Eye Movements During Reading Can Predict Deep Comprehension

Rosy Southwell, Caitlin Mills and Sidney D'Mello

July 12, 2020

Eye Movements During Reading Can Predict Deep Comprehension

Rosy Southwell

Institute of Cognitive Science, University of Colorado Boulder, CO, USA

Caitlin Mills

Department of Psycology, University of New Hampshire, NH, USA

Sidney D'Mello

Institute of Cognitive Science, University of Colorado Boulder, CO, USA

Author Note

The authors declare that there no conflicts of interest with respect to this preprint.

Correspondence should be addressed to Rosy Southwell. Email: roso8920@colorado.edu

Abstract

It is known that eye movements during reading reflect various reading processes as well as reader skill and attentiveness, but there is little work relating eye movements to reading comprehension outcomes. This work represents a novel step by showing that deep comprehension assessed by open-ended self-explanations during reading ($r = 0.39$, $p < 0.001$) can be predicted from eye movements in a person-independent manner. Our results have implications for theories of reading and for the design of real-time interventions.

Keywords: gaze tracking, reading, comprehension, machine learning

Introduction

Eye movements during reading are a rich source of information on reading strategies (Just & Carpenter 1980) and the cognitive state of the reader (e.g. whether they are skim-reading or mind-wandering). Indeed, recent work (D'Mello, Southwell & Gregg) has shown that rote comprehension as assessed by multiple-choice questions during reading can be predicted from global eye movement patterns. This indicates that successful encoding of factual information from the text is associated with particular eye movements during reading. However, to date, this link has not been demonstrated for deep comprehension. Whilst it is known that eye movements reflect localized features of the text (e.g. Rayner 1998), and that correct encoding of the latter is a prerequisite in many influential models of reading comprehension (e.g., Kintsch, 1998), the present work represents a novel step by asking whether reading comprehension as assessed by self-explanation has a significant relationship to eye movements. Self-explanation, where readers are asked to answer open-ended questions in their own words in writing (McNamara, 2007), renders the ongoing comprehension process more visible than, for instance, the multiple-choice questions used widely in comprehension research. Self-explanations often contain indicators of deep comprehension such as bridging inferences and elaboration (McNamara, 2004). We found that performance on open-ended, free-form questions interspersed during reading, targeting the crucial concepts in the text, can be predicted from the non-invasive measurement of eye movements in a manner that generalizes across individuals.

Methods

Procedure

Participants (N=106, mean age 21.1, 58% female) read a long, connected text whilst their eye movements were recorded. The text was non-fiction; a 6500-word excerpt from a book on surface tension in liquids (Flesh-Kincaid grade score 11.8) split into 15 sections, each representing particular concepts. Reading was self-paced, with the text split across 57 screens with an average of 115 words on each. Comprehension was assessed with open-ended self-explanations at the end of some of the 15 sections (mean 5 questions per participant; range 1-10). Prompts were structured to encourage readers to elaborate on and generalize concepts presented in the text. The self-explanations were scored by a human expert on a scale between 0 and 1; the mean score over subjects was 0.62 (standard deviation 0.24). Participants also completed multiple-choice assessments following reading the whole text and again at a 1-week delay. Each concept was assessed with a textbase-level and an inference-level item. For further details of the procedure, and examples of the comprehension assessment, see Mills et al. (2020), who present a analysis of an orthogonal manipulation in the same study.

Throughout the session, gaze was tracked using a Tobii TX300 gaze tracker sampling eye position of both eyes at a rate of 60-120 Hz. From this, fixations and saccades were extracted and summary metrics based on gaze were computed (see Mills et al., 2020 for details). Pages with fewer than 5 fixations or viewed for under 2 seconds were excluded as unread. Six gaze-based features chosen from the literature were used to build the models: number and mean duration of fixations, proportion of regression fixations (eye movements back to earlier parts of the text), mean saccade length, proportion of horizontal saccades, and fixation dispersion. These were averaged over all pages within a given conceptual section of the text, yielding 534 observations, and used to predict performance on the self-explanation item pertaining to that concept.

Models

We compare two types of machine-learnt regression models. 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’. Random forests are capable of modelling nonlinearities and interactivity between features. Linear regression models were also used, which are linear additive models. All models were fit with 4-fold participant-level cross validation, meaning that for each training iteration data from three-quarters of the subjects were used to train the model, but performance was assessed on the held-out portion. This ensures the final model generalizes to unseen participants. For each model this process was repeated over 100 runs, each with a different partitioning of subjects into folds. The median-performing model is reported for each model type.

To estimate the importance of the features used in the model, we fit a linear mixed-effects model to the same data, with features and condition (intervention versus control group) as fixed-effects and subjects as random-effects. To assess the degree to which the model captured within-subject variability in performance, we also fit models to a shuffled-surrogate dataset where the self-explanation scores were shuffled with respect to the concept-level gaze features within each subject. This maintained the subject-wise mean and variance in scores whilst breaking the link between scores on specific pages and gaze features. Predicted and actual scores were then averaged for each participant and the correlation was computed at the subject-level as our main measure of model performance.

Finally, the winning self-explanation model was used to generate predictions of self-explanation scores on all concepts ($n = 1631$) from the gaze features, not just those with a human-scored self-explanation. The model parameters from each of the four folds of the median-performing model were used to generate predictions which were then averaged. These predictions were then averaged at the subject-level and correlated with performance on the post-test items, for the 84 participants who completed the full study including the delay session.

Results

The linear regression model showed strong correlations ($r = 0.39$, $p < 0.001$) between predicted and observed subject-wise average comprehension scores. The random forest model significantly predicted comprehension ($r = 0.31$, $p = 0.001$) but correlation between model predictions and observed scores was lower than for the linear model; but not significantly so ($z = 1.02$, $p = 0.308$). The model trained on the shuffled-surrogate dataset also significantly predicted subject-level self-explanation scores ($r = 0.36$, $p < 0.001$). For the linear mixed-effect model the most diagnostic predictors were the reading time and saccade length, both positively correlated with comprehension scores.

Table 1 shows correlation between subject-level average self-explanation scores and performance on the post-test and delay multiple-choice items. The self-explanation model-generated predictions on all concepts correlated significantly with performance on the inference-level comprehension items presented after a 1-week delay ($r = 0.38$, $p < 0.001$). The hand-scored values also correlated with inference scores at delay, to a lesser degree ($r = 0.23$, p = 0.038), but this was not a significantly lower correlation (*Pearson & Filon's* $z = -1.44$, $p = 0.151$). No other correlations were significant.

Table 1

Pearson correlations between self-explanation scores and performance on multiple-choice items. Values in square brackets indicate the 95% confidence interval for each correlation.

	Self-explanation	
	Hand-scored	**Cross-validated predictions from linear regression model**
Textbase, Immediate	.07 [-.15, .28]	.19 [-.03, .39]
Inference, Immediate	.07 [-.14, .28]	.12 [-.10, .32]
Textbase, 1-week delay	.15 [-.07, .35]	.16 [-.06, .36]
Inference, 1-week delay	.23* [.02, .42]	.38** [.18, .55]

Discussion

We found that predictive models trained on eye movements significantly predicts the quality of readers' self-explanations. The best-performing model was a linear regression model, although a nonlinear random forest model performed similarly. This suggests that the relationship found between gaze features and reading comprehension is adequately captured by linear terms. The model performance was tested on subjects held out of the training set, indicating the link between eye tracking and comprehension score generalizes across individuals. The model fit was not significantly impacted when the features were shuffled within subjects, suggesting that the model was capturing mostly between-subjects patterns, although with so few observations per subject (on average self-explanations were completed for 5 concepts, but for some participants this was as few as 1) this may be unsurprising.

The model, when generalized to compute predicted self-explanation scores for concepts unseen during training, significantly predicted deep comprehension at 1-week delay. This suggests the model could be used as a measure of deep comprehension in an uninterrupted reading paradigm.

Our work contributes to theories of reading comprehension by integrating low-level models of eye movements (Rayner 1998) with higher-level models of comprehension (Kintsch 1998, Graesser et al. 1997). It also has implications for the potential of gaze-based tracking of ongoing reading comprehension as an alternative to intrusive learning assessments which have been shown to interfere with learning, either facilitating (e.g. Roediger and Karpicke 2006) or interrupting it (Foroughi et al 2015).

References

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*, 1–21. https://doi.org/10.1080/0163853x.2020.1739600

Foroughi, Cyrus & Werner, N & Barragan, D & Boehm-Davis, D. (2015). Interruptions Disrupt Reading Comprehension. Journal of Experimental Psychology. General. 144. https://doi.org/10.1037/xge0000074.

Graesser, A. C., Millis, K. K., & Zwaan, R. A. (1997). Discourse comprehension. *Annual review of psychology*, *48*(1), 163-189. https://doi.org/10.1146/annurev.psych.48.1.163

Just MA, Carpenter PA. A theory of reading: From eye fixations to comprehension. Psychological review. 1980 Jul;87(4):329. https://psycnet.apa.org/doi/10.1037/0033-295X.87.4.329

Kintsch, Walter. Comprehension: A paradigm for cognition. Cambridge university press, 1998.

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

McNamara, D. S. (Ed.). (2007). Reading comprehension strategies: Theories, interventions, and technologies. Lawrence Erlbaum Associates Publishers. https://doi.org/10.4324/9780203810033

Mills, C., Gregg, J., Bixler, R., & D'Mello, S. K. (2020). 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

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

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

Zwaan R.A., Radvansky G.A. Situation models in language comprehension and memory. Psychological bulletin. 1998 Mar;123(2):162. https://doi.org/10.1037/0033-2909.123.2.162