in the *low MAE* cases, where shuffling was less effective, the shuffled models achieved correlations similar to the main, unshuffled models (Table 5; shuffled and unshuffled correlations respectively, Study 1: rs = .311, .307; Study 2: rs = .400, .390). However, for the *high MAE cases* where shuffling was more effective, the shuffled model had a much lower (and non-significant) correlation (shuffled and unshuffled correlations respectively, Study 1: rs = .171, .304; Study 2: rs = .189, .313) . This indicates that the relationship between gaze and comprehension learnt by the main model captures within-subject variability rather than just between-subject individual differences.

5.1.2 Are the models biased by experimental condition?

To address the possible confounding effect of the experimental manipulations (Section 2.3), we separately computed model correlations for the Happy (520 instances, N=44) Neutral (554 instances, N=42) and Sad (538 instances, N=45) affect conditions for Study 1; and the Intervention (254 scored SEs; N=53) and Yoked Control (257 scored SEs; N=53) groups in Study 2. For Study 1, the main model only predicted SE scores for the Neutral (r = .453, p = .003) and Sad conditions (r = .334, p = .025); but not for the Happy condition (r = .139, p = .369). To address this, we re-fit the model on data from the Neutral and Sad conditions only (i.e., dropping the Happy condition), but because results were comparable with the full dataset, (r = .358, p = .001), we report the original model here. For Study 2, correlations were significant for the Intervention (r = .288, p = .037) and Control (r = .426, p<.001) groups, and the correlation did not differ between groups (z =-.80, p = .43), suggesting the main model was not simply fitting to differences between the two groups.

5.1.3 What are the predictive features?

Standardized coefficients of the linear regression models are shown in Table 6; these reflect the relative importance of each feature where one standard deviation change in the coefficient corresponds to a change in the SE score equal to the coefficient. Across both studies, increased saccade distance, reading time and fixations per word were all indicative of higher SE scores. Fixation duration was negatively

---

[1]Correlations were similar when restricted to concepts with human-scored SEs for Study 2 (r =0.403; <0.001; compared to main model z =-.832, p = .405) so we focus on all-concept predictions for Study 2 in the analyses.

[2]For Study 1, the mean absolute error (MAE) upon shuffling was .22 for the *low MAE* split and .44 for the *high MAE* split; the corresponding values for Study 2 were .08 and .38 respectively.

associated with SE score in Study 1 but positively to a small extent in Study 2. In terms of features capturing gaze path patterns, more horizontal saccades were associated with better predicted comprehension, and the proportion of regressions contributed positively to predicted SE score to a small degree. Fixation dispersion was positively associated with SE score for Study 1 only. We show example gaze paths in Figure 2, for two users reading the same page, but SE scores on the corresponding concept were 1 and 0 respectively; examples were chosen to be representative of the pattern of feature coefficients shown in Table 6. The user with a self-explanation score of 1 exhibited substantially more (but shorter) fixations despite a similar per-word reading time. This participant also fixated throughout the text and had longer saccades. Conversely, the user with an SE score of 0 appears to be reading thoroughly initially, but halfway down the page they exhibit a gaze pattern with sparse fixations that are not well aligned to the flow of the text and many words are skipped. The longer fixation duration indicates difficulties encoding information, and is often associated with mind-wandering.

Still, it is possible that additional gaze features – including those without theoretical precedent – could be informative for modeling deep comprehension. Therefore, we also extracted an additional 76 page-level summary metrics to give a total of 83 features. These included 79 global features comprising descriptive statistics such as median, standard deviation, skew and kurtosis of fixation-, blink- and saccade-level metrics (e.g. duration, velocity) and 4 content-specific features consisting of correlation strength between fixation durations and lexical properties (e.g. word length, frequency). However, these additional features did not improve model performance (linear regression: r = .204, .151 and random forest: r = .321, .283 for studies 1 and 2 respectively).

5.2 RQ2: How does the model compare with other gaze-based models of comprehension?

We compared the shallow and mind-wandering gaze-based predictive models to the main SE model (Table 4). First, we found that the mind-wandering model was negatively correlated with the SE model for Study 1 (r = -.414); but not for Study 2 (r = -.023). The shallow comprehension model correlated moderately strongly with the SE model (r = .242, r = .494 for the two studies respectively). In contrast, the mind-wandering and shallow models did not correlate with one another for either study (rs = -.004, .001). Overall, the pattern suggests that the three models are addressing related, but distinct aspects of reading comprehension.

Second, we found that the mind-wandering model negatively predicted SE scores, with a statistically comparable performance to the SE model (Study 1: z = .63, p = .53; Study 2: z = 1.19, p = .23). The shallow-online model did not significantly predict SE score for Study 1 but was comparable to the SE model for Study 2 (r = .329, p<.001; compared to the SE model z = .27, p = .79). To investigate the incremental predictive validity of the SE model net of these comparison models, we regressed actual SE scores on the predictions of all three models (main SE model, shallow online, mind-wandering). For Study 1, only the SE model predicted SE scores (Table 8) suggesting it was the best predictor, but all three models predicted SE scores for Study 2.

5.3 RQ3: To what extent does the model predict posttest measures of learning and comprehension?

We assessed comprehension via multiple-choice questions for each concept in immediate and delayed (1 week) posttests in both studies. Reading in Study 1 was uninterrupted, and SEs were collected after reading. However, in Study 2, participants were interrupted during reading with an initial SE prompt, upon which they constructed their responses, could re-read the text, and construct an improved SE. For this reason, we did not analyze the posttest performance further as the link between eye movements and later comprehension is disrupted by the multiple intervening processes.

Overall, we found that the SE model scores significantly correlated with all posttest assessments except shallow-immediate, with somewhat stronger correlation for the delayed, inference-level assessments (Table 7). The magnitude of the correlation was lower than with the actual human-scored SEs, which is what could be expected (Shallow-Immediate: $z = 3.84$, Inference-Immediate: $z = 4.64$, Shallow-Delayed: $z = 3.43$, Inference-Delayed: $z = 2.54$; all ps$<.05$). With respect to the comparison models, the mind-wandering model negatively correlated with all posttest scores except Shallow-Immediate and not significantly less than the SE model (z-statistic between .67and 1.31). The reading time only and shallow-online models did not significantly predict posttest scores.

Regression models were also fit to predict user-level posttest scores from the three competing comprehension models (Table 8). Based on the standardized coefficients in Table 8 (which are equivalent to correlation coefficients), we found that including the shallow-online model unsurprisingly reduced the correlation between the SE model and the Shallow-Immediate posttest from .12 to .05. The effect for the SE model also decreased a small amount for the other 3 posttests when including the comparison models, but not enough to render the SE model non-significant (.29 to .24, .31 to .28, and .35 to .29 for Inference-Immediate, Shallow-Delayed and Inference-Delayed respectively). This indicates the SE model uniquely predicts deep and delayed comprehension as intended.

5.4 RQ4: Does the model generalize to new users and reading contexts?

To test the generalizability of the SE model from each study, we generated predictions based on the gaze features from the other study. Specifically, the LR models from the median-performing run (as assessed by user-level correlation with the within-study test data) were used to generate a prediction based on the gaze features from the other study.

Results (see Table 9) indicated that cross-study model predictions correlated with actual SE scores (r = .274, p = .001 for Study 2 model and r = .439, p<.001 for Study 1 model) with a correlations statistically equivalent to the corresponding within-study models (z = 1.06, p = .29 for the Study 1 model and (z = .64, p = .52 for the Study 2 model). Critically, the SE model from Study 2 predicted the same three posttest scores on Study 1 data as the Study 1 SE model.

5.5 Comparison of linear regression versus random forest

We also fit random forest (RF) self-explanation models with detailed results in the Supplement. In Table 10, we provide an overview comparison among the two with respect to our research questions. For RQ1, both models successfully predicted SE scores, but the RF model had a somewhat higher correlation (r = .39) compared to the LR model (r = .32) for Study 1 (though not significantly so: z = -.918, p = .358); the two were equivalent for Study 2. The LR and RF models remained a significant predictor of SE scores after accounting for the alternate models (shallow online and mind-wandering) in both studies (Table 8; Supplement Table 4). In terms of correlations with posttests (RQ3), LR models predicted only three (compared to all four for RF) posttest outcomes, and both LR and RF remained significant predictors of all outcomes except shallow online after accounting for the comparison models. Finally, in terms of generalization across studies, there was an advantage of LR over RF, especially for the Study 2 model predicting Study 1 posttest scores. Finally, although not a quantitative metric, LR models are more interpretable than RF, which is another advantage.

6 Discussion

Our aim was to build automated models of reading comprehension based on summary gaze features that would be potentially feasible to incorporate into comprehension-aware reading interfaces. We extend previous work which has shown that shallow comprehension can be predicted from eye movements, extending this to deep comprehension.

6.1 Main findings

For our first research question (RQ1, *How accurately can self-explanations be modeled from eye gaze?*), we found that predictive models trained on summary statistics of eye movements computed over entire sections of text significantly predict deep comprehension as measured by self-explanations. Furthermore, we report significant model performance for users held out of the training set, indicating the link between eye tracking and comprehension score generalizes across users. Based on the analysis of a comparison model fit after shuffling features within participants, we concluded that our model does predict *within-user* variations in comprehension, which is a prerequisite for using the model to monitor reading comprehension fluctuations in real time.

For RQ2 (*How does a self-explanation model compare with other gaze-based models of comprehension?*), the predictions of the self-explanation model were moderately correlated with the mind-wandering and shallow online comprehension models. It also remained a significant predictor of SE scores and posttest outcomes after controlling for the comparison models. Overall, this provides evidence that the self-explanation model indexes related but distinct information held in eye movement features, beyond these comparison models.

For RQ3 (*To what extent does the model predict posttest measures of learning and comprehension?*), we found that the SE model predicted independent comprehension outcomes, specifically multiple-choice assessments of inference-level comprehension taken after reading and both shallow and inference comprehension after a one-week delay. It was a particularly strong predictor of inference-level comprehension after a delay, thereby achieving its stated aim.

Finally with respect to RQ4 (*Does the model generalize to new users and reading contexts?*), the Study 1 model trained on uninterrupted reading predicted self-explanation scores on Study 2, which was undertaken with a different reading paradigm where reading was interrupted by comprehension assessments after every few pages. Furthermore, the model trained on this interrupted-reading dataset predicted the self-explanation scores and posttest outcomes from Study 1.

We examined the model coefficients to better examine how eye gaze is associated with comprehension outcomes. This analysis comes with the caveat that little theoretical weight can, or should, be given to individual gaze features: it is only in tandem that the features are informative of comprehension, as evidenced by the very low correlations between individual features and self-explanation score (Table 3). Nevertheless, we tentatively identify patterns of effects across multiple variables which have previously been shown to correspond to particular process variables. In particular, previous research demonstrates that fewer, longer fixations indicate mind-wandering (Bixler & D'Mello, 2015; Faber et al., 2018; Reichle et al., 2010) which is the opposite pattern to what was found to predict good comprehension in Study 1, where more frequent but shorter fixations predicted comprehension, suggesting that the models are partly capturing attentional focus. In contrast, greater fixation duration was weakly associated with better comprehension in Study 2, but other features had a greater effect on comprehension outcomes.

In Study 2, we found that longer saccades coupled with a greater horizontal proportion were associated with better deep comprehension. This is the opposite pattern to D'Mello et al. (2020), where fewer fixations, longer saccades and greater horizontal saccade proportion were associated with worse shallow comprehension, a finding attributed to skim reading. Skimming has been shown to result in impaired comprehension (Rayner et al., 2012; Strukelj & Niehorster, 2018), but can be an adaptive strategy for

extracting the key meaning from text when reading under time pressure (Duggan & Payne, 2009) where skimming can improve gist comprehension at the expense of surface detail (Masson, 1982). This raises the question of whether those with better comprehension scores in our study were successfully deploying skim reading as a strategy, despite a lack of explicit time limit in the procedure. In particular, Study 2 participants may have utilized such a strategy in anticipation of the self-explanation prompts. In Study 1, self-explanations were collected after reading, where, prominently, the number of fixations per word were approximately double those for Study 2 (see Table 2), despite the same reading material.

Another important aspect pertains to whether the models captured within- or between- user variability. If the model were predominantly fitting between-user variance in comprehension, we would expect little impact on performance when disrupting the link between instance-level features and comprehension scores, while keeping user-level averages of features and labels the same. However, we found that where users had greater variability in their SE scores, shuffling scores in this way disrupted (but didn't completely destroy) the model's ability to predict comprehension, suggesting that the models do capture within-user patterns. Shuffling reduced the models' accuracy by 44% and 40% in the two studies respectively, but did not eliminate it altogether, suggesting that the models are to some extent indexing individual differences in reading fluency, working memory, and reading strategies, which comprise some of the many factors that can affect both eye movement properties and reading comprehension outcomes (Kuperman et al., 2018). For example, longer saccades (which were predictive of comprehension in our study) have been associated with both greater working memory span (Luke et al., 2018; an individual difference measure) but also with skim-reading for gist comprehension (Strukelj & Niehorster, 2018; a within-subject reading strategy). It is not possible to disentangle individual differences in reading ability from the dynamics of comprehension in the present data. It is also an important theoretical question as to whether the two can in fact be meaningfully disentangled outside of experimental paradigms specifically designed for this purpose.

In sum, our work contributes to theories of reading comprehension by showing that gaze features can predict deep comprehension, which is a step towards integrating low-level models of eye movements (Rayner, 1998) with higher-level models of comprehension (Kintsch, 1998).

6.2 Applications

This work is an important proof of concept that deep comprehension can be predictable from low-level, summary metrics of eye movements during reading, which has a number of applications. In the educational domain, deep comprehension could be tracked by gaze as an alternative to online learning assessments, given that interpolated assessments are known to affect comprehension more generally (Roediger and Karpicke 2006), and even specifically as in the case of self-explanations inducing inferencing during reading (Ozuru et al., 2010). Another use of the model could be in supplementing or replacing offline reading assessments, which can be anxiety-provoking and render deep comprehension more difficult to attain for some students (Calvo & Carreiras, 1993).

Beyond assessment, such models can inform real-time adaptations in textual interfaces contingent upon the user's understanding or comprehension ability. To this point, the linear regression model can be embedded in applications that require real-time responses because the two main steps – feature extraction and prediction generation – can be done in linear or constant time. Linear regression models can also be used in applications where model interpretability is critical.

We also found that the models explained approximately 10-20% of the variance in self-explanation scores (square of the correlation coefficients), which should be considered in designing effective interventions during reading. Specifically, given the possibility of prediction errors, any intervention should be carefully chosen to support comprehension generally and flexibly. One such example could be timing self-explanation prompts when the predicted SE score is low, which can trigger retrieval, error-monitoring and elaboration in the user (McNamara, 2004, Bielaczyc et al., 1995, Chi et al., 1994), which can be beneficial regardless of whether the model was correct in its assessment.

With respect to scaling up, it is advantageous to develop models which use eye-tracking features that can be obtained from a variety of equipment including commercial, off-the-shelf systems (Hutt et al., 2019), webcams (Robal et al., 2018), and even front-facing cameras in mobile devices (Krafka et al., 2016). At the same time, the quality of eye tracking data from these approaches varies widely (Niehorster et al., 2018; Robal et al., 2018). Therefore, we used a remote research-grade eye tracker in the present studies which does not constrain head movements to preserve ecological validity. We also only focused on a small set of six summary metrics of global eye gaze features (e.g., number of fixations; mean fixation durations) based on the empirical literature rather than individual fixations and saccades aligned to specific words, which are difficult to reliably compute using consumer-grade sensing. Thus, we expect our approach can be replicated with more cost-effective but less accurate eye tracking devices.

6.3 Limitations and future work

Like all studies, ours have limitations. For one, we provided evidence of model generalization between independent groups of users in two separate studies with differences in the reading procedure, yet the same expository text was used for both. We know that a multitude of text characteristics have a substantial effect on eye movements (Cook & Wei, 2019; Rayner & Duffy, 1986). Therefore, it would be desirable to test the model on different texts at different difficulty levels (Feng et al., 2013), genres (e.g., narrative versus expository texts), and also to use different presentation characteristics such as how the text is displayed (e.g., section-length effects; Forrin et al., 2018). This would be an imperative step before the comprehension model could realistically be deployed in real-world reading interfaces.

Broadening the training data to include multiple diverse texts would also likely increase the within-user variance in comprehension, thereby maximizing the chances of finding gaze-based signatures of comprehension net of individual differences. Relatedly, including quantitative measures of text characteristics in the model could boost performance and generalizability, by allowing the gaze-comprehension link to be conditioned on the reading content in a nuanced way. Only rarely do studies measure both text characteristics and individual differences in their influence on reading comprehension, but notable exceptions from Kuperman and colleagues (Kuperman et al., 2018; Kuperman & Dyke, 2011) found a benefit to including text properties. Similarly, prior knowledge on the topic of the text would be an interesting covariate to investigate, as it might alter strategic reading behaviors.

Another crucial issue to consider is the diversity of users included in the training sample. The present studies used university students as participants; this is a relatively uniform population of young, skilled readers despite the students being from three universities across the US. The linear relationships between gaze and comprehension may not extrapolate to less-skilled readers, or across the lifespan. The fact that we used comparison models (shallow, mind-wandering) trained on different participants may also contribute to their distinctness from the SE model and poorer capability to predict posttest scores

(although our cross-validation scheme aimed to avoid overfitting on our training data and we provide evidence of generalizability across studies).

Future work elaborating on the models may also consider a wider range of outcome assessments for training and evaluating the model. For example, we can envisage user-interface applications of a comprehension model where metacognitive and affective components of the reading process (such as ease, frustration, and self-assessed understanding) are more relevant outcomes than objective measures of comprehension (such as used here).

In terms of the type of machine-learned model, we ultimately focused on a linear regression model over a random forest. These models performed similarly, although the linear model generalized across studies notably better. It is likely that the relatively small size of the dataset on which the model was trained resulted in the random forest model overfitting to the data from each study. In addition, the fact that the response variable (SE score) is bounded between 0 and 1 results in a restriction of range which violates the assumption of normality of the error term, which can be a problem for performing hypothesis tests on the coefficients, although we did not perform such tests. Nevertheless, to address this issue, we did also run beta regression models (which are designed for responses bound by 0 and 1; Cribari-Neto & Zeileis, 2010), but these resulted in similar results to the linear regression models, so we did not report them here.

Finally, we used a research-grade eyetracker, albeit with a deliberate choice to use coarse features which do not rely on high spatial or temporal precision, but it is critical to verify that our findings replicate when using commercially available, inexpensive eye-tracking.

6.4 Conclusion

We show that gaze measures can be used to infer reading comprehension while users read a long, connected, and complex text. To our knowledge, this is the first work to demonstrate a reliable and generalizable link between eye gaze and self-explanations, which measures comprehension at deeper levels by encompassing multiple components of reading such as inferencing, elaboration, and prior knowledge activation (McNamara, 2004; McNamara et al., 2007, Ozuru et al., 2013). Crucially, we show that our model of eye movements during reading can still predict inference-level comprehension a week later. Our findings have implications for the potential of gaze-based tracking of ongoing reading comprehension, which could be used for developing personalized and adaptive reading interfaces.

Table 1. *Example self-explanation prompt and example responses given in the study alongside expert-judged scores*

Text: "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 eleven thousand holes in the bottom, any one of which is large enough to allow an ordinary pin to pass through. This experiment also illustrates how difficult it is to write real and perfect nonsense. You may remember one of the stories in Lear's book of Nonsense Songs. They went to sea in a sieve, they did, In a sieve they went to sea… They sailed away in a sieve, they did, in a sieve they sailed so fast, With only a beautiful pea-green veil, Tied with a rib and by way of a sail, To a small tobacco-pipe mast;" And so on. You see that it is quite possible to go to sea in a sieve--that is, if the sieve is large enough and the water is not too rough and that the above lines are now realized in every particular."

Prompt: How does "going to sea in a sieve" seem possible based on the authors demonstration?

Keywords: large, small, large enough, small enough, force, through

Example Responses	**Score**
"*If the holes of the sieve are small enough such that the elastic skin of water could form across each of them, and if the sea was fairly calm, and of course if the sieve itself was large enough to hold you (i.e. displace enough water such that the weight of the water is at least equal to the combined weight of you and the sieve).*"	1
"The water does not enter the sieve as long as it is not wetted or disturbed, so it should act like a normal boat."	0.5
"he said it works so ill just agree with him."	0

Table 2: *Mean (standard deviation) for concept-level gaze features and self-explanation scores. Posttest scores were averaged at the participant-level before summary statistics were computed.*

	Study 1	**Study 2**
Number of concepts	n=1612	n=1531
Fixations per word	2.21 (1.04)	1.06 (.65)
Fixation duration (ms)	264.33 (42.89)	261.68 (41.34)
Regression fixation proportion	.12 (.06)	.12 (.05)
Saccade distance (pixels)	254.51 (49.82)	258.57 (44.02)
Horizontal saccade proportion	.94 (.10)	.93 (.08)
Fixation dispersion	.44 (.07)	.43 (.05)
Reading time per word (s)	.25 (.08)	.27 (.14)
Self-explanation score	.41 (.39)	.62 (.36)
Shallow-Immediate	.56 (.50)	.69 (.37)
Deep-Immediate	.43 (.34)	.47 (.50)
Shallow-Delayed	.46 (.50)	.52 (.41)
Deep-Delayed	.49 (.50)	.41 (.49)

Table 3: *Pearson correlations for Study 1 and Study 2 [shown in brackets] among concept-level gaze features, and between features and SE scores, including concepts with no self-explanation score for Study 2. To save space, features are referred to by numbers in the header row, and the numbering can be read from the first column.*

Feature	SE score	1	2	3	4	5	6
1. Fixations per word	.11 [.21]						
2. Fixation duration	-.08 [.10]	.20 [.16]					
3. Regression fixation proportion	.05 [.05]	-.03 [.19]	-.09 [-.10]				
4. Saccade distance	.06 [-.01]	-.30 [-.29]	-.38 [-.34]	.20 [.06]			
5. Horizontal saccade proportion	.05 [.10]	.46 [.32]	.01 [-.13]	-.26 [.07]	-.15 [-.18]		
6. Fixation dispersion	.01 [-.17]	-.37 [-.26]	-.17 [-.17]	.01 [-.17]	.06 [.21]	-.33 [-.07]	
7. Reading time per word	.03 [.27]	.42 [.41]	.56 [.47]	.07 [.07]	-.41 [-.25]	.05 [-.18]	-.20 [-.38]

Table 4: *Correlations [with 95% confidence intervals] between SE scores and model predictions for the median iteration of the model.*

	Study 1 correlations (N=131)	**Study 2 correlations (N=106)**
SE models		
Main model – Linear regression	**.322*** [.159, .468]**	**.354*** [.176, .511]**
Human-scored concepts only	-	.403*** [.230, .551]
Reading time only	-.104 [-.270, .069]	.254** [.066, .424]
Alternative models of comprehension		
Shallow-online model	.063	.329***
Mind-wandering model	-.265**	-.200*

*$p < .05$. **$p < .01$. ***$p < .001$.

Table 5: Correlations computed separately for the lowMAE (shuffling ineffective) and highMAE (shuffling effective) splits for the main and shuffled models.

	Model	**Study 1**	**Study 2**
Low MAE	main	0.307* [0.068, 0.512]	0.390** [0.133, 0.597]
	shuffled	0.311* [0.073, 0.516]	0.400** [0.146, 0.605]
High MAE	main	0.304* [0.067, 0.509]	0.313* [0.047, 0.538]
	shuffled	0.171 [-0.074, 0.396]	0.189 [-0.086, 0.437]

*$p < .05$. **$p < .01$. ***$p < .001$.

Table 6: *standardized feature coefficients for the linear regression models. Coefficients and estimated 95% confidence intervals are shown for the median-performing model run.*

	Study 1		**Study 2**	
Feature	*Mean*	*95% CI*	*Mean*	*95% CI*
Fixations per word	.052	[.028, .076]	.024	[-.008, .056]
Mean fixation duration	-.044	[-.068, -.021]	.013	[-.018, .045]
Regression fixation proportion	.010	[-.010, .030]	.011	[-.021, .043]
Mean saccade distance	.032	[.010, .053]	.042	[.006, .078]
Horizontal saccade proportion	.011	[-.012, .033]	.040	[.011, .069]
Fixation dispersion	.022	[.001, .043]	.000	[-.031, .032]
Reading time per word	.032	[.007, .057]	.069	[.040, .098]

Figure 1: Density (left) and scatter plots (right) of predicted and actual participant-level scores for Study 1 (top) and Study 2 (bottom) respectively.

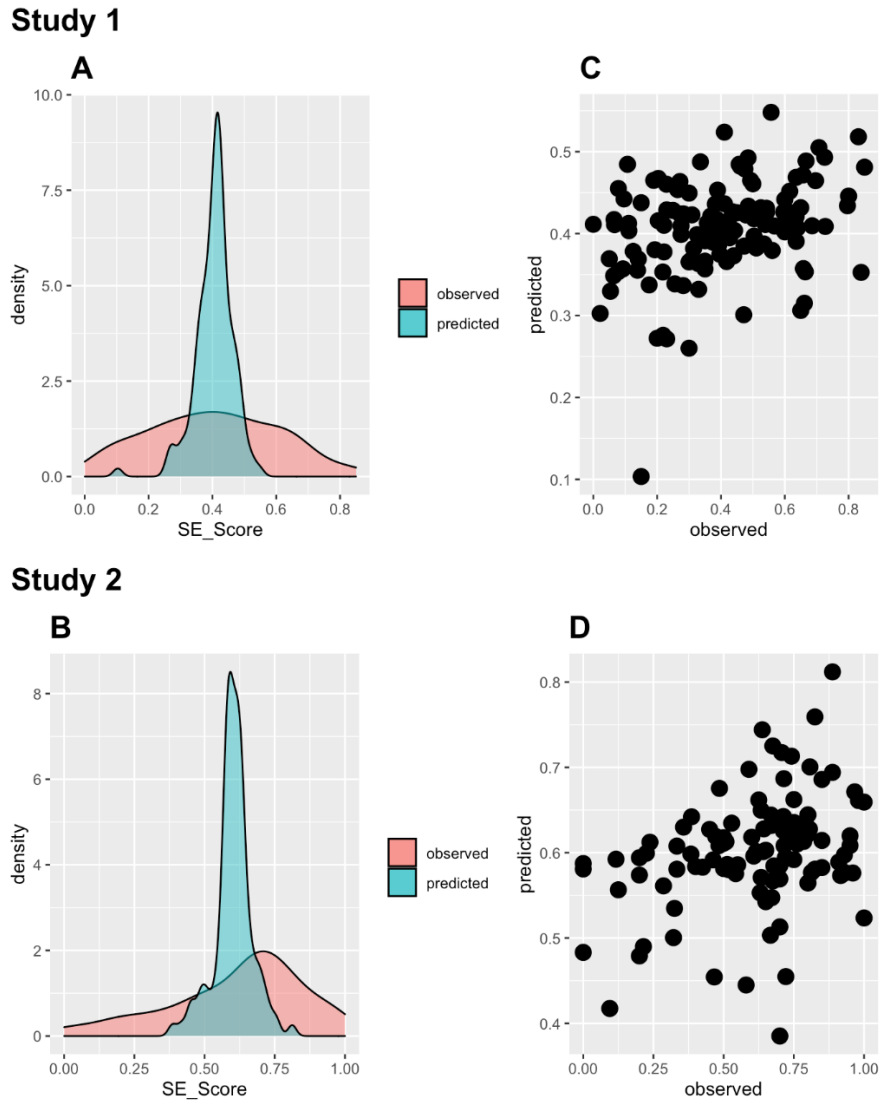
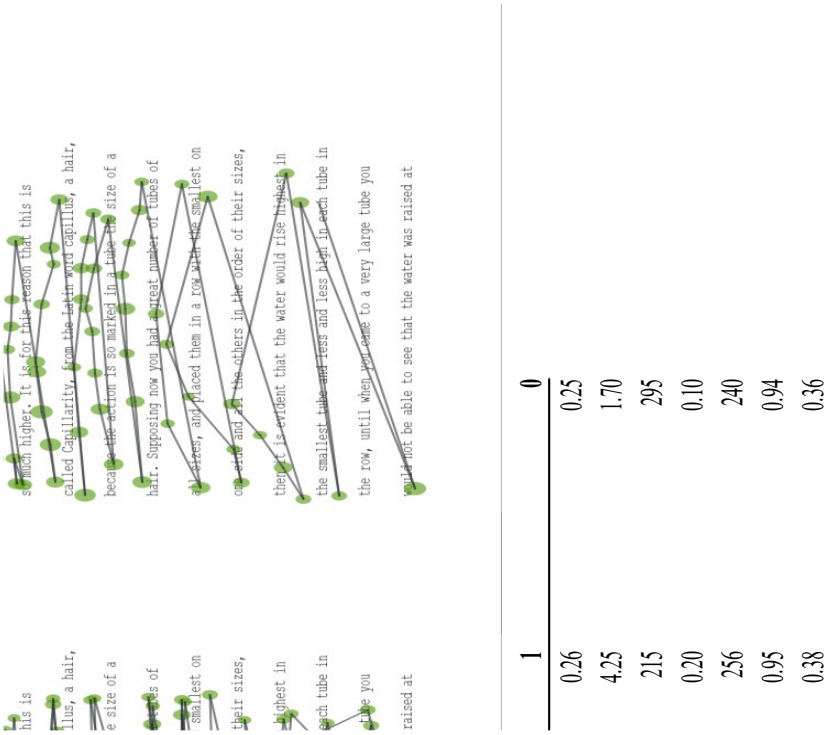


Figure 2. Gaze paths from two example participants reading the same page, with similar reading time per word, but different self-explanation scores. **a** SE score = 1; **b** SE score = 0. Fixations are shown in green, with the area proportional to fixation duration, and are connected with lines representing saccades.



1	0
0.26	0.25
4.25	1.70
215	295
0.20	0.10
256	240
0.95	0.94
0.38	0.36

Table 7: *Correlation between model predictions and post-test comprehension scores for Study 1. MW – mind-wandering. *p<.05 **p<.01 ***p<.001*

	SE score (actual)	SE model	Reading time model	Shallow-online model	MW model
Shallow-Immediate	.486***	.117	-.129	.106	-.135
Inference-Immediate	.675***	.288***	-.102	.038	-.225**
Shallow-Delayed	.603***	.307***	-.118	.061	-.187*
Inference-Delayed	.574***	.354***	.002	.132	-.244**

Table 8: *Standardized (Beta) coefficients for participant-level regression models controlling for other models (*p<.05 **p<.01 ***p<.001). The self-explanation model is shown in bold.*

	Study 1 (N=131 immediate, N=115 delay)					**Study 2 (N=106)**
	SE Score	Shallow-Immediate	Inference-Immediate	Shallow-Delayed	Inference-Delayed	SE Score
SE model	**.26****	**.05**	**.24***	**.28****	**.29****	**.25***
Shallow-online model	0	.09	-.02	-.01	.06	.21*
MW model	-.16	-.12	-.13	-.08	-.14	-.19*
R^2 / R^2 adjusted	.125 / .104	.031 / .008	.097 / .076	.100 / .076	.143 / .120	.195 / .171

Table 9: *Generalization of models between studies. (*p<.05 **p<.01 ***p<.001).*

		Cross-study correlation (Generalization)	Within-study correlation (for comparison)
Training Data	*Outcome*	**Study 2 model**	*Study 1 model*
Study 1	Self-explanation scores	**.274****	.322***
	Shallow-Immediate	**-.019**	.117
	Inference-Immediate	**.191***	.288***
	Shallow-Delayed	**.251****	.307***
	Inference-Delayed	**.324*****	.354***
		Study 1 model	Study 2 model
Study 2 scores	Self-explanation	**.439*****	.354***

Table 10: Comparison of random forest (RF) and linear regression (LR) models corresponding to the four research questions (RQs). Comparisons (<,> and =) are based on numerical comparison of multiple statistics capturing model performance on each metric; i.e. correlation coefficients for RQ1, RQ3 and RQ4; and coefficients for the SE model in the multi-model regressions for RQ2.

	RQ1: Convergent validity	**RQ2: Discriminant validity**	**RQ3: Predictive validity**	**RQ4: Generalization**
Study 1	RF>LR	LR=RF	RF>LR	LR>RF
Study 2	LR=RF	LR=RF	-	LR>RF

7 References

Agrawal, S., Norman, G. R., & Eva, K. W. (2012). Influences on medical students' self-regulated learning after test completion. *Medical Education*, *46*(3), 326–335. https://doi.org/1.1111/j.1365-2923.2011.04150.x

Ahn, S., Kelton, C., Balasubramanian, A., & Zelinsky, G. (2020). Towards Predicting Reading Comprehension From Gaze Behavior. *ACM Symposium on Eye Tracking Research and Applications*, 1–5. https://doi.org/10.1145/3379156.3391335

Alexander, P. A., & The Disciplined Reading and Learning Research Laboratory (2012). Reading Into the Future: Competence for the 21st Century. *Educational Psychologist*, *47*(4), 259–280. https://doi.org/10.1080/00461520.2012.722511

Amini, D., Anhari, M. H., & Ghasemzadeh, A. (2020). Modeling the relationship between metacognitive strategy awareness, self-regulation and reading proficiency of Iranian EFL learners. *Cogent Education*, *7*(01), 1787018. https://doi.org/10.1080/2331186x.2020.1787018

Andrews, S. (2012). Individual differences in skilled visual word recognition and reading. *Visual Word Recognition*, *2*, 151–172.

Ashby, J., Rayner, K., & Clifton, C. (2005). Eye movements of highly skilled and average readers: differential effects of frequency and predictability. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *58*(6), 1065–1086. https://doi.org/10.1080/02724980443000476

Bargary, G., Bosten, J. M., Goodbourn, P. T., Lawrance-Owen, A. J., Hogg, R. E., & Mollon, J. D. (2017). Individual differences in human eye movements: An oculomotor signature? *Vision Research*, *141*, 157–169. https://doi.org/10.1016/j.visres.2017.03.001

Bell, L. C., & Perfetti, C. A. (1994). Reading Skill: Some Adult Comparisons. *Journal of Educational Psychology*, *86*(2), 244–255. https://doi.org/10.1037/0022-0663.86.2.244

Biedert, R., Hees, J., Dengel, A., & Buscher, G. (2012). *A robust realtime reading-skimming classifier*. 123. https://doi.org/10.1145/2168556.2168575

Bielaczyc, K., Pirolli, P. L., & Brown, A. L. (1995). Training in Self-Explanation and Self-Regulation Strategies: Investigating the Effects of Knowledge Acquisition Activities on Problem Solving. *Cognition and Instruction*, *13*(2), 221–252. https://doi.org/10.1207/s1532690xci1302_3

Bixler, R., & D'Mello, S. (2015). Automatic gaze-based user-independent detection of mind-wandering during computerized reading. *User Modeling and User-Adapted Interaction*, *26*(1), 33–68. https://doi.org/10.1007/s11257-015-9167-1

Bondareva, D., Conati, C., Feyzi-Behnagh, R., Harley, J. M., Azevedo, R., & Bouchet, F. (2013). Artificial Intelligence in Education, 16th International Conference, AIED 2013, Memphis, TN, USA, July 9-13, 2013. Proceedings. *Lecture Notes in Computer Science*, 229–238. https://doi.org/10.1007/978-3-642-39112-5_24

Booth, R. W., & Weger, U. W. (2013). The function of regressions in reading: backward eye movements allow rereading. *Memory & Cognition*, *41*(1), 82–97. https://doi.org/10.3758/s13421-012-0244-y

Boulanger, D., & Kumar, V. (2019). An Overview of Recent Developments in Intelligent e-Textbooks and Reading Analytics. *ITextbooks@AIED*.

Boys, C. V. (1890). *Soap-bubbles and the forces which mould them*. Society for Promoting Christian Knowledge.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Britt, M. A., Richter, T., & Rouet, J.-F. (2014). Scientific Literacy: The Role of Goal-Directed Reading and Evaluation in Understanding Scientific Information. *Educational Psychologist*, *49*(2), 104–122. https://doi.org/10.1080/00461520.2014.916217

Broadway, J. M., Franklin, M. S., & Schooler, J. W. (2015). Early event-related brain potentials and hemispheric asymmetries reveal mind-wandering while reading and predict comprehension. *Biological Psychology*, *107*, 31–43. https://doi.org/10.1016/j.biopsycho.2015.02.009

Buscher, G., Dengel, A., & Elst, L. van. (2008). CHI '08 Extended Abstracts on Human Factors in Computing Systems, CHI EA '08. *Proceeding of the Twenty-Sixth Annual CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI '08*, 2991–2996. https://doi.org/10.1145/1358628.1358796

Calvo, M. G., & Carreiras, M. (1993). Selective influence of test anxiety on reading processes. *British Journal of Psychology*, *84*(3), 375–388. https://doi.org/10.1111/j.2044-8295.1993.tb02489.x

Carpenter, R. H. S. (2000). The neural control of looking. *Current Biology*, *10*(8), R291–R293. https://doi.org/10.1016/s0960-9822(00)00430-9

Chace, K. H., Rayner, K., & Well, A. D. (2005). Eye Movements and Phonological Parafoveal Preview: Effects of Reading Skill. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *59*(3), 209–217. https://doi.org/10.1037/h0087476

Chi, M. T. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. *Advances in Instructional Psychology*, *5*, 161--238.

Chi, M. T. H., Leeuw, N. D., Chiu, M.-H., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*(3), 439–477. https://doi.org/10.1016/0364-0213(94)90016-7

Chui, M., Manyika, J., Bughin, J., Dobbs, R., Roxburgh, C., Sarrazin, H., Sands, G., & Westergren, M. (2012). *The social economy: Unlocking value and productivity through social technologies*. Mckinsey Global Institute. https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-social-economy#

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A Dual Route Cascaded Model of Visual Word Recognition and Reading Aloud. *Psychological Review*, *108*(1), 204–256. https://doi.org/10.1037/0033-295x.108.1.204

Conati, C., Aleven, V., & Mitrovic, A. (2013). Eye-tracking for student modelling in intelligent tutoring systems. *Design Recommendations for Intelligent Tutoring Systems*, *1*, 227–236.

Cook, A. E., & Wei, W. (2019). What Can Eye Movements Tell Us about Higher Level Comprehension? *Vision*, *3*(3), 45. https://doi.org/10.3390/vision3030045

Copeland, L., & Gedeon, T. (2013). Measuring reading comprehension using eye movements. *2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)*, 791–796. https://doi.org/10.1109/coginfocom.2013.6719207

Copeland, L., Gedeon, T., & Caldwell, S. (2016). Effects of text difficulty and readers on predicting reading comprehension from eye movements. *2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 407–412. https://doi.org/10.1109/coginfocom.2015.7390628

Copeland, L., Gedeon, T., & Mendis, S. (2014). Predicting reading comprehension scores from eye movements using artificial neural networks and fuzzy output error. *Artificial Intelligence Research*, *3*(3). https://doi.org/10.5430/air.v3n3p35

Cribari-Neto F, Zeileis A (2010). Beta Regression in R. *Journal of Statistical Software*, 34(2), 1–24. https://doi.org/10.18637/jss.v034.i02.

Daley, S. G., Willett, J. B., & Fischer, K. W. (2014). Emotional Responses During Reading: Physiological Responses Predict Real-Time Reading Comprehension. *Journal of Educational Psychology*, *106*(1), 132–143. https://doi.org/10.1037/a0033408

Dijk, T. A. van, & Kintsch, W. (1983). *Strategies of Discourse Comprehension*. Academic Press.

D'Mello, S. K. (2016). Giving Eyesight to the Blind: Towards Attention-Aware AIED. *International Journal of Artificial Intelligence in Education*, *26*(2), 645–659. https://doi.org/10.1007/s40593-016-0104-1

D'Mello, S. K., Mills, C., Bixler, R., EDM, N. B., (2017). Zone out no more: Mitigating mind-wandering during computerized reading. *Pnigel.Com*.

D'Mello, S. K., & Mills, C. S. (2021). Mind wandering during reading: An interdisciplinary and integrative review of psychological, computing, and intervention research and theory. *Language and Linguistics Compass*, *15*(4). https://doi.org/10.1111/lnc3.12412

D'Mello, S. K., Southwell, R., & Gregg, J. (2020). Machine-Learned Computational Models Can Enhance the Study of Text and Discourse: A Case Study Using Eye Tracking to Model Reading Comprehension. *Discourse Processes*, *57*(5–6), 1–21. https://doi.org/10.1080/0163853x.2020.1739600

Dong, H. W., Mills, C., Knight, R. T., & Kam, J. W. Y. (2021). Detection of mind-wandering using EEG: Within and across individuals. *PLOS ONE*, *16*(5), e0251490. https://doi.org/10.1371/journal.pone.0251490

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Duggan, G. B., & Payne, S. J. (2009). Text Skimming: The Process and Effectiveness of Foraging Through Text Under Time Pressure. *Journal of Experimental Psychology: Applied*, *15*(3), 228–242. https://doi.org/10.1037/a0016995

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, *20*(6), 641–655. https://doi.org/10.1016/s0022-5371(81)90220-6

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: A Dynamical Model of Saccade Generation During Reading. *Psychological Review*, *112*(4), 777–813. https://doi.org/10.1037/0033-295x.112.4.777

Everdell, I. (2014). Eye Tracking in User Experience Design. *Section 3: Eye Tracking for Specific Applications*, 163–186. https://doi.org/10.1016/b978-0-12-408138-3.00007-8

Faber, M., Bixler, R., & D'Mello, S. K. (2018). An automated behavioral measure of mind-wandering during computerized reading. *Behavior Research Methods*, *50*(1), 1–17. https://doi.org/10.3758/s13428-017-0857-y

Feng, S., D'Mello, S., & Graesser, A. C. (2013). Mind wandering while reading easy and difficult texts. *Psychonomic Bulletin & Review*, *20*(3), 586–592. https://doi.org/10.3758/s13423-012-0367-y

Foroughi, C. K., Werner, N. E., Barragán, D., & Boehm-Davis, D. A. (2015). Interruptions disrupt reading comprehension. *Journal of Experimental Psychology: General*, *144*(3), 704. https://doi.org/10.1037/xge0000074

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. https://doi.org/10.1037/a0030217

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. https://doi.org/10.3758/s13423-011-0109-6

Gerrig, R. J., & O'Brien, E. J. (2005). The Scope of Memory-Based Processing. *Discourse Processes*, *39*(2–3), 225–242. https://doi.org/10.1080/0163853x.2005.9651681

Glass, A. L. (2009). The effect of distributed questioning with varied examples on exam performance on inference questions. *Educational Psychology*, *29*(7), 831–848. https://doi.org/10.1080/01443410903310674

Graesser, A. C., & Bertus, E. L. (1998). The Construction of Causal Inferences While Reading Expository Texts on Science and Technology. *Scientific Studies of Reading*, *2*(3), 247–269. https://doi.org/10.1207/s1532799xssr0203_4

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing Inferences During Narrative Text Comprehension. *Psychological Review*, *101*(3), 371–395. https://doi.org/10.1037/0033-295x.101.3.371

Graesser, A., Ozuru, Y., & Sullins, J. (2009). What is a Good Question? In M. G. M. L. Kucan & L. Kucan (Eds.), *Threads of coherence in research on the development of reading ability* (pp. 112–141).

Grainger, J., & Holcomb, P. J. (2009). Watching the Word Go by: On the Time-course of Component Processes in Visual Word Recognition. *Language and Linguistics Compass*, *3*(1), 128–156. https://doi.org/10.1111/j.1749-818x.2008.00121.x

Hastings, C., Mosteller, F., Tukey, J. W., & Winsor, C. P. (1947). Low Moments for Small Samples: A Comparative Study of Order Statistics. *The Annals of Mathematical Statistics*, *18*(3), 413–426. https://doi.org/10.1214/aoms/1177730388

Hutt, S., Mills, C., White, S., Donnelly, P. J., & D'Mello, S. (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*.

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. https://doi.org/10.1007/s11257-019-09228-5

Hutt, S., Mills, C., Bosch, N., Krasich, K., Brockmole, J., & Dmello, S. (2017). "Out of the Fr-Eye-ing Pan." Towards Gaze-Based Models of Attention during Learning with Technology in the Classroom. 94–103. https://doi.org/10.1145/3079628.3079669

Hyönä, J., Lorch, R. F., & Kaakinen, J. K. (2002). Individual Differences in Reading to Summarize Expository Text: Evidence From Eye Fixation Patterns. *Journal of Educational Psychology*, *94*(1), 44–55. https://doi.org/10.1037/0022-0663.94.1.44

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354. https://doi.org/10.1037/0033-295x.87.4.329

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*(2), 228–238. https://doi.org/10.1037/0096-3445.111.2.228

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, *8*(4), 441–480. https://doi.org/10.1016/0010-0285(76)90015-3

Just, M. A., & Carpenter, P. A. (1987). *The psychology of reading and language comprehension.* Allyn & Bacon.

Kendeou, P., & Broek, P. van den. (2010). The effects of prior knowledge and text structure on comprehension processes during reading of scientific texts. *Memory & Cognition*, *35*(7), 1567–1577. https://doi.org/10.3758/bf03193491

Kinnunen, R., & Vauras, M. (1995). Comprehension monitoring and the level of comprehension in high- and low-achieving primary school children's reading. *Learning and Instruction*, *5*(2), 143–165. https://doi.org/10.1016/0959-4752(95)00009-r

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 163–182. https://doi.org/10.1037/0033-295x.95.2.163

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press. https://doi.org/10.1016/s0378-2166(99)00090-9

Komogortsev, O., Karpov, A., & Holland, C. (2015). Oculomotor Plant Characteristics: The Effects of Environment and Stimulus. *IEEE Transactions on Information Forensics and Security*, *11*(3), 621–632. https://doi.org/10.1109/tifs.2015.2503263

Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., & Torralba, A. (2016). Eye Tracking for Everyone. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2176–2184. https://doi.org/10.1109/cvpr.2016.239

Krejtz, K., Sharif, B., Kelton, C., Wei, Z., Ahn, S., Balasubramanian, A., Das, S. R., Samaras, D., & Zelinsky, G. (2019). Reading detection in real-time. *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, 43. https://doi.org/10.1145/3314111.3319916

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v028.i05

Kuperman, V., & Dyke, J. A. V. (2011). Effects of individual differences in verbal skills on eye-movement patterns during sentence reading. *Journal of Memory and Language*, *65*(1), 42–73. https://doi.org/10.1016/j.jml.2011.03.002

Kuperman, V., Matsuki, K., & Dyke, J. A. V. (2018). Contributions of Reader- and Text-Level Characteristics to Eye-Movement Patterns During Passage Reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(11), 1687–1713. https://doi.org/10.1037/xlm0000547

Lai, S., Liu, J., Niu, B., Tian, H., & Wu, F. (2019). Combining Facial Behavioral Cues, Eye Movements and EEG-based Attention to Improve Prediction of Reading Failure. *2019 International Joint Conference on Information, Media and Engineering (IJCIME)*, *00*, 485–489. https://doi.org/10.1109/ijcime49369.2019.00103

Larsen, D. P., Butler, A. C., & III, H. L. R. (2013). Comparative effects of test-enhanced learning and self-explanation on long-term retention. *Medical Education*, *47*(7), 674–682. https://doi.org/10.1111/medu.12141

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766. https://doi.org/10.1016/j.jesp.2013.03.013

Loboda, T. D., Brusilovsky, P., & Brunstein, J. (2011). *Inferring word relevance from eye-movements of readers*. 175–184. https://doi.org/10.1145/1943403.1943431

Lou, Y., Liu, Y., Kaakinen, J. K., & Li, X. (2017). Using support vector machines to identify literacy skills: Evidence from eye movements. *Behavior Research Methods*, *49*(3), 887–895. https://doi.org/10.3758/s13428-016-0748-7

Luke, S. G., Darowski, E. S., & Gale, S. D. (2018). Predicting eye-movement characteristics across multiple tasks from working memory and executive control. *Memory & Cognition*, *46*(5), 826–839. https://doi.org/10.3758/s13421-018-0798-4

Makowski, S., Jäger, L. A., Abdelwahab, A., Landwehr, N., & Scheffer, T. (2019). *A Discriminative Model for Identifying Readers and Assessing Text Comprehension from Eye Movements*. 209–225. https://doi.org/10.1007/978-3-030-10925-7_13

Martínez-Gómez, P., & Aizawa, A. (2014). *Recognition of understanding level and language skill using measurements of reading behavior*. 95–104. https://doi.org/10.1145/2557500.2557546

Masson, M. E. (1982). Cognitive processes in skimming stories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8*(5), 400–417. https://doi.org/10.1037/0278-7393.8.5.400

Masson, M. E. J. (1983). Conceptual processing of text during skimming and rapid sequential reading. *Memory & Cognition*, *11*(3), 262–274. https://doi.org/10.3758/bf03196973

McDonald, S. A., & Shillcock, R. C. (2003). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science*, *14*(6), 648–652. https://doi.org/10.1046/j.0956-7976.2003.psci_1480.x

McNamara, D. S. (2004). SERT: Self-Explanation Reading Training. *Discourse Processes*, *38*(1), 1–30. https://doi.org/10.1207/s15326950dp3801_1

McNamara, D. S., Vega, M. de, & O'Reilly, T. (2007). Comprehension skill, inference making, and the role of knowledge. *Higher Level Language Processes in the Brain: Inference and Comprehension Processes.*, 233–251.

Meng, X.-L., Rosenthal, R., & Rubin, D. B. (1992). Comparing Correlated Correlation Coefficients. *Psychological Bulletin*, *111*(1), 172–175. https://doi.org/10.1037/0033-2909.111.1.172

Meseguer, E., Carreiras, M., & Clifton, C. (2002). Overt reanalysis strategies and eye movements during the reading of mild garden path sentences. *Memory & Cognition*, *30*(4), 551–561. https://doi.org/10.3758/bf03194956

Metzner, P., Malsburg, T., Vasishth, S., & Rösler, F. (2017). The Importance of Reading Naturally: Evidence From Combined Recordings of Eye Movements and Electric Brain Potentials. *Cognitive Science*, *41*(S6), 1232–1263. https://doi.org/10.1111/cogs.12384

Mills, C., Graesser, A., Risko, E. F., & D'Mello, S. K. (2017). Cognitive Coupling During Reading. *Journal of Experimental Psychology: General*, *146*(6), 872–883. https://doi.org/10.1037/xge0000309

Mills, C., Gregg, J., Bixler, R., & D'Mello, S. K. (2021). Eye-Mind reader: an intelligent reading interface that promotes long-term comprehension by detecting and responding to mind-wandering. *Human–Computer Interaction*, 1–27. https://doi.org/10.1080/07370024.2020.1716762

Miyata, H., Minagawa-Kawai, Y., Watanabe, S., Sasaki, T., & Ueda, K. (2012). Reading Speed, Comprehension and Eye Movements While Reading Japanese Novels: Evidence from Untrained Readers and Cases of Speed-Reading Trainees. *PLoS ONE*, *7*(5), e36091. https://doi.org/10.1371/journal.pone.0036091

Moort, M. L. van, Koornneef, A., & Broek, P. W. van den. (2020). Differentiating Text-Based and Knowledge-Based Validation Processes during Reading: Evidence from Eye Movements. *Discourse Processes*, 1–20. https://doi.org/10.1080/0163853x.2020.1727683

Mousavinasab, E., Zarifsanaiey, N., Kalhori, S. R. N., Rakhshan, M., Keikha, L., & Saeedi, M. G. (2018). Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments*, *29*(1), 1–22. https://doi.org/10.1080/10494820.2018.1558257

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

Niehorster, D. C., Cornelissen, T. H. W., Holmqvist, K., Hooge, I. T. C., & Hessels, R. S. (2018). What to expect from your remote eye-tracker when participants are unrestrained. *Behavior Research Methods*, *50*(1), 213–227. https://doi.org/10.3758/s13428-017-0863-0

Nilsson, M. (2012). *Computational Models of Eye Movements in Reading : A Data-Driven Approach to the Eye-Mind Link*. http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-167403

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. https://doi.org/10.1080/01638531003628809

Ozuru, Y., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing Comprehension Measured by Multiple-Choice and Open-Ended Questions. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *67*(3), 215–227. https://doi.org/10.1037/a0032918

Paquette, L., & Baker, R. S. (2019). Comparing machine learning to knowledge engineering for student behavior modeling: a case study in gaming the system. *Interactive Learning Environments*, *27*(5–6), 1–13. https://doi.org/10.1080/10494820.2019.1610450

Pattamadilok, C., Chanoine, V., Pallier, C., Anton, J.-L., Nazarian, B., Belin, P., & Ziegler, J. C. (2017). Automaticity of phonological and semantic processing during visual word recognition. *NeuroImage*, *149*, 244–255. https://doi.org/10.1016/j.neuroimage.2017.02.003

Price, C. J., Moore, C. J., Humphreys, G. W., & Wise, R. J. S. (1997). Segregating Semantic from Phonological Processes during Reading. *Journal of Cognitive Neuroscience*, *9*(6), 727–733. https://doi.org/10.1162/jocn.1997.9.6.727

R: A language and environment for statistical computing. (3.6.3). (2014). [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

Rajendran, R., Kumar, A., Carter, K. E., Levin, D. T., & Biswas, G. (2018). Predicting Learning by Analyzing Eye-Gaze Data of Reading Behavior. *International Educational Data Mining Society*.

Raney, G. E., Campbell, S. J., & Bovee, J. C. (2014). Using Eye Movements to Evaluate the Cognitive Processes Involved in Text Comprehension. *Journal of Visualized Experiments : JoVE*, *83*, 50780. https://doi.org/10.3791/50780

Rapp, D. N. (2006). The value of attention aware systems in educational settings. *Computers in Human Behavior*, *22*(4), 603–614. https://doi.org/10.1016/j.chb.2005.12.004

Rapp, D. N., & Broek, P. van den. (2005). Dynamic Text Comprehension: An Integrative View of Reading. *Current Directions in Psychological Science*, *14*(5), 276–279. https://doi.org/10.1111/j.0963-7214.2005.00380.x

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422. https://doi.org/10.1037/0033-2909.124.3.372

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye Movements as Reflections of Comprehension Processes in Reading. *Scientific Studies of Reading*, *10*(3), 241–255. https://doi.org/10.1207/s1532799xssr1003_3

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, *14*(3), 191–201. https://doi.org/10.3758/bf03197692

Rayner, K., Pollatsek, A., Ashby, J., & Jr, C. C. (2012). *Psychology of reading*.

Rayner, K., & Reichle, E. D. (2010). Models of the reading process. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 787–799. https://doi.org/10.1002/wcs.68

Rayner, K., Slattery, T. J., & Bélanger, N. N. (2010). Eye movements, the perceptual span, and reading speed. *Psychonomic Bulletin & Review*, *17*(6), 834–839. https://doi.org/10.3758/pbr.17.6.834

Rayner, K., & Well, A. (1996a). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review*, *3*(4), 504–509. https://doi.org/10.3758/bf03214555

Rayner, K., & Well, A. (1996b). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review*, *3*, 504–509. https://doi.org/10.3758/BF03214555

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. https://doi.org/10.1016/j.cogsys.2005.07.001

Reichle, E. D. (2015). Computational Models of Reading: A Primer. *Language and Linguistics Compass*, *9*(7), 271–284. https://doi.org/10.1111/lnc3.12144

Reichle, E. D., Rayner, K., & Pollatsek, A. (1999). Eye movement control in reading: accounting for initial fixation locations and refixations within the E-Z Reader model. *Vision Research*, *39*(26), 4403–4411. https://doi.org/10.1016/s0042-6989(99)00152-2

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E-Z Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, *26*(4), 445–476. https://doi.org/10.1017/s0140525x03000104

Reichle, E. D., Rayner, K., & Pollatsek, A. (2012). Eye movements in reading versus nonreading tasks: Using E-Z Reader to understand the role of word/stimulus familiarity. *Visual Cognition*, *20*(4–5), 360–390. https://doi.org/10.1080/13506285.2012.667006

Reichle, E. D., Reineberg, A. E., & Schooler, J. W. (2010). Eye Movements During Mindless Reading. *Psychological Science*, *21*(9), 1300–1310. https://doi.org/10.1177/0956797610378686

Reichle, E. D., & Reingold, E. M. (2013). Neurophysiological constraints on the eye-mind link. *Frontiers in Human Neuroscience*, *7*, 361. https://doi.org/10.3389/fnhum.2013.00361

Reichle, E. D., Warren, T., & McConnell, K. (2009). Using E-Z reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic Bulletin & Review*, *16*(1), 1–21. https://doi.org/10.3758/pbr.16.1.1

Reingold, E. M., Reichle, E. D., Glaholt, M. G., & Sheridan, H. (2012). Direct lexical control of eye movements in reading: Evidence from a survival analysis of fixation durations. *Cognitive Psychology*, *65*(2), 177–206. https://doi.org/10.1016/j.cogpsych.2012.03.001

Robal, T., Zhao, Y., Lofi, C., & Hauff, C. (2018). Webcam-based Attention Tracking in Online Learning: A Feasibility Study. *23rd International Conference on Intelligent User Interfaces*, 189–197. https://doi.org/10.1145/3172944.3172987

Roda, C., & Thomas, J. (2006). Attention aware systems: Theories, applications, and research agenda. *Computers in Human Behavior*, *22*(4), 557–587. https://doi.org/10.1016/j.chb.2005.12.005

Roediger, H. L., & Karpicke, J. D. (2006). Test-Enhanced Learning. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Roll, I., Aleven, V., McLaren, B. M., & Koedinger, K. R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction*, *21*(2), 267–280. https://doi.org/10.1016/j.learninstruc.2010.07.004

Sanches, C. L., Augereau, O., & Kise, K. (2018). Estimation of reading subjective understanding based on eye gaze analysis. *PLOS ONE*, *13*(10), e0206213. https://doi.org/10.1371/journal.pone.0206213

Scheiter, K., Schubert, C., Schüler, A., Schmidt, H., Zimmermann, G., Wassermann, B., Krebs, M.-C., & Eder, T. (2019). Adaptive multimedia: Using gaze-contingent instructional guidance to provide personalized processing support. *Computers & Education*, *139*, 31–47. https://doi.org/10.1016/j.compedu.2019.05.005

Schotter, E. R., Angele, B., & Rayner, K. (2012). Parafoveal processing in reading. *Attention, Perception, & Psychophysics*, *74*(1), 5–35. https://doi.org/10.3758/s13414-011-0219-2

Shute, V. & Zapata-Rivera, D. (2012). Adaptive Educational Systems. *Adaptive Technologies for Training and Education.* 7-27. https://doi.org/10.1017/CBO9781139049580.004.

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 - UIST '00*, 101–107. https://doi.org/10.1145/354401.354418

Sims, S. D., & Conati, C. (2020). *Proceedings of the 2020 International Conference on Multimodal Interaction*, 15–23. https://doi.org/10.1145/3382507.3418828

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. https://doi.org/10.1111/j.1749-818x.2010.00263.x

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. https://doi.org/10.3758/bf03194057

Smallwood, J., McSpadden, M., & Schooler, J. W. (2008). When attention matters: the curious incident of the wandering mind. *Memory & Cognition*, *36*(6), 1144–1150. https://doi.org/10.3758/mc.36.6.1144

Smilek, D., Carriere, J. S. A., & Cheyne, J. A. (2010). Out of mind, out of sight: eye blinking as indicator and embodiment of mind-wandering. *Psychological Science*, *21*(6), 786–789. https://doi.org/10.1177/0956797610368063

Southwell, R., Gregg, J., Bixler, R., & D'Mello, S. K. (2020). What Eye Movements Reveal About Later Comprehension of Long Connected Texts. *Cognitive Science*, *44*(10), e12905. https://doi.org/10.1111/cogs.12905

Steindorf, L., & Rummel, J. (2019). Do your eyes give you away? A validation study of eye-movement measures used as indicators for mindless reading. *Behavior Research Methods*, *52*(1), 162–176. https://doi.org/10.3758/s13428-019-01214-4

Stimpson, A. J., & Cummings, M. L. (2014). Assessing Intervention Timing in Computer-Based Education Using Machine Learning Algorithms. *IEEE Access*, *2*, 78–87. https://doi.org/10.1109/access.2014.2303071

Strukelj, A., & Niehorster, D. C. (2018). One page of text: Eye movements during regular and thorough reading, skimming, and spell checking. *Journal of Eye Movement Research*, *11*(1). https://doi.org/10.16910/jemr.11.1.1

Taylor, S. E. (1965). Eye Movements in Reading: Facts and Fallacies. *American Educational Research Journal*, *2*(4), 187–202. https://doi.org/10.3102/00028312002004187

Thomson, D. R., Smilek, D., & Besner, D. (2014). On the asymmetric effects of mind-wandering on levels of processing at encoding and retrieval. *Psychonomic Bulletin & Review*, *21*(3), 728–733. https://doi.org/10.3758/s13423-013-0526-9

Traxler, M. J., Long, D. L., Tooley, K. M., Johns, C. L., Zirnstein, M., & Jonathan, E. (2012). Individual Differences in Eye-Movements During Reading: Working Memory and Speed-of-Processing Effects. *Journal of Eye Movement Research*, *5*(1).

Tyner, K. (2014). *Literacy in a digital world: teaching and learning in the age of information*. Routledge. https://doi.org/https://doi.org/10.4324/9781410601971

Tzeng, Y., Broek, P. van den, Kendeou, P., & Lee, C. (2005). The computational implementation of the landscape model: Modeling inferential processes and memory representations of text comprehension. *Behavior Research Methods*, *37*(2), 277–286. https://doi.org/10.3758/bf03192695

Underwood, G., Hubbard, A., & Wilkinson, H. (1990). Eye Fixations Predict Reading Comprehension: The Relationships between Reading Skill, Reading Speed, and Visual Inspection. *Language and Speech*, *33*(1), 69–81. https://doi.org/10.1177/002383099003300105

Voßkühler, A., Nordmeier, V., Kuchinke, L., & Jacobs, A. M. (2008). OGAMA (Open Gaze and Mouse Analyzer): Open-source software designed to analyze eye and mouse movements in slideshow study designs. *Behavior Research Methods*, *40*(4), 1150–1162. https://doi.org/10.3758/brm.40.4.1150

Wallot, S., O'Brien, B. A., Coey, C. A., & Kelty-Stephen, D. (2015). Power-law fluctuations in eye movements predict text comprehension during connected text reading. *CogSci*.

Wallot, S., O'Brien, B. A., Haussmann, A., Kloos, H., & Lyby, M. S. (2014). The Role of Reading Time Complexity and Reading Speed in Text Comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(6), 1745–1765. https://doi.org/10.1037/xlm0000030

Yang, S.-N., & McConkie, G. W. (2001). Eye movements during reading: a theory of saccade initiation times. *Vision Research*, *41*(25–26), 3567–3585. https://doi.org/10.1016/s0042-6989(01)00025-6

Yeari, M., Broek, P. van den, & Oudega, M. (2015). Processing and memory of central versus peripheral information as a function of reading goals: evidence from eye-movements. *Reading and Writing*, *28*(8), 1071–1097. https://doi.org/10.1007/s11145-015-9561-4

Yuan, Y., Chang, K., Taylor, J. N., & Mostow, J. (2014). *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge, LAK '14*. 54–58. https://doi.org/10.1145/2567574.2567624

Zhan, Z., Zhang, L., Mei, H., & Fong, P. S. W. (2016). Online Learners' Reading Ability Detection Based on Eye-Tracking Sensors. *Sensors*, *16*(9), 1457. https://doi.org/10.3390/s16091457

Zwaan, R. A., & Radvansky, G. A. (1998). Situation Models in Language Comprehension and Memory. *Psychological Bulletin*, *123*(2), 162–185. https://doi.org/10.1037/0033-2909.123.2.162

Author Biographies

Rosy Southwell is a Postdoctoral Research Associate at the Institute of Cognitive Science at the University of Colorado Boulder, USA. She received her doctorate in Auditory Cognitive Neuroscience in 2019 from University College London, UK, and her MSci and BA in Natural Sciences from the University of Cambridge, UK. She is currently working on using sensing technologies (eye-tracking, physiology, EEG, fNIRS) and speech during complex learning to advance understanding of cognitive processes such as mind-wandering, attention, and comprehension; and building models of cognitive states from these data using deep and machine learning approaches.

Caitlin Mills is an Assistant Professor and leads the *Affect, Cognition, and Computation Lab* at the University of New Hampshire, USA. Her research group is primarily interested in understanding how thoughts arise and how they influence learning. The group uses a variety of techniques including eye-tracking and emotion inductions to build computational models incorporating how we think in predicting learning outcomes in real time. Caitlin has a Ph.D. in Cognitive Psychology from the University of Notre Dame, an M.A. in Psychology from the University of Notre Dame, and a B.A. in Psychology from Christian Brothers University.

Megan Caruso is a graduate student in computer science and cognitive science at the University of Colorado Boulder. She is currently pursuing her PhD in the Emotive Computing Lab under Sidney D'Mello where she investigates what eye gaze and brain measures can reveal about reading comprehension processes. She is currently using machine learning techniques to build models that can predict comprehension outcomes using multimodal input from these various sources, for eventual use in intelligent recommender systems.

Sidney D'Mello is a Professor at the Institute of Cognitive Science and the Department of Computer Science at the University of Colorado Boulder (since 2017). He was previously an Associate Professor in Psychology and Computer Science at Notre Dame (2012-2017). He received his PhD in Computer Science at the University of Memphis in 2009. D'Mello has published more than 300 articles, of which 16 received awards (4 others were finalists). He also directs the NSF National Institute for Student-Agent Teaming (iSAT). D'Mello received the 2018 Young Investigator Award from the Society for Text & Discourse and is a Scholar at the Student Experience Research Network.

This document contains supplementary material for the article, “Gaze-based predictive model of deep reading comprehension”

Supplemental Information

Random forest model results, equivalent to tables 4 and 7-9 in the main text, are included here.

Table S1: *Correlations between self-explanation scores and model predictions computed on participant-level averages of each. Self-explanation models (top) and competing comprehension models (bottom). For the self-explanation models trained in this study, the median and range of the correlation metric over all 100 runs is also shown. CI = confidence interval*

	Study 1 correlations (N=133)	**Study 2 correlations (N=106)**
Self-explanation models	*median [95% CI] over runs*	
Main model – RF	0.393***[0.237, 0.529]	0.357***[0.178, 0.513]
Human -scored concepts only		0.396***[0.222, 0.545]
Reading time only	0.217*[0.047, 0.374]	0.241* [0.053, 0.413]
Shuffled split comparison	0.006 [-0.236, 0.248]	0.341* [0.078, 0.560]
Alternative models of comprehension		
Shallow-online model	0.063	0.329***
Mind-wandering model	-0.265**	-0.200*

Table S2: *Correlation of model predictions and post-test comprehension scores. Predicted and observed scores were first averaged at the participant level. SE = self-explanation; MW – mind-wandering. Computed correlation used pearson-method with pairwise-deletion. *p<0.05 **p<0.01 ***p<0.001*

		SE score	SE model	Reading time model	Shallow-online model	MW model
Study 1	Shallow-Immediate	0.477***	0.203*	-0.148	0.122	-0.178*
	Inference-Immediate	0.676***	0.311***	0.193*	0.012	-0.243**
	Shallow-Delayed	0.600***	0.285**	0.155	0.048	-0.205*
	Inference-Delayed	0.592***	0.342***	0.099	0.105	-0.261**

Table S3: *Standardized (Beta) coefficients for participant-level regression models controlling for other models (*p<0.05 **p<0.01 ***p<0.001).*

	Study 1 (N=131 immediate, N=115 delay)					**Study 2 (N=106)**
	SE Score	Shallow-Immediate	Inference-Immediate	Shallow-Delayed	Inference-Delayed	SE Score
SE model	**0.34*****	**0.17**	**0.26****	**0.24***	**0.27****	**0.22***
Shallow-online model	0.04	0.1	0.02	0.04	0.11	0.21*
MW model	-0.12	-0.06	-0.11	-0.08	-0.13	-0.17
R^2 / R^2 adjusted	0.167 / 0.147	0.053 / 0.031	0.107 / 0.086	0.087 / 0.063	0.139 / 0.116	0.182 / 0.158

Table S4: *Generalization of Study 1 model to predict Study 2 outcomes, and vice versa. For comparison, the within-study model performance is also shown. All correlations are computed on the participant-level averages of predicted and observed scores. (*p<0.05 **p<0.01 ***p<0.001).*

		Generalization	Within-study performance
		Study 2 model	*Study 1 model*
Study 1 scores	Self-explanation	**0.195***	0.393***
	Shallow-Immediate	**0.013**	0.203*
	Inference-Immediate	**0.159**	0.311***
	Shallow-Delayed	**0.220***	0.285**
	Inference-Delayed	**0.053**	0.342***
		Study 1 model	*Study 2 model*
Study 2 scores	Self-explanation	**0.284****	0.357***