Affective and Pleasurable Design, Vol. 204, 2026, 62–72
https://doi.org/10.54941/ahfe1007332

Deep Learning of Latent Gaze Representations for Cognitive Ability and Mental State Estimation

Shunpei Kiuchi[1], Masami Matsushima[2,3], and Keiichi Watanuki[1,3]

[1]Graduate School of Science and Engineering, Saitama University 255 Shimo-okubo, Sakura-ku, Saitama-shi, Saitama 338-8570, Japan
[2]Je respire Co., Ltd. 8-8-10-101 Akasaka, Minato-ku, Tokyo 107-0052, Japan
[3]Advanced Institute of Innovative Technology, Saitama University 255 Shimo-okubo, Sakura-ku, Saitama-shi, Saitama 338-8570, Japan

ABSTRACT

This paper proposes a novel pipeline for estimating individual cognitive processing ability from eye-tracking data using the latent representations of a deep-learning model trained to predict trial-level correctness during problem-solving. Ten undergraduate participants completed four cognitive tasks: reading comprehension, visuospatial reasoning, memory recall, and focused attention. Binocular gaze data were recorded using a Tobii Pro eye-tracking device. A Transformer-based sequence model, optimized via a 200-trial Bayesian search using the silhouette score of the resulting latent space as the objective, was trained to classify each 100-timestep gaze sequence as correct or incorrect. The optimal architecture achieved a validation accuracy of 0.725 and produced a 32-dimensional latent representation per trial. Univariate logistic regression identified the three most cognitively informative latent dimensions (z_{0028}, z_{0022}, and z_{0031}), each achieving a classification accuracy of 0.700–0.712 independently. Within the resulting 3D subspace, the per-participant centroids of correct- and incorrect-trial embeddings exhibited consistent directional displacement along the primary cognitive axis, providing an interpretable subject-level index of the cognitive processing ability without any external standardized assessment. A supplementary longitudinal experiment further demonstrated that the session-level centroid shifted substantially toward and beyond the correct-trial region following a task-specific training intervention in a single participant, suggesting that the proposed representation is sensitive to training-induced cognitive changes. Although the latent space exhibited weak global cluster separation and the training experiment remains preliminary, these findings support the viability of gaze-based latent centroid tracking as a non-invasive biomarker for both static cognitive profiling and longitudinal cognitive change detection.

Keywords: Concept activation vector, Latent space, Eye tracking, Cognitive ability estimation, Transformer, Gaze-based assessment

INTRODUCTION

Quantifying human cognitive processing ability is essential for developing personalized learning support systems and designing efficient human–computer interfaces. The measurement of cognitive abilities has traditionally

Received April 14, 2026; Revised May 29, 2026; Accepted June 12, 2026; Available online July 20, 2026

relied on standardized paper-and-pencil tests or lengthy psychological evaluations (Institute of Medicine, 2015). However, these conventional methods impose significant physical and psychological burdens on the respondents and are ill suited for real-time or continuous assessment. In recent years, eye-tracking data have garnered significant attention as a powerful, non-invasive metric that reflects internal cognitive processes and attention allocation during task execution (Rayner, 1998).

Numerous studies have leveraged eye-tracking data alongside machine-learning and deep-learning techniques to predict learner comprehension, task correctness, or domain-specific expertise (Arnold et al., 2025; Shubi et al., 2024; Hosp et al., 2021). Despite these advancements, most existing approaches are limited to directly predicting specific task performances, such as whether a respondent will answer a particular question correctly. A significant gap remains in establishing a simple and versatile method for extracting and estimating a respondent's fundamental cognitive processing abilities from gaze data, which lie beneath the surface-level task performance.

To address this limitation, this study proposes a novel method for estimating fundamental cognitive processing abilities by utilizing the latent representations of a deep-learning model trained to predict task correctness from gaze data during problem-solving. Because a deep-learning model learns the correctness-prediction task, its high-dimensional latent representations capture condensed cognitive processing patterns that extend beyond mere physical gaze characteristics (Bengio et al., 2013). By leveraging this latent space, our approach aims to replace time-consuming traditional cognitive tests and estimate individual cognitive traits with ease by solely using task-based gaze data.

The objective of this study is to demonstrate the effectiveness of the proposed method and establish that latent representations of gaze data can serve as viable biomarkers for cognitive processing ability. The main contributions of this study are as follows: 1) We propose a novel pipeline that estimates essential cognitive processing abilities using the latent representations extracted from an eye-tracking-based correctness-prediction model. 2) We demonstrate that the proposed method enables the estimation of cognitive processing ability from gaze data alone, without relying on conventional cognitive assessments.

RELATED WORK

Eye tracking has been widely utilized in cognitive science and educational psychology to understand internal human states. Previous research has demonstrated that gaze metrics such as fixation duration, saccade velocity, and pupil dilation are closely correlated with cognitive load, attention, and reading comprehension (Gorin et al., 2024). For instance, Copeland et al. (2015) showed that specific gaze patterns during reading tasks can indicate the difficulty of the text and the reader's cognitive effort. Although

these studies have established a link between gaze and cognition, they have typically relied on handcrafted statistical features, rather than deep temporal representations.

With the advancement of deep learning, researchers have increasingly applied neural networks such as convolutional neural networks, recurrent neural networks, and long short-term memory to raw or sequence-based eye-tracking data. These models have shown high accuracy in predicting specific outcomes, such as whether a student will solve a problem correctly, or classifying a user's level of expertise (Shubi et al., 2024; Hosp et al., 2021). However, these models are generally optimized end to end for a single downstream task (e.g., the binary classification of correctness). The internal representations learned by these models have rarely been repurposed to evaluate broader cognitive traits.

In the broader field of deep learning, analyzing and utilizing latent representations (embeddings) have become standard practice for transfer learning and feature extraction. In domains such as natural language processing and computer vision, latent spaces can encode rich generalized semantic information that can be transferred to secondary tasks (Bengio et al., 2013; Yosinski et al., 2014). Within our context, we hypothesized that a model trained to predict problem-solving correctness inherently encodes the respondent's information-processing efficiency within its latent layers.

Although previous studies have successfully used gaze for task-specific predictions and latent spaces for transfer learning in other domains, the intersection of these concepts remains underexplored. Specifically, the extraction of generalized cognitive processing abilities from the latent space of a task-specific (correctness-prediction) gaze model is a novel approach. This study bridges this gap by demonstrating that deep latent representations contain sufficient intrinsic cognitive information to bypass traditional psychological assessments.

METHODOLOGY

Overview

Figure 1 depicts the overall pipeline of the proposed method, which consists of three stages: (1) collecting eye-tracking data during cognitive task execution; (2) training a Transformer-based correctness prediction model and extracting its 32-dimensional latent representations; and (3) identifying the most cognitively informative latent dimensions via logistic regression and estimating the individual cognitive processing ability from the resulting three-dimensional (3D) centroid coordinates. In addition to this primary pipeline, a supplementary longitudinal experiment was conducted with one participant to examine whether the proposed method can detect changes in cognitive processing ability induced by task-specific training.

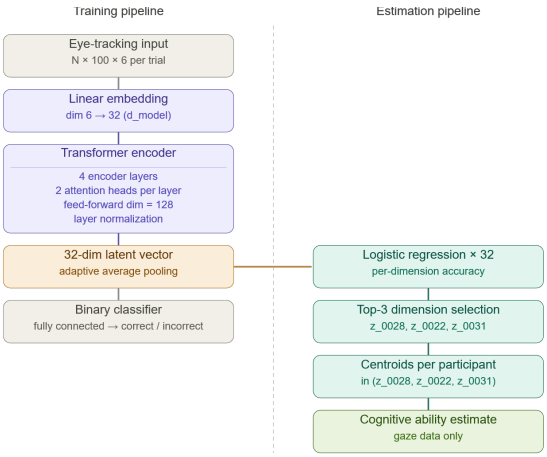

Figure 1: Overview of the proposed pipeline.

Participants and Experimental Design

Ten undergraduate students participated in the primary experiment. All participants had normal or corrected-to-normal vision and provided written informed consent prior to the experiment. Eye-movement data were recorded using Tobii Pro Fusion, which captured the binocular gaze coordinates and pupil diameter at a sampling rate of 120 Hz. The participants were seated at a fixed distance from the screen in a controlled laboratory environment to minimize head movement and ambient distraction. Figure 2 depicts the experimental environment.

Each participant completed a battery of four cognitive tasks designed to test distinct aspects of cognitive processing: (i) reading comprehension, targeting linguistic processing and semantic integration; (ii) figure and spatial recognition, targeting visuospatial reasoning; (iii) memory recall, targeting short-term memory retention; and (iv) focused attention, targeting sustained and selective attention. The binary correctness of each trial response (correct/incorrect) was recorded as the ground-truth label for the subsequent model training. Through this experimental procedure, a total of 1,041,868 timestep-level gaze recordings were collected across the cognitive task battery from 10 participants.

Eye-Tracking Data Preprocessing and Sequence Construction

Six gaze-related features were extracted for each timestep: the x- and y-coordinates of the left eye gaze point, x- and y-coordinates of the right eye gaze point, and pupil diameter of the left and right eyes, yielding a six-dimensional feature vector per sample. Missing values were imputed via linear interpolation along the time axis, with backward-f and forward-fill applied to handle the boundary segments.

The continuous gaze stream was then segmented into fixed-length sequences of $T = 100$ timesteps, resulting in sequences of shape (100, 6).

A single binary label was assigned to each sequence by majority voting over the per-timestep correctness labels within that sequence.

The full dataset was partitioned into training, validation, and test sets using a two-stage stratified random split to preserve the class balance across all subsets. In the first stage, 15% of all sequences was held out as the test set. In the second stage, the remaining 85% was further divided into a training set (70% of the total) and validation set (15% of the total). Stratification was applied at each stage with respect to the binary correctness label, and a fixed random seed was used throughout to guarantee reproducibility.

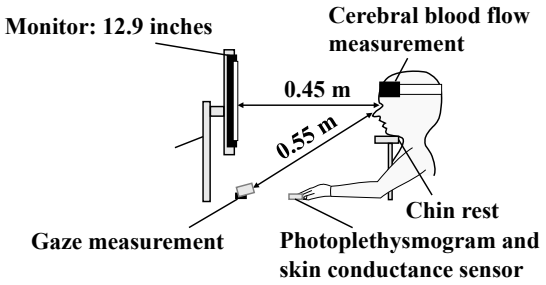

Figure 2: Experimental setup.

Correctness-Prediction Model

To extract cognitively meaningful latent representations, we trained a Transformer-based sequence model to predict the binary trial correctness from the preprocessed gaze sequences. The Transformer architecture was selected because of its capacity to model long-range temporal dependencies via multi-head self-attention, thereby capturing complex attentional and cognitive processing patterns that may not be adequately represented by recurrent architectures.

The model consisted of a positional encoding layer, four Transformer encoder blocks, each with two attention heads, feed-forward sublayers with a hidden dimension of 128, and a classification head. The penultimate layer produced a 32-dimensional latent representation vector for each trial through adaptive average pooling, which was subsequently passed to a fully connected layer for binary correctness prediction. The model was trained using the binary cross-entropy loss with the Adam optimizer, with dropout and early stopping applied for regularization.

The architectural hyperparameters were determined through Bayesian optimization over 200 trials, using the silhouette score of the resulting latent representations as the optimization objective, as opposed to using classification accuracy alone. This choice prioritizes the geometric interpretability of the latent space, ensuring that the selected architecture produces representations that are the most amenable to the downstream cognitive estimation stage.

Latent Dimension Selection via Logistic Regression

Following model training, the 32-dimensional latent representation vector was extracted for each trial across all participants. To identify the dimensions that most directly encoded cognitive processing ability, 32 independent univariate logistic regression models were constructed, each using a single latent dimension $z_i (i = 1, \ldots, 32)$ as the sole predictor of binary trial correctness. Each model was evaluated based on its classification accuracy on the held-out validation set, and the three dimensions yielding the highest accuracy (z_{0028}, z_{0022}, and z_{0031}) were selected as the cognitively informative subspace.

Cognitive Ability Estimation via 3D Latent Space

Individual cognitive processing ability was estimated from the centroid positions of correct- and incorrect-trial embeddings within the 3D subspace spanned by (z_{0028}, z_{0022}, z_{0031}). The two centroids for each participant p were computed as follows:

$$\grave{\mathrm{z}}_{\text{correct}}^{(p)} = \frac{1}{N_{\text{correct}}^{(p)}} \sum_{n \in \mathcal{C}^{(p)}} \left(z_{0028}^{(n)}, z_{0022}^{(n)}, z_{0031}^{(n)} \right)$$

$$\grave{\mathrm{z}}_{\text{incorrect}}^{(p)} = \frac{1}{N_{\text{incorrect}}^{(p)}} \sum_{n \in \mathcal{I}^{(p)}} \left(z_{0028}^{(n)}, z_{0022}^{(n)}, z_{0031}^{(n)} \right).$$

The spatial relationship between $\grave{\mathrm{z}}_{\text{correct}}^{(p)}$ and $\grave{\mathrm{z}}_{\text{incorrect}}^{(p)}$ — including the inter-centroid distance and directional displacement — was used as a continuous index of the individual's cognitive processing profile, requiring no external standardized cognitive assessment.

Supplementary Experiment: Detecting Training-Induced Cognitive Changes

A supplementary experiment was conducted with one participant to evaluate whether the proposed centroid-based representation can capture longitudinal changes in cognitive processing ability. This individual underwent a structured task-specific training intervention. Eye-tracking data were collected both before and after the training session under identical experimental conditions, and the corresponding latent representations were extracted using the pre-trained Transformer model without any fine-tuning.

For each session (pre- and post-training), the session-level centroid $\grave{\mathrm{z}}_{\text{session}}$, which was computed across all trials regardless of correctness, was projected onto the 3D subspace (z_{0028}, z_{0022}, z_{0031}). The directional displacement of $\grave{\mathrm{z}}_{\text{session}}$ between pre- and post-training was then evaluated relative to the positions of the population-level correct-trial centroid $\grave{\mathrm{z}}_{\text{correct}}$ and incorrect-trial centroid $\grave{\mathrm{z}}_{\text{incorrect}}$ to determine whether the post-training gaze representations shifted toward the region of the latent space associated with correct responses.

RESULTS

Correctness-Prediction Performance

The Transformer-based model achieved a validation accuracy of 0.725 on the binary correctness classification task, demonstrating that temporal gaze patterns during problem-solving carry meaningful predictive signals regarding a respondent's cognitive processing state.

Latent Space Geometry: Bayesian Optimization and Silhouette Score

Bayesian optimization over 200 trials identified four encoder layers and two attention heads as the optimal architecture, yielding a silhouette score of 0.13. Although this score indicated a weak geometric separation between the correct- and incorrect-trial clusters, it represented the highest value observed across all 200 trials, confirming that this configuration produced the most cognitively structured latent geometry achievable under the present experimental conditions.

Latent Dimension Selection via Logistic Regression

Table 1 presents the validation accuracies of the three selected dimensions of the 32 univariate logistic regression models. The selected dimensions correspond to the latent units z_{0028}, z_{0022}, and z_{0031}.

Table 1: Accuracy of the three selected dimensions.

Dimension	Validation Accuracy
z_{0028}	0.7124
z_{0022}	0.7030
z_{0031}	0.7002

All three dimensions approached the accuracy of the full Transformer model (0.725), and the narrow spread (< 0.02) suggests that they encoded complementary rather than redundant aspects of cognitive processing.

Centroid Positions in the 3D Latent Subspace

Table 2 reports the centroid coordinates for each group within the 3D subspace (z_{0028}, z_{0022}, z_{0031}).

Figure 3 shows the two-dimensional (2D) projections of all trial embeddings and centroids onto the XY (z_{0028}–z_{0022}), XZ (z_{0028}–z_{0031}), and YZ (z_{0022}–z_{0031}) planes. The incorrect-trial centroid (color 0) was positioned in the positive z_{0028} region, whereas the correct-trial centroid (color 1) was displaced toward the negative z_{0028} and more negative z_{0022}. The XY projection revealed the clearest directional axis separating the two groups, whereas the XZ and YZ projections showed more compressed distributions, indicating that z_{0031} contributed less to the primary axis of cognitive differentiation.

Table 2: Centroid coordinates for each group within the 3D subspace.

Group	z_{0028}	z_{0022}	z_{0031}
Incorrect trials	1.477	−2.053	0.941
Correct trials	−1.036	−3.904	0.660
Pre-training session	0.067	−2.817	0.676
Post-training session	−5.449	−8.610	−0.018

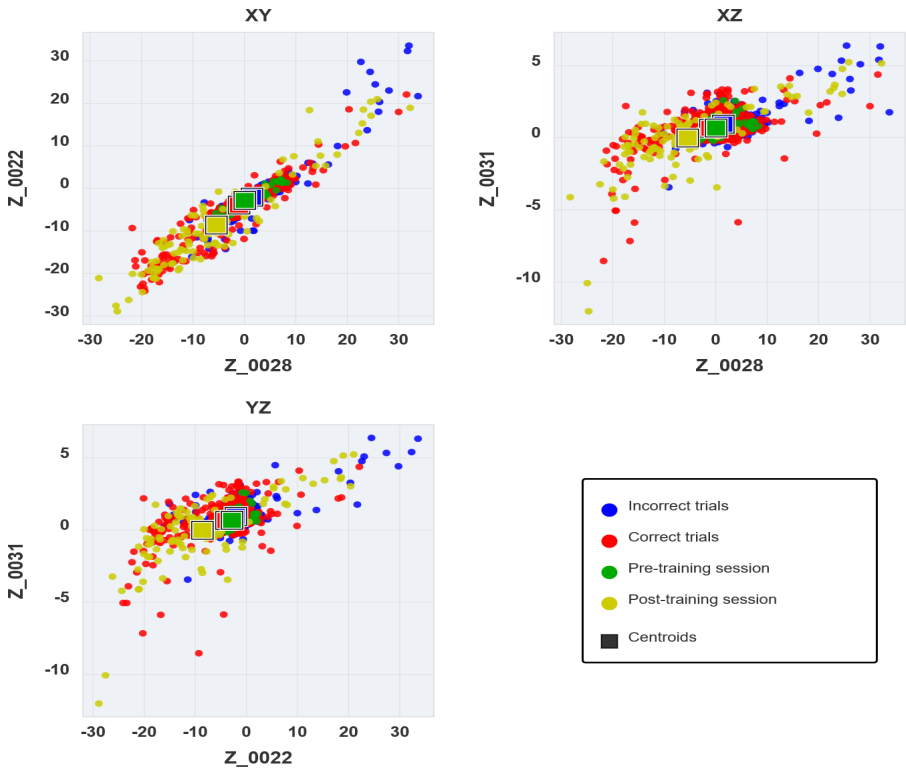

Figure 3: 2D projections of all trial embeddings.

Training-Induced Shift in Latent Centroid Position

The pre-training session centroid ($z_{0028} = 0.067$, $z_{0022} = -2.817$, $z_{0031} = 0.676$) was positioned in proximity to the incorrect-trial centroid ($z_{0028} = 1.477$, $z_{0022} = -2.053$, $z_{0031} = 0.941$), suggesting that the pre-training gaze dynamics were more closely associated with unsuccessful cognitive processing patterns. Following the training intervention, the session centroid shifted substantially to ($z_{0028} = -5.449$, $z_{0022} = -8.610$, $z_{0031} = -0.018$), moving well past the correct-trial centroid ($z_{0028} = -1.036$, $z_{0022} = -3.904$, $z_{0031} = 0.660$)

along the primary z_{0028}–z_{0022} axis. This directional displacement, which was consistently oriented toward and beyond the correct-trial region, indicated that the training intervention induced a marked reorganization of gaze-based latent representations in the direction associated with cognitive success.

DISCUSSION

Predictive Validity of Gaze-Based Latent Representations

The validation accuracy of 0.725 of the Transformer model confirms that temporal gaze sequences encode information that is predictive of trial-level cognitive outcomes. The three individual latent dimensions (z_{0028}, z_{0022}, z_{0031}) each achieved accuracies of 0.700–0.712, closely approaching the full-model performance, which demonstrates that a compact 3D subspace retains the majority of the cognitively relevant variance distributed across the 32-dimensional representation. This dimensional compactness is practically significant, because it renders the cognitive index geometrically interpretable and computationally lightweight.

Centroid Geometry as an Individual Cognitive Profile

At the individual level, the pre-training session centroid was positioned near the incorrect-trial region ($z_{0028} = 0.067$, close to the incorrect centroid at $z_{0028} = 1.477$), whereas the correct trial centroid was positioned in the opposite direction ($z_{0028} = -1.036$). This spatial relationship, in which proximity of the session centroid to either the correct or incorrect cluster reflects the prevailing cognitive processing tendency of the participant, provides an interpretable subject-level cognitive profile without requiring any external assessment.

Training Effect Visualization: Evidence of Cognitive Change Detection

The most striking finding of this study was the large post-training displacement of the session centroid toward the correct-trial region of the latent space. The pre-training centroid was proximal to the incorrect-trial centroid, whereas the post-training centroid ($z_{0028} = -5.449$, $z_{0022} = -8.610$) shifted substantially past the correct-trial centroid along the primary cognitive axis, representing a considerably greater displacement than the separation between the correct- and incorrect-trial centroids. This overshoot may reflect an accelerated reorganization of gaze dynamics following intensive training, although this interpretation must be treated with caution given the single-participant design. Importantly, this finding suggests that the proposed method is sensitive to not only stable individual differences but also dynamic training-induced improvements. If replicated at scale, this would establish gaze-based latent centroid tracking as a viable tool for longitudinal monitoring of cognitive changes in educational and clinical contexts.

Limitations and Future Work

This study has several important limitations. First, the training experiment involved only a single participant, rendering the observed centroid shift preliminary; controlled pre–post designs with larger samples are an immediate priority. Second, the silhouette score of 0.13 indicates weak latent space geometry, limiting the centroid-based index reliability under high within-class variability; contrastive loss functions or metric learning may improve the separability. Third, all four task types were pooled; the task-specific latent structures warrant further investigation. Finally, the construct validity against established measures, such as working memory capacity and fluid intelligence, remains to be established.

CONCLUSION

This study has proposed a novel method for estimating cognitive processing abilities from eye-tracking data by repurposing the 32-dimensional latent representations of a Transformer-based correctness-prediction model. The model achieved a validation accuracy of 0.725 and the three individual latent dimensions retained almost equivalent predictive power, enabling the derivation of a continuous subject-level cognitive profile from a compact 3D subspace without external psychological assessment. A supplementary longitudinal experiment further demonstrated that the session-level centroid shifted substantially toward the correct-trial region following a training intervention, suggesting that the proposed representation is sensitive to training-induced cognitive changes.

Key limitations include the single-participant training experiment, weak latent space cluster separation (silhouette score = 0.13), and absence of external cognitive reference standards for construct validation. Future work will address these issues through larger pre–post designs, contrastive learning objectives, task-specific model variants, and validation against established neuropsychological assessments. Despite these constraints, the findings provide initial empirical support for gaze-based latent centroid tracking as a non-invasive biomarker for both static cognitive profiling and longitudinal cognitive change detection, opening a promising direction toward lightweight, continuous cognitive assessment grounded in deep temporal gaze representations.

REFERENCES

Arnold, L., Aryal, S., Hong, B., Nitharsan, M., Shah, A., Ahmed, W., Lilani, Z., and Su, W. (2025). A systematic literature review of eye-tracking and machine learning methods for improving productivity and reading abilities. Applied Sciences, 15(6), 3308. https://doi.org/10.3390/app15063308.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

Copeland, L., Gedeon, T., and Caldwell, S. (2015). Effects of text difficulty and readers on predicting reading comprehension from eye movements. In Proceedings of the IEEE 6th International Conference on Cognitive Infocommunications (CogInfoCom) pp. 481–486. IEEE. https://doi.org/10.1109/CogInfoCom.2015.7390628.

Gorin, H., Patel, J., Qiu, Q., Merians, A., Adamovich, S., and Fluet, G. (2024). A review of the use of gaze and pupil metrics to assess mental workload in gamified and simulated sensorimotor tasks. Sensors, 24(6), 1759. https://doi.org/10.3390/s24061759

Hosp, B., Schultz, F., Kasneci, E., and Höner, O. (2021). Expertise classification of soccer goalkeepers in highly dynamic decision tasks: A deep learning approach for temporal and spatial feature recognition of fixation image patch sequences. Frontiers in Sports and Active Living, 3, 692526. https://doi.org/10.3389/fspor.2021.692526.

Institute of Medicine. (2015). Cognitive Tests and Performance Validity Tests. In Psychological Testing in the Service of Disability Determination. National Academies Press. https://www.ncbi.nlm.nih.gov/books/NBK305230/.

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

Shubi, O., Meiri, Y., Hadar, C. A., and Berzak, Y. (2024). Fine-grained prediction of reading comprehension from eye movements. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 3422–3441. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.198.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In Advances in Neural Information Processing Systems (Vol. 27). Curran Associates. https://arxiv.org/abs/1411.1792.