Gaze dynamics are sensitive to target orienting for working memory encoding in virtual reality

Candace E. Peacock Reality Labs Research; Redmond, WA, USA

Ting Zhang Reality Labs Research; Redmond, WA, USA

Brendan David-John Reality Labs Research; Redmond, WA, USA

T. Scott Murdison Reality Labs Research; Redmond, WA, USA

Matthew J. Boring Reality Labs Research; Redmond, WA, USA

Hrvoje Benko Reality Labs Research; Redmond, WA, USA

Tanya R. Jonker Reality Labs Research; Redmond, WA, USA

Numerous studies have demonstrated that visuospatial attention is a requirement for successful working memory encoding. It is unknown, however, whether this established relationship manifests in consistent gaze dynamics as people orient their visuospatial attention toward an encoding target when searching for information in naturalistic environments. To test this hypothesis, participants' eye movements were recorded while they searched for and encoded objects in a virtual apartment (Experiment 1). We decomposed gaze into 61 features that capture gaze dynamics and a trained sliding window logistic regression model that has potential for use in real-time systems to predict when participants found target objects for working memory encoding. A model trained on group data successfully predicted when people oriented to a target for encoding for the trained task (Experiment 1) and for a novel task (Experiment 2), where a new set of participants found objects and encoded an associated nonword in a cluttered virtual kitchen. Six of these features were predictive of target orienting for encoding, even during the novel task, including decreased distances between subsequent fixation/saccade events, increased fixation probabilities, and slower saccade decelerations before encoding. This suggests that as people orient toward a target to encode new information at the end of search, they decrease task-irrelevant, exploratory sampling behaviors. This behavior was common across the two studies. Together, this research demonstrates how gaze dynamics can be used to capture target orienting for working memory encoding and has implications for real-world use in technology and special populations.

Introduction

We use our eyes for many purposes, including finding things we want to use, inspecting objects, reading information, and understanding people and their intentions. For instance, we might fixate surfaces that are likely to contain the objects that we search for (Pereira & Castelhano, 2019), we might use regressive saccades to reread complex text (Booth & Weger, 2013), and we might fixate specific parts of people's faces to understand their emotions (Schurgin, Nelson, Iida, Ohira, Chiao, & Franconeri, 2014). In fact, given the tight coupling between gaze behaviors and our goals, Yarbus (1967) and others have used gaze behavior to understand an individual's task, even while the visual stimulus is held constant (Borji & Itti, 2014; Ellis & Stark, 1981; Henderson, Shinkareva, Wang, Luke, & Olejarczyk, 2013; Tatler, Wade, Kwan, Findlay, & Velichkovsky, 2010; Yarbus, 1967). Furthermore, gaze has also been used to detect future tasks and behaviors, including inferring the possible future targets of visual attention (Lengyel, Carlberg, Samad, & Jonker, 2021), anticipating upcoming lane changes (Doshi & Trivedi, 2009; Liu, 2001), and predicting ingredient selection

Citation: Peacock, C. E., Zhang, T., David-John, B., Murdison, T. S., Boring, M. J., Benko, H., & Jonker, T. R. (2022). Gaze dynamics are sensitive to target orienting for working memory encoding in virtual reality. *Journal of Vision*, *22*(1):2, 1–28, https://doi.org/10.1167/jov.22.1.2.

https://doi.org/10.1167/jov.22.1.2 Received June 4, 2021; published January 4, 2022 ISSN 1534-7362

(David-John, Peacock, Zhang, Murdison, Benko, & Jonker, 2021; Huang, Andrist, Sauppé, & Mutlu, 2015).

Although computational models of gaze dynamics have been used to infer an individual's goals in these domains (David-John et al., 2021; Doshi & Trivedi, 2009; Huang et al., 2015; Lengyel et al., 2021; Liu, 2001), they have not been used to detect when people find an object they want to encode into working memory (WM). WM is an important cognitive state as it critical for learning (Alloway, 2006; Titz & Karbach, 2014) and executive control (Gruber & Goschke, 2004; Poole & Kane, 2009; Titz & Karbach, 2014). Developing a computational model of target orienting for WM encoding would, therefore, be important, because people often execute search processes to find relevant information for WM encoding (e.g., searching for a written phone number to encode into WM). Furthermore, with such a model, a system could provide adaptive assistance (e.g., automated photo capture of an encoding target) to help people free up their cognitive resources.

There is some precedent in the literature that computational models composed of gaze-based features could be used to successfully predict when people transition from searching rich, complex information to targeting their visuospatial attention toward a target for WM encoding. For example, Malcolm and Henderson (2009) introduced a three-stage model to visual search, which described the onset of search (search initiation stage), the search process until a target is located (scanning stage), and the end of search after a target is located and compared with the target template (target verification stage) (Malcolm & Henderson, 2009). Here, empirical studies demonstrate that saccade amplitudes are longer and fixation durations are shorter during the scanning stage relative to the target verification stage (David, Beitner, & Võ, 2020), suggesting that there are gaze dynamics that reflect the stages of search. Other studies have found similar patterns of coarse-to-focal gaze behaviors from the onset of search to the offset of search irrespective of the three stages (Godwin, Reichle, & Menneer, 2014; Over, Hooge, Vlaskamp, & Erkelens, 2007) and during other scene viewing tasks (as described by the ambient–focal phenomenon; Unema, Pannasch, Joos, & Velichkovsky, 2005).

Although the differences in fixation durations and saccade amplitudes during the scanning and verification stages of visual search provide a compelling demonstration that a computational model of target orienting for WM encoding is feasible, it is unknown whether gaze dynamics reflect the anticipation of finding a search target for encoding. Specifically, gaze dynamics might change from the start to the end of scanning as people anticipate orienting to a target for WM encoding. Furthermore, there may be a broader set of gaze features beyond fixation durations and saccade amplitudes that capture the transition from scanning to target orienting for encoding. These features might also have unique time courses during which they become sensitive to the anticipation of a search target for encoding. The present study aims to uncover a novel, broader set of gaze features and the windows of time in which these features are most sensitive to target orienting for WM encoding.

The existing literature gives plausibility to this idea that gaze dynamics are indeed sensitive to the anticipation of task-relevant stimuli (McPeek, Han, & Keller, 2006; McPeek et al., 2003; Milstein & Dorris, 2007; Wilimzig, Schneider, & Schöner, 2006; Wu & Kowler, 2013). For example, people are faster to execute saccades that precede fixations on targets versus nontargets (Wu & Kowler, 2013), and they are faster to execute saccades when the anticipated saccade landing position might result in monetary reward versus not (Milstein & Dorris, 2007). Furthermore, stimulation studies provide causal evidence that, when the frontal eye fields and the superior colliculus can anticipate distractor information, saccades will curve away from distractors and toward targets (McPeek, 2006; McPeek et al., 2003). Together, these studies demonstrate that gaze dynamics reflect the anticipation of fixation on an important stimulus. Therefore, a reasonable extension from these studies is that gaze dynamics might also be sensitive to target orienting for WM encoding.

Although gaze dynamics reflect the anticipation of task-relevant stimuli, it is unknown whether gaze dynamics are related to successful WM encoding. However, several experimental studies suggest that, when visuospatial attention can anticipate (i.e., aligns with) the location of encoding, WM encoding (and subsequent memory retrieval) will be successful. Specifically, WM encoding benefits when a to-be-encoded object appears at cued locations associated with encoding (Schmidt, Vogel, Woodman, & Luck, 2002; Woodman, Vecera, & Luck, 2003). Furthermore, when an object is marked as a saccade target, its features are more likely to be encoded into WM than when an object is not a saccade target (Hanning, Jonikaitis, Deubel, & Szinte, 2015). This body of laboratory work suggests that, when visuospatial attention is aligned to (or can anticipate) the location of to-be-remembered information, encoding will likely be successful; in contrast, when visuospatial attention and gaze disengage from the location of encoding, WM will likely fail. Given the established relationship between visuospatial attention and WM encoding in the laboratory, it is possible that, as people search naturalistic environments, the spatial and temporal expectation of a target of WM encoding mediates how they move their eyes and that these gaze patterns are indicative of intentional visuospatial attention orienting for WM encoding.

Understanding whether gaze dynamics capture target orienting for WM encoding would have both scientific

and practical benefits. From a scientific perspective, by exploring gaze before the onset of encoding, we can gain insights into how gaze behaviors change from the beginning to the end of search as people orient toward a target they wish to encode in WM. From a practical perspective, if a system could anticipate an encoding target, then it could suggest applications that would allow people to externalize their memory load, such as a notepad app, a to-do list, an audio recorder, or even a camera for photo capture. These types of interventions would have usefulness for people in their everyday lives, and even greater benefits for those who suffer from WM deficits or other cognitive impairments.

For practical applications of the link between gaze, search processes, and WM encoding, it is critical to understand gaze behaviors in visually rich, naturalistic contexts that more closely represent the real world, because this practice has more ecological validity than laboratory studies that use simple stimulus arrays, such as colored squares on a blank background (Hanning et al., 2015; Schmidt et al., 2002; Woodman et al., 2003). Thus, in the present work, the goal was to explore whether gaze dynamics can capture target orienting for WM encoding in rich, complex, naturalistic environments in virtual reality (VR). The study was conducted in VR rather than using a traditional eye tracking setup with a monitor and fixed head position because VR allows for (1) the enhanced naturalism of scenes, (2) an expanded visual field with up to 360° for exploration, and (3) natural head motions and natural eye–head coordination. Although there are some limitations to using VR (e.g., the decreased sampling frequency of commercial trackers relative to research-grade trackers), the ability to deploy naturalistic environments and to capture natural head and eye movements was critical to understanding whether there is a set of gaze dynamics that are sensitive to target orienting for WM encoding in naturalistic environments.

In sum, the goal of the present work was to understand how temporal gaze dynamics unfold as people orient toward a target for encoding in complex, naturalistic settings that more closely emulate real-world encoding contexts than prior studies. To better understand the role of temporal gaze dynamics and how they unfold as people orient to a target of encoding, we used a computational model of gaze dynamics. By using a computational model, we were able to uncover novel gaze features and the unique time courses of when these gaze features were most sensitive to target orienting for encoding.

Present study

The primary goal of the present work was to explore whether there is a common set of gaze dynamics that are sensitive to target orienting for WM encoding in naturalistic environments. We chose to model both search target orienting and WM encoding anticipation together (rather than understand the unique gaze behaviors associated with each) because many real-world encoding scenarios will involve searching for information to encode (e.g., searching for a written phone number to encode into WM; searching for and encoding the size of a socket to see if it matches the size of a socket wrench). By using tasks that approximate real-world encoding scenarios (in which search precedes encoding), we were able to (1) understand from an ecologically valid perspective how gaze dynamics unfold in anticipation of a target of encoding and (2) produce a model that was more likely to generalize to new contexts which is important for practical applications where generalization is key. Furthermore, target finding and WM encoding go hand in hand, because focused visuospatial attention is required to orient to both visual search targets (Woodman & Luck, 2004) and visual WM encoding targets (Awh & Jonides, 2001; Gazzaley & Nobre, 2012; Schmidt et al., 2002). Given the overlap that exists between search target orienting and WM encoding both in terms of their natural co-occurrence in the world and the requirements of visuospatial attention to orient to both search and encoding targets, this study opted to analyze both processes in concert (termed "target orienting for encoding"), because there are likely gaze behaviors common to both.

Recently, we reported on an early exploration into the development of a model of target orienting for WM encoding using gaze data in naturalistic settings (i.e., VR; Peacock, David-John, Zhang, Murdison, Boring, Benko, & Jonker, 2021). We found that a model trained using sliding windows of gaze features detected when people oriented to a target for WM encoding using a commercial eye tracker in an immersive virtual apartment. This research provided a compelling proof-of-life example of the use of gaze data to detect target orienting for WM encoding in consumer settings. However, these models were trained individually for each participant, and they were only explored within a single task. As such, it is unclear whether these models captured stable and generalizable patterns of eye gaze as people oriented to a target for encoding. Although it was a compelling demonstration, it is possible that such within-subject models captured task-specific variance rather than a signal that was common to target orienting for WM encoding. As such, the primary goal of the present work was to uncover consistent gaze dynamics that occur before WM encoding and are robust across a range of people, tasks, and environments.

Therefore, this research addresses two hypotheses:

- H1: Gaze dynamics can be used to detect target orienting for WM encoding across people.

- H2: Gaze dynamics capture task- and environment-general variance in target orienting for WM encoding.

To address these hypotheses, eye tracking data were collected during two tasks in VR. In the first task, participants navigated and searched through a virtual apartment for objects they needed to encode for later recall (Experiment 1). In the second task, a new set of participants encoded nonwords corresponding to visual search target objects in a cluttered kitchen (Experiment 2). Sliding window logistic regression models were then trained to test whether gaze dynamics reflect target orienting for encoding.

To address H1, we trained a group model to predict target orienting for encoding and tested whether this model generalized to individual participants within the study, which allowed us to determine whether a consistent set of gaze features occur in the moments leading up to WM encoding.

From there, to address H2, we sought to explore whether the detected gaze dynamics captured a general set of orienting behaviors related to target orienting for encoding across tasks, environments, encoded stimuli, and people. To explore whether our gaze model could generalize to a new task, environment, stimulus set, and people, we applied the trained model from Experiment 1 to unseen data from Experiment 2. Finally, to deepen our understanding of how gaze behaviors relate to target orienting for encoding, we explored which specific gaze features generalized across tasks. To address this question, we trained a group model on each individual feature from Experiment 1 and tested it on data from Experiment 2, which would allow us to develop novel scientific insights into the specific features of gaze behavior that uniquely identify as people orient to encoding targets across different tasks.

Experiment 1

The goal of Experiment 1 was to determine whether a group model of gaze features could detect target orienting for encoding across individuals (H1). If, indeed, people tend to orient in consistent ways to search targets that they wish to encode, then natural gaze behaviors might reflect this behavior and could be used to predict and anticipate WM encoding.

Methods

Participants

Thirty-eight participants completed the study and were compensated for their participation. Informed consent was obtained and protocols were approved by the Western Institutional Review Board. Six participants were excluded from the dataset as they failed to complete the study due to discomfort or noise disruptions, resulting in a sample of 32 (mean age, 27.7 years; 16 females).

Apparatus and data collection

An HTC Vive headset with Tobii Pro binocular eye tracking (120 Hz) was used to render the VR tasks and collect eye tracking data. The HTC Vive controller was used for navigation around the apartment. At the beginning of the study, the built-in Tobii 5-point calibration protocol was used. Accuracy was verified before the start of the task and there was no recalibration throughout the task.

The experiment was implemented in Unity 2018.4.2f1, with the Tobii Pro Unity SDK (v1.4) handling the logging and synchronization of the eye tracking data. To synchronize the HTC Vive (90 Hz) to the eye tracker (120 Hz), the eye tracker queued timestamped data and the SDK retrieved the eye tracking data from the queue on each software frame.

In complex virtual environments, frames might be dropped; however, dropped frames did not affect the gaze data because they were logged on the tracker and the Unity refresh rate was dynamic to its load (i.e., it was designed to minimize dropped frames). Because dropped frames could still impact a user's view of the virtual environment, there could be an approximate 10-ms decrease in the reaction times recorded, particularly when participants move their head to look at a different target or area (i.e., the frame rate temporarily decreases to 45 Hz whenever 90 Hz cannot be maintained).

Stimuli

A high-resolution rendering of two virtual apartments were used as the experimental environments. Each apartment contained a bedroom, patio, kitchen, bathroom, and living room. The environments contained low clutter. The encoded stimulus was the semantic label corresponding to an object that was in each room.

Procedures

Participants were first outfitted with a VR headset. Participants completed a tutorial and practice trial in VR before the main trial sequence. Here, participants were seated and had free range of motion of their arms and head and used teleportation to navigate. During each trial, participants were spawned in a room of one of two virtual apartments. They received a text prompt to navigate to a specific room using point-and-teleport

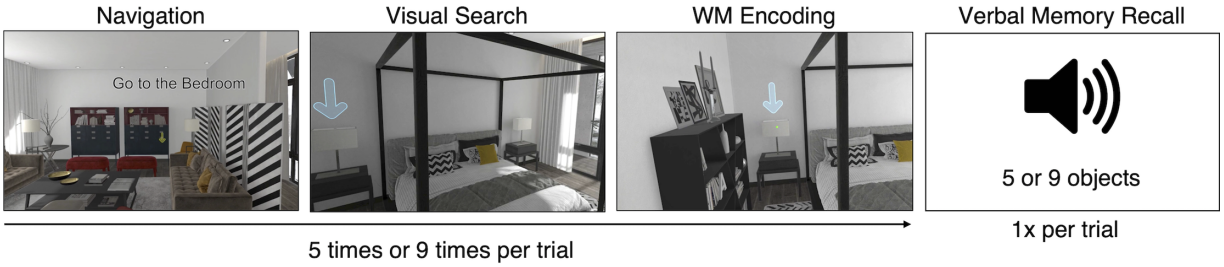

Figure 1. **Example of an experiment 1 trial.** During each trial, a text prompt was accompanied by a yellow arrow to guide navigation; a blue arrow was used to indicate the object to be recalled. The blue arrow later disappeared once participants' gaze intersected the target object; after encoding five or nine objects, participants were asked to verbally recall all the objects.

navigation (Figure 1a); they also received an arrow to indicate the room's location, which was visible through the walls to guide navigation. Upon arrival at the specified room, the navigation arrow and text prompt disappeared and participants were either prompted to navigate to another room or to find an object. In most cases, one of the objects in the room became marked with a blue arrow (Figure 1, second pane), and participants were to find the target object and remember its identity. The blue arrow existed in the room from the moment the participant arrived and, therefore, entered the periphery naturally as with any other object the participant viewed. After the initial gaze intersection on the object, the blue arrow would disappear and either (1) another object became marked in the same room or (2) participants were prompted to navigate to a new room. At the end of each trial, participants verbally recalled the objects and were then given an optional break. Participants first completed a tutorial and practice trial. Then, they completed 30 trials. Fifteen trials asked participants to remember 5 objects and the other 15 trials asked participants to remember 9 objects placed in various rooms in the apartment for a total of 210 encoded objects. At the end of each trial, they were then given the option of taking a break, if needed. One trial sequence was used for all participants.

Gaze feature preparation

Data segmentation: Encoding onsets were defined as the fixation after the first gaze ray intersection with either the arrow or object (whichever occurred first). To compute whether target objects were fixated or not, solid colliders were added to each target object. Then, a 30-m gaze ray was cast and any object currently intersected by the ray was reported on a sample-by-sample basis. Velocity-based event detection was also used on a sample-by-sample basis to detect when fixations occurred. Whenever the gaze ray intersected an object of interest while a fixation occurred, this was a fixation on the object.

For each trial, the time series eye tracking data were then segmented into clips, one for each encoded object. Each clip began when the blue arrow appeared and ended with the onset of encoding (i.e., fixation on the object). Clips did not include navigation between rooms.

Clips corresponding to forgotten items were discarded (14.43% data loss) to remove any instances of unsuccessful encoding (e.g., due to inattentiveness, fatigue, or distraction). To account for eye tracker error, which could result in an unparsed fixation on the to-be-encoded object, we used a strict criterion: only clips in which fixation occurred within 500 ms of the first gaze intersection on the object were used (M = 17.40% data loss). On average, clips had durations of 10,347.42ms (median = 7,221.11 ms).

Gaze feature computation

To compute the gaze features to be used in our modelling, the head orientation and normalized gaze unit vectors that were recorded were first temporally aligned using dplyr (version 1.0.5) in R (version 4.0.4). The normalized gaze unit vectors (i.e., eye-in-head coordinates) were then transformed by a rotation matrix in Python (version 3.6.10) to correct for head orientation, which transformed them into gaze-in-world coordinates (Diaz, Cooper, Kit, & Hayhoe, 2013). Gaze velocity was then computed as the angular distance between the gaze-in-world coordinates divided by the change in time. Numpy (version 1.18.2) and Pandas (0.25.2) were used to manipulate and store the data.

Filtering

The gaze data were filtered to remove noise and unwanted segments before event detection and feature extraction. Data from the practice trials and breaks were discarded before analysis, and we removed all gaze samples where the gaze velocity exceeded 800°/s, indicating unfeasibly fast eye movements (Dowiasch, Marx, Einhäuser, & Bremmer, 2015). The removed

values were then replaced through linear interpolation. When the eye tracker lost signal (e.g., when blinks occurred), the data output NaNs for the gaze-in-head samples corresponding to those time stamps. We linearly interpolated over these values to retain temporal continuity in the eye tracking data. Finally, a median filter with a width of seven samples was applied to the gaze velocity signal to smooth the signal and account for noise before event detection (Pekkanen & Lappi, 2017).

Event detection

The identification by velocity threshold was used for event detection (Salvucci & Goldberg, 2000). The identification by velocity threshold uses gaze velocity thresholds to segment fixations and saccades. A saccade was detected if the gaze velocity was greater than 70°/s for 12 to 300 ms (Diaz et al., 2013; George & Routray, 2016). A fixation was detected if the gaze velocity was less than 20°/s with a minimum fixation duration cutoff of 50 ms and a maximum fixation duration cutoff of 1500 ms (Diaz et al., 2013; Peacock et al., 2019).

Because models that incorporate scene semantics or knowledge of the environment might be impractical for real-world use with current technology, consumer-grade eye tracking does not track gaze locations with high precision and accuracy for everyone, and because systems identifying target objects at gaze coordinates require high-power cameras and computationally expensive computer vision models, the current model was developed using features that did not depend on gaze–environment interactions. Furthermore, to identify a broader set of gaze features that were sensitive to the offset of search (beyond fixation durations and saccade amplitudes) and the anticipation of WM encoding than what has been previously shown, sixty-one gaze features were computed (Supplementary Table S1 provides a full description of all 61 features and Supplementary Table S3 provides basic eye movement statistics related to these features). These features included gaze velocity and dispersion, which provided a continuous index of visual exploration, the k-coefficient which described the transition from ambient (i.e., short fixations, long saccades) to focal viewing (i.e., long fixations, short saccades) (Krejtz, Duchowski, Krejtz, Szarkowska, & Kopacz, 2016), and 58 event-based features that represented various statistics of fixation/saccade events (George & Routray, 2016).

Because features derived from fixations/saccades have missing values at time points when no fixation or saccade is occurring, linear interpolation was applied to produce a complete sequence of data. Each gaze feature except for the categorical features (i.e., saccade detection, fixation detection) was then z-scored within-participant to account for individual differences in baseline gaze behavior. Because Sklearn's logistic regression applies L2 regularization by default, gaze features with smaller magnitudes, which require larger regression coefficients, would have been unnecessarily penalized if z-scoring was not performed.

Data augmentation

Because the percentage of true classes (i.e., target orienting for WM encoding at the end of search) relative to null classes (i.e., no target orienting for encoding during early search) was very small (0.3% of the total data), we augmented the number of true data points to enhance true class signal by marking data occurring 20 ms before the encoding onset as a true class (i.e., blue band in Figure 2a). This increased the proportion of true class instances three-fold.

Sliding window framework

To create input samples for model training, a sliding window of N ms, which was determined through a hyperparameter search described in the Modeling section, was used for each feature (Figure 2a). As a sliding window for each feature slid along the clip, its class was determined by the class of the last sample in the window (green band in Figure 2c). Thus, there were both null classes (no encoding anticipation) and true classes (encoding anticipation) in each clip. Null classes averaged to the mean across the dataset because there was no consistency in the end sample. Because the data were z-scored, the variance of the null classes was approximately 1 and the mean was approximately 0. Conversely, because true classes consistently ended just before the encoding fixation, true classes differed from the null classes whenever a feature was sensitive to encoding.

To increase the computational efficiency and reduce the collinearity of the model, we down sampled by averaging every five samples (approximately 42 ms; Figure 2b). Downsampling did not change the interpretation of the window size, because it only changed how many beta values were added to the model. For example, downsampling an 83-ms window (approximately 10 samples) would average across time so that only two samples would be input to the model rather than all 10 (Figure 2b). The interpretation of the data would, however, remain the same. If the true classes were greater in value than the null classes when there were 10 samples in an 83-ms window, then the interpretation would not change when the data were downsampled.

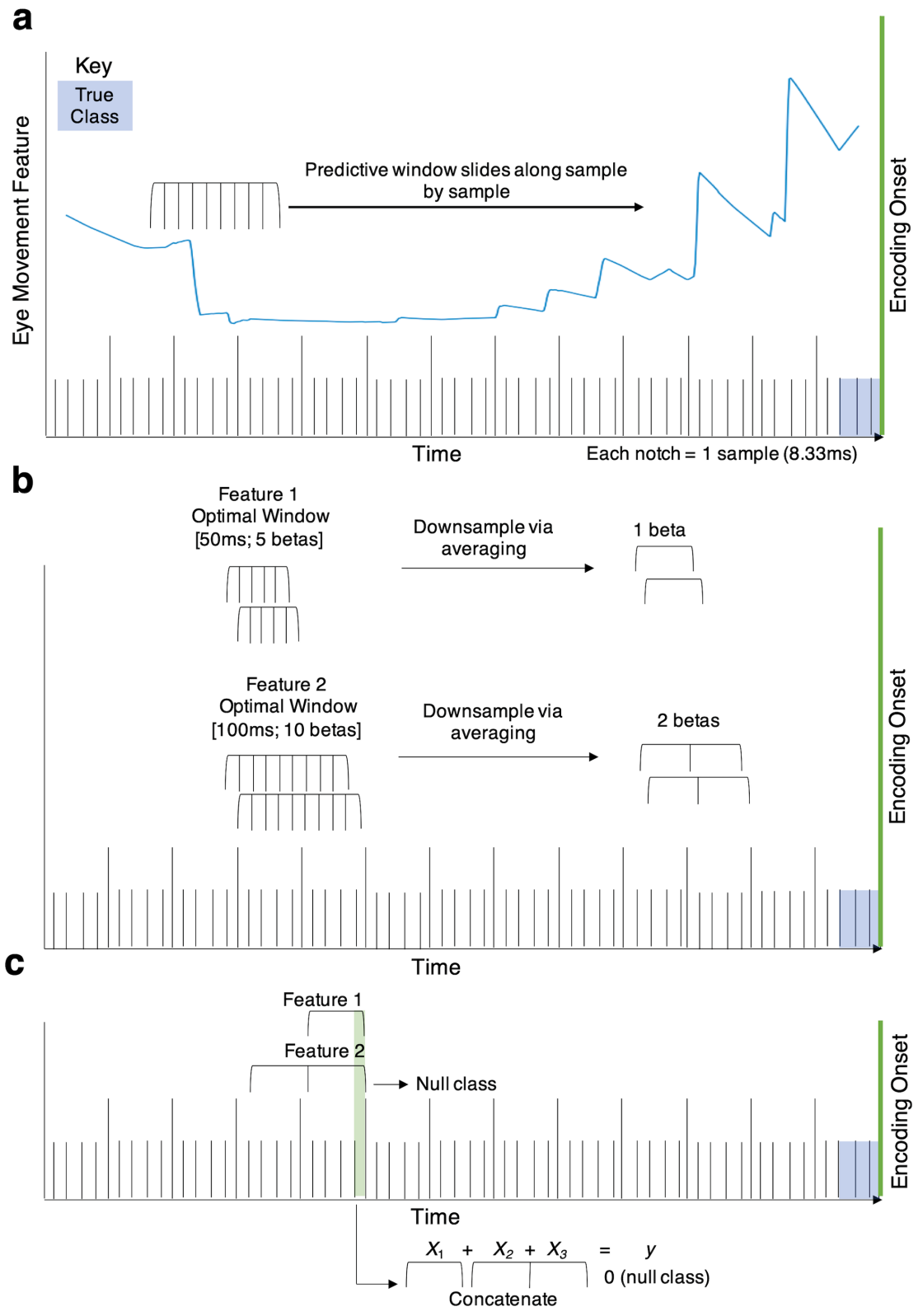

Figure 2. **Visualization of the sliding window framework.** (a) A hypothetical eye movement feature (blue line) that increases just before the onset of WM encoding. The predictive window for this feature slides along sample by sample to produce multiple windows of data. (b) An example of how features were temporally downsampled. If feature 1 (or 2) contains 5 (or 10) samples in the optimal predictive window, then these 5 (or 10) samples would be downsampled via averaging to generate 1 (or 2) beta parameter without sacrificing accuracy. (c) An example of how the feature concatenation was performed. The class would be determined by the *y* value of the last time stamp that was concatenated.

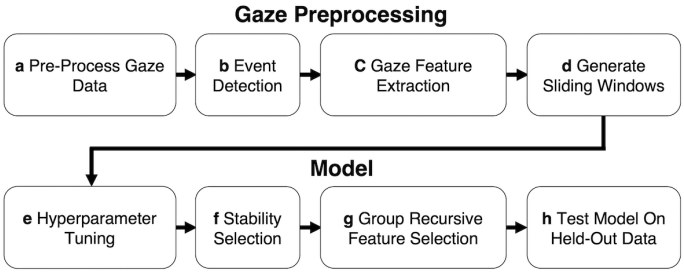

Figure 3. **The modeling framework.** The gaze data was first preprocessed (a–d) before modeling. Following the pre-processing, we then modeled the data (e–g) and tested it on a held-out test set (h).

Model

The modeling framework used sliding windows of size N as input and used hyperparameter tuning to find the optimal value for N. Feature selection was then performed to find the most predictive gaze features among the aforementioned 61 features. With these gaze features and their optimal window sizes, a group model was then trained and model performance was tested using two different metrics (Figure 3).

Model description

Logistic regression models were trained to learn the onset of WM encoding using Sklearn (version 0.24.2). The logistic regression models were selected because they were interpretable, lightweight, and they predict binary data. Due to an imbalanced dataset, which had 99.1% null samples and 0.9% true class samples, the class weights of the models were balanced by setting the weights to be inversely proportional to the number of samples for each class.

Evaluation metric

The area under the precision–recall curve (AUC-PR) was used for model tuning and evaluation, which is a better performance metric when data are heavily imbalanced (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015; Tatbul, Lee, Zdonik, Alam, & Gottschlich, 2018). The baseline value of the AUC-PR is derived from the chance rate of true examples, which can vary based on each individual and the size of the data window, making it difficult to compare model performance. To create a standardized chance rate for each individual, the null class was resampled for each of the training and test sets to ensure a fixed percentage of true classes (i.e., 0.9% was the average true class percentage across individuals). Although hyperparameter tuning and recursive feature selection were conducted using AUC-PR, the area under the receiver operator curve (AUC-ROC) is also reported because this metric is more commonly used and interpretable.

Generating stable hyperparameters and feature estimates

Hyperparameter tuning and stability selection was conducted using within-participant models to generate input for the group model and to obtain estimates for the stability selection procedure and to identify the most critical gaze features for classification. Here, within-participant models were used to ensure that the most stable estimates across individuals were selected, because prior research has found individual differences in these estimates (Peacock et al., 2021). To tune the model, each participant's data were split into 90% training and 10% test sets. Stratified 10-fold cross-validation with 3 repeats was used to compute a reliable estimate of the AUC-PR while preserving the percentage of samples for each class. The mean AUC-PR was computed across folds within participants to estimate the input to the group model. The performance of the within-participant models on the test set are not reported because these data were reported previously in Peacock et al. (2021).

Hyperparameter tuning

Hyperparameter tuning was first conducted using the training set to find the most stable and predictive window size for each feature. Hyperparameters are typically specified heuristically and then tuned for a given machine learning problem. Tuning allows one to build a model for each combination of hyperparameter values and select the best hyperparameter value based upon the one that provides the best results.

In this study, window sizes ranging from 42 to 1,000 ms were explored because it was hypothesized that 42 ms was the minimum amount of time people could anticipate encoding due to constraints with visual processing (Egeth & Yantis, 1997; Salthouse & Ellis, 1980), whereas 1,000 ms was the maximum amount of time it might take participants to identify the target (Thorpe, Fize, & Marlot, 1996). The 3- × 10-fold cross-validation was conducted independently for each window size, feature, and participant to select the optimal window size for each feature and participant. The average value across folds was used for each window size to determine the window size in which the true and null classes differed the most for each feature and participant. The mode (i.e., the majority vote) window size was computed across all participants as an index of the most stable window size (Table 1), which was used as input for the group model.

Feature name	Window size (ms)	Description
Fixation detection	1,000	An index of cognitive processing computed as a binary variable to describe if a fixation occurred (1) or not (0). If a fixation was detected, then a participant's eyes were stationary. If a fixation was not detected, then their eyes were moving.
Gaze velocity	916	An index of gaze exploration that was computed as the angular distance between two gaze samples divided by the change in time. Smaller gaze velocities indicated that the eyes were moving slower, whereas larger gaze velocities indicated that the eyes were moving faster.
Angular displacement between previous and current saccade centroid	125	A measure of the distance between subsequent saccade centroids that was computed as the smallest angle needed to rotate the centroid of a saccade (i.e., 3D gaze vector direction) overtop the previous saccade centroid. Saccade centroids were defined as the center position of all samples in a saccade. Smaller angular displacements indicated there were shorter distances between subsequent saccade centroids, whereas larger angular displacements indicated that subsequent saccade centroids were farther apart.
Angular displacement between previous and current saccade landing points	125	A measure of the distance between subsequent saccade landing points computed as the smallest angle needed to rotate the landing point of a saccade (i.e., 3D gaze vector direction) on top of the previous saccade landing point. Smaller angular displacements indicated that saccade landing points were closer together whereas larger angular displacement indicates that saccade landing points were farther apart. This feature is likely correlated with the angular displacement between saccade centroids when subsequent saccades are small because the centroids are close to landing points. Therefore, the angular displacement between saccade landing points provides a unique measure of the variance relative to the angular displacement between saccade centroids when one saccade was large and the next one was small because there would be less distance between the end points than the centroids of those saccades. Conversely, when one saccade was small and the next was large, there would be a greater distance between the end points than the centroids.
Angular displacement between previous and current fixation centroid	83	A measure of the distance between fixation centroids. This was computed as the smallest angle needed to rotate the centroid of a saccade (3D gaze vector direction) overtop the previous fixation centroid. Smaller angular displacements indicated that fixation centroids were closer together whereas larger angular displacements indicated that fixation centroids were further apart. The angular displacement between fixation centroids is a unique event compared with the angular displacement between saccade centroid/landing points and, therefore, the fixation centers would be in different locations than the saccade centroids/landing points.
Midlevel dispersion	916	An index of exploration that was defined as the maximum angular displacement of all the samples from the centroid of a 1000 ms period/window. Increased dispersion indicated that subsequent gaze samples were further apart (i.e., more exploration), whereas decreased dispersion indicated that subsequent gaze samples were closer together (i.e., less exploration).
Saccade velocity: standard deviation	1,000	The standard deviation of saccade velocity. Increased standard deviations of saccade velocity indicated that there was more variability in saccade velocities whereas decreased standard deviation in saccade velocity indicated that there was less variability in saccade velocities.
Midlevel K-coefficient	1,000	A coefficient that described the ambient focal phenomenon 500 ms before the current sample. This measure was derived by computing the z-score for each saccade amplitude and subtracting the z-score of fixation duration from the fixation that proceeded it. The coefficient corresponded to the average value over all saccades and fixations in the past 500 ms. Larger values resulted from large fixation durations with small saccade amplitudes (i.e., focal) whereas small or negative values indicated shorter fixation durations and larger saccade amplitudes (i.e., ambient).
Saccade acceleration: skew	125	A measure of how skewed the saccade acceleration distribution was. This measure was described by computing the skew of saccade acceleration samples. No skew indicated that the tails of the distribution were balanced, whereas positive skew indicated that the tail was on the right (i.e., faster saccade accelerations) and negative skew indicated that the tail was on the left (i.e., slower saccade accelerations). Skew in saccade accelerations typically occur when saccade decelerations are slower than saccade accelerations (Abrams et al., 1989; Chen et al., 2002; Opstal & Van Gisbergen, 1987).
Vertical component of saccade acceleration: skew	167	A measure of the skew of the vertical component of the saccade acceleration distribution. As the saccade acceleration skew feature measures both the horizontal and vertical angles in tandem, the vertical component of saccade acceleration skew only measured the vertical component of saccade acceleration. No skew indicated that the tails of the distribution were balanced, whereas positive skew indicated that the tail was on the right (i.e., faster saccade accelerations in the vertical component) and negative skew indicated that the tail was on the left (i.e., slower saccade accelerations in the vertical component).
Fixation duration	83	An index of cognitive processing that was defined as the end time minus the start time of a fixation. Longer fixation durations indicated increased cognitive processing whereas shorter fixation durations indicated decreased cognitive processing.
Vertical component of saccade velocity: mean	125	The mean of the vertical component of the saccade velocity. A larger mean indicated that saccades were directed more in the vertical direction whereas a smaller mean indicated that saccades were directed less in the vertical direction.

Table 1. A description of the features and the window sizes for those features that were retained from recursive feature selection.

For example, if the data from multiple participant models (i.e., the mode) agreed that a feature's optimal window size was 83 ms, then this feature would be most sensitive to target orienting and encoding anticipation within the 83 ms leading up to encoding. Thus, the true classes (i.e., target/encoding anticipation) would be the most different from the null classes (i.e., no target/encoding anticipation) within 83 ms of encoding.

Stability selection

Stability selection was then applied to the training set to identify the most stable gaze features across participants and to control for any false discoveries that might be made if the feature order was chosen arbitrarily. Initially, feature orders were randomized before stability selection for each participant. This step was undertaken because adding a weaker feature to the model could cause it to be selected erroneously over a

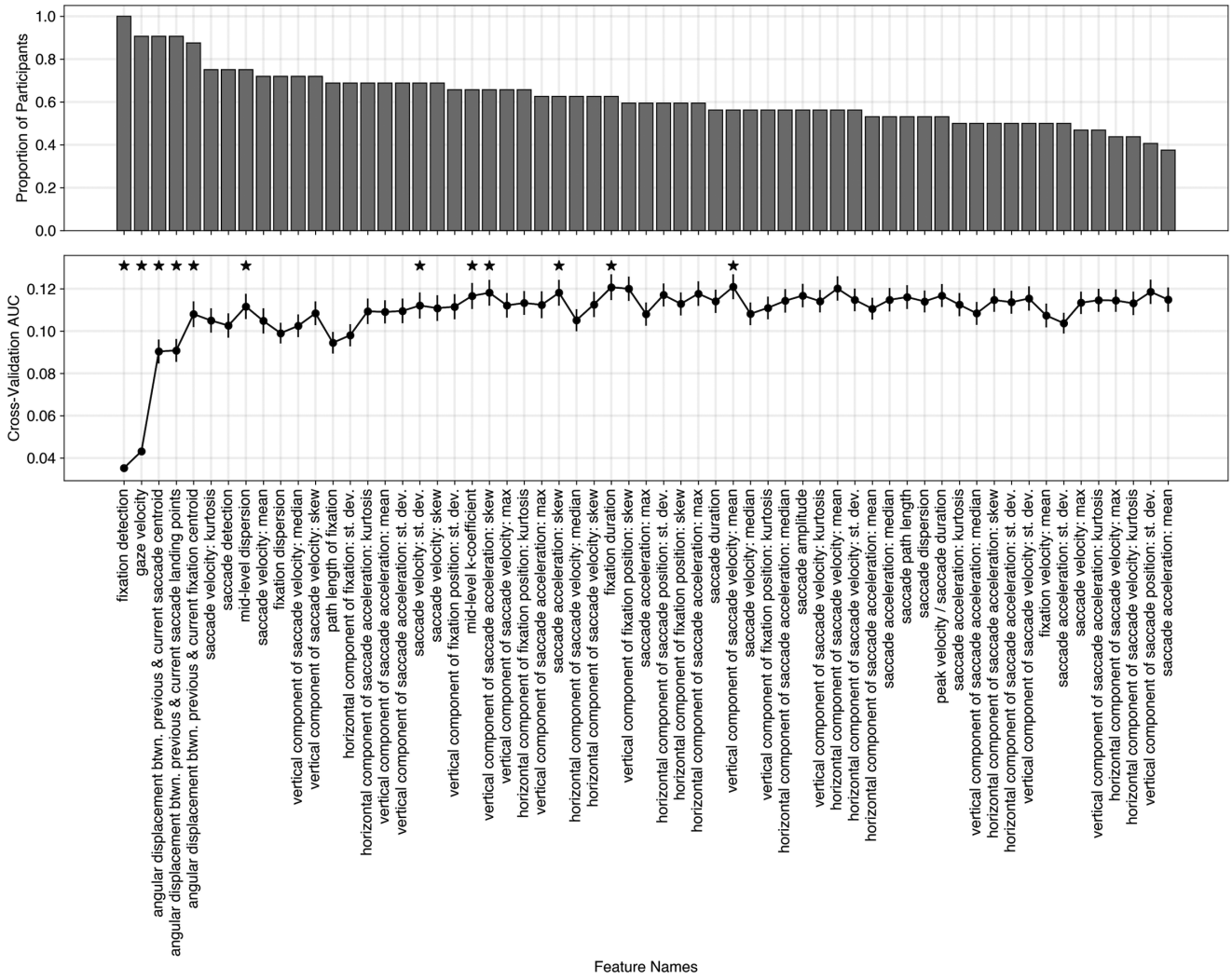

Figure 4. **Results of the stability selection and group feature selection processes.** The top panel depicts the proportion of participants that retained a given feature after the stability selection process. The bottom panel visualizes the features that were retained from the group recursive feature selection process as features were iteratively added, from most retained to least retained. Each point refers to the average AUC-PR from the cross-validation procedure and the error bars refer to 95% confidence intervals. Asterisks correspond with features that increased the AUC-PR relative to the previous benchmark and were thus used in the group models.

stable feature that was added later. If, instead, a weak feature was added to the model in a random order, then, it could be weeded out by stronger features that were added to the model before it. Similarly, if a strong feature was added to the model in a random order, then it could be used to explain the unique variance beyond the other features, irrespective of the order it was added. By adding features randomly to the model, only the best features were selected during stability selection and they were not selected on the premise of order. After randomization, the 3- × 10-fold cross-validation was performed for each feature that was added to the model for each participant. Features were retained if they increased the average AUC-PR across folds; otherwise, they were dropped. All features were concatenated into a single vector for input into the model, each using its optimal window size. The features were then rank ordered by the percentage of participants who retained a given feature (Figure 4a). The resulting feature order served as the input order for the group recursive feature selection process to ensure that noise did not eliminate a good feature.

Group model

The mode window size (Supplementary Table S2) and rank-ordered features (Figure 4a) were then used to build the group model. The group model was trained on a random continuous sequence of 20% of each participant's data. To ensure that the model retained

only useful features, recursive feature selection with stratified 10-fold cross-validation using 3 repeats was performed first. The mean AUC-PR was computed across folds to obtain the final features. The ranked features were input into a group recursive feature selection process one by one, from the most retained to the least retained (Figure 4). Features that increased the AUC-PR on the group training set, based on the average AUC-PR from the cross-validation procedure, were included in the final model. The features that decreased the AUC-PR were dropped from the final model (Figure 4b; Table 1). The retained features were then used to train the group model on the training set only.

Final test data partitions

The within-participant models were trained on 90% of the data with 10% of the data withheld for final model evaluation. The group models were trained on 20% of the data with 80% of the data withheld for final model evaluation. This resulted in no overlap of train or test samples for the within-participant or group models. To then analyze the performance of the group model at the individual level (H1), we pulled from the group model test data (80%). However, we did not use the remaining 80%. Instead, for a fair comparison of the group model performance against the within-participant model performance, we wanted to ensure that the samples in our test set for this analysis met two conditions: (1) They were independent from the model training and cross-validation for both the within-participant models and the group models, and (2) they occurred in both the within-participant model test set and the group model test set. To this end, we isolated the group model test samples that matched the within-participant model test samples for each participant and tested the model on that. This process resulted in test samples that were independent from the training sets. See Figure 5 for a visual depiction of the train and test partitions.

Model testing

H1 sought to determine whether a group model could detect target orienting for encoding across people. If gaze dynamics could reflect as people orient to a target of encoding beyond what is expected by chance, then gaze dynamics alone would be sensitive to target orienting for WM encoding across individuals.

To explore whether the model could exceed chance performance, one-sample *t*-tests were used to evaluate whether the AUC-PR and AUC-ROCs were significantly greater than chance (0.009 and 0.5, respectively). Cohen's D (i.e., the difference between the mean score and the chance rate divided by the standard deviation) (Cohen, 1969, 1992) was also computed for each of the *t*-tests as a measure of effect size. Based on the effect size measures, the loss from training to testing was also computed (Supplementary Analysis 1).

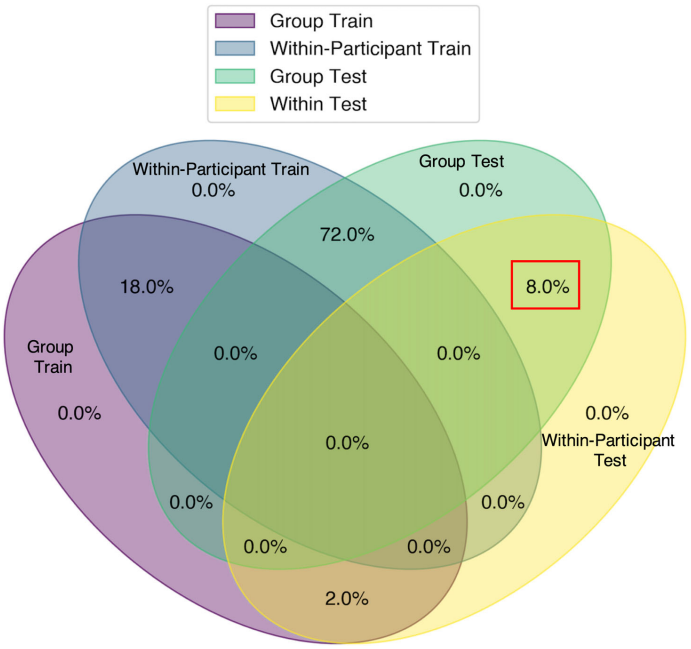

Figure 5. **Partitioning of the train and test sets for an example participant.** The group model was trained on 20% of the data (purple circle) with 80% of the data held out (green circle). The within-participant model was trained on 90% of the data (grey circle) with 10% of the data held out (yellow circle). Samples that overlapped between the group and within-participant test sets were then identified (red box) for the final model evaluation. This process produced test samples that were independent of both training sets.

Results

H1 addressed whether a group model could detect target orienting for WM encoding across people. If gaze dynamics can anticipate target orienting for encoding beyond what is expected by chance, then this work suggests that gaze dynamics alone are indeed sensitive to target orienting for encoding across individuals. To explore whether the model could exceed chance performance, we evaluated whether the AUC-PR and AUC-ROCs were significantly greater than chance (0.009 and 0.5, respectively) via a one-sample *t*-test.

Overall, the results demonstrated that the group model performed significantly better than chance, $M = 0.10$, $SD = 0.07$, $t(31) = 7.94$, $p < 0.001$, $d = 1.3$, when considering the AUC-PR, suggesting that gaze dynamics reflect target orienting for encoding (Figure 6). The same outcome was found when

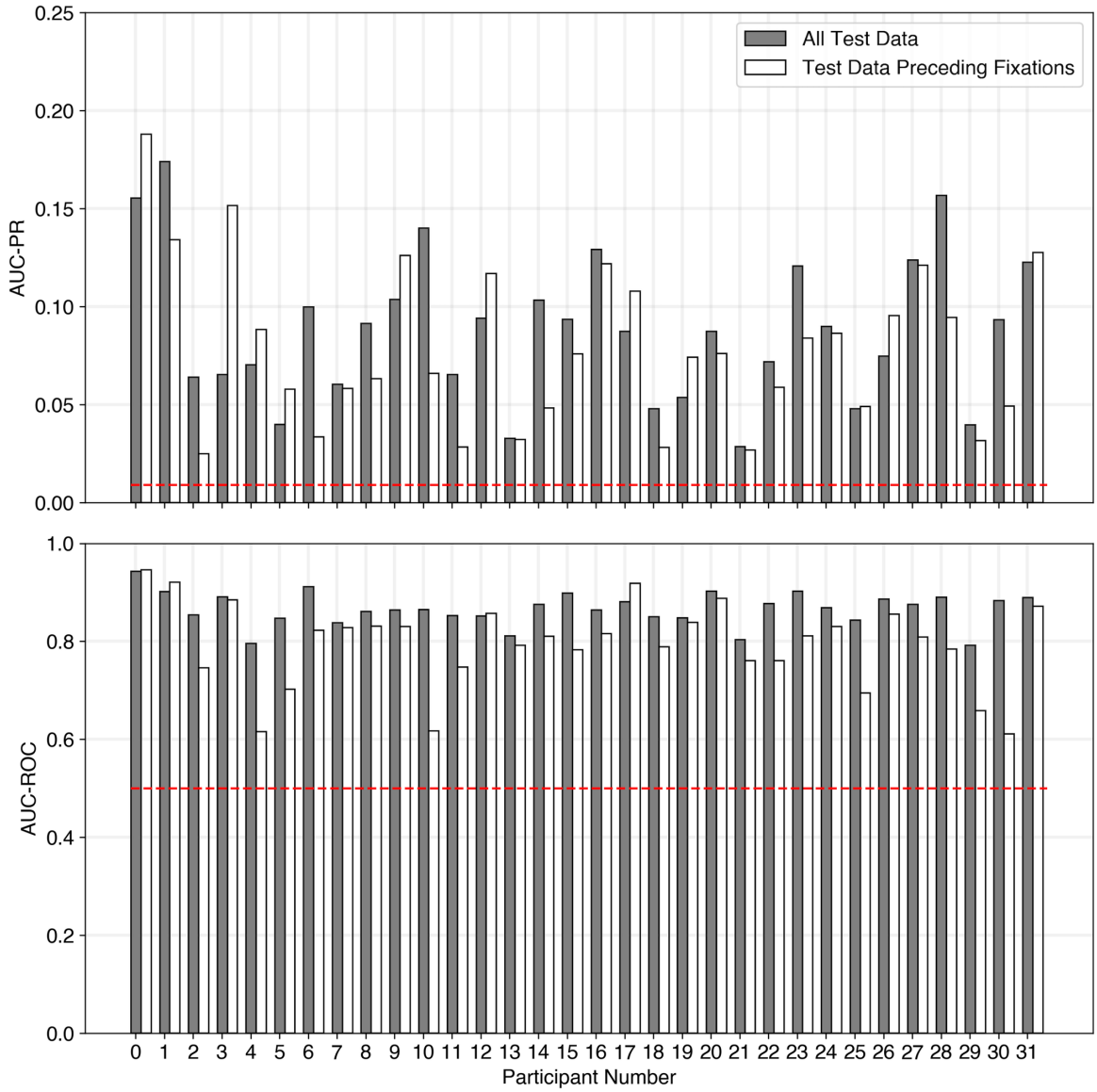

Figure 6. **The group model tested on individual participants.** The gray bars represent the H1 results, and the white bars represent the results corresponding to the filtered test data preceding fixations. The top panel represents the AUC-PR results, and the bottom panel represents the AUC-ROC results. The red, dashed line represents chance performance.

AUC-ROC was used as the evaluation metric, $M = 0.87$, $SD = 0.05$, $t(31) = 42.41$, $p < 0.001$, $d = 7.4$. Note that both effect sizes are large.

One potential reason for the chance performance of the model could be that the model had simply learned to detect the onset of fixations. Given that the true class samples were epoched by determining the onset of fixation on the target object, it is possible that the model learned to detect all fixation onsets, which occurred in every instance of the true class and only occasionally in the null class set. For example, gaze velocity decreases before all fixations so it could be the case that our model was simply a fixation anticipator, rather than an anticipator of a target object. To rule out this possibility, all null class samples that immediately preceded a fixation were identified. This subset matched null classes with true classes by ensuring that all samples immediately preceded the onset of a fixation. These filtered test cases were then resampled to match the standardized chance rate (0.009) (Figure 7).

If the model had learned to detect the onset of fixations rather than encoding target onsets, it should have confused these filtered null classes with true classes, thereby resulting in chance performance. Overall, the H1 model performed significantly above chance when considering AUC-PR, $M = 0.12$, $SD = 0.14$; $t(31) = 4.67$, $p < 0.001$, $d = 0.79$, suggesting that the model could discriminate encoding target onsets specifically from other fixation onsets (Figure 6). The same result was found for the AUC-ROC, $M = 0.79$, $SD = 0.10$;

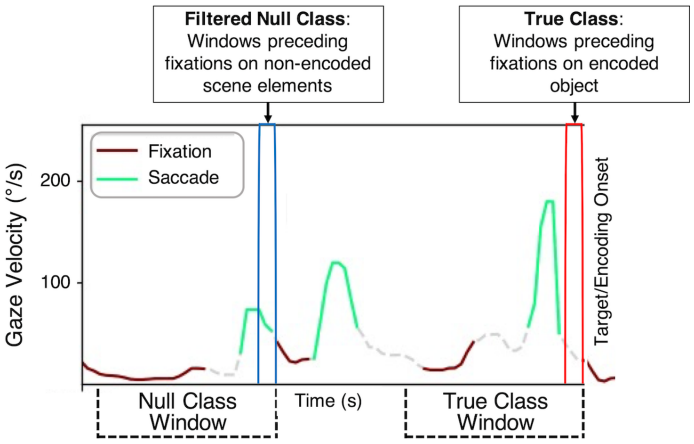

Figure 7. **Fixation detector analysis description.** The time series of gaze velocity corresponds with the samples where saccades (green) and fixations (brown) were detected. To test whether the model was simply detecting fixations, the null class windows that ended just before the onset of a fixation (blue) were filtered. This temporally matched the filtered null classes to the true classes which also ended in no fixation and were followed by the onset of a fixation on the target object (red).

$t(31) = 15.85$, $p < 0.001$, $d = 2.90$. Both effect sizes are again large.

Discussion

Overall, this experiment found that the model was able to detect target orienting for encoding significantly better than would be expected by chance. Although the H1 results are a compelling demonstration of the model's ability to detect target orienting for encoding using gaze, an outstanding issue is that the model has only been tested on a single task in which participants navigated through a single environment. Given that eye movements are influenced by both task (Henderson et al., 2013; Srivastava, Newn, & Velloso, 2018; Tatler et al., 2010; Yarbus, 1967) and the environment (Antes, 1974; Henderson & Hayes, 2017; Loftus & Mackworth, 1978; Nuthmann, 2017; Nuthmann & Einhäuser, 2015; Peacock et al., 2019), it is unknown whether the model trained on Experiment 1 generalizes to other tasks or environments. Therefore, it is important to evaluate whether the model captures general variance in detecting target orienting for encoding.

Experiment 2

A model of WM encoding is only useful if it can detect target orienting for encoding across a range of participants, tasks, and stimuli. As such, it is important to identify and understand the gaze dynamics that capture task- and environment-general variance in target orienting for encoding. To address this open question, we tested the model from Experiment 1 on an entirely novel task completed by a new group of participants from Experiment 2 (Table 2). This task was designed to be quite different in its visual features and task demands; participants were to find previously encoded objects in a cluttered room and encode an associated nonword.

Experiment 2 was a strong test of whether a gaze-based model of WM encoding can generalize to a new set of participants, a new environment, a new task, and new encoding stimuli.

If the model does indeed generalize, Experiment 2 will also allow us to gain insight into the specific gaze features that reflect target orienting for WM encoding in naturalistic environments. Specifically, we explored a subset of models that were trained on individual features on Experiment 1, and then applied these models to Experiment 2, which allowed us to identify eye tracking features that do indeed generalize across tasks and environments.

Methods

Participants

Twenty-seven participants completed the study. Informed consent was obtained, and the protocols were approved by the Western Institutional Review Board. Two participants were excluded due to excessive noise that was observed in the eye movement data (only 1%

Features	Experiment 1	Experiment 2
Participant set	Unique group of participants	Unique group of participants
Room type	Bedroom, patio, kitchen, bathroom, living room	Kitchen
Clutter	Low	High
Task	Find cue and encode target object 5 or 9 times	Find 1 or 3 target objects and encode 1 or 3 associated nonwords
Encoded stimulus	Semantic label of target object	Associated nonword
Memory test	Verbal recall	Nonword recognition

Table 2. Design Differences between Experiments 1 and 2.

of fixations were detected; calibration issues) resulting in a sample of 25 (mean age, 28.45 years; 7 females).

Apparatus and data collection

The same apparatus used in Experiment 1 was used in Experiment 2; however, the experiment was implemented in Unity 2017.1.1f1. Data synchronization and dropped frames were handled in the same manner as Experiment 1.

Stimuli

Instead of using a virtual apartment environment, this experiment used a virtual kitchen that was randomly cluttered with common objects. Some of these objects had nonwords attached to them. The encoded stimulus that would need to be remembered was the nonword attached to the target object.

Procedure

Participants were first outfitted with a VR headset. Participants then completed a tutorial and a practice trial to familiarize themselves with the task. During the task, participants were seated and had free range of motion of their arms and head. They did not navigate throughout the environment.

During each trial in the main trial sequence, participants were first asked to maintain central fixation and press the trigger on the HTC Vive hand controller when they were ready to start the trial. At this point, either one or three common objects (e.g., a soccer ball, gold bar, milk carton) appeared at central fixation for up to 8,000 ms with the option to terminate encoding early to move on to the next phase, which was visual search. Participants were asked to remember these object(s) because they would have to find them in a cluttered environment (Figure 8). Then, participants were spawned into the doorway of a virtual kitchen, which was randomly cluttered with numerous common objects that spanned across 180° from their starting position. The objects were cluttered randomly by dropping them into the scene with several constraints to prevent occlusion (e.g., an invisible triangle was placed above the sink to reroute any falling objects to the counter rather than into the sink).

Participants were to find the target objects and remember the associated nonword. The nonword was a four-letter pseudoword (e.g., oped, umms, tuds, yoms). In each trial, 12 total objects had a randomly assigned nonword, which provided several lures to make search difficult (Figure 8). The trial timed out after 15 seconds if they were not able to complete the search. Participants were given the option to terminate search early if they had encoded the nonword before the 15-s period ended.

Once the participants had found their items, they pressed a key to begin nonword recognition. During the memory test, the target object(s) were shown in the middle of the array with 12 nonwords encircling them (Figure 8). Participants used the HTC Vive Controller to select the associated nonwords. In total, there were 13 one-object trials and 13 three-object trials, resulting in 52 nonwords encoded throughout the entire experiment. Each participant was exposed to a random and unique combination of memory objects and nonwords.

Data preparation

Data segmentation: During Experiment 2, participants may have scanned over target objects or located them before trying to remember the associated nonwords. As such, there may have been cases where there were multiple gaze intersections with objects and nonwords during a trial. Thus, encoding onsets were determined using the fixation following the last gaze intersection with each nonword. The timeseries for one trial was thus segmented into one clip for each encoded nonword. The onset of the first clip was the beginning of visual search up until the last time the first encoded nonword was fixated. The clips used for the second encoded object started after encoding ended on the first nonword up until the last fixation on the second nonword. The clips used for the third encoded object started after

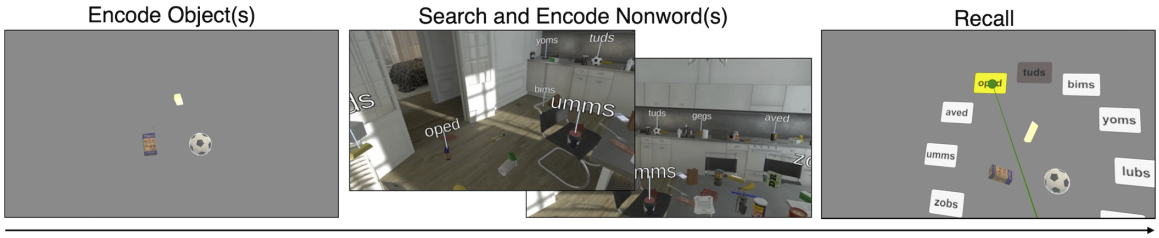

Figure 8. **Trial sequence for Experiment 2.** One or three to-be-remembered objects were presented on a gray background; participants were to find these objects in a cluttered room and then remember the nonword that was located above the target object. After all the objects were found, participants were asked to recall the associated nonwords from an array of target and distractor nonwords.

encoding ended on the second nonword up until the last fixation on the third nonword. Consistent with the data segmentation used in Experiment 1, clips where the time between the gaze intersection and fixation was longer than 500 ms were discarded ($M = 5.31\%$ data loss). Clips corresponding to forgotten items were also discarded (i.e., one-object condition $M = 6.27\%$ data loss; three-object condition $M = 17.00\%$ data loss). If one word was forgotten in a three-word trial, the clips corresponding to the remembered objects were still used. On average, the clip durations were 1,253.85 ms (median = 1,131.70 ms; $SD = 384.30$ ms).

Model

To evaluate whether the group model generalized to a different task and environment, the group model trained on Experiment 1 was tested on the test data from individual participants in Experiment 2.

Evaluation metric

Similar to Experiment 1, both the AUC-PR and AUC-ROC were computed for Experiment 2. A standardized chance rate of 0.9% was also used for AUC-PR in Experiment 2.

Model testing

The group model from Experiment 1 was tested on the gaze data corresponding to each participant from Experiment 2. As per Experiment 1, one-sample t-*tests* were used to evaluate whether the AUC-PR and AUC-ROCs were significantly greater than chance (0.009 and 0.5, respectively). Cohen's D was also computed for each of the *t*-tests as a measure of effect size.

Results

Given that Experiment 2 was vastly different from Experiment 1 in featural information, environment, encoded stimulus, and task, it served as a strong test of whether gaze reflects target orienting for encoding in diverse settings (H2). Despite the differences between Experiments 1 and 2 (Table 2), however, the model from Experiment 1 performed significantly better than chance in Experiment 2 with the AUC-PR, $M = 0.06$, $SD = 0.11$; $t(24) = 2.53$, $p = 0.02$; $d = 0.46$, suggesting that gaze dynamics do capture target orienting for encoding, even when the settings are different. One participant (14) was an outlier in performance, performing significantly better than all others. However, even when removing this superior performer, the results were still significant, AUC-PR $M = 0.05$, $SD = 0.06$, $t(23) = 2.96$, $p = 0.007$; $d = 0.68$ (Figure 9a). Similar results were also found for the AUC-ROC, $M = 0.70$, $SD = 0.16$; $t(24) = 5.99$, $p < 0.001$, $d = 12.5$ (Figure 9b).

We then conducted permutation testing to test for how many subjects of Experiment 2 the model performed significantly above the chance level. To this end, we shuffled the *y*-labels for each participant 10,000 times and compared the true model predictions with the shuffled predictions using *p* values computed from the permutation. To compute these *p* values, we computed the number of permutations with equal or higher accuracy than the AUC for each participant and divided this value by the number of permutations. If the model had learned behaviors true to target orienting for encoding, then the performance on shuffled labels should be significantly lower than the model's performance on the true, nonshuffled labels. Using the *p* values from the permutation ($\alpha = 0.05$), we found that the model performed better relative to the nonshuffled labels for 14 of the 25 participants when considering AUC-PR. When considering the AUC-ROC metric, the model performed better on the nonshuffled labels for 18 out the 25 participants (Supplementary Analysis 2).

Individual feature analysis

To assess which features best generalized between Experiments 1 and 2, we trained a group model on Experiment 1 using each individual feature from the selected features. The features were 12 that were selected from the group-level recursive feature selection procedure. We then tested how well each feature detected target orienting for encoding on each participant's data from Experiment 2 and evaluated it using one-sample *t*-tests to contrast the results against chance. All p-values were corrected using the false discovery rate correction (Benjamini & Hochberg, 1995) in Python using the statsmodels package (version 0.12.2). Overall, 6 of the 12 features were found to significantly detect target orienting for encoding in Experiment 2 (Figure 10; Table 3).

Fixation detection

The first feature that generalized across experiments was fixation detection (Figure 11; see Supplementary Figure S1 for visualizations of the features that did not generalize). There was a higher fixation probability as people oriented to a target for encoding at the end of search (i.e., true classes) than during earlier search (i.e., null classes). Given that prior work suggests that attention and gaze must spatially align with an encoding target for successful WM encoding (Hanning et al., 2015; Schmidt et al., 2002; Woodman et al., 2003), increased fixation probability before encoding is consistent with a hypothesis that the oculomotor system maintains fixation to detect and precisely orient

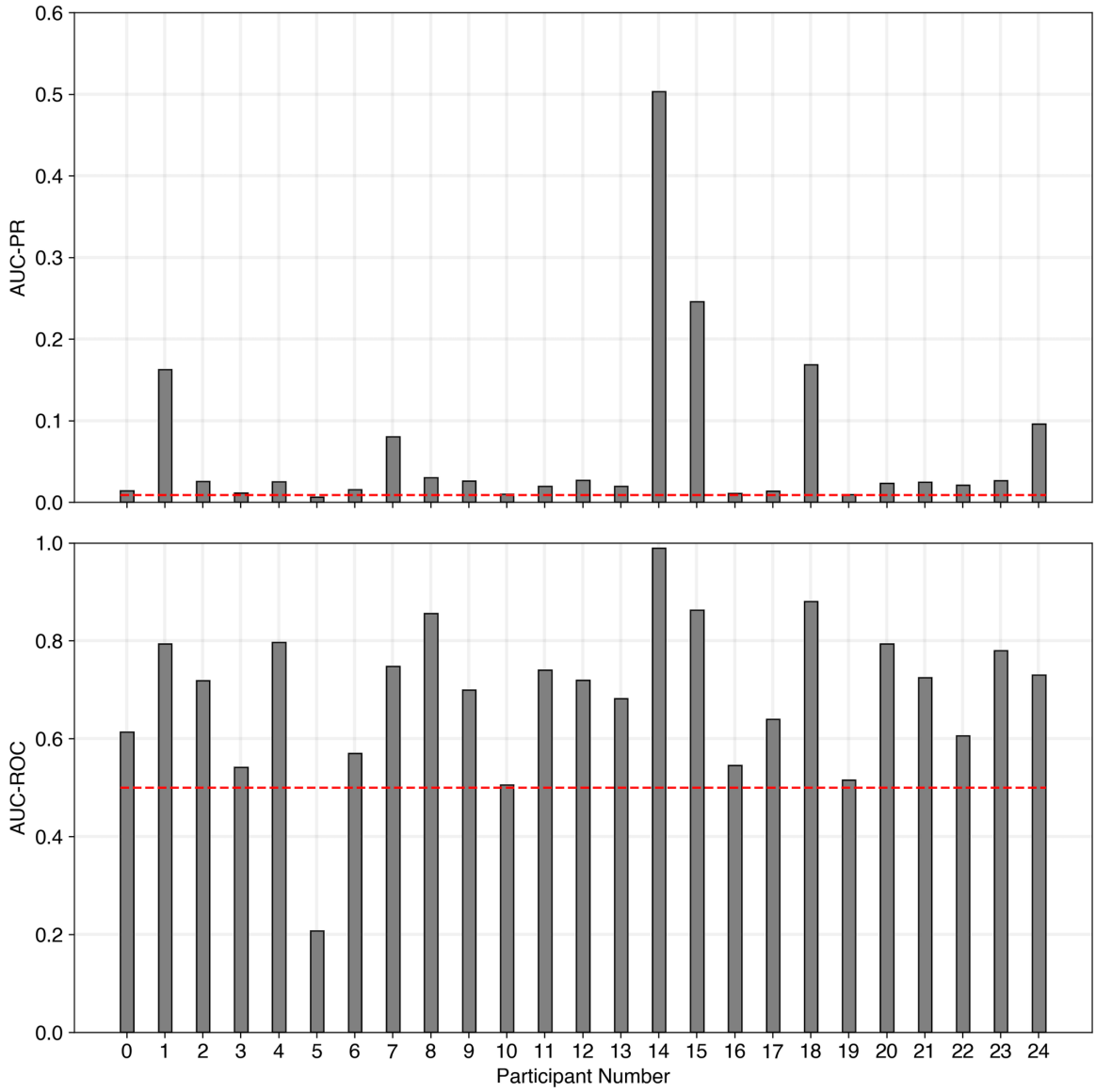

Figure 9. **Group model tested on individual participants from Experiment 2.** The gray bars represent the AUC-PR (top panel) or AUC-ROC (bottom panel) and the red, dashed line represents chance in Experiment 2.

to the encoding target in the periphery (Chen, Acharya, Oulasvirta, & Howes, 2021; Schuetz, Murdison, MacKenzie, & Zannoli, 2019). This, in turn, would likely decrease any visual interference that might occur if a saccade overshot the target.

Angular displacement features

Second, the analysis demonstrated that there was decreased angular displacement between saccade centroids, saccade landing points, and fixation centroids as participants oriented to the target for encoding at the end of search. A decrease in angular displacement indicates that there was a decreased distance between the centroid/landing point of the current event and the centroid/landing point of the prior event, that is, the fixations or saccades were shorter. Angular displacement was computed as the smallest angle needed to rotate the centroid/landing point of a current event (i.e., 3D gaze vector direction) overtop the previous centroid/landing point of the previous event. Here, we observed that the angular displacement decreased as people oriented to the target for encoding. This finding is consistent with a hypothesis that the oculomotor system executes long, orienting saccades to the target of encoding and then produces small, fine-tuning saccades to precisely orient to, and focus on, the encoding target.

The angular displacement between saccade centroids and saccade landing points are likely to be correlated with one another because they are derived from different locations from the same saccades. Given their

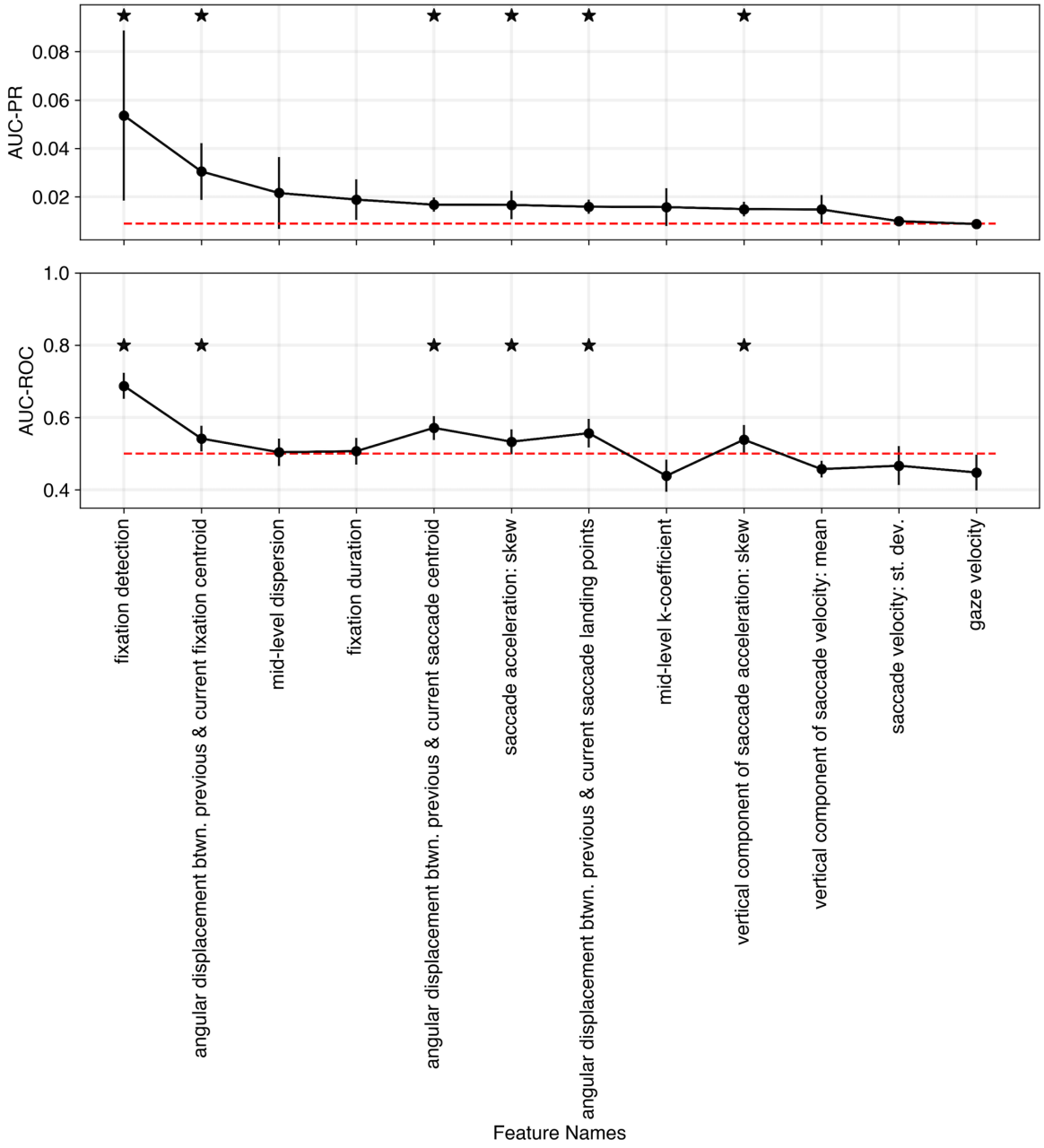

Figure 10. **Features ranked from the most generalizable to the least generalizable.** The top panel visualizes the features ranked based upon AUC-PR, which was the metric used to rank feature performance. The models were trained on the data collected during Experiment 1 and tested on each participant's data collected during Experiment 2. Each data point represents the mean AUC-PR. The bottom panel represents the corresponding mean AUC-ROC for each ranked feature. Asterisks depict whether the false discovery rate–corrected *p* value for each feature was significant ($p < 0.05$). The red, dashed lines represent chance. Error bars represent 95% confidence intervals.

similarity, why, then, were both measures selected during the recursive feature selection? Although the angular displacement between saccade landing points is correlated with the angular displacement between saccade centroids when subsequent saccades are small (because centroids would be close to landing points), the angular displacement between saccade landing points will provide unique variance relative to the angular displacement between saccade centroids when one saccade is large and the next one is small, because there will be less distance between the end points than the centroids of those saccades. Conversely, when one saccade is small and the next is large, there will be a greater distance between the end points than the centroids. Because both measures were selected from recursive feature selection and were smaller as

	AUC-PR					AUC-ROC				
Feature names	M (SD)	T	False discovery rate *p* value	Cohen's D	Generalized?	M (SD)	t	False discovery rate *p* value	Cohen's D	Generalized?
Fixation detection	0.054 (0.09)	2.49	0.04	0.50	Yes	0.69 (0.09)	10.20	<0.001	2.11	Yes
Angular displacement between previous and current fixation centroid	0.030 (0.03)	3.59	0.004	0.70	Yes	0.54 (0.09)	2.31	0.06	0.44	No
Midlevel dispersion	0.022 (0.04)	1.66	0.13	0.33	No	0.50 (0.10)	0.20	0.84	0.00	No
Fixation duration	0.019 (0.02)	2.30	0.05	0.50	No	0.51 (0.10)	0.36	0.79	0.10	No
Angular displacement between previous and current saccade centroid	0.017 (0.01)	5.16	<0.001	0.80	Yes	0.57 (0.08)	4.22	0.002	0.88	Yes
Saccade acceleration: skew	0.017 (0.02)	2.53	0.04	0.40	Yes	0.53 (0.09)	1.88	0.10	0.33	No
Angular displacement between previous and current saccade landing points	0.016 (0.01)	4.67	0.001	0.70	Yes	0.56 (0.10)	2.83	0.03	0.60	Yes
Midlevel K-coefficient	0.016 (0.02)	1.69	0.13	0.35	No	0.44 (0.11)	2.69	0.03	−0.54	No
Vertical component of saccade acceleration: skew	0.015 (0.01)	3.82	0.003	0.60	Yes	0.54 (0.10)	1.87	0.10	0.40	No
Vertical component of saccade velocity: mean	0.015 (0.02)	1.91	0.10	0.30	No	0.46 (0.06)	3.61	0.006	−0.66	No
Saccade velocity: standard deviation	0.010 (0.003)	1.43	0.18	0.33	No	0.47 (0.14)	1.21	0.29	−0.21	No
Gaze velocity	0.009 (0.003)	0.43	0.67	0.00	No	0.45 (0.13)	2.07	0.09	−0.38	No

Table 3. Statistics about the generalizability of individual features for the AUC-PR and AUC-ROC metrics.

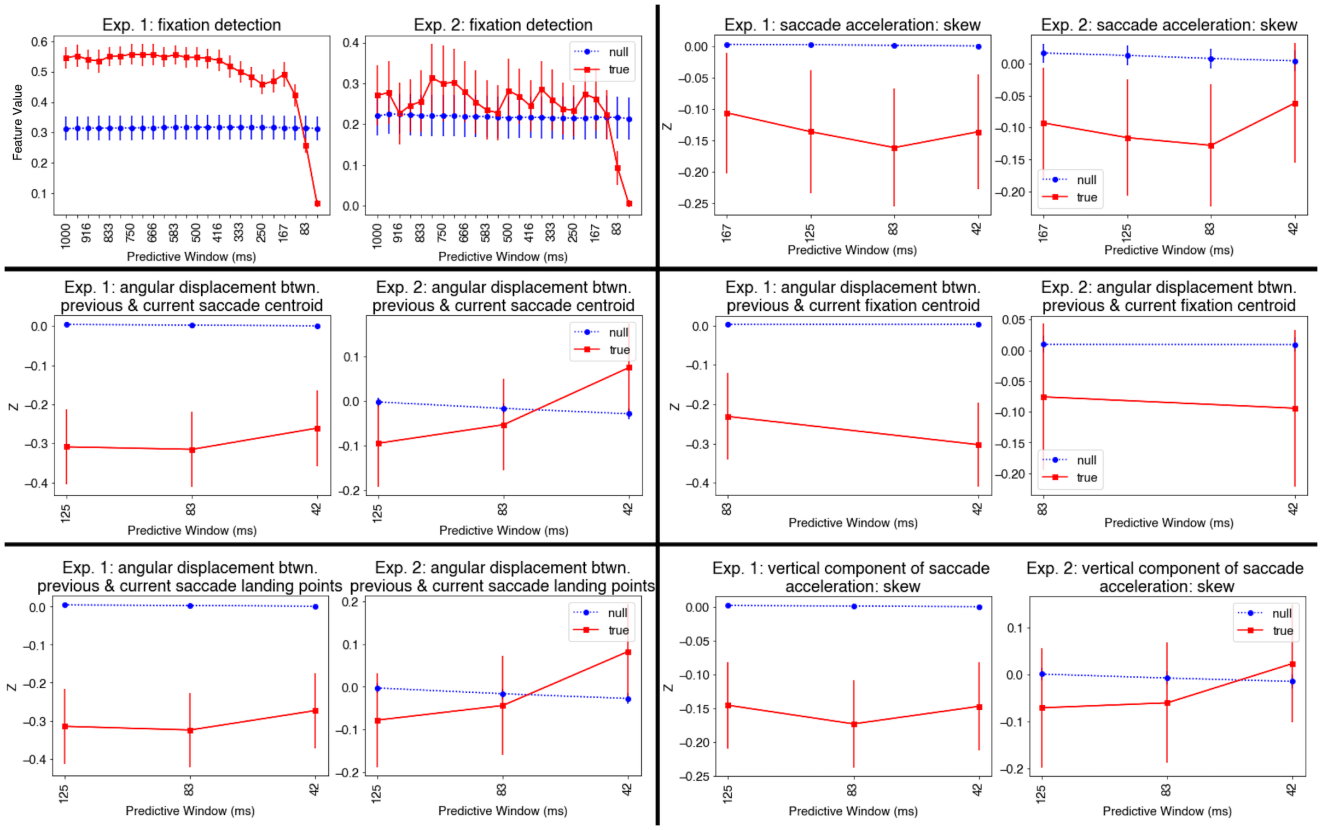

Figure 11. **Comparison of best-generalizing features in Experiments 1 and 2.** This figure depicts the value of the features that generalized relative to the encoding onset across Experiments 1 and 2 based on the AUC-PR results. The red line (solid line, square dots) represents true classes, and the blue line (dotted line, circles) represents null classes. The error bars represent 95% confidence intervals.

people oriented to the target, this finding suggests that subsequent saccades were more variable in size during early search periods than at the end of search as people oriented to the target resulting in unique variance explained by the landing point measure relative to the centroid measure. Indeed, the standard deviation of saccade amplitudes was significantly greater earlier in search (Exp 1: $M = 1.00$, median $= 1.00$, $SD = 0.01$; Exp 2: $M = 0.99$, median $= 0.99$, $SD = 0.05$) than at the end of search as people oriented to the target for encoding (Exp. 1: $M = 0.91$, median $= 0.91$, $SD = 0.28$, $W = 152.0$, $z = –2.09$, $p = 0.04$; Exp. 2: $M = 0.84$, median $= 0.80$, $SD = 0.21$, $W = 73.0$, $z = –2.79$, $p = 0.005$) according to a Wilcoxon signed-rank test that was performed in lieu of a t-test because the distributions were not normally distributed. Together, these results suggest that saccades were longer and more variable in size during early search and were smaller and more consistent in size at the end of search as people oriented to the target.

Finally, fixation centroids were found to differ from the saccade centroid/landing point measures, most likely because they represent different events, and the fixation centroids occur in different locations than the saccade landing points and saccade centroids. Here, the fixation centroids were closer together as people oriented to the encoding target, which suggests that the oculomotor system prioritized the region of encoding for visual analysis.

Overall, the results from these angular displacement features suggest that, at early stages of search, participants made long orienting saccades as they sampled the environment toward a target of interest and then produced small fine-tuning saccades to focus on the target of encoding. This result might also be consistent with a hypothesis that at the end of search, the visual system samples information in the region of encoding to precisely orient to the encoding target which reduces visual interference from encoding-irrelevant regions. During early search, however, it is more beneficial to sample information in different regions to explore the environment.

Saccade acceleration

Finally, the results found negative skew in overall saccade acceleration as people oriented to the target for encoding. Saccade skew typically arises when

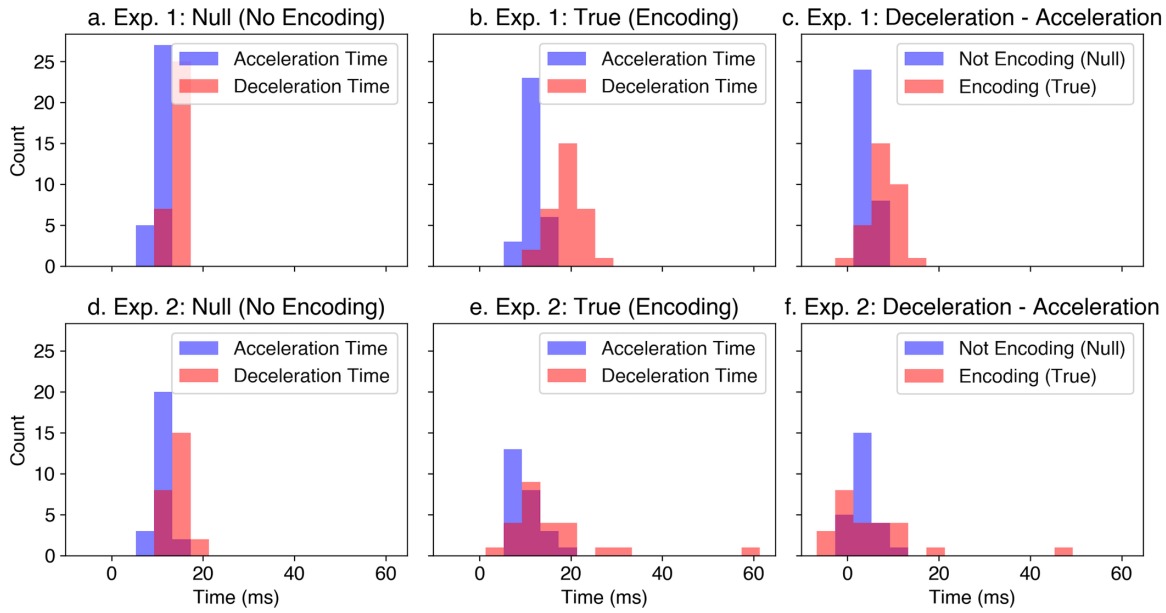

Figure 12. Saccade deceleration versus acceleration in at the end of search when people oriented to the target (i.e., true classes) and early during search (null classes). The saccade deceleration time (red) and acceleration time (blue) during null classes in Experiment 1 (a) and Experiment 2 (d). The saccade deceleration time (red) and acceleration time (blue) during true classes in Experiment 1 and (b) Experiment 2 (e). The difference between deceleration and acceleration for the true classes (red) versus the null classes (blue) for (c) Experiments 1 (c) and 2 (f).

saccade deceleration is slower than saccade acceleration (Abrams, Meyer, & Kornblum, 1989; Chen, Lin, Chen, Tsai, & Shee, 2002; Opstal & Van Gisbergen, 1987), suggesting that, at the end of search, people gradually slow their saccades as they approached an encoding target. To test whether this was the case in our data, we computed saccade deceleration time (i.e., time of saccade end minus time of saccade peak velocity) and saccade acceleration time (i.e., time of saccade peak velocity minus time of saccade start) for the true and null classes (Figure 12). The saccade deceleration time was then subtracted from the saccade acceleration time, resulting in positive values, which indicated slower deceleration and negative values, which indicated slower acceleration. Decelerations were found to be significantly slower as people oriented to the encoding target late in search ($M = 7.92$, $SD = 3.26$) compared with early in search ($M = 3.90$, $SD = 1.59$) in Experiment 1, as evidenced by a paired-samples t-test, $t(31) = 8.12$, $p < 0.001$. A similar pattern was observed in Experiment 2, with slower decelerations before encoding ($M = 5.64$, $SD = 10.26$) than baseline ($M = 3.63$, $SD = 2.82$), although this difference was not statistically significant, $t(24) = 0.91$, $p = 0.37$.

Saccades with higher amplitudes, durations, and velocities (and therefore faster accelerations) have been found to have greater standard deviations in their saccadic end points (Abrams et al., 1989), which suggests that faster saccades might be more likely to undershoot or overshoot a target. Therefore, at the end of search, participants might have slowed their saccade decelerations to avoid overshooting task-relevant encoding information. The negative skew in the vertical component of saccade acceleration was also observed as well, which suggests that as participants anticipated encoding information below the arrow or above the object, they avoided overshooting the target in the vertical direction as well.

Overall, the results from the individual feature analyses suggest that when people orient to an encoding target, the oculomotor system reduces exploratory sampling behaviors that might visually interfere and focuses on precisely orienting to the target. Here, fewer, smaller, and slower eye movements are executed to avoid overshooting the target (Abrams et al., 1989; Chen et al., 2021; Schuetz et al., 2019). This, in turn, might result in reduced visual interference from task-irrelevant information.

Discussion

Experiment 2 explored whether gaze dynamics capture when people orient to a target for encoding in a novel task, which would support the claim that gaze dynamics capture task-general signatures of target orienting for encoding. First, we tested whether the model trained on Experiment 1 would generalize to a new, unseen dataset that contained new participants, a new task, a new virtual environment, and a new type of

encoded stimuli. We hypothesized that our model had learned task-general gaze dynamics, which would allow it to perform above chance even on a task it had never seen before. Indeed, we found that the model detected when people orient to a target for encoding significantly above chance, suggesting that gaze dynamics capture task-general variance in target orienting for encoding.

Second, we explored which specific features generalized to new tasks, participants, and environments. Six of the 12 features were predictive WM encoding onsets in Experiment 2. An examination of the direction of the true classes relative to the null classes suggested that as people orient toward a search target for encoding, they change their oculomotor behavior to avoid overshooting the target. This, in turn, might decrease visual interference from task-irrelevant information.

General discussion

The primary goal of this research was to explore whether people produce characteristic gaze behaviors as they orient toward a target for WM encoding. In service of this goal, in two studies, participants searched for target objects in virtual apartments to encode (Experiment 1) or searched for target objects and had to encode associated nonwords (Experiment 2) into WM. We then explored whether gaze behavior contained consistent patterns that could be used by systems to anticipate users' behavior and intentions while searching for an encoding target. By training sliding window logistic regression models that only used gaze-based features, we found that these models could detect when people oriented to a target for encoding well above chance. Additionally, a follow-up analysis confirmed that the model was detecting target orienting for WM encoding rather than the onset of a fixation, thereby supporting the conclusion that people produce characteristic gaze dynamics as they orient toward a search target for WM encoding. Last, when the model that had been trained on Experiment 1 was tested on the data from Experiment 2, which contained an entirely different task, different stimuli, a novel virtual environment, and new participants, the identified gaze dynamics were generalizable. This methodology provided a strong indication that the model was not simply learning task-specific gaze behaviors.

Prior work has demonstrated that visuospatial attention is required for successful WM encoding (Schmidt et al., 2002; Woodman et al., 2003). That is, to successfully encode novel information into WM, one must orient their spatial attention to the location of encoding. This prior work has demonstrated a temporal relationship between visuospatial attention and WM encoding, but the link has focused on WM processing once people have already oriented their attention to the relevant visuospatial location. To date, it has not yet been explored how people orient their visual attention when they search through naturalistic environments for a target that they will need to remember. This research demonstrated, for the first time, that people produce consistent gaze dynamics as they orient to a target for encoding in naturalistic settings. Furthermore, our modeling framework unveiled several, novel gaze features that were sensitive to the anticipation of task-relevant target information.

Generalizability and the relationship between gaze and target orienting

Beyond establishing that gaze dynamics can be used to detect when people orient to targets for encoding in rich naturalistic tasks, this research also demonstrated that this phenomenon is generalizable across tasks. This was an important exploration because, in prior work, gaze has been found to be affected by both task (Henderson et al., 2013; Srivastava et al., 2018; Tatler et al., 2010; Yarbus, 1967) and the environment (Antes, 1974; Henderson & Hayes, 2017; Loftus & Mackworth, 1978; Nuthmann, 2017; Nuthmann & Einhäuser, 2015; Peacock et al., 2019). Given that a task and environment can substantially alter gaze patterns, it is possible that any model built on gaze data within a single task is in fact capturing task-specific gaze patterns. This is indeed a critique that could be made for any neuro- or bio-sensing device and model (for a similar argument, see Boring, Ridgeway, Shvartsman, & Jonker, 2020). For this reason, we applied our gaze-dynamics model to a novel dataset to explore its generalizability and found that gaze dynamics contain valuable markers of task-general target orienting for encoding behaviors, a finding that has both theoretical and practical value in demonstrating that gaze dynamics can generally capture target orienting for encoding in rich, naturalistic settings.

The model likely generalized well because it was grounded in a core set of gaze features that captured consistent orienting behaviors across the two studies, that is, fixation detection, the angular displacement between the current and previous fixation centroid, the angular displacement between the current and previous saccade centroid, the angular displacement between the current and previous saccade landing positions, vertical component of saccade acceleration skew, and saccade acceleration skew (Figure 11). When it comes to the probability of fixation, there was a greater probability of fixation as people oriented to a target of encoding, suggesting that people fixate as they anticipate orienting toward a target of encoding. This finding is consistent with the hypothesis that the oculomotor system

maintains fixation so that people can detect and precisely orient themselves toward encoding targets in their periphery (Chen et al., 2021; Schuetz et al., 2019). Additionally, the angular displacement between saccade centroids and saccade landing points was smaller as people oriented to the memory target, which suggests that gaze was focused in the region of the encoding target at the end of search and was more exploratory during early search periods. Furthermore, saccades were more variable in size during early search, which suggests that gaze patterns are more diverse when people search without the intent to orient to a target of encoding. Together, these findings suggest that the oculomotor system executes long, orienting saccades during early search and then executes small, fine-tuning saccades as it anticipates precisely orienting to, and encoding, a target (Chen et al., 2021; Schuetz et al., 2019). The angular displacement between fixation centroids was also smaller when people oriented to an encoding target. This finding provides converging evidence that the oculomotor system focused on the general region of the encoding target for visual analysis. Finally, a negative skew in the vertical and total saccade acceleration of gaze samples was observed, meaning that saccade decelerations were slower than accelerations as people oriented to the target. Prior work finds that saccades with faster accelerations have greater standard deviations in their saccadic end points (Abrams et al., 1989), thereby suggesting that faster saccades might be subject to undershoot or overshoot a target (Plamondon & Alimi, 1997). Therefore, saccade decelerations might have been slower to avoid overshooting or undershooting task-relevant target information (Plamondon & Alimi, 1997). This result provides converging evidence that, as people orient toward a target of encoding, they limit exploratory sampling behaviors. Here, they execute fewer, smaller, and slower eye movements to avoid overshooting the target of encoding.

Some of these gaze behaviors might capture another strategy, beyond the precise targeting of a to-be-encoded object. Additionally, some of these gaze behaviors might in fact reflect attempts to decrease visual interference from task-irrelevant information as people focus their attention on encoding into WM. This interpretation is consistent with previous work examining the relationship between visuospatial attention and WM encoding. Indeed, moving one's eye after objects are encoded into visual WM can disrupt memory for those items (Cronin & Irwin, 2018; Tas, Luck, & Hollingworth, 2016). The same has been demonstrated with the spatial allocation of eye movements before encoding: if attention or gaze do not spatially align with the encoding target, WM encoding will fail (Hanning et al., 2015; Schmidt et al., 2002; Woodman et al., 2003). As such, some of the gaze behaviors observed in the present study are also consistent with a strategy to decrease interference before WM encoding. Specifically, fewer eye movements were executed, eye movements were smaller, and saccade decelerations were slower as people oriented to the encoding target, which might suggest that people restrict their oculomotor behavior to precisely fixate the target (and therefore decrease visual interference) as they anticipate encoding.

Ambient focal phenomenon

The ambient focal phenomenon defines the change from exploratory analysis behaviors (i.e., large saccades and short fixations) to focused viewing behaviors (i.e., small saccades and longer fixations) (Krejtz et al., 2016; Unema et al., 2005). While saccade amplitudes, fixation durations, and the k-coefficient (which are core elements of the ambient focal hypothesis) did not generalize in the present study, it could be the case that our data-driven approach unveiled other features that are relevant to the ambient-focal phenomenon than have been previously documented. For example, there was increased fixation probability and reduced distance between saccade centroids as people oriented toward the target at the end of search relative to early search. Furthermore, given that features related to the ambient focal phenomenon generalized despite our use of VR and head-corrected gaze features (rather than gaze features alone as is typical with traditional eye-tracking studies), this suggests that even in complex settings, our results are analogous to those found with traditional eye tracking studies. Our results do not solely reflect the ambient focal phenomenon, however, because there were also features that generalized that were inconsistent with ambient focal. For example, there was greater skew in saccade acceleration at the end of search when the target and encoding were anticipated than at the beginning of search which, does not fit the ambient focal hypothesis. This finding suggests that, although there might be some ambient focal behaviors involved in target orienting for encoding (which might be consistent with the visual interference hypothesis noted in the previous section), the ambient focal hypothesis does not solely explain the present pattern of results.

Target orienting during search

Prior work in the visual search literature has demonstrated that gaze dynamics differ between the scanning and verification stages of search, with longer saccade amplitudes and shorter fixation durations during the scanning stage relative to the target verification stage (David et al., 2020). It is, however, unknown how gaze dynamics unfold throughout scanning as people anticipate orienting to a target.

The present work found changes in gaze dynamics leading up to target acquisition, suggesting that gaze dynamics reflect the anticipation of a target at the end of scanning. Furthermore, because we used a broader set of gaze features than those typically used to study the search stages, this allowed us to glean deeper insights regarding how the eyes move from the start to the end of scanning. For example, the finding that there was greater skew in saccade acceleration at the end of scanning as people oriented to the target than during early scanning suggests that not all saccade targets are created equal during the scanning phase of search.

Our computational modeling framework also allowed us to determine the window size in which each feature was most sensitive to the anticipation of a target. Here, we found that these window sizes ranged from 83 to 1,000 ms, suggesting that gaze behaviors do not unfold in a parallel fashion as a target is anticipated. Indeed, within one second of target acquisition, the probability of fixation decreases, which might be consistent with prior work demonstrating that fixation durations transition from short to long as search progresses (Godwin et al., 2014; Over et al., 2007). However, the angular displacement between saccade centroids and landing points was most sensitive within 125 ms of target acquisition. Although prior work typically models changes in fixation and saccade dynamics throughout the course of search and viewing in parallel (David et al., 2020; Godwin et al., 2014; Krejtz et al., 2016; Over et al., 2007; Unema et al., 2005), the present study suggests that the time courses of these features should be studied independently, because the gaze behaviors related to these features might unfold in an asynchronous manner.

Target orienting versus target orienting for encoding

The primary goal of the present work was to explore whether there is a common set of gaze dynamics that are sensitive to target orienting for encoding. To this end, we modeled both search target orienting and WM encoding together, as these processes co-occur. By better approximating real-world encoding scenarios in which people orient to a search target for encoding, we were able to (1) understand how gaze dynamics unfold as people orient to a target for encoding in complex, ecologically valid settings and (2) build a model that was more likely to generalize to new contexts. In our view, we were successful in both goals. Here, we found that people produced consistent gaze dynamics as they oriented to a target for memory retention across two vastly different studies. Furthermore, our model learned a novel set of task-general gaze behaviors as they relate to how people naturally orient to a search target for encoding.

Because encoding was coupled with search, the model likely captured gaze features that were sensitive to encoding anticipation (i.e., target orienting for encoding) and search target orienting. Supplementary Analysis 3 provided suggestive evidence that the model was significantly more sensitive to encoding anticipation than orienting to a search target, but the above chance performance of the AUC-ROC metric suggests that there are likely gaze behaviors that are common to both processes. Indeed, focused visuospatial attention is required to orient to a search target (Woodman & Luck, 2004) and to successfully orient to and encode a WM target (Awh & Jonides, 2001; Gazzaley & Nobre, 2012; Schmidt et al., 2002). Given the overlap that exists between search target orienting and WM encoding in terms of their co-occurrence in the world and the requirements of visuospatial attention to orient to both search and encoding targets, it will be important for future work to disentangle which gaze features are unique to each process while also having the understanding that some gaze features may be common to both.

The present modeling framework provides an exciting opportunity to tease apart which temporal gaze dynamics are common and unique to successful search target orienting versus WM encoding anticipation. Future work may wish to use the present framework to train a model on a search task without the WM encoding component and test which features generalize to a search task that ends with WM retention. From a theoretical perspective, this would allow for more nuanced conclusions regarding the roles of gaze dynamics to search target orienting and WM encoding. From a practical perspective, if a model could distinguish between successful search target orienting versus WM encoding anticipation, this would allow for different types of cognitive aids.

Implications for memory

In the present study, WM was operationalized as a resource that encodes and maintains information for later use. Here, we used a traditional WM paradigm in which people encoded semantic information from visual perception, retained that information, and later recalled it. Our models were then trained to detect when people orient to a target for encoding, providing support for the hypothesis that gaze dynamics can be used to detect when people orient to a target of encoding. It is important to note, however, that there are many definitions of WM (Baddeley, 2012; Cowan, 2017; Nee & Jonides, 2013) and, as such, it is unclear which cognitive subprocesses the model has learned to detect. For instance, the central executive model of WM

proposes that WM is necessary for both maintaining and *manipulating* information while performing a range of tasks (Baddeley & Hitch, 1974). Our model was trained to detect how people orient to targets of WM encoding with regard to the retention of information, but it is unclear which of the many possible cognitive subprocesses are driving this effect, including encoding for manipulation within the central executive, or other possible cognitive processes.

It could also be the case that the nature of our tasks indexed both WM and long-term memory. Indeed, participants encoded five or nine objects at a time, resulting in some items that were maintained for a longer period than others. We posit a mechanism in which items were initially encoded into WM and then some were transferred to a longer term store depending on various factors, such as duration of rehearsal (Atkinson & Shiffrin, 1968; Hartshorne & Makovski, 2019). Although some items that were maintained longer were likely transferred to a longer term store, there is some evidence that our tasks also indexed WM maintenance strategies. Indeed, participants likely needed to actively rehearse the encoded target objects and nonwords so they would not forget them, which is consistent with a WM-based account (Cowan, 2008). In addition, both tasks probed verbal WM, which has a capacity limit of approximately seven plus or minus two items (Cronin & Irwin, 2018; Miller, 1956). It could also be the case that participants chunked related items as the semantic relatedness of objects was not controlled for. Although there is evidence for both WM and long-term memory stores, a hybrid account could also exist in which some items were maintained in a longer term store and others were maintained in WM. Here, it could be the case that some items were at the focus of attention, whereas others may have been maintained in a longer term store that was readily accessible by attention. This finding would be consistent with the tricentric account of WM (Nee & Jonides, 2013; Oberauer, 2002).

People must first orient to and encode information before they can maintain or manipulate it. Therefore, it stands to reason that these gaze features would reflect target orienting for encoding regardless of the context, as it is necessary to ensure that information is attended to and encoded correctly (with a high fidelity) so that it can be operated on in cognition. Although it is important to understand how various gaze dynamics map onto distinct cognitive subprocesses, this was beyond the scope of this article. This article demonstrated that gaze dynamics carry signatures of people's intentions to orient to targets for encoding, laying the groundwork for future research on specific cognitive subprocesses, which would enable the development of models that can provide more specific predictions about people's transitions between various cognitive states in naturalistic environments.

Future directions

Although the present study chose a large set of features to demonstrate that a model can be used to detect target orienting for encoding, there might be other features that are sensitive to target orienting for encoding. Future work might investigate whether people look at less cluttered regions before encoding to reduce visual interference. It might also be interesting to determine whether saccade landing positions differ for encoded objects versus nonencoded objects. Furthermore, given the top-down nature of the tasks that were used, future work should evaluate whether regressive saccades or relative saccade angles influence encoding. Finally, although head direction was not used as an independent feature in this study, it might be the case that observers better encode objects that require more effort because moving the head is energetically expensive (Draschkow, Kallmayer, & Nobre, 2020; Droll & Hayhoe, 2007; Solman & Kingstone, 2014).

An additional outcome of the present research is that the modeling approach and results provide a framework to explore models of target orienting for encoding in ecologically valid contexts. The results demonstrate that simple models can capture changes in cognitive states in naturalistic environments using only gaze dynamics. This approach has the potential to be integrated within real-time systems, which could be used to help people with WM encoding by deploying adaptive interventions to help decrease memory load. For example, these real-time systems could suggest applications that would allow people to externalize their memory load, such as a notepad app or a camera. Such applications would have implications for people in their everyday lives and even greater benefits to those who suffer memory impairments.

Conclusion

The sliding window logistic regression models presented here allowed for an exploration into the degree to which gaze behaviors can be used to anticipate target orienting for WM encoding across a variety of participants, encoding stimuli, virtual environments, and tasks. Together, these findings provided theoretical insights regarding the role of a broad set of temporal gaze dynamics in target orienting for WM encoding and provide important practical implications that should be considered when using a model that can detect encoding onsets in technology and special populations.

Keywords: working memory, intent prediction, cognitive state, eye movements, decoding

Acknowledgments

The authors thank Taylor Bunge, Aria Fereydouni, and Anjali Misra for assisting with the data collection and memory test scoring, and Brian Hecox for software support. They also thank Deborah Cronin, Marian Berryhill, Lisa Johnson, and Kevin Jones for lending their expertise on WM and for their helpful references. Finally, they thank Michelle Annett for providing copyediting resources.

Commercial relationships: none.
Corresponding author: Tanya R. Jonker.
Email: tanya.jonker@fb.com.
Address: Reality Labs Research, 9845 Willows Rd, Redmond, WA 98052, USA.

References

Abrams, R. A., Meyer, D. E., & Kornblum, S. (1989). Speed and accuracy of saccadic eye movements: Characteristics of impulse variability in the oculomotor system. *Journal of Experimental Psychology: Human Perception and Performance, 15*(3), 15.

Alloway, T. P. (2006). How does working memory work in the classroom? *Educational Research and Reviews, 1*(4), 134–139.

Antes, J. R. (1974). The time course of picture viewing. *Journal of Experimental Psychology, 103*(1), 62–70, https://doi.org/10.1037/h0036799.

Atkinson, R. C., & Shiffrin, R. M. (1968). *Human memory: A proposed system and its control processes* (Vol. *2*). Cambridge, MA: Academic Press.

Awh, E., & Jonides, J. (2001). Overlapping mechanisms of attention and spatial working memory. *Trends in Cognitive Sciences, 5*(3), 119–126.

Baddeley, A. D. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology, 63*, 1–29.

Baddeley, A. D., & Hitch, G. (1974). Working Memory. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. *8*, pp. 47–89). Cambridge, MA: Academic Press, https://doi.org/10.1016/S0079-7421(08)60452-1.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, 57*(1), 289–300, https://doi.org/10.2307/2346101.

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

Boring, M. J., Ridgeway, K., Shvartsman, M., & Jonker, T. R. (2020). Continuous decoding of cognitive load from electroencephalography reveals task-general and task-specific correlates. *Journal of Neural Engineering, 17*(5), 056016, https://doi.org/10.1088/1741-2552/abb9bc.

Borji, A., & Itti, L. (2014). Defending Yarbus: Eye movements reveal observers' task. *Journal of Vision, 14*(3), 29–29, https://doi.org/10.1167/14.3.29.

Chen, X., Acharya, A., Oulasvirta, A., & Howes, A. (2021). An adaptive model of gaze-based selection. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, (virtual) May 8–13, 2021*; 1–11, https://doi.org/10.1145/3411764.3445177.

Chen, Y.-F., Lin, H.-H., Chen, T., Tsai, T.-T., & Shee, I.-F. (2002). The peak velocity and skewness relationship for the reflexive saccades. *Biomedical Engineering: Applications, Basis and Communications, 14*(02), 71–80, https://doi.org/10.4015/S1016237202000115.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences.* New York, NY: Academic Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? *Progress in Brain Research, 169*, 323–338, https://doi.org/10.1016/S0079-6123(07)00020-9.

Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review, 24*, 1158–1170.

Cronin, D. A., & Irwin, D. E. (2018). Visual working memory supports perceptual stability across saccadic eye movements. *Journal of Experimental Psychology: Human Perception and Performance, 44*(11), 1739–1759, https://doi.org/10.1037/xhp0000567.

David, E., Beitner, J., & Võ, M. L.-H. (2020). Effects of transient loss of vision on head and eye movements during visual search in a virtual environment. *Brain Sciences, 10*(11), 841, https://doi.org/10.3390/brainsci10110841.

David-John, B., Peacock, C. E., Zhang, T., Murdison, T. S., Benko, H., & Jonker, T. R. (2021). Towards gaze-based prediction of the intent to interact in virtual reality. *ACM Symposium on Eye Tracking Research and Applications, Stuttgart. Germany, 2021*; 7.

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves.

Proceedings of the 23rd International Conference on Machine Learning - ICML '06, Pittsburgh, PA, June 25–29; 233–240, https://doi.org/10.1145/1143844.1143874.

Diaz, G., Cooper, J., Kit, D., & Hayhoe, M. (2013). Real-time recording and classification of eye movements in an immersive virtual environment. *Journal of Vision, 13*(12), 5–5, https://doi.org/10.1167/13.12.5.

Doshi, A., & Trivedi, M. M. (2009). On the roles of eye gaze and head dynamics in predicting driver's intent to change lanes. *IEEE Transactions on Intelligent Transportation Systems, 10*(3), 453–462, https://doi.org/10.1109/TITS.2009.2026675.

Dowiasch, S., Marx, S., Einhäuser, W., & Bremmer, F. (2015). Effects of aging on eye movements in the real world. *Frontiers in Human Neuroscience, 9*, 46, https://doi.org/10.3389/fnhum.2015.00046.

Draschkow, D., Kallmayer, M., & Nobre, A. C. (2020). When natural behavior engages working memory. *Current Biology, 31*, 869–874.e5, https://doi.org/10.1016/j.cub.2020.11.013.

Droll, J. A., & Hayhoe, M. M. (2007). Trade-offs between gaze and working memory use. *Journal of Experimental Psychology: Human Perception and Performance, 33*(6), 1352–1365, https://doi.org/10.1037/0096-1523.33.6.1352.

Egeth, H. E., & Yantis, S. (1997). Visual attention: Control, representation, and time course. *Annual Review of Psychology, 48*(1), 269–297, https://doi.org/10.1146/annurev.psych.48.1.269.

Ellis, S. R., & Stark, L. (1981). In *Pilot scanning patterns while viewing cockpit displays of traffic information*. NASA Ames Research Center & U.C. Berkely Department of Physiological Optics, https://doi.org/10.1093/oxfordhb/9780199539789.013.0033

Gazzaley, A., & Nobre, A. C. (2012). Top-down modulation: Bridging selective attention and working memory. *Trends in Cognitive Sciences, 16*(2), 129–135, https://doi.org/10.1016/j.tics.2011.11.014.

George, A., & Routray, A. (2016). Real-time eye gaze direction classification using convolutional neural network. *2016 International Conference on Signal Processing and Communications (SPCOM), Bangalore, India, June 12–15*; 1–5, https://doi.org/10.1109/SPCOM.2016.7746701⟨/bib>

Godwin, H. J., Reichle, E. D., & Menneer, T. (2014). Coarse-to-fine eye movement behavior during visual search. *Psychonomic Bulletin & Review, 21*(5), 1244–1249, https://doi.org/10.3758/s13423-014-0613-6⟨/bib>

Gruber, O., & Goschke, T. (2004). Executive control emerging from dynamic interactions between brain systems mediating language, working memory and attentional processes. *Acta Psychologica, 115*(2), 105–121, https://doi.org/10.1016/j.actpsy.2003.12.003⟨/bib>

Hanning, N. M., Jonikaitis, D., Deubel, H., & Szinte, M. (2015). Oculomotor selection underlies feature retention in visual working memory. *Journal of Neurophysiology, 115*(2), 1071–1076, https://doi.org/10.1152/jn.00927.2015⟨/bib>

Hartshorne, J. K., & Makovski, T. (2019). The effect of working memory maintenance on long-term memory. *Memory & Cognition, 47*(4), 749–763, https://doi.org/10.3758/s13421-019-00908-6⟨/bib>

Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour, 1*, 743–747, https://doi.org/10.1038/s41562-017-0208-0⟨/bib>

Henderson, J. M., Shinkareva, S. V., Wang, J., Luke, S. G., & Olejarczyk, J. (2013). Predicting cognitive state from eye movements. *PloS One, 8*(5), e64937, https://doi.org/10.1371/journal.pone.0064937⟨/bib>

Huang, C.-M., Andrist, S., Sauppé, A., & Mutlu, B. (2015). Using gaze patterns to predict task intent in collaboration. *Frontiers in Psychology, 6*, 1049, https://doi.org/10.3389/fpsyg.2015.01049.

Krejtz, K., Duchowski, A., Krejtz, I., Szarkowska, A., & Kopacz, A. (2016). Discerning ambient/focal attention with coefficient *K*. *ACM Transactions on Applied Perception, 13*(3), 11:1–11:20, https://doi.org/10.1145/2896452.

Lengyel, G., Carlberg, K., Samad, M., & Jonker, T. R. (2021). Predicting visual attention using the hidden structure in eye-gaze dynamics. *EMICS '21: ACM CHI '21 Workshop on Eye Movements as an Interface to Cognitive State, Yokoyama, Japan, May 8–16*;9.

Liu, A. (2001). *Modeling and prediction of human driver behavior*. *9th International Conference on Human-Computer Interaction, New Orleans, LA, August; 5*.

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance, 4*(4), 565–565.

Malcolm, G. L., & Henderson, J. M. (2009). The effects of target template specificity on visual search in real-world scenes: Evidence from eye movements. *Journal of Vision, 9*(11), 8–8, https://doi.org/10.1167/9.11.8.

McPeek, R. M. (2006). Incomplete suppression of distractor-related activity in the frontal eye field results in curved saccades. *Journal of Neurophysiology, 96*(5), 2699–2711.

McPeek, R. M., Han, J. H., & Keller, E. L. (2003). Competition between saccade goals in the superior colliculus produces saccade curvature. *Journal of Neurophysiology, 89*(5), 2577–2590.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81–97.

Milstein, D. M., & Dorris, M. C. (2007). The influence of expected value on saccadic preparation. *Journal of Neuroscience, 27*(18), 4810–4818, https://doi.org/10.1523/JNEUROSCI.0577-07.2007.

Nee, D. E., & Jonides, J. (2013). Trisecting representational states in short-term memory. *Frontiers in Human Neuroscience, 7*, 796–796.

Nuthmann, A. (2017). Fixation durations in scene viewing: Modeling the effects of local image features, oculomotor parameters, and task. *Psychonomic Bulletin & Review, 24*(2), 370–392, https://doi.org/10.3758/s13423-016-1124-4.

Nuthmann, A., & Einhäuser, W. (2015). A new approach to modeling the influence of image features on fixation selection in scenes. *Annals of the New York Academy of Sciences, 1339*(1), 82–96, https://doi.org/10.1111/nyas.12705.

Oberauer, K. (2002). Access to information in working memory: Exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 411–421.

Opstal, A. J. V., & Van Gisbergen, J. A. M. (1987). Skewness of saccadic velocity profiles: A unifying parameter for normal and slow saccades. *Vision Research, 27*, 731–745.

Over, E. A. B., Hooge, I. T. C., Vlaskamp, B. N. S., & Erkelens, C. J. (2007). Coarse-to-fine eye movement strategy in visual search. *Vision Research, 47*(17), 2272–2280, https://doi.org/10.1016/j.visres.2007.05.002.

Peacock, C. E., David-John, B., Zhang, T., Murdison, T. S., Benko, H., & Jonker, T. R. (2021). Gaze signatures decode the onset of working memory encoding. *EMICS '21: ACM CHI '21 Workshop on Eye Movements as an Interface to Cognitive State, Yokahama, Japan, May 8–16; 6*.

Peacock, C. E., Hayes, T. R., & Henderson, J. M. (2019). Meaning guides attention during scene viewing even when it is irrelevant. *Attention, Perception, and Psychophysics, 81*(1), 20–34, https://doi.org/10.3758/s13414-018-1607-7.

Pekkanen, J., & Lappi, O. (2017). A new and general approach to signal denoising and eye movement classification based on segmented linear regression. *Scientific Reports, 7*(1), 17726, https://doi.org/10.1038/s41598-017-17983-x.

Pereira, E. J., & Castelhano, M. S. (2019). Attentional capture is contingent on scene region: Using surface guidance framework to explore attentional mechanisms during search. *Psychonomic Bulletin & Review, 26* (4), 1273–1281, https://doi.org/10.3758/s13423-019-01610-z.

Plamondon, R., & Alimi, A. (1997). Speed/accuracy trade-offs in target-directed movements. *Behavioral and Brain Sciences, 20*, 279–303; discussion 303, https://doi.org/10.1017/S0140525X97001441.

Poole, B. J., & Kane, M. J. (2009). Working-memory capacity predicts the executive control of visual search among distractors: The influences of sustained and selective attention. *Quarterly Journal of Experimental Psychology, 62*(7), 1430–1454, https://doi.org/10.1080/17470210802479329.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One, 10*(3), e0118432, https://doi.org/10.1371/journal.pone.0118432.

Salthouse, T. A., & Ellis, C. L. (1980). Determinants of eye-fixation duration. *American Journal of Psychology, 93*(2), 207–234, https://doi.org/10.2307/1422228.

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. *Proceedings of the 2000 Symposium on Eye Tracking Research and Applications, New York, NY, November 6–8*; 71–78.

Schmidt, B. K., Vogel, E. K., Woodman, G. F., & Luck, S. J. (2002). Voluntary and involuntary attentional control of visual working memory. *Perception and Psychophysics, 64*(754–763).

Schuetz, I., Murdison, T. S., MacKenzie, K. J., & Zannoli, M. (2019). An explanation of Fitts' law-like performance in gaze-based selection tasks using a psychophysics approach. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Scotland, UK, May 4–9*; 1–13, https://doi.org/10.1145/3290605.3300765.

Schurgin, M. W., Nelson, J., Iida, S., Ohira, H., Chiao, J. Y., & Franconeri, S. L. (2014). Eye movements during emotion recognition in faces. *Journal of Vision, 14*(13), 14–14, https://doi.org/10.1167/14.13.14.

Solman, G. J. F., & Kingstone, A. (2014). Balancing energetic and cognitive resources: Memory use during search depends on the orienting effector. *Cognition, 132*(3), 443–454, https://doi.org/10.1016/j.cognition.2014.05.005.

Srivastava, N., Newn, J., & Velloso, E. (2018). Combining low and mid-level gaze features for desktop activity recognition. *Proceedings*

of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2, 1–27, https://doi.org/10.1145/3287067.

Tas, A. C., Luck, S. J., & Hollingworth, A. (2016). The relationship between visual attention and visual working memory encoding: A Dissociation between covert and overt orienting. *Journal of Experimental Psychology: Human Perception and Performance, 42*, 1121–1138.

Tatbul, N., Lee, T. J., Zdonik, S., Alam, M., & Gottschlich, J. (2018). Precision and Recall for Time Series. *32nd Conference on Neural Information Processing Systems, Montreal, Quebec, Canada, December 2–8*; 11.

Tatler, B. W., Wade, N. J., Kwan, H., Findlay, J. M., & Velichkovsky, B. M. (2010). Yarbus, eye movements and vision. *I-Perception, 1*(1), 7–27.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*, 520–522.

Titz, C., & Karbach, J. (2014). Working memory and executive functions: Effects of training on academic achievement. *Psychological Research, 78*(6), 852–868, https://doi.org/10.1007/s00426-013-0537-1.

Unema, P. J. A., Pannasch, S., Joos, M., & Velichkovsky, B. M. (2005). Time course of information processing during scene perception: The relationship between saccade amplitude and fixation duration. *Visual Cognition, 12*(3), 473–494, https://doi.org/10.1080/13506280444000409.

Wilimzig, C., Schneider, S., & Schöner, G. (2006). The time course of saccadic decision making: Dynamic field theory. *Neural Networks: The Official Journal of the International Neural Network Society, 19*, 1059–1074, https://doi.org/10.1016/j.neunet.2006.03.003.

Woodman, G. F., & Luck, S. J. (2004). Visual search is slowed when visuospatial working memory is occupied. *Psychonomic Bulletin & Review, 11*(2), 269–274, https://doi.org/10.3758/BF03196569.

Woodman, G. F., Vecera, S. P., & Luck, S. J. (2003). Perceptual organization influences visual working memory. *Psychonomic Bulletin & Review, 10*(1), 80–87, https://doi.org/10.3758/BF03196470.

Wu, C.-C., & Kowler, E. (2013). Timing of saccadic eye movements during visual search for multiple targets. *Journal of Vision, 13*(11), 11–11, https://doi.org/10.1167/13.11.11.

Yarbus, A. L. (1967). Eye movements during perception of complex objects. In *Eye Movements and Vision* (pp. 171–211). New York: Springer, https://doi.org/10.1007/978-1-4899-5379-7_8.