Discourse Processes

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hdsp20

Machine-Learned Computational Models Can Enhance the Study of Text and Discourse: A Case Study Using Eye Tracking to Model Reading Comprehension

Sidney K. D’Mello , Rosy Southwell & Julie Gregg

DISCOURSE PROCESSES
2020, VOL. 57, NOS. 5–6, 420–440
https://doi.org/10.1080/0163853X.2020.1739600

Check for updates

Machine-Learned Computational Models Can Enhance the Study of Text and Discourse: A Case Study Using Eye Tracking to Model Reading Comprehension

Sidney K. D'Mello, Rosy Southwell, and Julie Gregg

Institute of Cognitive Science, University of Colorado Boulder

ABSTRACT
We propose that machine-learned computational models (MLCMs), in which the model parameters and perhaps even structure are learned from data, can complement extant approaches to the study of text and discourse. Such models are particularly useful when theoretical understanding is insufficient, when the data are rife with nonlinearities and interactivity, and when researchers aspire to take advantage of "big data." Being fully instantiated computer programs, MLCMs can also be used for autonomous assessment and real-time intervention. We illustrate these ideas in the context of an eye movement–based MLCM of textbase comprehension during reading along connected text. Using a dataset where 104 participants read a 6,500-word text, we trained Random Forests models to predict comprehension scores from six eye movement features. The models were highly accurate (area under the receiver operating characteristic curve = .902; r = .661), robust, and generalized across participants, suggesting possible use in future studies. We conclude by arguing for an increased role of MLCMs in the future of discourse research.

Introduction

> Camilla: *"You, sir, should unmask."*
> Stranger: *"Indeed?"*
> Cassilda: *"Indeed it's time. We have all laid aside disguise but you."*
> Stranger: *"I wear no mask."*
> Camilla: *(Terrified, aside to Cassilda.) "No mask? NO MASK!* [emphasis added]
> (Chambers, 1985)

This short dialog from "The King in Yellow and Other Horror Stories" by Robert W. Chambers illustrates the power of discourse in the hands of a gifted writer. In just 33 words Chambers presents a complex narrative involving three characters embedded in a rather macabre interaction with surprise and impending terror. His brevity teases our fascination, temping us to imagine what came before this strange meeting and what horror might occur next. Indeed, the power of discourse is not in the words themselves but in what lies beneath.

Given the complexity of discourse, which increases by orders of magnitude when moving from written prose to spoken dialogs and even further for multiparty conversations, it might seem foolish to suggest that computational methods can contribute anything of value beyond crunching data. But this is exactly what we are suggesting. More so, it is precisely in contexts of immense complexity where their merits truly lie—in a computer's ability to efficiently sift through and identify non-obvious patterns in vast quantities of data. We argue that computational methods are a complementary and, in some cases, essential companion to existing approaches to studying

CONTACT Sidney K. D'Mello sidney.dmello@colorado.edu 594 UCB, Boulder, CO 80309, USA.

discourse, including qualitative analyses, code-and-count methods, reaction time studies, eye tracking, brain imaging, experimental methods, and so on. The secret is in the type of computational method advocated: machine-learned computational models (MLCMs). We illustrate these points in the context of a case study involving the use of an eye-gaze–based MLCM of text comprehension during reading.

What Is an MLCM?

A model is a representation of a thing (Frigg & Hartmann, 2018), be it a phenomenon (e.g., reading comprehension; McNamara & Magliano, 2009), data (e.g., reading times; A. C. Graesser et al., 1980), or a theory (e.g., landscape model of reading; Van den Broek et al., 1999). Models can be physical (e.g., robotic model of eye movements; Villgrattner & Ulbrich, 2010), symbolic (e.g., equations governing saccadic control; Tatler et al., 2017), or a fictional entity like the phonological loop in Baddeley's working memory model (Baddeley, 1992). A computational model is a specific type of model whose representations are in silico, that is, performed on a computer or simulated by a computational device. An MLCM is a computer model *learned* from data or experience. Simply put, it is a program learned from data (Domingos, 2012). As we elaborate below, the degree and type of learning involved distinguishes MLCMs from "traditional" models and computer programs.

An MLCM has three main components: (1) a structure, such as an equation, a decision tree, an artificial neural network, or a graph; (2) feature representations, which pertain to higher order abstractions of data (e.g., fixation durations extracted from raw gaze points in eye tracking); and (3) parameters (or coefficients or weights) and (optionally) hyperparameters, which control the learning process itself. For example, in the following linear regression model, *comprehension* = *5* × *total reading time* + *2*, the equation is the model, comprehension is the outcome (predicted variable), total reading time is the feature, 5 and 2 are the parameters, and there are no hyperparameters. Generally, learning an MLCM from a dataset consists of adjusting the model parameters until the discrepancy between predicted and target values is minimized.

Table 1 provides a coarse-grained comparison of different computational models along a number of dimensions. The key distinguishing feature between the MLCMs (last three rows) versus traditional computational models (first row) is that the latter are more or less mathematical realization of a theory. They have a fixed structure, fixed feature representations, set parameters, and no learning. The parameters might, and probably should, be obtained from prior data, but they seldom change. Models of eye movements during reading, such as E-Z Reader (Reichle et al., 2003) and SWIFT (Engbert et al., 2005), are pertinent examples of such models. Traditional computational models are theory-heavy and data-light, whereas MLCMs are data-heavy and their theoretical commitments vary but are not as extensive as traditional models.

Table 1. Comparison of different types of computational models.

Structure	Feature Representations	Parameters/ Coefficients	Functions Learned	Theoretical Commitments	Data Required	Model Type	Examples
Fixed	Fixed	Fixed	None	Most	Fewest	Mathematical models	EZ Reader; SWIFT
Fixed	Fixed	Learned	Linear classifiers	More	Fewer	Standard regression modeling	Generalized linear models (e.g., linear and logistic regression)
Fixed	Fixed	Learned	Linear and nonlinear	Fewer	More	Standard machine learning	Random forest; support vector machines; shallow neural networks
Fixed	Fixed/learned	Learned	Very complex nonlinear functions	Fewest	Most	Deep neural learning	Convolutional neural networks; long short-term memory neural networks[a]

[a]Although subsumed under deep learning methods, these models can be considered shallow if only a single hidden layer is used.

The simplest MLCMs are standard regression models, such as linear and logistic regressions, and their variants (e.g., ridge regression). Multicollinearity and model fitting concerns with these models often preclude the use of too many features for a given size of training dataset; they typically have a handful (usually under 10) of prespecified features whose coefficients are learned from data. Given the small number of parameters to be learned, this approach requires some but not a substantial amount of data (a few hundred cases). The advantage of regression models lies in their simplicity and interpretability, but they are limited in their ability to model more complex data (such as nonlinear interactions among features). In contrast, standard machine learning methods, such as neural networks, support vector machines (Cortes & Vapnik, 1995), and Random Forest (Breiman, 2001) can model nonlinearities and interactions in the data. Describing each of these types of model in detail is beyond the scope of the present article, but see Mitchell (1997) and Witten and Frank (2005) for a primer on machine learning. These models typically have tens to a few hundred features, so they require more data (several hundred to a few thousand cases) to reliably estimate the parameters. Standard machine learning models can be judiciously constructed to be consistent with theory but to a lesser extent than standard regression models. This is because there are many more free parameters to fit, and these parameters themselves determine the nature of the interaction between features and/or the nature of the function mapping features to outcomes.

Deep neural learning or deep learning (Goodfellow et al., 2016; Le Cun et al., 2015) is a different class of models that have gained prominence over the past decade. These models are constructed by combining multiple "layers" of artificial neural networks, consisting of an input layer, one or more "hidden" (intermediate) layers, and an output layer, where hidden layers learn useful intermediate representations of the data by combining input features. These deep neural networks can model extremely complex phenomena. They are also capable of representation learning in that they can learn features themselves by extracting patterns from raw data instead of requiring prespecification of features like in the other modeling approaches. Deep learning models are extremely complex with the number of free parameters in the tens to hundreds of thousands, so they require copious amounts of training data and their interpretability is low. They also have very few theoretical commitments and are basically very powerful prediction machines. As Table 1 illustrates, there is a tradeoff between theoretical commitments, explanation versus prediction, and the amount of data needed to train viable models. In most cases we recommend experimenting with standard regression modeling and standard machine learning as these two approaches appropriately balance these tradeoffs.

A curious reader might ask what distinguishes standard regression models, which are extensively used in virtually all areas of science, from an MLCM regression model. The traditional case focuses on the significance of the model coefficients, whereas the MLCM approach focuses on the accuracy of the model predictions to "new" data. Thus, the former approach favors explanation and description of the entire dataset, whereas the latter favors prediction; see Yarkoni and Westfall (2017) for a detailed discussion on this issue and an enthusiastic call for psychology to engage in more predictive modeling. From a methodological standpoint, instead of building a model on the entire data set and examining p values of the coefficients and perhaps the goodness of fit, the MLCM approach constructs the model from a subset of the data and then computes fit statistics on the held-out data, a process called cross-validation. This process lessens the extent to which the model is influenced by the idiosyncrasies of individual data points (i.e., overfitting), ensuring the model is capturing general patterns across observations (where "observations" in the context of a discourse MLCM might be readers, texts, or utterances).

How MLCMs can enhance the study of discourse

Before delving into the potential benefits of MLCMs, let us address a common criticism that machine learning is an atheoretical fishing expedition that produces spurious results. Although it is easy to find examples where this criticism applies, rejecting the entire field on these grounds is no different from rejecting all experimental approaches to the study of discourse based on the existence of confounded experiments (which are abundant) and careless (or unethical) data analysis methods (e.g., *p*-hacking; Head et al., 2015). It is similar to rejecting qualitative approaches as being insufficiently rigorous or thought experiments as they are not scientific.

We believe it is more productive to develop well-designed MLCMs that are scientifically rigorous *and* useful computational tools. In our view well-designed MLCMs of discourse should be adequately grounded in theory but should not be overly constrained by theory. This is because discourse is complex: Although our understanding of it is growing, it is still limited. Many of our theories apply only to particular contexts, and computational instantiations of theory are likely to fail when taken out of controlled experimental paradigms into the messiness of the real world. An MLCM is probably not needed if a phenomenon is sufficiently understood that it can be computationally instantiated with high fidelity under realistic conditions. Instead, an MLCM is most beneficial when there is some theoretical understanding but not enough to instantiate a mechanistic model of the theory.

What can an MLCM do? It can guide theory by determining whether the right ingredients are in place or if something fundamental is missing. All things considered, a model that fails to generalize or generate accurate predictions might suggest some missing components. For example, Bartlett et al. (2014) aimed to model expressions of pain from facial movements automatically extracted from video. They found that the temporal dynamics of the facial expressions were critical in discriminating real from posed expressions of pain. Similarly, an MLCM can ascertain which features are more important than others and what minimalistic feature set is sufficient to model the phenomenon of interest. Using the same example, a single facial movement, mouth opening, provided the most information in that the duration and variance of mouth openings and the interval between consecutive mouth openings was lower for faked versus genuine pain expressions. The structure of the model itself can provide insights into how the various components interact, for example, when one component moderates, often nonlinearly, the influence of another on the outcome prediction.

Finally, an MLCM is a computational tool that can be used for measurement, to provide feedback, to drive reflection, and for intervention. For example, Jensen et al. (2020) developed an MLCM that automatically assessed the quality of teacher discourse in real-world classrooms. They used the model to provide feedback to teachers to help them improve their discourse. D'Mello, Mills et al. (2017) and Mills et al. (in press) used a previously developed eye-gaze–based MLCM of mind wandering during reading (Faber et al., 2018) to deliver real-time interventions consisting of comprehension questions and self-explanations aimed at re-engaging attention and correcting any comprehension deficits associated with mind wandering.

We have developed multiple MLCMs that use linguistic, paralinguistic, behavioral, and physiological signals with the goal of understanding and/or facilitating cognitive, noncognitive, socio-affective-cognitive, and life outcomes. Such work includes a range of discourse scenarios: rhetorical, expository, pedagogical, dialogic, and collaborative discourse collected in individual, small group, multiparty, and human–computer interactions in the lab and in the wild (e.g., Bosch & D'Mello, in press; Bosch et al., 2016; D'Mello & Graesser, 2010; Faber et al., 2018; Grafsgaard et al., 2018; Hutt et al., 2019; Kelly et al., 2018; Stewart et al., 2019; Stone et al., 2019). We have also used these models for assessment (Faber et al., 2018; Jensen et al., 2020) and real-time intervention (Aslan et al., 2019; D'Mello, Mills et al., 2017; Mills et al., in press). We have provided descriptions and tutorials of the MLCM approach along with examples in different research areas, specifically measurements of emotion (D'Mello et al., 2018) and engagement (D'Mello, Dieterle et al., 2017). In the remainder of this article we illustrate the use of MCLMs to the study of discourse by presenting an unpublished study where we developed an MLCM of reading comprehension from eye movements.

Illustrative example: modeling reading comprehension from eye movements

Reading for understanding is a complex process that requires low-level text processing; active construction and maintenance of representations, retrieval, and integration of information from long-term memories; and generation of predictions and inferences (A. Graesser et al., 1994; Kintsch, 1988; Lesgold & Perfetti, 1978; Rayner et al., 2012). With this in mind, accurately measuring reading comprehension is critical to understanding the real-time dynamics of text processing, such as whether and when readers generate elaborative inferences about the text (e.g., A. Graesser et al., 1994; McKoon & Ratcliff, 1992) or whether readers are attending to text at all (e.g., Feng et al., 2013).

Reading comprehension is typically assessed using comprehension questions presented alongside or after the text. These questions can take various forms (multiple choice, short response, self-explanation) and can assess comprehension at different levels, for example, probing factual content from the reading (textbase-level) or deeper, inference-level understanding of the text (McNamara & Magliano, 2009). Here, we examine the question of whether an MLCM of eye movements can generate accurate, real-time, and generalizable predictions of comprehension during reading. This knowledge, in turn, would contribute to theories of eye movements during reading, and the model itself can be used to assess reading comprehension as it unfolds or to trigger interventions when signs of comprehension difficulty emerge.

Background and research linking eye movements and reading comprehension

Given that reading requires processing of fine-grained visual stimuli (i.e., letters and words), eye movements are fundamentally linked to the cognitive processes underlying reading. Decades of research has capitalized on this insight, termed the eye–mind link (Just & Carpenter, 1976), by using eye tracking to investigate how readers extract coherent and even rich representations of meaning from these abstract visual stimuli. This research has demonstrated that eye movements are sensitive to text properties from the word to text levels, including word frequency (Inhoff & Rayner, 1986), lexical and syntactic ambiguity (Duffy et al., 1988; Frazier & Rayner, 1987), and text difficulty (Rayner et al., 2006), and has made great strides toward characterizing how eyes move during reading generally (see Rayner, 2009; Rayner & Reichle, 2010 for reviews). However, in spite of decades of progress in understanding how the eyes move during reading, limited research has leveraged these insights to measure comprehension in real time.

Why might this be the case? One reason is that existing work has struggled to establish consistent links between comprehension and eye movement features. For instance, 10% to 25% of eye movements are regressive to an earlier part of the text (Rayner et al., 2012). These regressions have long been interpreted as a corrective response when the reader has difficulty integrating the current word with prior context (Frazier & Rayner, 1987; Meseguer et al., 2002). However, some studies positively link regressions with accurate comprehension (Inhoff et al., 2018; Metzner et al., 2016; Schotter et al., 2014), whereas others show null (Christianson et al., 2016; Wallot et al., 2015) or even negative (Kemper et al., 2004) associations. Relatedly, longer fixation durations have been linked to both effortful reading (Rayner et al., 2006) and its opposite, mind wandering (Faber et al., 2018). On this point, research also suggests that people self-report rereading the previous one or two lines of text after a mind-wandering episode (Varao-Sousa et al., 2017), likely reflecting re-engagement with the text in an attempt to repair comprehension. Thus, although both re-engagement and mental-model repair may involve regressing to earlier parts of the text, repair might also occur covertly (i.e., resolved in working memory), typically resulting in longer fixation durations but not necessarily a regression (Meseguer et al., 2002).

Why is establishing consistent links between eye movements and comprehension so challenging? One possibility is that eye movements primarily reflect local text processing (e.g., word identification, syntactic parsing), which is almost always successful in skilled readers, and thus does not predict later comprehension. Another possibility, however, is that mappings between eye movements and comprehension may not be consistent because they are influenced by reader- and text-specific factors. Thus, the same eye movement features may reflect different cognitive processes in different contexts. For instance, longer fixations may reflect mind-wandering (Faber et al., 2018; Foulsham et al., 2013), which is a negative predictor of comprehension (D'Mello, 2019; Randall et al., 2014), but may alternatively signal efforts to repair inaccurate or poor-quality text representations, presumably leading to better comprehension outcomes (Frazier & Rayner, 1982). There are also inconsistencies in the literature. For example, some studies found fewer fixations associated with mind-wandering (Bixler & D'Mello, 2016; Faber et al., 2018, in press; Smilek et al., 2010), whereas others found the opposite, which might be attributable to methodological differences (Faber et al., in press).

Can an MLCM help resolve the lack of consistency between eye movements and comprehension outcomes? The answer relies on the observation that unique eye movement signatures may emerge when multiple features are considered in conjunction, whereas consistent mappings might not be evident when eye movement features are considered individually, as in most studies. For instance, although longer fixations may indicate either mental model repair or mind-wandering, the two could be differentiated using other features (e.g., number of fixations and regressions). Further, processes underlying successful comprehension (e.g., motivation, attention) fluctuate over time and may alter corresponding eye movements. When skimming or mind-wandering, for instance, eye movements exhibit less systematic correspondence with the underlying text compared with focused reading (Foulsham et al., 2013; Reichle et al., 2010). Thus, examining multiple features in context of one another, in particular those that capture alignment between eye movements and the text, could illuminate systematic relationships between eye movements and comprehension. Here, we test whether an MLCM can capture these complex relationships.

Previous work on MLCMs of comprehension during reading

Research in the human–computer interaction domain has developed MLCMs of reading comprehension, assessed alongside or immediately after reading short passages, from eye-movement features. For example, Copeland et al. (2014) (also see Copeland & Gedeon, 2013) recorded participants' eye movements while they read a nine-slide (~400 words per slide) tutorial on conducting a web search. They trained artificial neural networks to predict performance on quiz (comprehension) questions presented alongside or immediately after each slide from slide-level eye movement features (number of fixations, mean fixation duration, total text fixation duration, number of regressions, regression fixation proportion, mean forward saccade length) and eye movement–derived features (these included the ratios of number/duration of fixations to words and answer-seeking behavior, which was defined by saccades between the texts and the questions when the two were side by side). This approach yielded accurate predictions of performance on quiz questions, although comparisons with chance were not reported. See Copeland et al. (2015, 2016) for additional studies that further investigate factors that influence prediction accuracy (e.g., text difficulty and whether or not the text was in the reader's first language).

Other studies have examined whether MLCMs of eye movements can be used to predict general language skill. For example, Martinez-Gomez and Aizawa (2014) examined whether participants' level of understanding, as well as English language skill, could be predicted based on their eye movements during reading. Participants read two short (~450 word) educational texts while their eye movements were recorded. They answered eight questions to assess their understanding after each text. Random forest models with leave-one-participant-out cross-validation (see below) yielded significantly above-baseline ($p < .001$; ~50% error reduction) predictions of binarized (high vs. low) text understanding, where the baseline consisted of simply always predicting the majority class. Notably, eye movement features were more discriminative than linguistic features in these models. However, models predicting continuous comprehension scores performed at chance levels. The researchers also obtained above-baseline ($p = .015$) performance when training an MLCM to predict participants' English skill as measured by their scores on standardized English tests (Test of English for International Communication or Test of English as a Foreign Language). Similarly, using support vector machines, Lou et al. (2017) could discriminate high versus low literacy participants with 80.3% accuracy (as assessed by a Chinese standardized test similar to the Scholastic Aptitude Test;

baseline accuracy was not reported for comparison) using text-mapped eye-movement features (e.g., fixation times on section headers).

Although these studies provide encouraging evidence that MLCMs of eye movements could be plausibly used for diagnosis and intervention in the face of reading difficulties, there are some important limitations to consider. If eye movement–based models are to serve as a viable mechanism for monitoring comprehension, they must be able to generate accurate predictions for previously *unseen* individuals. However, only a few studies have examined generalizability of models to new readers by using person-independent cross-validation, which entails testing models on data from participants not used for model training. Specifically, the model reported in Copeland and Gedeon (2013) showed poor performance on held-out participants. Remaining work used data-point-level cross-validation (Copeland et al., 2014, 2015), as opposed to participant-level cross-validation, which has been acknowledged as a limitation (Copeland, 2016). Data-point-level cross-validation means that features at the level of individual observations (e.g., single pages) are randomly held out of the training data, but as a result the same participant's data, albeit from different pages, will recur in the training and test partitions, potentially jeopardizing the generalizability of the resulting model. Participant-level cross-validation, where model performance is assessed using data from participants who did not appear in the training data, is required to demonstrate that the model generalizes to unseen participants.

Further, while the model in Martinez-Gomez and Aizawa (2014) successfully predicted the comprehension levels for the best and worst performers, they excluded data from participants with intermediate comprehension levels, which are possibly the more difficult cases. Their validation method also did not ensure generalizability to new participants. Whereas Lou et al. (2017) did use appropriate participant-level cross-validation, they focused on predicting binarized (high vs. low) literacy skills not comprehension outcomes. Thus, existing models have only been able to make very coarse-grain predictions about new participants using eye movements. This could be useful for diagnostic purposes but may not provide sufficient granularity to inform theory or to deploy real-time interventions during reading.

Design considerations for a gaze-based MLCMs of reading comprehension

Our aim was to examine whether eye movements can be used to train an MLCM of reading comprehension in a way that is generalizable across readers. The choice of our modeling approach was guided by theory and empirical research on eye movements during reading (cited above) and was based on a number of design considerations. First, we know that relationships between eye movement features and comprehension can be interactive, that is, features may have different relationships with comprehension depending on reader- and text-specific factors. There is also the conflicting goal of balancing prediction accuracy with model explainability (Molnar, 2019) and generalizability. With these considerations in mind, we selected Random Forest (Breiman, 2001) for our classifier. This is a classifier architecture based on decision trees. A decision tree can be thought of as a flowchart describing possible paths to a decision based on binary decisions determined by the values of particular features, where the outcome of each decision defines which "branch" to progress to next (e.g., if number of fixations < 5, then look at number of regressions). Decision trees capture nonlinearity and interactivity between features in an interpretable fashion. For instance, a decision tree could "branch" based on number of regressions being greater or lesser than some threshold amount and make different predictions about comprehension accuracy based on other features within each branch (e.g., few regressions could predict accurate comprehension with short vs. long reading times, i.e., the branching factor). Decision trees are relatively interpretable because their structure consists of a sequence of readable "if, then" rules. A random forest consists of an ensemble of different decision trees, each using random subsets of training examples selected with replacement for each tree (called bagging), and within each tree random subsets of features at each branch point. Due to these properties, random forests are more likely to generalize: by randomly leaving out fractions of the full dataset, both in terms of features and training examples, the final model comprising the full "forest" is less prone to overfit to the data.

Second, to ensure that the learned relationships generalize to "new" readers, we applied a person-independent cross-validation technique by testing the model using data from held-out participants who were not used in model training. If the learned relationships are too specific to the training participants (i.e., the model is overfit), then the model will perform poorly when generating predictions for previously unseen readers. However, if performance on the held-out participants is high, this would indicate generalization to new readers, albeit with data collected in similar contexts.

Third, we restricted our choice of eye gaze features to six that we largely based on prior literature. This was an important design consideration to balance the tradeoff between alignment with theory and allowing room for new discovery. We intentionally selected a smaller feature set so that the resultant models could be interpreted and to avoid the criticism of engaging in unbridled exploration. Further, for this approach to be applicable to real-world applications (i.e., eye movement–based interventions), it would need to be easily applied to new texts and robust to routine eye-tracking errors. To this end we focused on global eye movement features (e.g., number of fixations and mean fixation duration on a page) that are less affected by calibration errors, less reliant on positional information (e.g., which word is fixated), and do not need to be mapped to local text properties (e.g., the frequency of a fixated word). Our final set of features included the number of fixations, mean fixation duration, regression fixation proportion, mean saccade length, horizontal saccade proportion, and fixation dispersion.

Fourth, we aimed for the model to be fine-grained in that it can predict comprehension on individual items. Thus, rather than predicting passage- or person-level differences in comprehension (e.g., Copeland & Gedeon, 2013; Lou et al., 2017; Martinez-Gomez & Aizawa, 2014), we predicted comprehension at the page level (where a page refers to the text presented on a computer screen). Generating page-level predictions would allow for real-time automated assessment of comprehension, which could be used to deploy interventions and to study comprehension processes on-line.

In what follows we discuss the steps toward building the aforementioned MLCM, beginning with collection of training data for machine learning. This entailed interrupting the reader with online comprehension assessments, a required step to collect training data to build the model. If successful, the model can then be used to generate the assessments for new readers without interruptions, summarized as follows:

Training (eye gaze features + **comprehension scores**) → *computational model*
Deployment (eye gaze features + *computational model*) → **comprehension scores**

Introduction

Data

We leveraged data from a previous eye-tracking study that collected assessments of reading comprehension during a computerized reading task (D'Mello, Mills et al., 2017). At the time of writing, the eye-tracking data collected in this study were not previously published.

Participants

Participants were 104 students at a private Midwestern university in the United States who participated in exchange for course credit. Participants signed a written informed consent form before participating, and the study was approved by the university's Institutional Review Board.

Materials and procedure

Participants read a 6,500-word excerpt from a book about the surface tension of liquids, *Soap Bubbles and the Forces which Mold Them* (Boys, 1895). The excerpt was taken from the first 35 pages of the book and was modified to remove images and associated references in text, which were not necessary for comprehension. The text was divided into 57 pages (screens; 115 words average per screen) and presented on a computer screen in 35-point Courier New font. Sentences but not words could be split across page boundaries. Left and right eye movements were recorded using the Tobii TX300 (Tobii Pro, Stockholm, Sweden) remote eye tracker sampling at 120 Hz. Head position was unrestrained, so participants could select a comfortable position for reading. Reading was self-paced, and participants advanced through the text one page at a time via a key press. However, they could not return to a previously read page.

Comprehension was assessed during reading using four-option multiple-choice questions that tapped page-specific, textbase-level (i.e., factual) content of the text. Below is an example of a sentence of text and associated comprehension assessment.

Text: "Plateau in his famous work, Statique des Liquides, quotes a passage from a book by Henry Berthoud, to the effect that there is an Etruscan vase in the Louvre in Paris in which children are represented blowing bubbles from a pipe."

Question: "The suggestion that there is an Etruscan vase in the Louvre that depicts children blowing bubbles from a pipe was put forth by:"

Choices: (a) Lord Rayleigh; (b) Van der Mensbrugghe; (c) Millais; (d) Plateau [**correct answer**].

Questions could occur after reading any of the 57 pages (apart from the first 2), but the number of questions (M = 15, SD = 4) and exact pages on which questions appeared differed by participant. There were two groups of participants in this study, and whether a question was asked on a given page was determined as follows. For the experimental group the computer interface presented comprehension questions as determined by another eye movement–based MLCM of mind wandering (Faber et al., 2018). This mind-wandering MLCM uses eye movement features to generate a probability that the participant is currently mind-wandering. If this probability exceeded a threshold, this triggered the reading interface to display a comprehension question. A different set of yoked-control participants received identical interventions to the experimental participants (i.e., comprehension questions occurring on the same pages, irrespective of mind-wandering). Participants were allowed to advance if they answered the first two questions correctly or after the second question regardless. Participants were given the option to reread the preceding page before answering the second question (only data from the first read of a page were included in analyses). Our present focus is training an MLCM to predict accuracy on the first comprehension question on a page from eye gaze recorded during the first read of that page.

Participants completed a post-test with 38 questions from the same pool immediately after reading, which averaged 34 minutes after reading began. These questions are not analyzed here as the current work focuses on modeling comprehension during reading.

Modeling approach

Figure 1 depicts an overview of the modeling approach, which focused on training and validating supervised machine learning models that predict comprehension from eye movements. We examined differences in eye movement features and comprehension across conditions (intervention vs. yoked control). No significant differences were observed, so we merged across conditions to obtain the maximum amount of data for machine learning and to improve model generalizability. We also modeled these conditions separately to ascertain whether the mind-wandering intervention confounded the link between eye movements and comprehension (i.e., whether the comprehension MLCM performance varied by experimental condition).

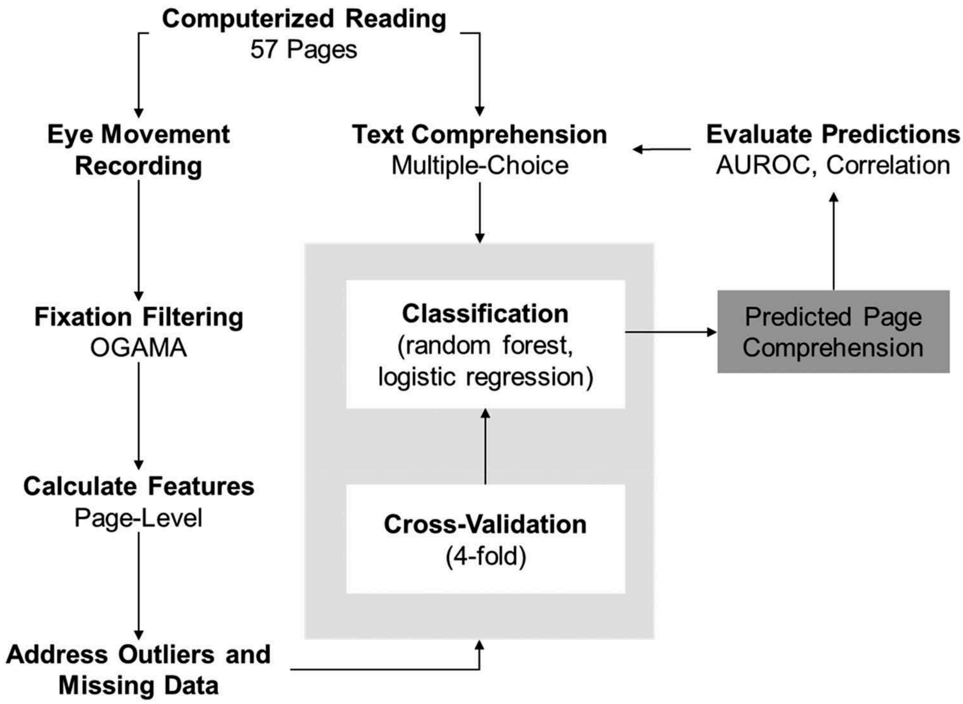

Figure 1. A schematic of the machine learning approach. Models were trained to predict page-level multiple-choice comprehension question performance (correct vs. incorrect) from the eye movement features of the associated page using data from computerized reading studies. "OGAMA" = Open Gaze and Mouse Analyzer.

Eye movement features

Eye data from both eyes were averaged and then fixations and saccades estimated using a dispersion-based fixation filter from Open Gaze and Mouse Analyzer (Voßkühler et al., 2008). Fixations were defined as consecutive eye movement samples within a range of 57 pixels (approximately 1 degree of visual angle), and saccades were computed from the fixations. We examined first-read eye movements and answer correctness for the first question (during reading) from the 1,618 pages with accompanying questions.

Eye movement features were chosen based on prior literature and the principle that page-level fluctuations in comprehension would be best captured by features that may index alignment between eye movements and the text. Four eye movement features were literature-based, including *number of fixations* on the page and their *mean fixation duration* in ms. Further, we selected *regression fixation proportion*, which is the proportion of all fixations on a page that were preceded by a regression (defined as any fixations on a word with an index lower than that of the previous word that was fixated on), and *mean saccade length*, which is the average number of pixels between two subsequent fixations. These features have been used in similar modeling efforts (Copeland, 2016; Copeland & Gedeon, 2013; Copeland et al., 2015, 2014; Martinez-Gomez & Aizawa, 2014) and have been empirically linked to factors that influence text comprehension, including mind-wandering (Reichle et al., 2010; Uzzaman & Joordens, 2011) and text difficulty (Rayner et al., 2006). This literature-driven choice of features illustrates how MLCMs can be at least in part constrained by theory.

The remaining two eye movement features, horizontal saccade proportion and fixation dispersion, have yet to be examined in the literature but were included for their potential to capture overall alignment of eye movements with the text, which may be linked to comprehension (Wallot et al.,

Table 2. Descriptive and correlation of eye gaze features and comprehension accuracy.

	Mean (SD)	Mean Fix Duration	N Fix	Reg Fix Prop	Mean saccade Length	Horizontal saccade Prop	Fix Disp
Mean fixation duration, ms	267 (33)	—					
No. of fixations	100 (36)	0.323	—				
Regression fixation proportion	0.136 (0.037)	0.180	0.269	—			
Mean saccade length, pixels	237 (29)	−0.429	−0.419	−0.102	—		
Horizontal saccade proportion	0.891 (0.073)	−0.194	0.170	−0.210	0.030	—	
Fixation dispersion	0.397 (0.050)	−0.276	−0.440	−0.274	0.245	0.211	
Comprehension accuracy	0.684 (0.169)	0.228	0.372	0.200	−0.260	−0.430	−0.421

Means and SDs in column 2 are computed over participants.

2015). Specifically, *horizontal saccade proportion* was calculated as the proportion of saccades with an angle no more than 30 degrees above or below the *x*-axis and *fixation dispersion* as the root-mean-square of the distance of each fixation to the average fixation. Thus, these two features may capture especially sparse or erratic eye movements that could signal a disconnect between eye movement patterns and the text on the page. The use of this highly constrained feature set also minimized multicollinearity, which is a concern for some machine learning models (see Table 2 for between-feature correlations).

Because head movement was not restrained to allow for naturalistic reading, participants could shift position leading to eye-tracking disruptions. To compensate, we excluded pages that were clearly unread (reading time < 1 s; 7 pages) and those without recorded eye movement data (i.e., one fixation or fewer on the page; 81 pages). Only 5% of pages (88 of 1,618) were discarded as a result of these criteria, leaving 1,530 pages for modeling. To address outliers we replaced eye gaze feature values greater than 2.5 median absolute deviations with the highest observed value of that particular feature within these bounds. As elaborated above, we further minimized the impact of eye tracking errors by focusing on global eye movement features, which are based on relative rather than absolute eye position (i.e., word-specific features like gaze duration and dwell time were not included).

Supervised classification and validation

Random Forest classification models were implemented in R with the caret package (Kuhn, 2008). We used the default parameters, which were varying the number of features to split at each node, then selecting the value that optimizes model fit; and building 500 decision trees. We compared random forests, which capture interactivity and nonlinearity, with logistic regression, which is a linear additive model. This also illustrates how an MLCM can be used to answer a pertinent question on the nature of the relationship between eye movements and comprehension. Finally, we examined whether model performance relied on systematic correspondence between eye movements and comprehension by comparing the random forest model to shuffled surrogates, created by shuffling the comprehension scores.

We used a participant-level fourfold cross-validation procedure to ensure generalizability to new participants. Specifically, data were split into four subsets at the participant level, trained on three of the subsets, and tested on the remaining subset. This process was repeated four times so that predictions were generated for all four test subsets. To assess stability of results, the entire process was repeated 100 times. There was very little variability across runs (see Table 3 for stability of model performance across the 100 runs) so we focused on the median-performing model (based on the correlation metric; see below) in our analyses.

Table 3. Summary of classification model performance.

Model	Page Level (AUROC)	Participant Level (Correlation)
Median models	AUROC [95% CI]	Pearson's *r* [95% CI]
Random forest	0.902* [0.885, 0.919]	0.661*** [0.537,0.757]
Logistic regression	0.879* [0.860, 0.899]	0.594*** [0.453,0.706]
Shuffled random forest	0.475 [0.494,0.556]	−0.019 [−0.211,0.174]
Across 100 iterations	Mean [min, max]	Mean [min, max]
Random forest	0.902 [0.89, 0.91]	0.663 [0.602, 0.705]
Logistic regression	0.878 [0.866, 0.883]	0.592 [0.548, 0.617]
Shuffled random forest	0.463 [0.437, 0.486]	−0.024 [−0.193, 0.125]

Significant AUROCs (bootstrapped 95% CIs nonoverlapping with chance [0.5]) and correlations are marked with asterisks (*). For correlations, *$p < .05$, **$p < .01$, ***$p < .001$.

Model evaluation

The model output is a probability that the question on a given page was answered correctly (i.e., a continuous variable between 0 and 1, termed the probability of the correct class). We computed page-level accuracy as the area under the receiver operating characteristic curve (AUROC; chance = 0.5) between the class probability of the correct class (between 0 and 1) and observed scores (1 or 0). We also averaged both predicted (0 or 1, after applying a threshold at 0.5 to the class probabilities) and observed (0 or 1) comprehension scores over all pages for each subject and then computed the correlation between the predicted and observed performance at the participant level. Note that the model was *trained* based on its performance at the page level for individual participants, but through averaging its predictions at the page level we use the same MLCM to make predictions at the participant level. Whereas page-level accuracy evaluates the models' ability to make fine-grained predictions on individual comprehension items based on eye movements from corresponding pages, participant-level accuracy is a coarser measure focusing on between-subject variability. AUROCs were compared using the roc.test function in the pROC package (Robin et al., 2011), and correlations were compared using tests of dependent correlations as implemented in the cocor package (Diedenhofen & Musch, 2015).

Modeling results

Model accuracy

The results are summarized in Table 3, and ROC curves for all models are depicted in Figure 2. The random forest model generated highly accurate predictions of both page- and participant-level comprehension (page-level AUROC = 0.902; participant-level $r = 0.661$, $p < .001$). As illustrated in Figure 3, the predicted distribution of comprehension scores closely aligned with the observed distribution (top left), and the model accurately captured participant- (top right) and page-level (bottom) variation in comprehension. This indicates that eye movements can be reliably linked to comprehension during reading.

To examine whether interactivity of features improved model performance, we compared the random forest model with a logistic regression (an additive) model trained on the same features. The logistic regression model also generated above-chance predictions of page- and participant-level comprehension (page-level AUROC = 0.879, participant-level $r = 0.594$, $p < .001$). Interactivity significantly improved page-level model performance ($z = 3.07$, $p = .002$) but less so for participant-level performance ($z = 1.82$, $p = .07$), suggesting a slight improvement.

Robustness checks

We examined whether above-chance predictions resulted from systematic correspondence between eye movements and comprehension. The shuffled model performed at or near chance on the page

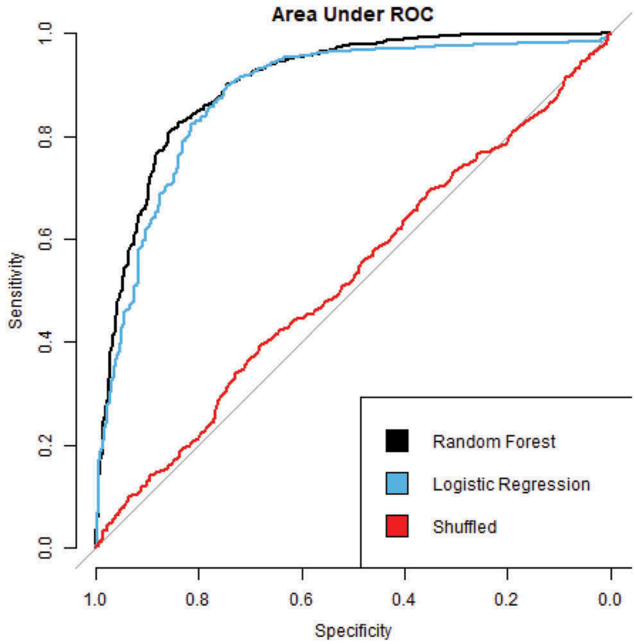

Figure 2. Receiver operating characteristic (ROC) curves for the median models.

and participant levels (AUROC = 0.48; $r = -0.02$, $p = .85$) and reliably worse than its nonshuffled counterpart (page level: $z = 20.70$, $p < .001$; participant level: $z = 5.81$, $p < .001$).

Next, to examine the impact of outlier treatment on model performance, we compared the performance of the random forest model (with outliers addressed as described above) with an identical model in which outliers were not addressed. Results were equivalent at both the page level (model with outliers: AUROC = 0.904; comparison: $z = -0.60$, $p = .55$) or participant levels (model with outliers: $r = 0.663$, $p < .001$; comparison: $z = -0.15$, $p = .87$), suggesting that similar model performance may be achieved without outlier treatment. Results were also similar and not significantly different ($z = -1.35$, $p = .177$) when no pages were removed and missing feature values were replaced with zeroes (zero imputation; AUROC = 0.864; $r = 0.633$, $p < .001$)

Checks for confounds

It was critical to determine whether model performance could be attributed to the contingency between eye movements and interventions that were triggered by mind-wandering (in the experimental condition). Thus, we analyzed whether predictions were dependent on experimental condition (intervention vs. yoked control) by modeling each condition separately. Both models performed above chance on the page and participant levels (intervention: AUROC = 0.882, $r = 0.652$, $p < .001$; yoked-control: AUROC = 0.892, $r = 0.664$, $p < .001$), and there was no reliable difference in performance across the models (participant level $z = -0.100$, $p = .921$; page level $z = 0.522$, $p = .601$). Thus, model performance did not rely on the mind-wandering–contingent interventions.

Relatedly, to what extent do associations between eye movement features and comprehension reflect lower-level processes, such as whether participants were attending to the text? To examine this we computed correlations between participants' average mind-wandering likelihood (derived from the eye movement–based MLCM used to trigger the interventions) for 100 participants with available data and their predicted comprehension scores from the median random forest model. As expected, we observed a significant negative correlation between mind-wandering and predicted

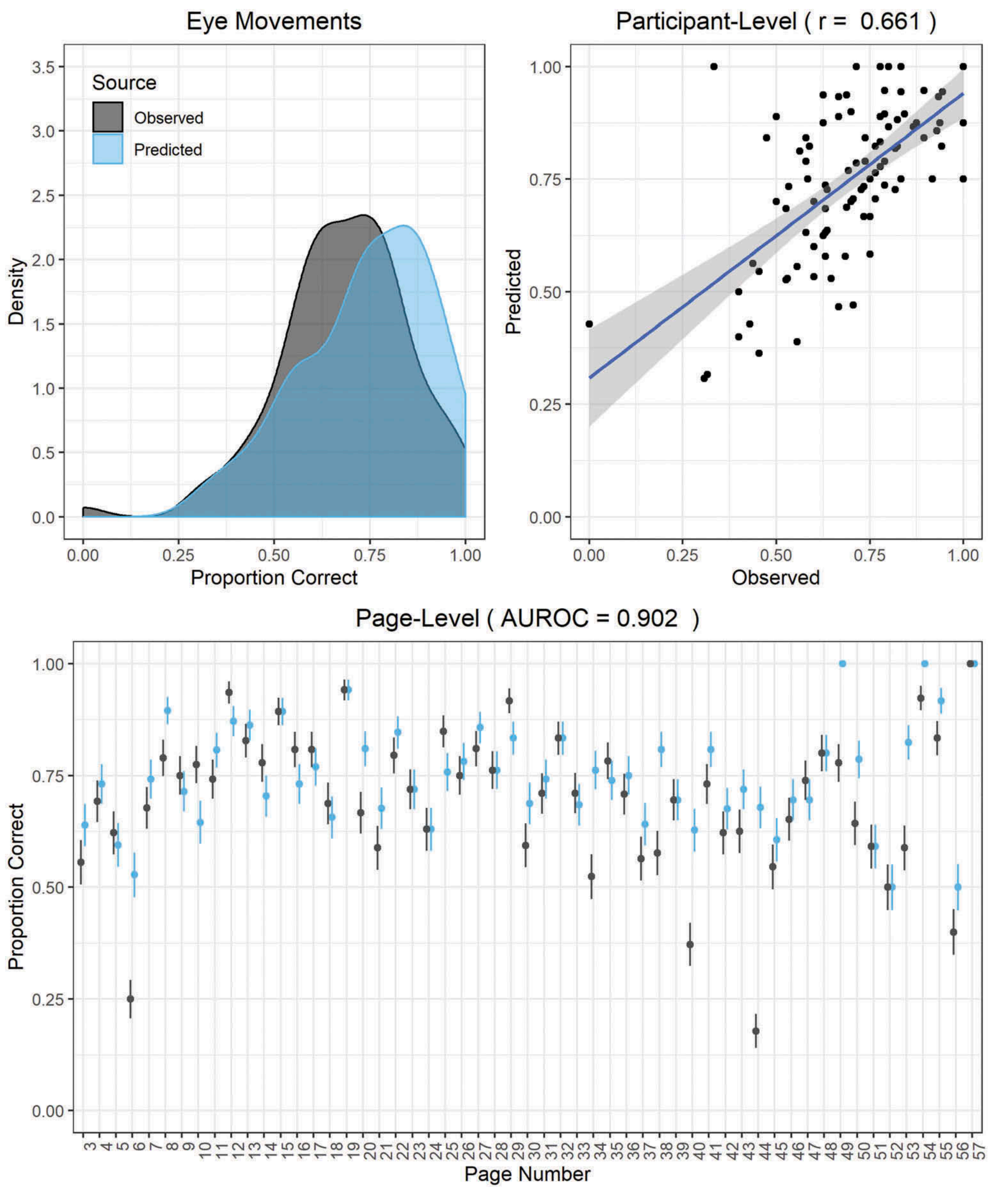

Figure 3. Visualization of page- and participant-level performance for the median random forest model predicting comprehension from eye movements. Top left: Predicted and observed participant-level mean comprehension accuracy. Top right: Correlation between predicted and observed participant-level mean comprehension accuracy. Bottom: Mean page-level accuracy; predicted and observed. Error bars are 95% confidence intervals computed across participants.

comprehension (r = −0.21, p = .04), although mind-wandering accounted for only 4.4% of the variance in predictions (R^2 = 0.044). This suggests that the model captured higher-level reading processes in addition to whether or not the reader was attending to the text.

Predictive features

How do eye movement features relate to comprehension? To investigate this question, we analyzed the models further. We opted to focus on the logistic regression model in lieu of the Random Forest model because the former, albeit slightly less accurate, is more interpretable; this is because feature values are linearly combined in logistic regression such that the coefficients are readily interpretable as the contribution of each feature to model performance. One way to quantify coefficient importance is to examine the model coefficients by computing the mean and variability across the cross-validated folds. The alternate approach, which we used here, is to focus on significance of the coefficients. Accordingly, we fit mixed effects logistic regression models to quantify coefficient significance using the lme4 library in R (Bates & Maechler, 2015) with participant as an intercept-only random effect. To address outliers, feature values were median absolute deviation-scaled before modeling (scaled values were truncated at ±2.5 median absolute deviation).

Model coefficients are presented in Table 4. We found that more fixations, shorter saccades, fewer horizontal saccades, and lower fixation dispersion were associated with better comprehension scores. Why might this be the case? To illustrate, Figure 4 depicts example eye movements from two participants corresponding to correct (left) and incorrect (right) responses to a comprehension question immediately after reading a page. Longer saccades and fewer fixations are indicative of skimming (Rayner et al., 2012), which is the inverse of what predicted accurate comprehension in our model. Further, higher horizontal saccade proportion may result from eye movements that too rigidly align with the text, which could also reflect skimming (as in the right compared with the left panel of Figure 4). Finally, higher fixation dispersion may correspond with sparse and erratic eye movements with poor correspondence to the text. Thus, lower horizontal saccade proportion and fixation dispersion, in combination with more fixations and shorter saccades, may indicate that the reader was sufficiently engaged to construct high-quality representations of the text.

Table 4. Coefficients from a generalized linear mixed model with eye movement features as fixed effects and participants as an intercept-only random effect.

	Comprehension Accuracy		
Predictors	Coefficient	95% CI	*p*
(Intercept)	1.34	1.06–1.63	<.001
Mean fixation duration	−0.05	−0.27 to 0.17	.642
No. of fixations	1.22	0.97–1.47	<.001
Regression fixation proportion	−0.01	−0.21 to 0.19	.948
Mean saccade length	−0.47	−0.68 to −0.27	<.001
Horizontal saccade proportion	−1.57	−1.78 to −1.36	<.001
Fixation dispersion	−0.49	−0.68 to −0.30	<.001

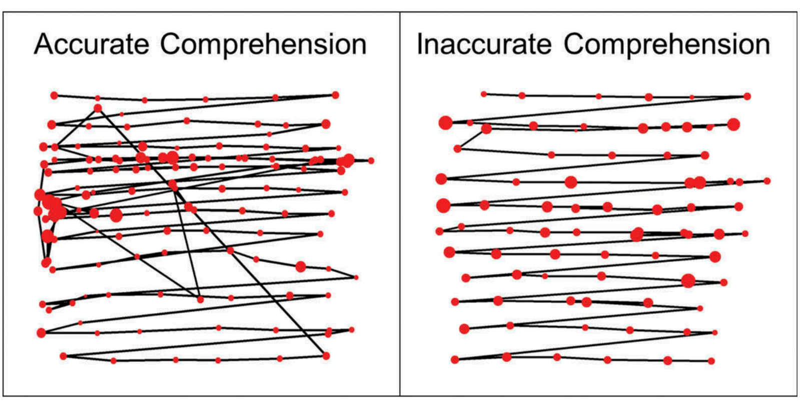

Figure 4. Comparison of eye movements from two participants reading the same page. Circles represent fixations and are scaled by duration, and lines represent saccades. The eye movements in the left and right panels preceded correct and incorrect responses to a subsequent comprehension question, respectively.

Note that the marginal R^2 for the mixed-effects model was 0.566 and the conditional R^2 was 0.680. This corresponds to marginal r = 0.752 and conditional r = 0.824. This is indeed a highly accurate model, with correlation coefficients higher than for both the MLCMs (random forest and logistic regression). However, this can be attributed to the difference in the way these models were fit. The mixed model is an explanatory model, fit to the entire dataset as a whole. Therefore, the marginal and conditional R^2 do not reflect the predictive accuracy of the model. In contrast, the same model fit with the MLCM approach has a correlation of 0.627 when fit with cross-validation. Put differently, this reduction in correlation provides an index of the extent to which the mixed modeling approach overfits to the data. This model, if it were used to predict comprehension on unseen participants, would likely underperform as the model will have overfit to the training data.

Discussion of modeling approach and results

Our goal was to demonstrate how eye movements could be used to develop an MLCM of reading comprehension. Our results showed that our models produced highly accurate predictions of page- and participant-level comprehension, suggesting this approach yields predictions of sufficient accuracy and granularity (AUROC of .90; r = .66) to be used for research as well as intervention and diagnostic purposes. Further, we demonstrated that the success of these models relies on the page-level correspondence between eye movements and comprehension, specifically as illustrated by comparisons with a random surrogate model.

This current model offers several crucial improvements over prior work. First, it extends previous research predicting overall comprehension (Copeland, 2016; Copeland & Gedeon, 2013; Copeland et al., 2014, 2015; Martinez-Gomez & Aizawa, 2014) to real-time, page-level comprehension during reading. This is critical because real-time assessment of comprehension is needed to better our theoretical understanding of short-lived, dynamic phenomena during reading. For example, as suggested by the bottom panel of Figure 3, comprehension may fluctuate throughout a reading as attention waxes and wanes, or if particular text regions are especially difficult to integrate. By leveraging such a real-time, eye movement–based measure of comprehension, we anticipate that researchers will be better poised to study text comprehension dynamics without interrupting the reader.

This work also provides encouraging evidence for the generalizability and scalability of this approach. Namely, using cross-validated models we achieved highly accurate predictions on held-out participants, suggesting that the models can be used predict comprehension for new readers in a similar context. In further support of this, the current modeling approach generalized across two experimental conditions (intervention vs. yoked-control) and were highly stable across 100 runs (see the lower half of Table 3). In addition, by using global eye movement features (e.g., number and duration of fixations on a page), the current approach minimizes the need for human coding of the text (e.g., defining text regions of interest and defining linguistic properties for those regions) and minimizes the impact of routine eye-tracking and calibration errors that would be expected in real-world settings (e.g., classrooms).

In addition, our results provide further insights on how eye movements are linked to comprehension. We hypothesized based on previous research that modeling interactivity among eye movements is key toward capturing the link between eye movements and comprehension. In line with this, when we compared random forest (interactive) and regression (additive) algorithms, we observed that modeling interactivity slightly improved page-level model performance, but the improvement was quite minor, suggesting that an additive model would suffice. Indeed, a benefit of the regression model over the random forest model is that the contribution of individual features to model performance was more easily quantified. The tradeoff between the ability to model more complex data and the ability to interpret the structure of the resulting model in terms of its individual features is indeed an important consideration in the choice of specific model to use with an MLCM approach. Accordingly, we found that more fixations along with shorter saccades (the inverse of eye movements typically observed during skimming) were predictive of accurate comprehension. Further, features that indexed spatial alignment between eye movements and the text (horizontal saccade proportion and fixation dispersion) also predicted comprehension. Collectively, these findings suggest that eye movements during reading capture higher-level processes (e.g., motivation), which can be used to accurately diagnose comprehension in real-time.

Notably, models of eye movements during reading assume the reader is attending and that local text processing is generally successful (e.g., E-Z Reader model: Reichle et al., 1998). The current research suggests that such assumptions might be reconsidered when the goal is to model naturalistic reading. Instead, a comprehensive model of eye movements during reading of longer, connected texts must integrate attentional and motivational processes (e.g., mind-wandering) that are neglected in current models. How could this be accomplished? One approach could be to develop models which detect the unique eye movement signatures of reading behaviors and associated mental states (e.g., mind-wandering, skimming, and repair) that interfere with or facilitate comprehension.

Encouragingly, an MLCM approach has been applied to detect mind-wandering (Bixler & D'Mello, 2016; Faber et al., 2018) and skimming (Biedert et al., 2012). However, much work still needs to be done to model what occurs during motivated, attentive reading, some of which might be covert (e.g., developing models to predict covert comprehension repair and inferencing). These models of the reading process could also offer improvement over the current results by integrating predictions from several submodels (e.g., of skimming, mind-wandering, and repair) in an ensemble-like fashion to generate highly robust predictions of comprehension accuracy.

Despite the success of the present MLCM that models eye movements from eye gaze, many open questions remain. For example, the current approach models only one dataset with a single text and only textbase-level assessments of comprehension. Thus, the extent to which the current findings generalize to other contexts (e.g., texts) and whether eye movement measures capture deeper text processing remains to be investigated. A critical test of the current MLCM would be to test its performance on gaze data from uninterrupted reading. If this model successfully predicts multiple-choice item performance *after* reading, then it could be usable as a fine-grained, page-level comprehension measure without interrupting the participant. Further, we did not explicitly measure the underlying reading behaviors and processes involved in reading (such as skimming, inferencing, and error-monitoring), which would be a critical step toward building more comprehensive models of real-time reading comprehension and make more explicit the link between the model structure and theory-derived constructs. An additional avenue for future work using this MLCM would be to assess whether it generalizes to reading different texts and under different contexts.

Keith Rayner, perhaps the most influential scholar of eye movements and reading of the past 40 years, and his colleagues (Rayner et al., 2006) imagined that once eye-tracking technology became more portable and affordable, eye movements could be used as a diagnostic and intervention tool for comprehension difficulties. With consumer-off-the-shelf eye-tracking technology becoming more cost-effective and accurate (Gibaldi et al., 2017), that day has seemingly arrived from a technological standpoint. Whereas we use an expensive research-grade eye tracker in this work, other studies have successfully developed MLCM using consumer-off-the-shelf eye trackers in authentic environments (Hutt et al., 2019). For example, Hutt et al. (2019) successfully developed MLCMs of mind-wandering while interacting with a learning technology using data collected with consumer-off-the-shelf eye-trackers in a classroom context. The researchers also found the model to be comparably accurate with models based on data collected in a laboratory setting with consumer-off-the-shelf eye-trackers (Hutt et al., 2016). Thus, recent advances in eye-tracking technology coupled with the MLCM approach advocated here take us a step toward bringing Keith Rayner's predictions about the future of eye movement–based diagnostic and intervention tools from the realm of science fiction to plausible reality.

Concluding remarks

It is generally accepted that computational analyses of discourse can complement other methods including think-alouds, code and count, experimental methods, and the like. Here we suggest that MLCMs provide a unique, yet complementary, approach to the study of discourse. To make our case we distinguished MLCMs from traditional computational models, highlighted different types of MLCMs, discussed their potential benefits, and demonstrated the overall idea by developing and validating an MLCM model of reading comprehension from eye movements. Despite the encouraging success of these models, there is still a long way to go. Even the most sophisticated MLCM is unlikely to deeply understand, for example, the excerpt from *The King in Yellow* reproduced in the opening lines of this article. We do not know the answer yet, except to advocate for a future of discourse research that incorporates computational models in its existing arsenal of theoretical, observational, and experimental methods. In this future MLCMs are rapidly instantiated to analyze experimental data while also serving as measurement and intervention tools. Qualitative inspection of their parameters and behavior provide insights into the underlying theories, which in turn can be tested via different instantiations of the models. Thus, computational models and theory development go hand in hand.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was supported by the National Science Foundation (DRL 1235958, IIS 1523091, and DRL 1920510). Any opinions, findings, conclusions, or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of National Science Foundation.

References

Aslan, S., Alyuz, N., Tanriover, C., Mete, S. E., Okur, E., D'Mello, S. K., & Esme, A. A. (2019). Investigating the impact of a real-time, multimodal student engagement analytics technology in authentic classrooms *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI 2019)*. ACM.

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556–559. https://doi.org/10.1126/science.1736359

Bartlett, M. S., Littlewort, G. C., Frank, M. G., & Lee, K. (2014). Automatic decoding of facial movements reveals deceptive pain expressions. *Current Biology*, *24*(7), 738–743. https://doi.org/10.1016/j.cub.2014.02.009

Bates, D. M., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., ... & Bolker, M. B. (2015). Package 'lme4'. *Convergence*, *12*(1), 2. http://CRAN.R-project.org/package=lme4

Biedert, R., Hees, J., Dengel, A., & Buscher, G. (2012). *A robust realtime reading-skimming classifier* [Paper presentation]. Proceedings of the symposium on eye tracking research and applications, Santa Barbara, USA.

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

Bosch, N., D'Mello, S., Baker, R., Ocumpaugh, J., & Shute, V. (2016). Using video to automatically detect learner affect in computer-enabled classrooms. *ACM Transactions on Interactive Intelligent Systems*, *6*(2), 17.11–17.31. https://doi.org/10.1145/2946837

Bosch, N., & D'Mello, S. K. (in press). Automatic detection of mind wandering from video in the lab and in the classroom. *IEEE Transactions on Affective Computing*.

Boys, C. V. (1895). *Soap bubbles, their colours and the forces which mold them*. Society for Promoting Christian Knowledge.

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

Christianson, K., Luke, S. G., Hussey, E. K., & Wochna, K. L. (2016). Why reread? Evidence from garden-path and local coherence structures. *The Quarterly Journal of Experimental Psychology*, *0218*(May), 1–51. https://doi.org/10.1080/17470218.2016.1186200

Copeland, L. (2016). Eye tracking to support eLearning. PhD Thesis, The Australian National University. https://openresearch-repository.anu.edu.au/handle/1885/108880

Copeland, L., & Gedeon, T. (2013). Measuring reading comprehension using eye movements. *2013 IEEE 4th International Conference On Cognitive Infocommunications (CogInfoCom)* (pp. 791–796), Gyor, Hungary. https://doi.org/10.1148/radiol.12120134

Copeland, L., Gedeon, T., & Caldwell, S. (2015). *Effects of text difficulty and readers on predicting reading comprehension from eye movements* [Paper presentation]. *6th IEEE Conference on Cognitive Infocommunications*. CogInfoCom 2015.

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), 35–48. https://doi.org/10.5430/air.v3n3p35

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

D'Mello, S. K. (2019). What do we think about when we learn? In K. Millis, J. Magliano, D. Long, & K. Wiemer (Eds.), *Deep comprehension: Multi-disciplinary approaches to understanding, enhancing, and measuring comprehension* (pp. 52–67). Routledge.

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

D'Mello, S. K., & Graesser, A. (2010). Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-adapted Interaction*, *20*(2), 147–187. https://doi.org/10.1007/s11257-010-9074-4

D'Mello, S. K., Kappas, A., & Gratch, J. (2018). The affective computing approach to affect measurement. *Emotion Review*, *10*(2), 174–183. https://doi.org/10.1177/1754073917696583

D'Mello, S. K., Mills, C., Bixler, R., & Bosch, N. (2017). Zone out no more: Mitigating mind wandering during computerized reading. In X. Hu, T. Barnes, A. Hershkovitz, & L. Paquette (Eds.), *Proceedings of the 10th international conference on educational data mining* (pp. 8–15): International Educational Data Mining Society.

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PloS One*, *10*(4), e0121945. https://doi.org/10.1371/journal.pone.0121945

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87. https://doi.org/10.1145/2347736

Duffy, S. A., Morris, R. K., & Rayner, K. (1988). Lexical ambiguity and fixation times in reading. *Journal of Memory and Language*, *27*(4), 429–446. https://doi.org/10.1016/0749-596X(88)90066-6

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, *112*(4), 777. https://doi.org/10.1037/0033-295X.112.4.777

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

Faber, M., Krasich, K., Bixler, R., Brockmole, J., & D'Mello, S. K. (in press). The eye-mind wandering link: identifying gaze indices of mind wandering across tasks. *Journal of Experimental Psychology: Human Perception and Performance*.

Feng, S., D'Mello, S., & Graesser, A. (2013). Mindwandering while reading easy and difficult texts. *Psychonomic Bulletin & Review*, *20*(1), 586–592. https://doi.org/10.3758/s13423-012-0367-y

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

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension. *Cognitive Psychology*, *14*(2), 178–210. https://doi.org/10.1016/0010-0285(82)90008-1

Frazier, L., & Rayner, K. (1987). Resolution of syntactic category ambiguities: Eye movements in parsing lexically ambiguous sentences. *Journal of Memory and Language*, *26*(5), 505–526. https://doi.org/10.1016/0749-596X(87)90137-9

Frigg, R., & Hartmann, S. (2018). Models in science. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Summer 2018 ed.). Metaphysics Research Lab, Stanford University.

Gibaldi, A., Vanegas, M., Bex, P. J., & Maiello, G. (2017). Evaluation of the Tobii EyeX Eye tracking controller and Matlab toolkit for research. *Behavior Research Methods*, *49*(3), 923–946. https://doi.org/10.3758/s13428-016-0762-9

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Graesser, A., 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. C., Hoffman, N. L., & Clark, L. F. (1980). Structural components of reading time. *Journal of Verbal Learning and Verbal Behavior*, *19*(2), 135–151. https://doi.org/10.1016/S0022-5371(80)90132-2

Grafsgaard, J. F., Duran, N. D., Randall, A. K., Tao, C., & D'Mello, S. K. (2018). Generative multimodal models of nonverbal synchrony in close relationships In S. K. D'Mello, L. Yin, L. P. Morency, & M. Valstar (Eds.), *Proceedings of the 13th IEEE conference on automatic face and gesture recognition (FG'18)* (pp. 195–202). IEEE.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, *13*(3), e1002106. https://doi.org/10.1371/journal.pbio.1002106

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

Hutt, S., Mills, C., White, S., Donnelly, P. J., & D'Mello, S. K. (2016). The Eyes have it: Gaze-based detection of mind wandering during learning with an intelligent tutoring system. In T. Barnes, M. Chi, & M. Feng (Eds.), *Proceedings of the 9th international conference on educational data mining (EDM 2016)* (pp. 86–93): International Educational Data Mining Society. https://doi.org/10.1177/1753193415617756

Inhoff, A. W., Gregg, J., & Radach, R. (2018). Eye movement programming and reading accuracy. *Quarterly Journal of Experimental Psychology*, *71*(1), 3–10. https://doi.org/10.1080/17470218.2016.1226907

Inhoff, A. W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception {\&} Psychophysics*, *40*(6), 431–439. https://doi.org/10.3758/BF03208203

Jensen, E., Dale, M., Donnelly, P., Stone, C., Kelly, S., Godley, A., & D'Mello, S. K. (2020). Toward automated feedback on teacher discourse to enhance teacher learning *Proceedings of the ACM CHI conference on human factors in computing systems (CHI 2020)*.

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

Kelly, S., Olney, A. M., Donnelly, P., Nystrand, M., & D'Mello, S. K. (2018). Automatically measuring question authenticity in real-world classrooms. *Educational Researcher*, *47*(7), 451–464. https://doi.org/10.3102/0013189X18785613

Kemper, S., Crow, A., & Kemtes, K. (2004). Eye-fixation patterns of high- and low-span young and older adults: Down the garden path and back again. *Psychology and Aging*, *19*(1), 157–170. https://doi.org/10.1037/0882-7974.19.1.157

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

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Le Cun, Y., Bengio, Y., & Hinton, G. E. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lesgold, A. M., & Perfetti, C. A. (1978). Interactive processes in reading comprehension. *Discourse Processes*, *1*(4), 323–336. https://doi.org/10.1080/01638537809544443

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

Martinez-Gomez, P., & Aizawa, A. (2014). *Recognition of understanding level and language skill using measurements of reading behavior* [Paper presentation], Haifa, Israel. *IUI*.

McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological Review*, *99*(3), 440–466. https://doi.org/10.1037/0033-295X.99.3.440

McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model of comprehension. *Psychology of Learning and Motivation*, *51*, 297–384. https://doi.org/10.1016/S0079-7421(09)51009-2

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., von der Malsburg, T., Vasishth, S., & Rösler, F. (2016). The importance of reading naturally: Evidence from combined recordings of eye movements and electric brain potentials. *Cognitive Science*, *41*(S2), 1–32. https://doi.org/10.1111/cogs.12384

Mills, C., Gregg, J., Bixler, R., & D'Mello, S. K. (in press). Eye-mind reader: An intelligent reading interface that promotes long-term comprehension by detecting and responding to mind wandering. *Human-Computer Interaction*.

Mitchell, T. (1997). *Machine learning*. Mc-Graw-Hill.

Molnar, C. (2019). Interpretable machine learning. A Guide for Making Black Box Models Explainable, 2019. https://christophm.github.io/interpretable-ml-book/.

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

Rayner, K. (2009). Eye movements in reading: Models and data. *Journal of Eye Movement Research*, *2*(5), 1–10.

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

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

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, *105*(1), 125. https://doi.org/10.1037/0033-295X.105.1.125

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The EZ 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., Reineberg, A. E., & Schooler, J. W. (2010). Eye movements during mindless reading. *Psychological Science*, *21*(9), 1300–1310. https://doi.org/10.1177/0956797610378686

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77. https://doi.org/10.1186/1471-2105-12-77

Schotter, E. R., Tran, R., & Rayner, K. (2014). Don't believe what you read (Only once): Comprehension is supported by regressions during reading. *Psychological Science*, *25*(6), 1218–1226. https://doi.org/10.1177/0956797614531148

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

Stewart, A., Vrzakova, H., Sun, C., Yonehiro, J., Stone, C., Duran, N., Shute, V., & D'Mello, S. K. (2019). I say, you say, we say: using spoken language to model socio-cognitive processes during computer-supported collaborative problem solving. *Proceedings of the ACM on Human-Computer Interaction, Volume 3(CSCW)*, *39*, 1–19. https://doi.org/10.1145/3359296

Stone, C., Quirk, A., Gardener, M., Hutt, S., Duckworth, A. L., & D'Mello, S. K. (2019). Language as thought: using natural language processing to model noncognitive traits that predict college success. *Proceedings of the 9th international learning analytics and knowledge conference (LAK'19)*.

Tatler, B. W., Brockmole, J. R., & Carpenter, R. H. (2017). LATEST: A model of saccadic decisions in space and time. *Psychological Review*, *124*(3), 267–300. https://doi.org/10.1037/rev0000054

Uzzaman, S., & Joordens, S. (2011). The eyes know what you are thinking: Eye movements as an objective measure of mind wandering. *Consciousness and Cognition*, *20*(4), 1882–1886. https://doi.org/10.1016/j.concog.2011.09.010

Van den Broek, P., Young, M., Tzeng, Y., & Linderholm, T. (1999). The landscape model of reading: Inferences and the online construction of a memory representation. *The Construction of Mental Representations during Reading*, *13* (5), 71–98.

Varao-Sousa, T. L., Solman, G. J., & Kingstone, A. (2017). Re-reading after mind wandering. *Canadian Journal of Experimental Psychology/Revue Canadienne De Psychologie Expérimentale*, *71*(3), 203–211. https://doi.org/10.1037/cep0000123

Villgrattner, T., & Ulbrich, H. (2010). Optimization and dynamic simulation of a parallel three degree-of-freedom camera orientation system *2010 IEEE/RSJ international conference on intelligent robots and systems* (pp. 2829–2836). IEEE.

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., Brien, B. A. O., Coey, C. A., & Kelty-Stephen, D. G. (2015). *Power-law fluctuations in eye movements predict text comprehension during connected text reading.* Cognitive Science Society. Annual Conference. Proceedings (Online), 2015, 2583–2588.

Witten, I., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd ed.). Morgan Kaufmann.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393