Webcam-Based Eye Tracking to Detect Mind Wandering and Comprehension Errors

Stephen Hutt[1,5], Aaron Wong[2], Alexandra Papoutsaki[4], Ryan S. Baker[1], Joshua I. Gold[1], and Caitlin Mills[2,3]

[1]University of Pennsylvania

[2]University of New Hampshire

[3]University of Minnesota

[4]Pomona College

[5]University of Denver

Author Note

Due to ethical concerns and participant data agreements, supporting data cannot be made openly available. Details of the data and how to request access are available from the Corresponding Author.

Supplemental material is available at this link: https://upenn.box.com/s/4pq2fwnhvd9g2pbsjgfbe31t173nkiv1

Correspondence concerning this article should be sent to Stephen Hutt, 3700 Walnut Street, Philadelphia, PA, 19104, USA. Email: Stephen.hutt@du.edu

Abstract

Recent advances in computer vision have opened the door for scalable eye tracking using only a webcam. Such solutions are particularly useful for online educational technologies, in which a goal is to respond adaptively to students' ongoing experiences. We used Webgazer, a webcam-based eye-tracker, to automatically detect covert cognitive states during an online reading-comprehension task related to task-unrelated thought and comprehension. We present data from two studies using different populations: 1) a relatively homogenous sample of university students ($N = 105$), and 2) a more diverse sample from Prolific ($N = 173$, with < 20% White participants). Across both studies, the webcam-based eye-tracker provided sufficiently accurate and precise gaze measurements to predict both task-unrelated thought and reading comprehension from a single calibration. We also present initial evidence of predictive validity, including a positive correlation between predicted rates of task-unrelated thought and comprehension scores. Finally, we present slicing analyses to determine how performance changed under certain conditions (lighting, glasses, etc.) and generalizability of the results across the two datasets (e.g., training on the data Study 1 and testing on data from Study 2, and vice versa). We conclude by discussing results in the context of remote research and learning technologies.

Keywords: eye tracking, comprehension, task-unrelated thought

Webcam-Based Eye Tracking to Detect Mind Wandering and Comprehension Errors

Eye-tracking technology has progressed substantially in ease and extent of use over the last few decades. Early systems were often intrusive, like contact lenses fitted with search coils, and used in only a small number of specialized laboratories. More recent systems typically use non-invasive, video-based technologies and are used extensively in psychology, neuroscience, marketing, education, and other fields. These technological advances have helped to support major advances in our understanding of extensive relationships between gaze and cognition (Eckstein et al., 2017). For example, eye-gaze behaviors can be used to detect certain cognitive processes that can affect learning, such as *mind wandering* (referred to here as task-unrelated thought; TUT), and predict learning outcomes, such as *comprehension* (D’Mello et al., 2020; Faber et al., 2017; Hutt et al., 2016; Hutt, Mills, et al., 2017; Mills et al., 2016). These kinds of gaze-based detectors hold promise both as a basic-research tool for understanding the cognitive factors that relate to learning and for developing adaptive interventions to support more-effective student learning (D’Mello et al., 2012; Hutt et al., 2021; Mills et al., 2020).

However, to date, the practical application of gaze-based approaches to monitor and affect learning in the real world has been severely limited by the cost and availability of appropriate eye-tracking systems. For example, past studies that used eye tracking to infer task-unrelated thought and comprehension were limited to highly controlled laboratory settings and/or expensive eye-tracking hardware that can cost upwards of $40,000. Thus, relatively few individuals and schools have been able to take advantage of these promising technologies. Limiting accessibility to students in wealthy districts is not a viable path forward, because it would only further perpetuate the inequities that already exist in education. Instead, new tools are needed to bring the potential power of gaze-based detectors of cognitive states to more real-world contexts with broad and equitable accessibility.

The primary goal of the current work was thus to provide the first evidence of a scalable option for detecting *comprehension* and *task-unrelated thought*, in real-time, using webcam-based eye tracking embedded within a web browser. We focused on the use of systems requiring no specialized hardware beyond a web camera built into most laptops and other computers and publicly available software (Papoutsaki et al., 2016). This approach makes it possible to extend the benefits of detector-based measurement and automated personalization to a broader, more economically diverse population of individuals. As a secondary goal, we aimed to show that this system can be used to reproduce and build on previous findings that quantified relationships between these learning-related cognitive constructs and gaze.

Theoretical Background and Related Work

Eye movements have long been viewed as a window into the cognitive processes that unfold during reading (Rayner, Chace et al., 2006; Rayner, Reichle et al., 2006; Reichle et al., 2012). Although a complete account of the “eye-mind” link is outside the scope of this paper, it is relevant to mention that eye gaze is considered a real-time index of the information-processing priorities of the visual system. For example, visual information is acquired primarily during periods when the eye remains relatively stable, known as fixations. In contrast, visual input is suppressed during saccades, which are ballistic movements of the eyes between fixations (Campbell & Wurtz, 1978; Irwin & Carlson-Radvansky, 1996; Matin, 1974; Zuber & Stark, 1966). Therefore, ongoing task goals are often best served when patterns of fixation ensure that central gaze, and therefore visual attention, is allocated to the most important visual information within the environment. This idea is particularly relevant to reading: fixation patterns are sensitive to both features of text being read and the reader’s understanding of that text (Rayner, Chace et al., 2006; Rayner, Reichle et al., 2006). Below we briefly summarize past work relating gaze patterns to reading comprehension and TUT and describe the specific contributions of the present work.

Reading comprehension. From a theoretical perspective, reading comprehension is often understood in terms of the Construction-Integration model (CI model). This model proposes that the mental model constructed while reading a text consists of three primary levels (Kintsch, 1998; McNamara & Magliano, 2009). The first, and most basic, level is the *surface code*. This level reflects the verbatim wording and structure of the text. This level fades quickly from memory but is used to identify semantic and syntactic relationships. The second level, which is constructed from the first, is the *textbase*. This level preserves the key fact-level information that is necessary to eventually represent the "gist" of the text. The third level, which builds on the textbase with information from the reader's prior knowledge to construct a more elaborate mental representation of the text's meaning, is the *situation model*. This level contains all inferences generated to establish connections amongst ideas in the text and prior knowledge. It may be helpful to consider textbase comprehension as fact-based memory, whereas situation-model comprehension can be seen as an overall conceptual model of the text.

Our understanding of reading comprehension via CI and other models has benefitted greatly from the use of eye tracking (Rayner, 2009). For example, eye movements, such as regressions (moving backwards through the text) and longer fixations, have been linked to difficulties in constructing a situation model and consequently comprehension (Rayner et al., 2006; Schotter, Tran, & Rayner, 2014). In addition, eye movements can be sensitive to text characteristics such as difficulty (Rayner et al., 2006) and genre (Kraal et al., 2019). In recent years, attempts have been made to use these kinds of gaze-tracking metrics to predict comprehension (Ahn et al., 2020; D'Mello et al., 2020; Wallot et al., 2015). Historically these predictions have been largely unsuccessful in terms of accuracy and generalizability. For example, in certain naturalistic reading contexts (e.g., text not altered for stimuli presentation) standard global features such as fixation duration and number of eye movements were not predictive of comprehension (Wallot et al., 2015). Likewise, in another study, fixation times and overall reading times were also not predictive of long-term memory and comprehension on their own (Yeari et al., 2015; Dirix et al., 2020).

Nevertheless, more recent research indicates that comprehension prediction from eye gaze may be possible. For example, one gaze-based model was able to explain ~40% of the variance in comprehension on a [describe test/condition] (r = 0.661, D'Mello et al., 2020). These kinds of eye-gaze-based models can also predict text-based comprehension and are generalizable across multiple datasets (Southwell et al 2020).Despite this progress, there is still a need to: 1) extend these predictive models to situation model comprehension, as a way to assess whether students have a deep level of understanding, as opposed to simply recalling factual details of the text (i.e., build a person-independent predictive model of whether a correct or incorrect inference is made about a text as it unfolds in real-time; while also 2) finding more scalable solutions, given that the models mentioned above were all trained using a high-cost research-grade eye-tracker with a high sampling rate and high-fidelity data (Tobii TX Pro 300).

Task-unrelated thought (TUT). One construct that has been closely linked to the disruption of comprehension is TUT (D'Mello & Mills, 2021; Phillips et al., 2016; Smallwood, 2011), commonly referred to as mind-wandering. TUT is defined as the act of shifting from an external task (e.g., reading) to internal thoughts about something unrelated to the current task (Smallwood & Schooler, 2015). TUT is ubiquitous in both everyday life and during reading, with estimates ranging from 20–40% of the time on average (D'Mello & Mills, 2021; Killingsworth & Gilbert, 2010; Klinger & Cox, 1987). Critically, TUTs are consistently negatively related to measures of performance in cognitively demanding tasks including reading comprehension (D'Mello & Mills, 2021; Randall et al., 2014).

TUT is thought to be a barrier to building an accurate mental model of a text because of its downstream effects on processing. For example, the cascade model of inattention (Smallwood, 2011) suggests that "perceptual decoupling" occurs during TUT, leading to slowed or diminished processing at lower levels of encoding (i.e., the surface code). This decoupling then causes breakdowns in the ability to integrate information across the multiple levels, from processing the individual words to the meaning of a sentence. As such, interactive learning software that can adaptively respond to TUT improves students' deep comprehension (Mills et al., 2020), but reliable detection is a necessary first step.

A growing body of research suggests that changes in eye movements can be indicative of when people are off-task. For example, this relationship has been used to build gaze-based TUT detectors during reading (Bixler & D'Mello, 2014, 2016; Hutt, Hardey et al., 2017). Commonly, supervised classification models are trained to discriminate between responses to embedded mind-wandering probes ("yes, I was off task" versus "no, I was on task") using global (i.e., not context-specific) gaze features (such as average fixation duration, fixation dispersion, saccade frequency, angle, etc.). The models are then validated by testing their generalizability to unseen individuals.

As with comprehension, much of the work in this space has leveraged research-grade eye tracking in the laboratory. However, some recent work supports the idea that lower-fidelity eye tracking can be used to automatically detect TUT. For example, Hutt and colleagues (2016, 2019) demonstrated that TUT detection could also be achieved with a COTS eye-tracker, which retails for $100–150 USD. Though this tracking system uses a lower sampling frequency and provides less accurate and precise gaze measurements than more expensive systems, successful TUT detection was still possible and was later used to deliver learning interventions that benefited learners with low prior knowledge (Hutt et al., 2021). Though these eye-trackers present a more affordable approach, they still require additional, specialized hardware, thus limiting overall scalability.

Overview and Novelty of Current Work

To overcome the limitations in scalability inherent to using expensive and/or specialized equipment, we focused on webcam-based eye-trackers that are beginning to be used in research and other settings (Degen et al., 2021; Semmelmann & Weigelt, 2018; Yang & Krajbich, 2020). A known limitation of these webcam systems is that they tend to be less accurate and precise than many specialized video-based systems (Zhang et al., 2019), particularly when they are deployed in real-world conditions in which lighting, head position, and other factors are not as controlled as typically are in laboratory settings. Thus, a major, open question is whether webcam systems provide a sufficiently reliable estimate of gaze position to be useful for monitoring gaze-sensitive cognitive states during reading.

A few studies have shown promise in this regard. For example, an unsupervised classification method has been used to derive areas of interest (AOIs) from gaze data collected with a webcam as users interacted with a communication task (Tran et al. 2019). In that study, gaze points were clustered to model users' interpersonal behavior and ultimately improve interactions. Though AOIs present a slightly coarser-grain analysis than may be needed to monitor cognitive states (D'Mello et al., 2020), this work demonstrates that webcam-based gaze tracking is still picking up on a valid interaction signal between eye movements and comprehension. Particularly encouraging is recent work comparing webcam-based eye movements to data collected from the Tobii Pro Glasses 2 (Valliappan et al., 2020). Across four tasks, data from the standard camera embedded in a smartphone was comparable to data collected from the Tobii glasses. Though the Tobii glasses are not necessarily a "gold standard" PCCR tracker, with sampling rates lower than that of the EyeLink and other lab-based trackers, this work presents an important comparison between PCCR approaches and methods that utilize the RGB webcam. Our work builds upon these successes (though using a different gaze-tracking system) to examine if webcam data is sufficient for real-time modeling of TUT and comprehension.

Finally, almost all of the work reviewed above has used predominantly White samples to build detectors of TUT and comprehension, limiting its generalizability and potential scalability. Here we intentionally collected data from two different populations (Study 1: predominantly white university students; Study 2: mostly non-White adults recruited on the online platform Prolific) to see how our models generalize across these populations, i.e., to check for algorithmic bias in the eye-tracking technology.

Methods

Below we describe our general data collection method for two different studies, noting any (minor) differences between the two.

Participants

In Study 1, 105 University of New Hampshire students participated in the experiment (ages 18–25, 77 self-identifying as female, 27 as male, 1 as non-binary; 83.1% White) for course credit in their Psychology-related courses. In Study 2, one hundred seventy-three participants (ages 18–52, 130 self-identifying as female, 40 as male, 3 as non-binary) were recruited through Prolific, an online data collection platform that allows individuals to sign up and receive compensation for participating in research studies. Participants were paid $4 for completing the study. To create a more diverse sample, we used the Prolific selection criteria to oversample participants of color. See Table 6 for a complete breakdown of participants by race.

The location of both studies was at the participant's own discretion (wherever they chose to complete the online study), without a researcher present, and no video (other than for gaze tracking) was recorded. As a result, we have no structured way to evaluate when tracking error is a fault of the tracker and/or when it might be a context issue (e.g., the participant is looking down, or covering their face with their hand etc.).

Materials

Task. The study used a narrative anticipation task which involved reading 65 narrative stories taken from Cranford and Moss (2018). The goal of this task was for participants to make an inference about the ending of a story, based on information given, which is a common exercise for teaching and/or developing reading comprehension skills. Each story consisted of three sentences and had three possible endings. Each ending is initially plausible, but there is only one appropriate ending after reading all three sentences. An example story is: "Larry always wanted to know what it was like to live in a foreign country. He went to read at his favorite store on main street. The steam rose from the cup as Larry brought it to his lips and slowly…". The three ending options were: 1) "sipped coffee", 2) "bought muffin", and 3) "rolled marble". As the story unfolds, the incorrect options become less plausible until the reader can make the inference that the only appropriate ending is "sipped coffee." We note that this task does not reflect reading long, extended texts, but rather a reading comprehension skill-building exercise common in English-language learning, standardized tests (and test-prep), and other K-12 learning platforms.

Webcam-based gaze tracking. Gaze locations were collected using Webgazer (Papoutsaki et al., 2016). Webgazer is an online, webcam-based eye-tracker written in Javascript that can be integrated in any website to infer gaze locations in real time using the user's webcam. Webgazer initially uses facial and eye detection algorithms to detect pupil locations and represents the eye as an image patch. It then maps pupil locations and eye features to gaze locations using a ridge-regression model. Webgazer uses all eye features within a temporal interval of 500 ms when determining the onscreen x- and y-coordinates.

Based on user interactions such as clicks and mouse movements that normally occur during web navigation, Webgazer is also able to continually self-calibrate to maintain mapping accuracy. In a lab study, Webgazer achieved 4.17° gaze accuracy (Papoutsaki et al., 2016). As a point of comparison, commercial eye-trackers achieve <1° gaze accuracy. It should be noted that because Webgazer runs on the client side, sampling rate cannot be guaranteed and varies as a result of available resources.

Procedure

After participants provided consent and remote connection to the webcam had been established by the software, participants completed Webgazer's calibration process. Participants were directed to look at a red dot as it moved to 20 different locations around the screen. In Study 2, a pseudocalibration was additionally used to help account for possible calibration drift over time. The pseudocalibration did not affect Webgazer's calibration but created an adjustment that could be applied to the gaze locations reported by Webgazer. In the pseudocalibration, participants looked at red dots at four locations which corresponded to the three options and center of the screen. The pseudocalibration was performed before the first main trial and after any trial in which no pseudocalibrated gaze locations were located in the option locations during the choice screen. Aside from the population differences and pseudo-calibration, nothing else was changed across the two datasets. For this initial feasibility study and to ensure that the two datasets were comparable, the pseudocalibration was not used to correct any data collected in the second study

After calibration, participants completed the narrative anticipation task. The studies used a single between-participants manipulation with participants randomly assigned to one of two conditions related to how the stimuli were delivered. This manipulation is not relevant for the current research, the goal of which is to build a generalizable detector that works in either of the two conditions. We nevertheless describe the design in full here in case others wish to replicate our work. The two conditions were: 1) “audio”, in which participants heard a reading of each sentence; and 2) “visual”, in which participants read each sentence presented on the screen. Figure 1 shows the timeline for a trial. For each trial, participants were initially presented with the three endings for a story and asked to familiarize themselves with the on-screen location of each option before progressing. Participants controlled when to move onto the next sentence, but in the audio condition they could not progress until the reading of the sentence was finished. In both conditions, the three answer options were displayed on the screen at all times, making it possible to collect gaze data relative to the positions of those opinions throughout each trial.

Comprehension assessment. After reading or hearing all three sentences, participants were asked to click on the option that best completes the story. Participants completed 5 practice trials and 60 main trials. After completion of the narrative anticipation task, participants completed a demographic survey. The experiment lasted ~40 minutes.

Task-unrelated thought (TUT) probes. Detectors of TUT have almost exclusively relied on self-reports from participants to determine the ground truth data labels. The gold-standard in the field is to use a probe-caught method (Varao-Sousa & Kingstone, 2019; Weinstein, 2018), whereby participants are interrupted periodically to report on whether they are off-task (thinking about something else) or on-task at the current moment. Previous work has vetted the probe-caught method in a variety of ways, showing consistent results and reliable correlations with eye-gaze, pupillometry, reaction times, and performance (Foulsham et al., 2013; Franklin et al., 2013; McVay & Kane, 2012; Randall et al., 2014).

We used this method to probe participants on half of the stories. On these probe trials, participants were asked to report whether they were thinking about the story (on-task) or something else (off-task). The probes' timing was balanced across sentences to prevent predictability, occurring for thirty stories with ten after the first sentence, ten after the second sentence, and ten after the third sentence. Timing of the probe – both in terms of which story probes occurred and at what location within the story (sentence 1, 2, or 3)— was randomly assigned.

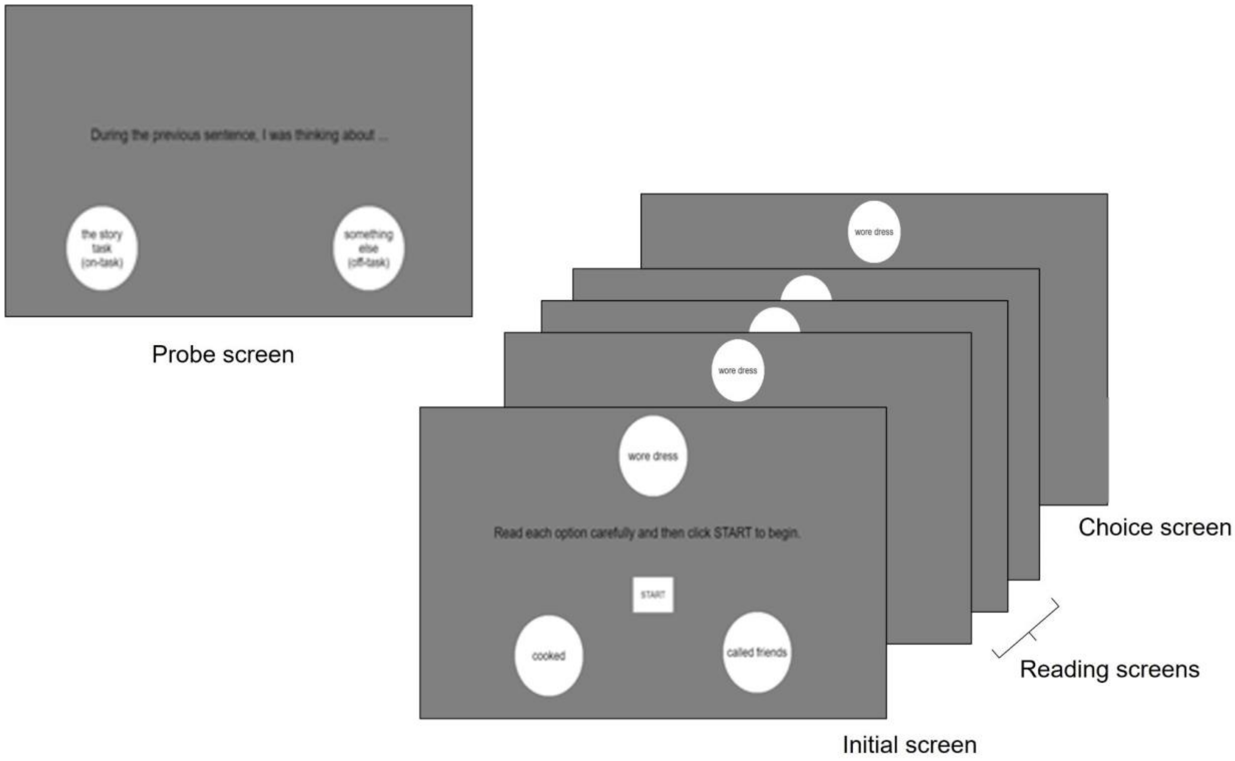

During the previous sentence, I was thinking about ...
the story task (on-task)
something else (off-task)
Probe screen
wore dress
wore dress
wore dress
Read each option carefully and then click START to begin.
START
cooked
called friends
Choice screen
Reading screens
Initial screen

Figure 1. Sample trial sequence. The probe screen could occur after any reading screen**.**

Feature Engineering

For TUT models, we calculated gaze features from the time of screen up until the probe occurred to avoid using any data from after the probe. Because the probe could appear at different times throughout the trials, only the previous screen was used to ensure a consistent amount of data per instance. For example, if the probe occurred just after screen two, the gaze data from screen two would be used to predict TUT; if the probe occurred just after screen three, then screen three would be used etc. No eye-tracking data were used from the probe screen itself. For the comprehension models, the question had a consistent placement at the end of the reading (three screens), allowing us to use more data and still have consistent data volume across instances. We calculated features from all three screens of reading prior to the user being presented with the question. If a probe occurred during a trial, the gaze data from the probe screen was excluded. No data from the choice screen were used.

We converted the raw gaze data into features to use in our prediction models. Based on previous work (Bixler & D'Mello, 2016; Hutt et al., 2019), we investigated both global and local gaze features. We did not consider contextual/interaction features such as response time, although some of these features are implicitly encoded in the gaze features; e.g., more samples likely correspond to longer response times. The specific global and local feature categories that we used are described below.

Global gaze features. Global gaze features focus on general gaze patterns and are independent of the content on the screen. The global features we used were selected to relate to previous studies of comprehension and TUT, while allowing for the reduced accuracy that was expected from Webgazer compared to in-laboratory systems. Specifically, for each sentence/screen we calculated: 1) the number of gaze samples, which is a measure of how much valid gaze data there was during a sentence, giving an overview of how much the participant was looking at the screen; 2) the number of unique gaze samples, which is a measure of the number of distinct screen locations where a user looked. This value, though correlated with feature 1, removes duplicate screen locations, so gives a measure how much gaze may be moving around the screen; this is then further extended by 3) the dispersion of gaze points, which we quantified as the root mean square of the distances of each fixation to the average gaze position and is a measure of how spread out the gaze was. Because our data collection was based on variable and unknown sampling rates, we did not attempt to calculate fixation durations or identify saccades, as has been done in previous work. Additionally, our metrics do not correct for sampling rate, instead evaluating the robustness of gaze tracking just from the raw data. Future work should address this limitation.

Local gaze features. In contrast to the global features, local features encode where the gaze is fixated and thus were based on both gaze location and the relative screen content. To calculate these features, we first defined the three option locations and the center of the screen (sentence location) as areas of interest (see Figure 2). For each page/screen, we then calculated the time spent on each AOI. Finally, we added context to these actions, based on the task (see Table 1). As this was an initial experiment, the features selected represent more fundamental local features that though relevant to this work and reading literature, again ackbowledged the expected reduced accuracy of the gaze tracker. We did not include features such as gaze on individual words that we considered to be too task-specific, given that our goal was to obtain a more general understanding of the potential for generalizable gaze tracking in this domain.

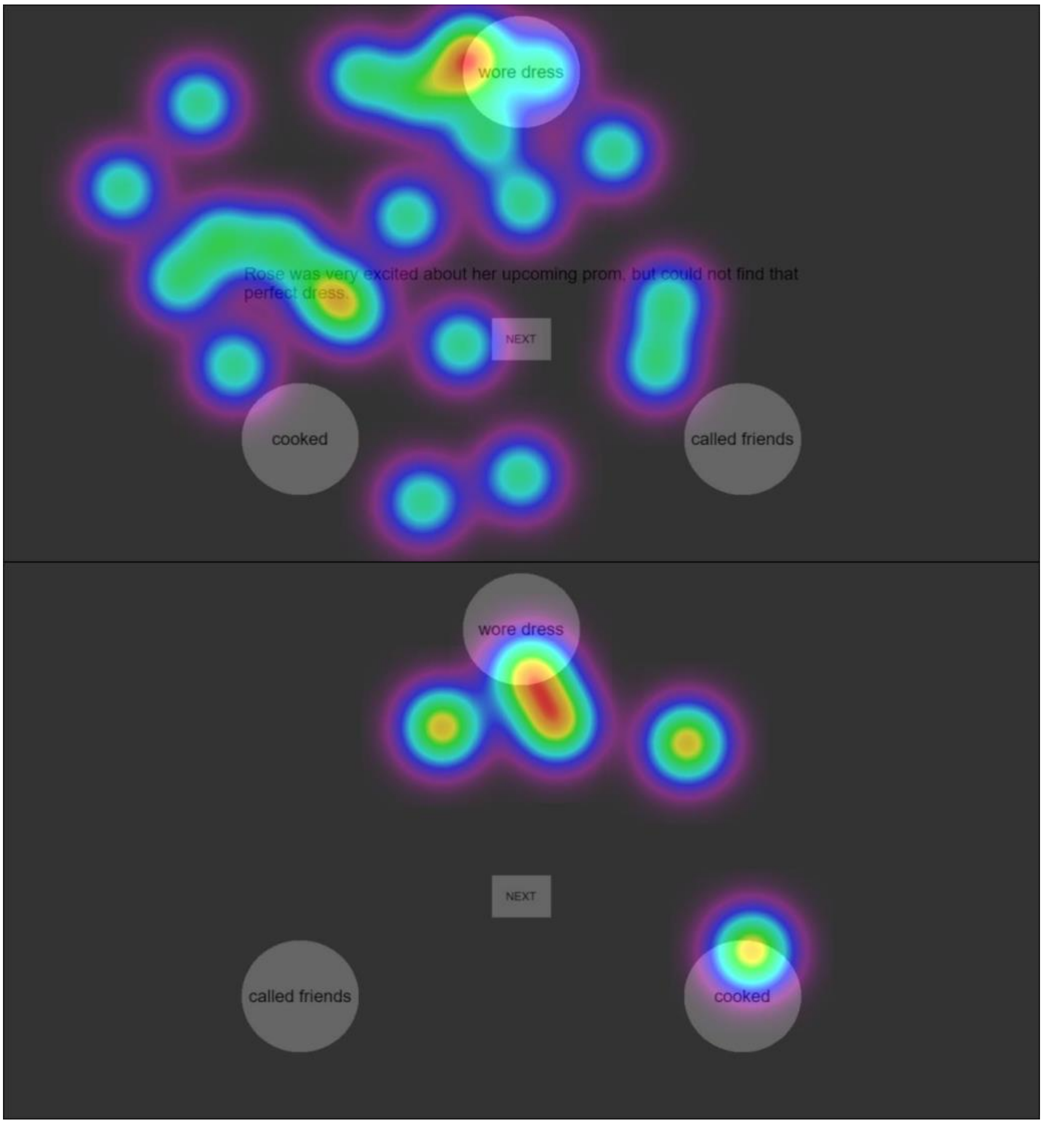
Table 1. Local Feature Description Per Page

Feature	**Description**
Option 1 Gaze	Number of Gaze points on Option 1
Option 2 Gaze	Number of Gaze points on Option 2
Option 3 Gaze	Number of Gaze points on Option 3
All Options Gaze	Sum of three above features
Sentence/Center Gaze	Number of Gaze points on the Sentence/Center *
Correct Answer Gaze	Number of Gaze points on the correct option
Incorrect Answer Gaze	Number of Gaze points on incorrect options

*Note. In the case where the sentence was not shown, this feature is gaze in the center of the screen.

Figure 2 shows example heatmaps of one participant's eye gaze during a reading screen, and before providing their answer. Each of the three options were shown in circles with diameters equal to 20% of the screen height (e.g., if the screen height was 100 pixels, the diameter of the stimuli would be 20 pixels), each of which was associated with a slightly larger AOI (with diameters equal to 30% of the screen height). In this case, the gaze followed some expected patterns, including higher gaze densities around the AOIs on the screen.

However, calibration drift was also evident, for example in the top image where calibration has likely drifted to the left. Subsequent analyses that used the eye-gaze data to predict cognitive states (TUT and comprehension) thus allowed a margin for error in AOI calculations.



Rose was very excited about her upcoming prom, but could not find that perfect dress.

Figure 2. Heatmap overlay showing a participant's eye gaze during a reading page (Top) and the choice page (bottom) of the task. Red indicates high concentration of fixations, purple low concentration of fixations.

Classification Models and Validation

To relate global and local gaze features to the TUT probes and reading-comprehension scores, we used scikit-learn (Pedregosa et al., 2011) to implement five classifiers (Logistic

Regression, Random Forest, Gradient Boosted, Support Vector Machine, and Decision Tree). We also implemented XGBoost with a separate library (Chen & Guestrin, 2016). Where appropriate, hyperparameters were tuned on the training set using scikit-learn's cross-validated grid search (Pedregosa et al., 2011). Because of the limited volume of data and feature space, we did not consider neural networks or deep learning approaches at this time (see Future Work, below).

We validated the models with a participant-level, 10-fold cross-validation scheme. This process ensures that no instances of any individual participant could appear in both the training and test sets within a fold. All features were z-scored by condition (visual or audio) within each fold. We used the training data to calculate the statistics needed for z-scoring (mean, standard deviation, max, min), which were then subsequently applied to testing sets.

For both TUT and comprehension, we observed substantial imbalance between the classes (i.e., many more instances were not TUT than were). Class imbalance can present challenges, because supervised learning methods tend to bias predictions towards the majority class label. To compensate for this concern, we used the SMOTE algorithm (Chawla et al., 2002) to create synthetic instances of the minority class by interpolating feature values between an instance and its randomly chosen nearest neighbors until the classes were equated. SMOTE was applied only on the training sets. The original class distributions were maintained in the testing sets to ensure the validity of the results.

Evaluation. The analyses are described below, first for TUT and then for comprehension (correct/incorrect answers). Given the class imbalance in our data, we report precision, recall, and F_1 scores as metrics for each class. To support easier comparison with previous work, we also report kappa value (Landis & Koch, 1977) to correct for chance. Precision, which provides detail on how accurate the model is for a specific class, was calculated as the number of true positives divided by the total number of true instances (in ground truth). Thus, for example, 40 correct predictions about of 100 instances of a given class X corresponds to precision of 0.4. Recall, sometimes also called true positive rate, or sensitivity, refers to how many instances of class X were predicted correctly. Both metrics are informative about our model, and can present a trade-off; for example, if you over predict class X you may increase recall but decrease precision. F_1 is defined as the harmonic mean between precision and recall, in order to combine the two scores into one meaningful evaluation. F_1 was calculated as:

$$2\frac{precision \ \times recall}{precision + recall}$$

The highest possible F_1 score is 1 (indicating perfect precision and recall), and the lowest possible score is 0 (indicating that either precision or recall is 0). We report the individual metrics along with the combined metric in acknowledgement that whereas F_1 weights precision and recall equally, in practice different types of misclassification can be more or less important (Hand & Christen, 2018).

To support easier comparison with previous work, we also report kappa values (Landis & Koch, 1977). The kappa metric is similar to F_1 score in that it can be viewed as a combination of precision and recall. However, unlike F_1 score, the kappa metric attempts to correct for chance. Kappa values > 0 indicate improvement over chance, whereas a kappa value of 1 indicates perfect classification.

Baseline/Chance Models for Comparison. We included two different "baselines" for model-comparison purposes. This first chance baseline was generated using the Dummy Classifier in scikit-learn. The Dummy Classifier randomly assigns a label based on the base rate of the training sample; e.g., if 25% of the training data was TUT, then there is a 25% chance the dummy classifier will predict TUT.

As an additional baseline model, we trained a model using interaction data that is separate from the webcam data. Specifically, we used the time that participants spent on either the page before the probe (for TUT) or all three pages (for comprehension), because reading time can be correlated with TUT and comprehension (Mills et al., 2017). The main purpose of these baselines was to determine whether the eye tracking was providing information beyond what we could get from less complex means, such as basic log files. By comparing to these baseline models, we can assess if, and by how much, adding gaze data can help to improve predictions of TUT and comprehension beyond reading time. Put even more simply, we can ask, Is the eye tracking worth the trouble?

Results

Before considering the results of our predictive models, we first consider the gaze itself. Though it is not possible in this experimental design to statistically evaluate the quality of the gaze recognition, or indeed the precision of the points recorded, we are able to anecdotally analyze the data through the heatmaps generated. As noted above, we observed drift in the recordings, and that in the reading task, gaze did not always appear to be on the sentence being read. This is somewhat to be expected, prior work has consistently reported lower gaze precision for webcam-based approaches (Zhang et al., 2019). This challenge has been addressed in the past by looking at relative changes in gaze patterns rather than specific gaze locations (D'Mello & Mills, 2021; Hutt et al., 2016; Mills et al., 2020), an approach we adopted as well.

To evaluate the predictive models, we begin by presenting results using a combined dataset with all participants from both studies in a single model. This approach allowed us to test the feasibility of the webcam-based eye-tracker for detecting TUT and comprehension with the largest, most diverse dataset under realistic conditions likely to introduce multiple, uncontrolled sources of error. After examining the combined dataset results, we then present model performance for each individual study, as well as cross-training results (train on Study 1, test on Study 2, and vice versa). Finally, we conclude with a slicing analysis to determine if model performance changes under various environmental conditions or across racial/ethnic subgroups.

Overall Models

Correctness. Across all participant and conditions, correct answers were given 88% of the time. Across detection methods, gaze patterns measured via the webcam-based eye-tracker could be used to predict correct responses (indicating comprehension) better than chance (Table 2). The best performance was achieved by models using only Local features (the time spent in the three AOIs corresponding to the locations of the alternatives on the screen; kappa = 0.57; there was no reliable difference in performance when the models were tested separately for conditions in which the stimuli were presented as text or audio, t-test *p*=0.19). Models using only Global features (the number of gaze points, number of unique gaze points, and dispersion of gaze points) were also above chance but only marginally outperformed the interaction feature baseline (kappa = 0.15 vs 0.11). Combining these two feature sets produced worse performance than just Local features, potentially due to increased noise in the dataset. Thus, under these (admittedly limited) conditions, our results are encouraging and suggest that webcam-based eye tracking can be useful for assessing constructs like comprehension in online environments.

In terms of individual label values (correct or incorrect), the models also showed an increase in F_1 when predicting incorrect responses relative to chance, with higher values for both precision and recall. Thus, the model could be used to identify when someone did not understand a text as well as when they did. This result is likely relevant for future applications, where the ability to diagnose when someone makes an incorrect inference can be a possible place for real-time interventions.

Table 2. Results for predicting Correctness and Incorrectness

				Correct			Incorrect		
Features	**Classifier**	**SMOTE**	**Kappa**	F_1	**Precision**	**Recall**	F_1	**Precision**	**Recall**

Chance Baseline		N	0.00	0.90	0.91	0.90	0.11	0.11	0.11
Reading Time Baseline	Gradient Boosted	Y	0.11	0.86	0.84	0.92	0.13	0.20	0.05
Global	Gradient Boosted	Y	0.15	0.91	0.90	0.92	0.21	0.25	0.19
Local	XGBoost	Y	0.57	0.96	0.93	0.99	0.60	0.94	0.44
Global + Local	XGBoost	Y	0.54	0.96	0.94	0.98	0.57	0.79	0.46

Given that there were two conditions for how the stimuli was delivered in the experiment (though these were not explicitly examined here) we compared participant level accuracy of the best performing model across condition with a t-test. Results showed no significant difference (p=.19).

TUT. Gaze patterns from the webcam-based eye-tracker were also able to be used to predict TUT (Table 3). Specifically, the results indicate that: 1) all models outperformed the chance-baseline, and 2) the combined Global + Local model had the best performance (which was similar for text or audio conditions across participants, p=0.36). Our finding that the most effective TUT prediction for this task relies on a mixture of general (Global) and context-specific (Local) features differs from past results. In general, global features have tended to be most predictive of TUT, whereas context-sensitive (local) features have provided a variety of results across different domains and tasks (Bixler & D'Mello, 2016; Hutt et al., 2019; Hutt, Hardey et al., 2017). For example, local features provided improved TUT detection in reading from extended texts versus global features alone (Bixler & D'Mello, 2016). Our work shows that this benefit of local features can also be found in the kind of brief comprehension exercises we used.

These results are somewhat modest in magnitude but are in line with previous studies using commercial eye-trackers (Bixler et al., 2015) that reported kappa values of ~0.20 (Bixler et al., 2015; Blanchard et al., 2014; Mills et al., 2016). They thus demonstrate that even with lower-quality, scalable sensing, we can still harness the so-called "eye-mind link" and detect TUT with webcam eye-tracking data. The eye-gaze models notably outperform the baseline response time-only model, indicating that there is a valid signal being detected. We note from the precision and recall scores indicate that a false positive is less likely than a false negative. Though there is still inaccuracy, this is potentially useful if triggering interventions and shows potential for future work.

We again compared participant level accuracy across the two stimuli conditions in the experiment (whether the sentence was presented as text, or audio) with a t-test. We observed no significant difference in model accuracy (*p* = .36)

Table 3. Results for predicting TUT

Feature Set	Classifier	SMOTE	Kappa	Weighted F1 Score	F1_TUT	Prec. TUT	Recall TUT
Chance Baseline		N	0.00	0.53	0.18	0.18	0.18
Reading Time Baseline	Logistic Regression	Y	0.04	0.60	0.10	0.25	0.04
Global	Gradient Boosted	Y	0.07	0.65	0.24	0.32	0.20
Local	Gradient Boosted	Y	0.12	0.67	0.22	0.32	0.17
Global + Local	XG Boost	Y	0.15	0.69	0.25	0.36	0.21

Convergent validity. We explored the convergent validity of our models by calculating a set of correlations derived from the model's predictions and the ground-truth data. Each correlation was calculated at the participant level in order to avoid violating independence assumptions. We calculated four participant level values: TUT Ground Truth (proportion of probes to which participants reported TUT), Correctness Ground Truth (proportion of correct inferences made), TUT predicted (the average TUT prediction for that participant) and Correctness Prediction (the average Correctness prediction for that participant. The resulting correlation matrix is shown in Table 4.

The models' predictions of TUT and correctness were each positively correlated with their respective ground-truth labels. Specifically, actual and predicted TUT were weakly correlated (Spearman's *rho* = 0.27), whereas actual and predicted correctness were strongly correlated (*rho* = 0.77, which is comparable to correlations reported in D'Mello et al., 2020, with higher-quality equipment). For a few of the participants, the model-predicted rate for question correctness was identical to the ground-truth rate, contributing to this high correlation.

Moreover, the models' predicted rates of TUT were negatively correlated with ground-truth correctness (average question score). The magnitude of this negative correlation was somewhat similar when using predicted TUT (*rho* = -0.23) or participant-level ground-truth TUT rate (measured as the average of probe responses; *rho* = -0.11). This test of convergent validity is based on the consistent negative relationship between self-reported instances of TUT and reading comprehension scores in the literature, with an average reported effect size of *r* = -0.28 (D'Mello & Mills, 2021).

Table 4. Correlation Matrix for Student Level TUT and Correctness Rates, Both Ground Truth Values and Predicted Values from the Best Reported Models.

	TUT Ground Truth	**Correctness Ground Truth**	**TUT Predicted**
Correctness Ground Truth	-0.109		
TUT Predicted	0.277*	-0.272*	
Correctness Predicted	-0.126*	0.774*	-0.231*

*Indicates p < 0.05

Feature Importance. To evaluate the importance of each gaze-based feature to our predictive models, we calculated SHapley Additive exPlanations (SHAP) values (Lundberg & Lee, 2017) using the shap library in python. For the two best models of TUT and Correctness reported above, we computed the mean absolute SHAP value of each feature, per fold, and then averaged across folds to generate one value per feature (each between 0 and 1).

TUT. The variance in feature importance was low (*SD* = 0.01), implying that an ensemble of features was necessary for effective prediction. The top three features were all global features, characterizing the number of gaze points on a given page. This result aligns with earlier research using eye gaze for TUT detection, which has shown that the number of fixations is highly predictive (Bixler & D'Mello, 2016; Hutt et al., 2019). The most predictive feature was number of gaze points on the third page. This result indicates that the last page provided more predictive power than the previous two but could reflect the proximity of the third page to the probe.

Correctness. The variance in feature importance was also very low (*SD* = 0.01), although slightly larger than for TUT. The most important features for this model were related to the number of gaze points on the answer options across all three sentences and the time spent looking at incorrect options. It should be noted that feature importance values for the three options were very close to each other (a range of 0.003), indicating that readers were most likely to be correct if they had spent time considering all options rather than focusing on one answer (even if it was the correct one). In both cases, we note the low variance between SHAPley values. Additional feature engineering and refinement may provide a more detailed insight into the relationships between individual eye movements and these two constructs.

Individual Studies and Cross-training Analyses

The above analyses combined both datasets to: 1) increase the amount of data available for training the model, and 2) avoid overfitting to a particular sample at the outset. Below we report analyses that treated each study as a separate source of data to further probe the reliability and generalizability of our models that use webcam-based eye tracking to predict TUT and comprehension. We examined different combinations of training and testing sets (see Table 5). In cases where the model was trained and tested on data from the same study, the same cross validation approaches described above were employed. In cases where the training and test sets were from different studies, models were trained on the entire training data set and tested on the entire testing data set. We interpret the results in terms of whether there are algorithmic biases manifesting as a degradation in prediction from one sample to another; i.e., does training the model on the predominantly White sample generalize to a more diverse sample, and vice versa?

Individual dataset models. Results from training and testing on the individual datasets (i.e., train on Study 1 or 2, test on the same Study) were similar to those from the combined datasets (Table 5; also see Supplementary materials for full details on individual dataset results). We did not observe major changes in the kappa values for the individual datasets (e.g., train on Study 1, test on Study 1) compared to the combined dataset results presented above. Study 1 slightly outperformed the combined data for TUT (0.15 for combined compared to 0.19 in Study 1), whereas Study 2 slightly outperforms the combined dataset for correctness (0.55 for combined compared to 0.58 in Study 2). These findings suggest that there were no strong and systematic biases in the eye-tracking system that might have affected its ability to collect interpretable gaze data within each of the two study populations.

Cross-training models. When models were trained on one dataset and tested on the other dataset (keeping them completely independent), there was a slight degradation in performance. However, all models still performed above the respective chance baselines. Moreover, the degradation was bidirectional: in all cases, training on one sample (Study 1 or Study 2) led to a degradation in performance when tested using the other sample. Although these results indicate some level of generalizability between the two datasets, the differences should be noted. These results serve as a reminder of the importance of context and eventual use case when collecting/selecting training data.

Table 5. Kappa values from cross training experiments.

Model		**Test Study 1**	**Test Study 2**
Correct Inference	**Train Study 1**	0.55	0.41
	Train Study 2	0.30	0.58
TUT	**Train Study 1**	0.19	0.15
	Train Study 2	0.09	0.12

Slicing Analyses

To examine the data in finer detail, we used slicing analyses (Gardner et al., 2019) to identify if and how the predictiveness of the webcam data for reading comprehension and TUT differed for particular subpopulations. The best performing models for each construct (tables 2 and 3, respectively) were evaluated in the slicing analysis, as well as models trained on each individual study. We considered four relevant subpopulations: 1) whether or not the participant wore glasses, 2) the lighting of the room the participant was in, 3) whether they reported having ever received treatment for a neurological problem, and 4) race/ethnicity. For each of the four categories, we relied on participant self-report and self-identification. For each subpopulation, we calculated model performance (kappa value) using just the instances from that subpopulation. For example, in Study 1, 15 participants wore glasses, so the model is then evaluated on instances only from those 15 participants.

Results of the slicing analysis are shown in Table 6. For both correctness and TUT, the results were relatively robust to wearing glasses or not, and to lighting changes. We did, however, notice a slight decrease in performance for individuals who self-identified as having a neurological condition, amounting to a 7% reduction in accurately predicting correctness and a 5% reduction for TUT.

The results for correctness were also relatively robust to differences in race/ethnicity, despite some variation across different racial categories. The overall kappa was 0.57, with the kappas for Asian/Pacific Islander, Black/African American, and Hispanic/Latinx in a comparable range of 0.56 to 0.59. There was a slight drop (~6%) for Other-identifying participants at 0.51. There was a more substantial drop for the group of participants who identified as Native American (kappa = 0.26); however, this group contained only two students and thus should be interpreted with caution and followed up with future studies using larger samples (see Future Work). Overall, we interpret these results to be encouraging, providing the first (to our knowledge) glimpse into how gaze-based detectors of TUT/Comprehension may differ across these moderators.

This relative stability in correct inference prediction implies that the cause of the variation in detector performance may not be caused entirely by the quality of the eye tracking, or a potential bias in the tracking. In addition, the variation may result from noise in the self-reports and how participants responded (e.g., how comfortable participants are reporting being off-task), patterns in the simple gaze features used or the algorithm, or other factors. Identifying these factors will require further study, but it is nevertheless an important result to know that such variability in detection is occurring, which is an important step towards fixing it (Baker & Hawn, in press).

In contrast, TUT prediction was less stable across race/ethnicity. There was a reduction in performance for TUT detection for participants that identified as Black/African American (kappa = 0.04) versus those that identified as White (kappa = 0.17). This variation is the difference between a functioning (albeit modest) detector and chance prediction levels, suggesting that some aspect of the data or detector was different for these participants (see Discussion). However, the same reduction was not observed when predicting correct inferences for the Black-identifying participants (kappa = 0.57 for all students, 0.59 for White participants versus 0.56 for Black participants in the model trained on all data). Though there is still variation for predicting correct inferences across race/ethnicity, especially when training on just study 2 data (kappa = 0.67 for White students and 0.51 for Black students), the relative change is lower, and the resulting detector would still be considered effective (as opposed to chance level for TUT).

In general, these results suggest that this tracking methodology and the detection that it facilitates are acceptably robust for the task of correctness prediction. However, it is necessary to conduct additional work before including TUT prediction, especially with the variability across race/ethnicities. For example, our analyses was underpowered for Native American students (*N*=2), corresponding to an ineffective model. Additional analysis (perhaps less quantitative) is required before we may draw any general conclusions. If pursuing a more robust predictive model from the data alone, additional training data would be required before the models can be adequately tested on this population. Thus, although these analyses are important to conduct, they are not intended to be conclusive or prescriptive for who and when webcam-based eye tracking will work.

Table 6. Slicing Analysis by different moderators

		Participants (N)			**Correct Inference kappa**			**TUT kappa**		
Moderator	**Value**	**All**	**Study 1**	**Study 2**	**All**	**Study 1**	**Study 2**	**All**	**Study 1**	**Study 2**
	All Participants	278	105	173	0.57	0.55	0.58	0.15	0.19	0.12
Glasses	Wore glasses	66	15	51	0.57	0.61	0.56	0.15	0.17	0.12
	Did not wear glasses	212	90	122	0.58	0.57	0.58	0.15	0.20	0.09
Lighting	Well Lit	186	63	123	0.56	0.54	0.56	0.13	0.17	0.11
	Dimly Lit	84	38	46	0.60	0.58	0.61	0.17	0.15	0.09
	Lights Off	8	4	4	0.56	0.49	0.50	-0.20	0.02	-0.04
Neurological Condition	Yes	19	14	5	0.51	0.51	0.43	0.10	0.19	0.10
	No	259	91	168	0.58	0.58	0.58	0.15	0.09	0.12
Race/ Ethnicity	White/Caucasian	125	88	37	0.59	0.56	0.67	0.17	0.19	0.18
	Latinx/Hispanic	49	7	42	0.58	0.63	0.58	0.12	0.08	0.12

Asian/Pacific Islander	44	6	38	0.56	0.49	0.61	0.12	0.16	0.11
Black/African American	42	3	39	0.56	0.71	0.51	0.04	0.15	0.02
Native American	2	0	2	0.26		0.14	-0.20		-0.20
Other	16	1	15	0.51	0.00	0.55	0.22	0.11	0.23

General Discussion

The idea that eye gaze behaviors provide a window into the mind has led to important research discoveries over many decades (Huey, 1908; Rayner, 1998; Rayner, Chace et al., 2006). However, the high cost of eye-trackers has severely limited the scalability of existing approaches to detect cognitive states in real-time. Here we attempt to address this issue by integrating Webgazer into an online educational task in order to build models of TUT and comprehension during reading, with the goal of showing a proof-of-concept method for scalable eye-tracking.

Main findings

This work demonstrates the feasibility of using webcams for modeling internal states during learning. Though this data stream is of a lower fidelity than a typical PCCR eye-tracker (e.g., measured at 30 Hz rather than >60 Hz, with reduced precision and accuracy), this work demonstrates that with appropriate calibration, Webgazer can be sufficient in most cases for user modeling. Despite having only a single calibration for a 40-minute task, our models still made predictions at above chance levels. Our models also performed better (in terms of kappa values for the combined dataset, as well as for each dataset individually) than a model trained on participant response time alone, demonstrating that using webcams for this task was a useful augmentation. Moreover, both TUT and comprehension models performed comparably (as measured by kappa values) to prior work using research-grade equipment, for most groups of learners and tasks. This result is particularly encouraging given the poorer quality of our gaze data, which nonetheless was sufficient to model users at above chance rates and leverage this cheaper, more accessible technology to provide cognitive insights.

Throughout this work, we have used a chance baseline as a comparison point, with models performing above chance being considered successful. By this metric, our findings are roughly comparable to others using research-grade eye-tracking (Bixler et al., 2015, whose highest reported kappa value for gaze was 0.15) or EEG signals (Dong et al., 2021, whose MCC was .206). No such comparable values exist for comprehension, but given that we outperform the conventional rates for TUT, we speculate that webcam-based eye-tracking may also be suitable for real-world applications. In general, the limited degree to which our model is better than chance suggests that if used for intervention, it should be applied in "fail-soft" interventions (see discussion below). More generally, more nuanced definitions of "successful" will require taking into account the intended application and risk

Our slicing analyses indicated that our detectors were largely robust to individual differences of race/ethnicity (with one key exception; see discussion below), and whether participants wore eyeglasses, as well as conditional differences such as lighting. We also found that models trained were generalizable between the two datasets, though did experience some performance degradation in both directions (e.g., train Study 1, test on Study 2, and train Study 2, test on Study 1 both saw a drop off in performance).

The one notable evidence of variability in our slicing analysis was the drop in performance for TUT detection across race/ethnicity, with lower performance for participants who identified as Black/African American or Native American. This reduction in performance could be for many reasons, such as differences in the self-reporting of TUT or in the simple eye-movement features. It seems unlikely to be a result of egregious computer-vision issues with the tracking, such as contrast issues or biases in the facial detection, given that the gaze tracking was sufficient for successful for correct inference prediction. However, more analyses are needed to rule out more minor differences in how the eyes are tracked. Future research is also needed to determine why we observed such differences, as well as how webcam-based eye-tracking and TUT detection can be improved for *all* participants. Given the broad, robust nature of the tracking for correct inference prediction, our work provides some initial evidence of feasibility for the webcam-based method in general. More evaluation is needed to determine which tasks this approach can be used for, without concerns of bias.

This caveat notwithstanding, our work serves as a proof-of-concept for a future real-time detector that leverages webcam data. All data used in the models came from interactions prior to the prediction point and could be gathered in real-time (either the previous page of reading or the previous three pages). Similarly, the model accuracy is within the ranges of detectors previously used in the literature for real-time intervention. This work adds to a growing body of research examining the feasibility of webcam-based eye tracking and adds further credence to their use a proxy for PCCR gaze tracking. Furthermore, it offers potential to scale up decades of research examining links between comprehension, TUT, and eye gaze, taking these experiments into new, ecologically valid environments.

Applications

It is perhaps easier to start with how this approach should *not* be used. Sensor technologies such as webcams hold great potential but also pose great risk. This method should not be used to monitor students (or anyone) without their permission, or without transparency as to how their data is being collected/stored. Any future application should clearly inform the user of what data is being collected and how it is being used.

Assuming careful consideration of privacy and transparency, these methods have many possible applications in software development. Eye tracking has consistently been used to identify interaction patterns and improve software development (Jacob, 1995; Kukkonen, 2005). Being able to monitor constructs such as attention in a cheap and scalable way can improve this process and help developers understand when materials or software is not engaging the audience (Toreini et al., 2020).

More specific to educational contexts, our work sets the foundation for improving the scalability of modeling techniques with the end goal of improving research methods, learning technologies, and student experiences. For example, our results show that webcam-based detectors provide more accurate detection than response time detector and are thus likely to be more useful for real-time intervention techniques that correct deficits “in the moment.” A student who is unlikely to answer a comprehension question correctly could be advised to read the text again before attempting the question or be given hints about which parts of the story are the most critical.

It is also important to consider that any intervention must rely on detection, which is inherently imperfect, especially in the case of TUT detection. False alarms (predicting someone is off-task when they are not) and misses (missing an instance of TUT) are both possible and must be accounted for in any application. In our view, detection does not need to be perfect to be useful. Indeed, prior work has used imperfect detection to trigger meaningful interventions for TUT, using a probabilistic approach (e.g., if the likelihood of TUT is 70%, then there is a 70% chance of an intervention) (Mills et al., 2020). Any interventions should also be designed to “fail-soft” in that there are no harmful effects to learning if delivered incorrectly. For example, an intervention may ask students to provide a self-explanation of what they had just read if they have been detected to be off-task. If a student is not off-task, this will reinforce what they already know without damaging the experience too greatly. A student that is off-task will be prompted to realize that they are missing details and go back.

The comprehension detector, has higher precision than recall for students who have not understood the text, meaning that a miss (predicting comprehension when the student has not understood the text) is more likely than a false positive (predicting a student has not understood the text when they have). In this case, the confidence in an individual prediction can be high, which is useful for most applications and reduces the need for a “fail-soft” approach, but more refinement is needed to reduce the number of misses in the model, to guarantee that detector is supporting all students.

Given these implementation considerations, the detectors presented in this work provide proof-of-concept for potential real-time integration using webcam eye-tracking solutions. Though there are known inaccuracies, these inaccuracies can, in principle, be accounted for to provide valuable real-time information and adaptation. Though we cannot directly measure the inaccuracies in this work due to not having an appropriate comparison set, previous work with WebGazer has already been evaluated to have error rates of up to 4 degrees. It is thus encouraging that despite this, our anecdotal evaluation shows that gaze is adaptive to the stimuli, and falls on expected AOIs in many cases.

Limitations and Future Work

There were several limitations of this work. Firstly, our study was designed to test low-cost eye tracking, by using the webcams included in devices. However, webcams have limited resolution and accuracy compared to research-grade eye-trackers. These limitations govern what can be derived from this data and the subsequent strength of any conclusions we can draw relative to the broader eye-tracking literature. Research-grade eye tracking will remain the gold standard for gaze-based research, even though webcam-based approaches have great promise in the real world. We have shown that despite low-quality tracking, we are able to model complex constructs in a manner that is scalable and easy to implement.

Second, though we have taken steps to improve ecological validity through webcam use, future work should focus on using other tasks. Given that our results likely depend strongly on exactly how we presented our stimuli, future work should consider alternate presentations of text, or alternate tasks. For example, the same approach could be implemented but using longer passages and with page-by-page reading. Although many psychological experiments routinely use word-by-word or sentence-by-sentence paradigms when studying reading comprehension, it is important to test the boundary conditions of the webcam-based eye-tracker, particularly as areas of interest become more difficult to outline (e.g., small, single-spaced typeface). This limitation also extends to our high baseline comprehension accuracy rate (average accuracy was 88% correct), resulting in unbalanced class labels. Given this imbalance, it is not surprising that the chance classifier performs much better for correctness than for incorrectness. Future work may also consider alternative forms of assessing comprehension.

We also note the better performance of Local versus Global features. This finding implies that for correctness, where a participant is looking is more important than their more general gaze patterns. This result makes sense given the layout of the stimuli, with the three answers located at different locations on the screen. This result also testifies to the general accuracy of the eye tracking we used; were it highly inaccurate, it is unlikely the local features would have been as effective for prediction. However, other formats of stimuli/material presentation would be helpful in future work. Similarly, we should consider more complex feature sets and deep learning approaches as additional data is collected. As is often the case with human participant data, the current dataset is not large enough for effective deep learning, however the scalability of this approach presents the opportunity to collect vast datasets. Now that the initial feasibility has been shown, future work should consider collecting a larger dataset that enable more complex data mining and machine learning.

Future work should also consider a more in-depth feature-engineering process, with additional global and local features. The features included in this work, though theoretically relevant, provide a baseline for future gaze-feature development. Other features may include more detailed logging of reading regressions, for example, or word-level features considering how long a participant spends on each word of a sentence. Some features may also require that tracking accuracy and/or resolution is first improved before they can be calculated.

Though local features provided valuable predictive information for comprehension and TUT, we have not explored the mechanistic relationship between eye gaze and these constructs in this work. We argue that webcam-based gaze tracking is perhaps not suited to this kind of fine-grained analysis of reading behaviors, but we are encouraged that the data-driven models presented here are able to identify the cognitive events considered. Future work could consider how gaze mechanisms identified with research-grade tracking systems translate to this more accessible tracking option, which could help determine if and how such mechanisms can still be detected using webcam-based systems.

This work is further limited by our use of thought probes. Thought probes require users to be mindful of their unrelated thoughts and respond honestly. Although this methodology has been previously validated (Franklin et al., 2013; Randall et al., 2014), it is still limited due to the reliance on self-reports. Unfortunately, there is no clear alternative to track a highly internal state like TUT outside of measuring brain activity directly, which is also limited in many respects. Indeed, this too is part of the motivation for automated detectors. Future work should focus on validating our detectors so that thought probes are no longer necessary for measuring TUT, though fully automated TUT detections may well be a long way in the future.

Concluding Remarks

In sum, we provide evidence for a scalable solution for detecting attention and comprehension using only stock web cameras. Although there is still much room for improvement, the possibility to reach more individuals in more real-world settings—

particularly those who are historically-underrepresented—creates important opportunities for improving learning supports, and we hope to continue development along these lines.

Declarations

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

All research studies and procedures listed here were reviewed and approved by the appropriate University institutional review boards.

As noted above, supporting data cannot be made openly available due to ethical concerns and participant data agreements. Details of the data and how to request access are available from the Corresponding Author.

Open Practices Statement

The study reported in this article was not formally preregistered. The data have not been made available on a permanent third-party archive under the guidance of our Institutional Review Board; requests for the data can be sent via email to the corresponding author.

References

Ahn, S., Kelton, C., Balasubramanian, A., & Zelinsky, G. (2020). Towards predicting reading comprehension from gaze behavior. *ACM Symposium on Eye Tracking Research and Applications*, 1–5. https://doi.org/10.1145/3379156.3391335

Bixler, R., Blanchard, N., Garrison, L., & D'Mello, S. K. (2015). Automatic Detection of Mind Wandering During Reading Using Gaze and Physiology. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 299–306. https://doi.org/10.1145/2818346.2820742

Bixler, R., & D'Mello, S. K. (2014). Toward Fully Automated Person-Independent Detection of Mind Wandering. In V. Dimitrova, T. Kuflik, D. Chin, F. Ricci, P. Dolog, & G.-J. Houben (Eds.), *User Modeling Adaptation and Personalization* (pp. 37–48). Springer. https://doi.org/10.1007/978-3-319-08786-3_4

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

Blanchard, N., Bixler, R., Joyce, T., & D'Mello, S. K. (2014). Automated Physiological-Based Detection of Mind Wandering during Learning. In S. Trausan-Matu, K. Boyer, M. Crosby, & K. Panourgia (Eds.), *Intelligent Tutoring Systems* (pp. 55–60). Springer International Publishing. https://doi.org/10.1007/978-3-319-07221-0_7

Campbell, F. W., & Wurtz, R. H. (1978). Saccadic omission: Why we do not see a grey-out during a saccadic eye movement. *Vision Research*, *18*(10), 1297–1303. https://doi.org/10.1016/0042-6989(78)90219-5

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*(1), 321–357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Cranford, E. A., & Moss, J. (2018). Mouse-tracking evidence for parallel anticipatory option evaluation. *Cognitive Processing*, *19*(3), 327–350. https://doi.org/10.1007/s10339-017-0851-4

Degen, J., Kursat, L., & Leigh, D. D. (2021). Seeing is believing: Testing an explicit linking assumption for visual world eye-tracking in psycholinguistics. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43).

Dirix, N., Vander Beken, H., De Bruyne, E., Brysbaert, M., & Duyck, W. (2020). Reading Text When Studying in a Second Language: An Eye-Tracking Study. *Reading Research Quarterly*, *55*(3), 371–397. https://doi.org/10.1002/rrq.277

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

D’Mello, S. K., Olney, A., Williams, C., & Hays, P. (2012). Gaze Tutor: A Gaze-reactive Intelligent Tutoring System. *International Journal of Human-Computer Studies*, *70*(5), 377–398. https://doi.org/10.1016/j.ijhcs.2012.01.004

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*. https://doi.org/10.1080/0163853X.2020.1739600

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

Eckstein, M. K., Guerra-Carrillo, B., Singley, A. T. M., & Bunge, S. A. (2017). Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? *Developmental Cognitive Neuroscience*, *25*, 69–91. https://doi.org/10.1016/j.dcn.2016.11.001

Faber, M., Bixler, R., & D’Mello, S. K. (2017). An automated behavioral measure of mind wandering during computerized reading. *Behavior Research Methods*, 1–17. https://doi.org/10.3758/s13428-017-0857-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), Art. 1. psyh. https://doi.org/10.1037/a0030217

Franklin, M. S., Broadway, J. M., Mrazek, M. D., Smallwood, J., & Schooler, J. W. (2013). Window to the wandering mind: Pupillometry of spontaneous thought while reading. *Quarterly Journal of Experimental Psychology*, *66*(12), 2289–2294. https://doi.org/10.1080/17470218.2013.858170

Gardner, J., Brooks, C., & Baker, R. S. (2019). Evaluating the Fairness of Predictive Student Models Through Slicing Analysis. *Proceedings of the 10th Conference on Learning Analytics and Knowledge*, 10. https://doi.org/10.1145/3303772.3303791

Hand, D., & Christen, P. (2018). A note on using the F-measure for evaluating record linkage algorithms. *Statistics and Computing*, *28*(3), 539–547. https://doi.org/10.1007/s11222-017-9746-6

Huey, E. B. (1908). *The Psychology and Pedagogy of Reading: With a Review of the History of Reading and Writing and of Methods, Texts, and Hygiene in Reading*. The Macmillan company.

Hutt, S., Hardey, J., Bixler, R., Stewart, A., Risko, E., & D'Mello, S. K. (2017). Gaze-based Detection of Mind Wandering during Lecture Viewing. *10th International Conference on Educational Data Mining*, 226–231.

Hutt, S., Krasich, K., Brockmole, J. R., & D'Mello, S. K. (2021). Breaking Out of the Lab: Mitigating Ming Wandering with Gaze-Based Attention-Aware Technology in Classrooms. *ACM SIGCHI: Computer-Human Interaction*, 1–14. https://doi.org/10.1145/3411764.3445269

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

Hutt, S., Mills, C., Bosch, N., Krasich, K., Brockmole, J. R., & D'Mello, S. K. (2017). 'Out of the Fr-Eye-ing Pan': Towards Gaze-Based Models of Attention During Learning with Technology in the Classroom. *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, 94–103. https://doi.org/10.1145/3079628.3079669

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.), *The 9th International Conference on Educational Data Mining* (pp. 86–93).

Irwin, D. E., & Carlson-Radvansky, L. A. (1996). Cognitive suppression during saccadic eye movements. *Psychological Science*, *7*(2), 83–88. https://doi.org/10.1111/j.1467-9280.1996.tb00334.x

Jacob, R. J. (1995). Eye tracking in advanced interface design. *Virtual Environments and Advanced Interface Design*, *258*, 288. https://doi.org/10.1093/oso/9780195075557.003.0015

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge university press.

Kukkonen, S. (2005). Exploring eye tracking in design evaluation. *Joining Forces*, 119–126.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*. https://doi.org/10.2307/2529310

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777.

Matin, E. (1974). Saccadic suppression: A review and an analysis. *Psychological Bulletin*, *81*(12), 899–917. https://doi.org/10.1037/h0037368

McNamara, D. S., & Magliano, J. (2009). Toward a Comprehensive Model of Comprehension. In *Psychology of Learning and Motivation* (Vol. 51, pp. 297–384). Elsevier. http://linkinghub.elsevier.com/retrieve/pii/S0079742109510092

McVay, J. C., & Kane, M. J. (2012). Drifting from slow to ‘D’oh!’: Working memory capacity and mind wandering predict extreme reaction times and executive control errors. *Journal of Experimental Psychology: Learning Memory and Cognition*, *38*(3), 525–549. https://doi.org/10.1037/a0025896

Mills, C., Bixler, R., Wang, X., & D’Mello, S. K. (2016). Automatic gaze-based detection of mind wandering during film viewing. In T. Barnes, M. Chi, & M. Feng (Eds.), *The 9th International Conference on Educational Data Mining.* (pp. 30–37).

Mills, C., Gregg, J., Bixler, R., D'Mello, S. K., & 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*, *00*(00), 1–27. https://doi.org/10.1080/07370024.2020.1716762

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). WebGazer: Scalable Webcam Eye Tracking Using User Interactions. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 3839–3845.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://doi.org/10.1007/s13398-014-0173-7.2

Phillips, N. E., Mills, C., D'Mello, S., & Risko, E. F. (2016). On the influence of re-reading on mind wandering. *Quarterly Journal of Experimental Psychology*, *69*(12), Art. 12.

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

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading*, *10*(3), Art. 3.

Rayner, K., Reichle, E. D., Stroud, M. J., Williams, C. C., & Pollatsek, A. (2006). The effect of word frequency, word predictability, and font difficulty on the eye movements of young and older readers. *Psychology and Aging*, *21*(3), Art. 3.

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

Semmelmann, K., & Weigelt, S. (2018). Online webcam-based eye tracking in cognitive science: A first look. *Behavior Research Methods*, *50*(2), 451–465. https://doi.org/10.3758/s13428-017-0913-7

Smallwood, J. (2011). Mind-wandering While Reading: Attentional Decoupling, Mindless Reading and the Cascade Model of Inattention. *Language and Linguistics Compass*, *5*(2), 63–77. https://doi.org/10.1111/j.1749-818X.2010.00263.x

Smallwood, J., & Schooler, J. W. (2015). The science of mind wandering: Empirically navigating the stream of consciousness. *Annual Review of Psychology*, *66*, 487–518. https://doi.org/10.1146/annurev-psych-010814-015331

Toreini, P., Langner, M., & Maedche, A. (2020). Using eye-tracking for visual attention feedback. In *Information Systems and Neuroscience* (pp. 261–270). Springer.

Tran, M., Sen, T., Haut, K., Ali, M. R., & Hoque, M. E. (2019). Are you really looking at me? A Feature-Extraction Framework for Estimating Interpersonal Eye Gaze from Conventional Video. *ArXiv Preprint ArXiv:1906.12175.*

Valliappan, N., Dai, N., Steinberg, E., He, J., Rogers, K., Ramachandran, V., Xu, P., Shojaeizadeh, M., Guo, L., Kohlhoff, K., & Navalpakkam, V. (2020). Accelerating eye movement research via accurate and affordable smartphone eye tracking. *Nature Communications*, *11*(1), 4553. https://doi.org/10.1038/s41467-020-18360-5

Varao-Sousa, T. L., & Kingstone, A. (2019). Are mind wandering rates an artifact of the probe-caught method? Using self-caught mind wandering in the classroom to test, and reject, this possibility. *Behavior Research Methods*, *51*(1), Art. 1.

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

Weinstein, Y. (2018). Mind-wandering, how do I measure thee with probes? Let me count the ways. *Behavior Research Methods*, 1–20.

Yang, X., & Krajbich, I. (2020). *Webcam-based online eye-tracking for behavioral research*.

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

Zhang, X., Sugano, Y., Fritz, M., & Bulling, A. (2019). MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(1), 162–175. https://doi.org/10.1109/TPAMI.2017.2778103

Zuber, B. L., & Stark, L. (1966). Saccadic suppression: Elevation of visual threshold associated with saccadic eye movements. *Experimental Neurology*, *16*(1), 65–79. https://doi.org/10.1016/0014-4886(66)90087-2