Beyond Accuracy: Leveraging Eye Tracking and Machine Learning to Enhance Diagnostic Precision and Adaptivity in Computer-Based Testing

Traditional Computerized Adaptive Testing (CAT) methods—typically grounded in Item Response Theory (IRT)—focus solely on response accuracy and item difficulty to estimate learner ability. While efficient, such systems overlook rich behavioral data that can capture engagement, cognitive effort, and test-taking strategy.

Recent advances in eye tracking and machine learning (ML) enable the integration of real-time behavioral signals (e.g., fixations, saccades, pupil dilation) into test delivery systems. These signals can provide insight into how a person engages with an item—not just whether they answer correctly. This project investigates whether such features can:

* improve diagnostic resolution of learner states,
* inform adaptive item selection, and
* mitigate guessing or disengagement.

**2. Objectives**

This study aims to:

Identify eye movement patterns that distinguish high- vs. low-performing participants.

Use behavioral features to personalize item delivery in an ML-driven adaptive testing environment.

Detect and counteract guessing behavior using gaze-based indicators.

[Optional] Develop a framework to classify cognitive strategies (e.g., piecemeal, heuristic, and potential guessing or disengagement) based on these metrics.

[Optional] Apply machine learning models to predict strategy type and performance from gaze behavior.

**3. Research Questions**

Q1: What eye movement patterns best distinguish high- vs. low-performing participants across item types? Compare eye-tracking metrics (e.g., fixation duration, scanpath entropy, blink rate) between groups.

Q2: How can machine learning models personalize item delivery based on momentary attentional states? Train supervised models to predict attentional engagement using gaze patterns. Use model outputs to dynamically adapt item sequencing or difficulty.

Q3: Is adaptive testing based on both behavioral (gaze) and performance data more robust to guessing or gaming strategies? Identify markers of guessing (e.g., rapid response + shallow gaze). Simulate or detect strategic behaviors and test robustness of behavioral-based adaptation.

**4. Methodology**

***Participants***

80 young children participated in lab-based spatial reasoning tasks (e.g., mental rotation).

***Materials***

Eye-tracker (e.g., Tobii Pro or EyeLink)

***Analysis***

Feature engineering: Collect and extract metrics as below:

| **Field/Variable** | **Description** | **Category** | **Included** |
| --- | --- | --- | --- |
| **Fixation** |  |  |  |
| AVERAGE\_FIXATION\_DURATION | Avg. duration (ms) of all fixations in the trial (selected interest period) | Fixation | Yes |
| FIXATION\_COUNT | Total number of fixations in the trial (selected interest period) | Fixation | Yes |
| FIXATION\_DURATION\_MAX | Longest fixation duration (ms) in the trial | Fixation | Yes |
| FIXATION\_DURATION\_MIN | Shortest fixation duration (ms) in the trial | Fixation | Maybe |
| FIXATION\_DURATION\_MAX\_TIME | Start time (EDF file time) of the longest fixation in the trial | Fixation | No |
| FIXATION\_DURATION\_MIN\_TIME | Start time (EDF file time) of the shortest fixation in the trial | Fixation | No |
| MEDIAN\_FIXATION\_DURATION | Median fixation duration (ms) in the trial (selected interest period) | Fixation | Yes |
| SD\_FIXATION\_DURATION | SD of fixation durations in the trial (selected interest period) | Fixation | Test |
| - |  |  |  |
| **Saccade** |  |  |  |
| AVERAGE\_SACCADE\_AMPLITUDE | Avg. amplitude (deg of visual angle) of all saccades in the trial (selected interest period) | Saccade | Yes |
| MEDIAN\_SACCADE\_AMPLITUDE | Median saccade amplitude (deg of visual angle) in the trial (selected interest period) | Saccade | Yes |
| SACCADE\_COUNT | Total number of saccades in the trial | Saccade | Yes |
| SD\_SACCADE\_AMPLITUDE | SD of saccade amplitudes in the trial (selected interest period) | Saccade | Test |
| - |  |  |  |
| **Pupil** |  |  |  |
| PUPIL\_SIZE\_MEAN | Average pupil size in the trial (selected interest period) | Pupil | Yes |
| PUPIL\_SIZE\_MAX | Largest pupil size in the trial (selected interest period) | Pupil | Test |
| PUPIL\_SIZE\_MIN | Smallest pupil size in the trial (selected interest period) | Pupil | Test |
| PUPIL\_SIZE\_MAX\_TIME | EDF file time of sample with maximum pupil size | Pupil | No |
| PUPIL\_SIZE\_MIN\_TIME | EDF file time of sample with minimum pupil size | Pupil | No |
| PUPIL\_SIZE\_MAX\_X, PUPIL\_SIZE\_MAX\_Y | Gaze X/Y at maximum pupil size sample | Pupil | No |
| PUPIL\_SIZE\_MIN\_X, PUPIL\_SIZE\_MIN\_Y | Gaze X/Y at minimum pupil size sample | Pupil | No |
| - |  |  |  |
| **Blink** |  |  |  |
| AVERAGE\_BLINK\_DURATION | Avg. duration (ms) of all blinks in the trial (selected interest period) | Blink | Yes |
| BLINK\_COUNT | Total number of blinks in the trial (selected interest period) | Blink | Yes |

Strategy labeling: We can consider two methods:

1. Use item-level accuracy and person-level latent ability (IRT)
2. Use rule-based heuristics (and optionally clustering) to assign tentative strategy labels.

Machine learning:

Unsupervised: K-means, HMMs, hierarchical clustering

Supervised: Random Forest, SVM, logistic regression, Neural

Interpretation: Identify features most predictive of strategy type and performance.

**5. Expected Contributions**

A validated set of gaze-based markers for performance and engagement.

A prototype behavior-aware adaptive testing system.

Evidence that eye tracking improves robustness and personalization of learning assessments.

**6. Theoretical Framework**

This study is grounded in:

Lohman’s theory of cognitive strategies in spatial reasoning

Hegarty’s empirical findings on eye movement patterns and strategy use

Just & Carpenter’s eye–mind hypothesis linking visual attention and cognitive processing

**7. Implications**

Advances adaptive learning technology for personalized education and formative assessment.

Informs design of robust tests that adapt to both knowledge and cognitive state.

Bridges psychometrics and affective computing.