2026 ASEE ANNUAL CONFERENCE June 21–24, 2026 | Charlotte Convention Center, NC
Host Sponsor UNC Charlotte
Paper ID #52297

Beyond Examinations: A Multi-modal Approach to Measure Learning Through Physiological Biomarkers

Dr. Brainerd Prince, Plaksha University

Brainerd Prince is the Associate Professor and the Director of the Centre for Thinking, Language and Communication at Plaksha University. He teaches courses such as Reimagining Technology and Society, Ethics of Technological Innovation, and Art of Thinking for undergraduate engineering students and Research Design for PhD scholars. He completed his PhD on Sri Aurobindo's Integral Philosophy from OCMS, Oxford – Middlesex University, London. He was formerly a Research Tutor at OCMS, Oxford, and formerly a Research Fellow at the Oxford Centre for Hindu Studies, a Recognized Independent Centre of Oxford University. He is also the Founding Director of Samvada International Research Institute which offers consultancy services to institutions of research and higher education around the world on designing research tracks, research teaching and research projects. His first book The Integral Philosophy of Aurobindo: Hermeneutics and the Study of Religion was published by Routledge, Oxon in 2017. For more information, please visit: https://plaksha.edu.in/faculty-details/dr-brainerd-prince

Dr. SIDDHARTH SIDDHARTH, Plaksha University

Dr. Siddharth leads the Human-Technology Interaction Lab at Plaksha University. He recently came to Plaksha from the Human Factors Center of Excellence at Microsoft USA where he worked as Senior Human Factors Researcher. He is passionate about working on research themes with practical applications in the technology industry.

At Plaksha, Dr. Siddharth is investigating and modeling human physiological responses to improve cognitive well-being, and enhance human safety, productivity, and creativity. By integrating domain-specific knowledge from the fields of signal processing, computational neuroscience, computer vision, and machine learning, his research group is utilizing human physiological responses to track various cognitive states such as emotions, stress, awareness, mental load, and more. His research and recommendations have practical applications in the technology industry for the development of safer and more user-centered designs that promote cognitive well-being.

In the past, Dr. Siddharth led the Microsoft Human Factors Center of Excellence's efforts to design personalized user experiences through neurophysiological insights into users' preferences enabling them to be more productive and establish meaningful connections with each other. He has also worked at Facebook Reality Labs, Samsung Research America, French National Center for Scientific Research (CNRS), and National University of Singapore. His research has been awarded fellowships and research grants from various institutions including the National Science Foundation, Army Research Lab, and the Kavli Institute of Brain and Mind. It has also been published in many international conferences and reputed journals including IEEE Transactions and Nature Scientific Reports and has spawned multiple patents.

Dr. Siddharth received his Ph.D. from the University of California San Diego where he also served as the Graduate Fellow Editor of the Prospect Journal of International Affairs and Graduate Student Member of the South Asia Initiative. His first book, Founding Generations: Democracy's Origins and Parallels in America and India, was released recently. For more up-to-date information about him, please visit https://ssiddharth.in/about-me

Mr. Utkarsh Agarwal, Plaksha University

Utkarsh Agarwal is an undergraduate engineering student pursuing a B.Tech in Computer Science and Artificial Intelligence at Plaksha University. His academic interests lie at the intersection of multimodal systems, human cognition, and applied machine learning, with a particular focus on leveraging sensor data and physiological signals to better understand learning processes. In this work, we explore a multimodal approach to assessing examinations by integrating behavioral and physiological markers, aiming to move beyond traditional evaluation methods toward more holistic and data-driven measures of learning.

Mr. Yash Sangtani, Plaksha University

Yash Sangtani is a final-year undergraduate student in Computer Science and Artificial Intelligence at Plaksha University, building data-driven solutions at the intersection of machine learning, systems thinking, and real-world impact. His work spans reinforcement learning, predictive modeling, and LLM-powered pipelines — from training autonomous racing agents on AWS DeepRacer, to developing ML models that predict stubble-burning incidents and agricultural yield from satellite imagery. He has also applied ML to sports analytics, modeling expected possession value in basketball to extract actionable insights from in-game dynamics.

Beyond Examinations: A Multi-modal Approach to Measure Learning Through Physiological Biomarkers

Abstract: Research has shown that traditional student assessment methods, ranging from subjective essays to standardized tests, face notable limitations as they can induce stress and may be susceptible to plagiarism, particularly with the rise of generative AI tools. Thus, there is a growing need for real-time measures of learning that not only reduce student stress but also efficiently detect genuine cognitive engagement as markers of learning. By leveraging the recent advances in bio-sensing, signal processing, and machine learning, we aim to uncover physiological signatures that may correlate with learning. This fusion of sensor modalities represents a departure from reliance on verbal or written student outputs, offering a potential window into learning processes in real time. In this pilot study (n=6), eye-tracking features often showed larger effect-size correlations with learning gains than facial-expression or ECG features; however, no single biomarker remained significant after correcting for multiple comparisons. The most robust result was a cross-modal interaction in which focused medium-duration, low-dispersion fixations interacted with heart-rate variability (RMSSD) to predict learning gain (FDR $q < 0.05$). These results are exploratory and motivate larger studies to validate single-feature and cross-modal biomarkers. This exploratory work examines assumptions about education assessments, shifting the focus from assessing learning using traditional methods that measure the current state of knowledge through oral or written submissions to capturing the change in knowledge (delta) in the student during the course intervention. These preliminary findings suggest that real-time physiological biomarkers may open up new avenues for personalized education, though further research with larger samples is needed to validate these initial observations.

Keywords:

Computer Vision, Cognitive learning, Bio-sensing, Affective states of learners, Human-centric AI

1. Introduction

Educational assessment is fundamental for measuring student learning, but traditional items like standardized Multiple Choice Question (MCQs) tests and essay-based evaluations are increasingly having limitations in their scope. Traditional exams, while once considered objective, remain susceptible to evaluator bias. The Global shift towards digital and hybrid learning contexts has magnified integrity issues: students can easily access online resources, leverage advanced generative AI tools and circumvent traditional plagiarism-detection methods.

Furthermore, our research identifies a significant limitation in traditional student assessments. They typically measure only the current state of knowledge of the student at the time of assessment rather than the change in the state of knowledge in the student caused by learning. To clarify this distinction: the "current state of knowledge" refers to what a student knows at a single point in time, typically captured through end-of-unit exams or standardized tests administered after instruction has concluded. In contrast, the "change in state of knowledge" refers to the actual learning gain—the difference between what a student knew before an instructional intervention and what they know afterwards. Traditional assessments capture only the former; a student may perform well on an exam due to prior knowledge, natural aptitude, or external resources, without the assessment revealing whether genuine learning occurred during the instructional period.

Conventional assessment tools reflect only a snapshot of existing knowledge without truly capturing the progression or improvement over time.

This limitation has significant implications for educational practice. Without measuring the change in knowledge, educators cannot distinguish between students who learned substantially during instruction and those who simply possessed prior knowledge. Moreover, traditional post-hoc assessments cannot identify moments of confusion or engagement during the learning process itself, missing opportunities for timely intervention. Our approach addresses this gap by measuring physiological indicators during the learning event, enabling the quantification of knowledge change (delta) rather than merely assessing endpoint knowledge.

We propose a novel methodology utilizing physiological and behavioral biomarkers to quantify authentic changes in knowledge at the time of learning. Preliminary findings of our research indicate that facial expressions and heart rate data are weak indicators of learning, while eye-tracking metrics show promising effect-size associations in this pilot; however, statistical significance is limited after multiple-comparison correction. Interaction effects between modalities (e.g., gaze × HRV) appear especially promising.

2. Literature Survey

2.1 An Evolution of understanding "learning" in Engineering Education

Historically, engineering learning has been rooted in a "shop culture" characterized by practical skill-development and hands-on apprenticeship. Following World War II and the establishment of the National School Foundation, there was a decisive shift from "shop culture" to "school culture". Learning was redefined in terms of mathematical and scientific skills over practical knowledge. [1] [2] By the late 20th century, driven largely by accreditation changes like ABET EC2000, the understanding of learning shifted from what is taught to what students can do. The shift in understanding of learning necessitated an evolution in tools that would help educators assess this learning.

2.2. Evolution of learning assessment tools

Engineering assessment has been focused on "input method", i.e., what material was taught rather than what students learnt. Traditional assessment models, therefore, rely on narrow definitions centered around either recall of content knowledge or recipe-style application of skills. [3] The traditional approach to designing assessments begins with clear learning objectives, which are later aligned with Bloom's taxonomy or the revised Bloom's taxonomy (2001) to create the test that needs to be written. This approach has led to a limiting definition of what is to be assessed. [4] Historically, standardized tests emerged to provide systematic and scalable metrics of performance [5], but scholars have long noted their drawbacks, including stress-induced skewing of results and questionable validity in truly capturing higher-order thinking. Essay-based evaluations, though more open-ended and reflective, are subject to rater-bias [6]. Furthermore, the challenge of evaluating open-ended and resource-intensive tasks has driven engineering educators to integrate technologies into assessment. Classroom Response systems, Computerized Adaptive Testing, and Computer-Based Tasks are a few examples of advancements in this area. [7] [8]

2.3. Evolution of physiological markers as measures in assessment

Recent surges in plagiarism and the easy availability of essay-writing tools, particularly LLMs, have eroded trust in traditional assessments, exposing new vulnerabilities in proctoring and test design. [9] Against this backdrop, educational neuroscience and affective computing researchers have begun proposing physiological signals like pupil dilation, heart rate variability, eye gaze patterns, and facial expression as robust biomarkers of cognitive load and engagement. [10] [11] [12] These approaches mark a departure from conventional assessments, capturing the learning process as it unfolds rather than relying on end-point artefacts. As the field of neuroscience emerged, assessment tools evolved to visualize brain activity directly. Techniques such as functional Magnetic Resonance Imaging (fMRI), Electroencephalography (EEG), and Magnetoencephalography (MEG) have allowed researchers to measure the time course of brain activity with high temporal accuracy. These methods are valuable for measuring "implicit behaviors", including visual attention, concentration, and metacognitive learning strategies. [13]

Building on these foundational studies, our work integrates multiple sensing modalities of ECG, pupillometry, and computer vision-based facial expression tracking and comprehensively applies them to measure student learning. Unlike prior research that often focuses on a single modality or a narrowly defined context, this multimodal framework offers a potentially broader and more nuanced picture of cognitive and emotional states, promising new directions in both educational evaluation and the design of more responsive, adaptive learning ecosystems.

3. Methodology

3.1. Research Question

The primary question addressed by this research is to what extent there is a quantifiable link between student physiology and knowledge acquisition by correlating physiological and behavioral indicators. Specifically, we operationalize "knowledge acquisition" as the change in test performance (delta) between pre-video baseline assessments and post-video assessments, directly addressing the gap identified in traditional assessments that measure only endpoint knowledge. By simultaneously recording physiological signals during the learning event (video viewing) and measuring the resulting knowledge change, we aim to identify biomarkers that correlate with genuine learning gains rather than static knowledge states.

3.2. Experimental Design

For this study, we recruited six freshmen undergraduate engineering students of Plaksha University. We selected participants based on their academic stage and existing curricular exposure to ensure consistency and relevance of content presented during the experiment. The experiment was assessed and approved by the ethics committee of the institute, and safeguards were implemented for data anonymization and storage.

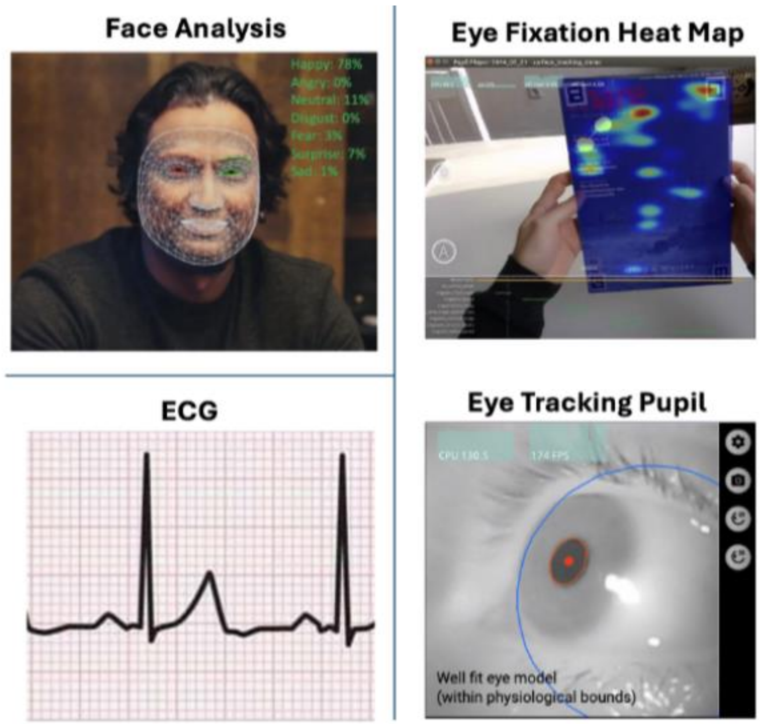


Figure 1: Multi-modal bio-sensing modalities used in this study and a snapshot of data processing for each modality.

Facial expressions are categorized into seven emotion classes by tracking changes in expressions using a deep learning neural network, eye gaze tracking provides fixation heatmap and pupil dilation, and ECG data provides heart-rate and heart-rate variability.

Theme	**Videos**	**Difficulty Level**
Genetic and Molecular Biology	1. Genes, DNA and Chromosomes	Easy
Machine Learning	2. Supervised vs. Unsupervised Learning 3. K-Means clustering	Easy Easy
Deep Learning	4. U-Net and Image Segmentation	Difficult
Fluid Dynamics	5. Velocity Potential and Stream Function	Difficult
Quantum Computing	6. Partial trace in Quantum Mechanics	Difficult

Table 1: Mapping of Themes to Videos and Difficulty Levels

We carefully designed the experiment by curating a set of six videos with two distinct levels of difficulty—three videos that would be easily accessible by the chosen demographic and three intentionally chosen to represent advanced topics beyond the participants' current academic knowledge. The first category aimed at building upon foundational concepts the students had previously encountered during high school or introductory coursework. For instance, one video, lasting approximately 2 minutes and 15 seconds, covered the topic "Gene vs. DNA vs.

Chromosome", aiming to clarify and extend basic biological knowledge the students previously held. The second video, approximately 3 minutes in length, introduced concepts in Machine Learning, specifically the differences between supervised and unsupervised learning, clearly distinguishing the nature of labeled and unlabeled datasets. The third video of this category, also approximately 3 minutes long, explained the K-Means clustering technique, relying on visualizations to illustrate clustering through centroid calculations. These videos were specifically selected to ensure that meaningful learning occurred, given the participants' existing understanding of biology, statistics, and data science.

The second set of three videos was strategically selected to cover complex topics anticipated to exceed the students' current understanding, providing a contrasting scenario to evaluate the experimental approach effectively. One video, around 3 minutes and 20 seconds in duration, explained the UNET architecture in deep learning for image segmentation, involving advanced concepts such as backpropagation, convolutional neural networks, and skip connections—concepts unlikely to be familiar to first-year students without advanced coursework. Another challenging video, lasting approximately 4 minutes and 20 seconds, delved deeply into mathematical principles in fluid dynamics, such as velocity potential, flow fields, and stream functions. These advanced mathematical constructs require foundational knowledge beyond typical first-year engineering curricula. Lastly, an approximately 5-minute video introduced complex topics from quantum computing, specifically the concept of partial traces, which builds upon sophisticated foundational concepts like quantum entanglement, density operators, and tensor products.

Prior to commencing the experiment, participants were equipped with a head-mounted eye-tracking device and an ECG monitoring chest strap. A webcam integrated in the laptop remained active throughout the experiment, capturing participants' facial expressions continuously as shown in Fig 2.

Participants interacted with a custom-built experimental software platform. Initially, they reviewed and provided voluntary consent through an electronic consent form. Following consent, participants completed a brief questionnaire measuring their prior knowledge and familiarity with topics in computational and data sciences, biological sciences, and theoretical/applied mathematics using a 5-point Likert scale. This preliminary step aimed to identify any potential anomalies arising from prior exposure beyond the standard curriculum. Each experimental session included one video, preceded by five baseline questions and followed by seven post-video questions: five MCQs for content comprehension and two Likert-scale self-assessments on cognitive load and understanding improvement. Participants had 1 minute 15 seconds for baseline questions and 1 minute 30 seconds for post-video questions. To reduce artifacts, instructions emphasized minimal head movement, reduced distractions, and focused screen attention. A 30-second rest interval separated sessions to capture non-learning periods. The order of six videos was randomized per participant to ensure data reflected video content difficulty rather than sequence effects.

3.3. Assessment Design and Validation

To ensure meaningful measurement of learning gains, baseline and post-video questions were carefully designed to be matched in both difficulty and content coverage. Each question pair targeted the same underlying concept from the video content, with post-video questions rephrased to assess genuine comprehension rather than surface-level memorization. For example, if a baseline question asked about the definition of a target variable, the corresponding post-video question assessed the same concept through application or identification in a different context. All MCQs were reviewed by subject matter experts to ensure appropriate difficulty calibration and content alignment between baseline and post-video assessments.

Regarding MCQ reliability, items were constructed following established principles of educational measurement, including clear stem construction, plausible distractors, and single correct answers. While formal psychometric validation (such as item discrimination indices) was not conducted due to the pilot nature of this study with a limited sample size, the questions were iteratively refined based on expert feedback to maximize clarity and discriminatory potential. Future studies with larger samples will enable comprehensive item analysis to further establish measurement reliability.

The 30-second rest interval designated as the "non-learning baseline" was selected based on prior research in psychophysiological measurement, which suggests that brief rest periods of 20-60 seconds are sufficient for physiological signals to return toward baseline levels following cognitive tasks. During this interval, participants were instructed to relax and look at a neutral screen without any learning content. This duration was chosen as a practical compromise that allowed adequate physiological stabilization while maintaining experimental efficiency and participant engagement across the six video sessions. By capturing physiological data during these rest periods, we established a within-participant reference point against which learning-phase biomarkers could be compared, thereby isolating changes specifically associated with active learning engagement.

3.4. Operationalizing Knowledge Change

Central to our methodology is the explicit measurement of knowledge change rather than endpoint knowledge alone. The delta (Δ) metric—calculated as the difference between post-video and baseline test scores—directly operationalizes the "change in state of knowledge" concept introduced earlier. This design choice addresses the limitation of traditional assessments by ensuring that our outcome measure reflects genuine learning gains attributable to the instructional intervention (video content), rather than prior knowledge or other confounding factors. By correlating physiological signals recorded during the learning event with this delta metric, we establish a direct link between real-time biomarkers and the actual learning process, rather than merely associating physiology with static performance levels.

3.5. Data Collection

All participants provided informed consent prior to the commencement of data collection. Six undergraduate volunteers (4 males and 2 females), aged between 17 and 19, enrolled in a first-year engineering program at Plaksha University, participated in this study.

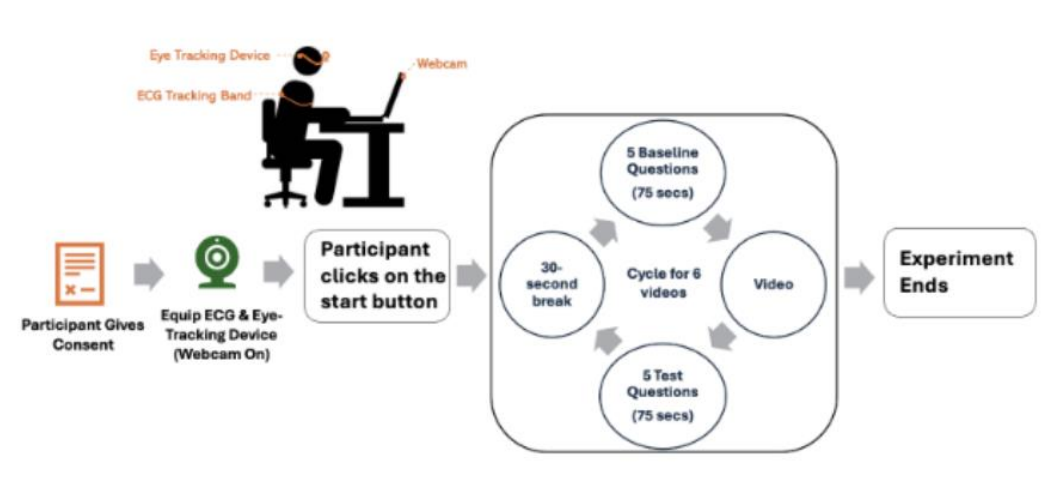


Figure 2: Experiment Protocol for each participant

Data was systematically collected throughout the experiment using a combination of questionnaires and sensor-based monitoring. Participants underwent two primary types of assessments: pre-video baseline questionnaires and post-video tests. The pre-video assessments included multiple-choice questions intended to establish baseline knowledge of participants regarding the video's content. For instance, prior to the video on supervised learning, participants answered a question, "Which term refers to the predicted outcome the model is trained to estimate?" with options including "Target variable", "Feature", "Cluster", and "Category", with "Target variable" as the correct response.

Following the videos, participants completed post-video tests comprising similar multiple-choice questions, slightly rephrased to evaluate true understanding rather than memorization.

Throughout each experimental session, physiological and behavioral data were simultaneously recorded using multiple monitoring techniques. Eye-tracking data were captured using a head-mounted eye-tracking device (Pupil Core, Pupil Labs GmbH, Germany). This device collected gaze patterns, pupil dilation, fixation duration, blink frequency, and provided a first-person scene video recording for contextual visual reference (Figure 1). Additionally, computer vision-based analysis was conducted through webcam recordings to track facial expressions. This enabled automatic identification of emotional states, providing additional insights into the cognitive and emotional responses of the participants. Table 3 shows the list of all the physiological responses that were collected using these sensors.

Electrocardiogram (ECG) data were also collected using a wearable ECG device (MAX-ECG-MONITOR). This provided detailed cardiac measurements, including heart rate, R-R intervals to analyse heart rate variability, and motion data to identify potential movement artifacts affecting signal quality. Additionally, temperature data provided supplementary physiological context.

Participants minimized movement and distractions while maintaining screen focus to improve data quality. They self-rated prior knowledge in computational/data sciences, biology, and mathematics on a five-point scale, helping researchers control for background differences. All sensor data, questionnaire responses, and interaction times were recorded and synchronized for comprehensive analysis of engagement, cognitive load and learning outcomes.

Measurement	Category	Definition
Blinks	Short Medium Long	< 200 ms 200 to 400 ms > 400 ms
Pupil Diameter	Small Medium Large	< 2 mm 2 to 4 mm > 4 mm
Fixations	Short Medium Long	< 130 ms 130 to 190 ms > 190 ms
Dispersion	Low High	< 0.75 degrees 0.75 to 1.5 degrees

Table 2: Eye tracking Measurement Definitions

3.6. Data Analysis

For every video, we quantified the prevalence of each Facial Expression and Eye Tracking response (e.g., anxiety, calm, etc.) by converting raw counts into percentages, calculated as:

(count of response indicator/ total category count) × 100

In contrast, for Heart Activity, we computed average values (e.g., mean heart rate or mean SDNN) instead of percentages. Because missing entries were primarily due to sensor dropouts with no valid samples ("no_samples"; 12/54 = 22.2% of eye-tracking/blink/pupil/fixation observations and 2/54 = 3.7% of facial-emotion observations), we kept them as NA and computed percentages/correlations using only valid samples rather than imputing zeros.

Sensing Modality	Category	Physiological Response	Response Indicator
Facial Expressions	Primary Emotion	Anxiety/Stress Calm/Disengaged Aversion/Discomfort Frustration/Disappoint. Unexpected Moments Positive Engagement Irritation/Resistance	Fear Neutral Disgust Sad Surprise Happy Anger
	Grouped Emotion	Positive Emotions Negative Emotions Emotional Diversity	Happy, Surprise Fear, Sad, Disgust, Anger Non-neutral
Eye Tracking	Blink Duration	Sustained Attention Typical Blinking Fatigue/Reduced Focus	Short blinks Medium blinks Long blinks
	Fixation Duration	Detailed Attention Rapid Scanning Steady Focus Moderate Exploration Deep Concentration Prolonged Search	Short fix., low disp. Short fix., high disp. Med. fix., low disp. Med. fix., high disp. Long fix., low disp. Long fix., high disp.
	Pupil Diameter	Low Effort/Arousal Typical Arousal High Cog. Load/Interest	Small pupil diam. Med. pupil diam. Large pupil diam.
Heart Activity	Heart Rate	Autonomic Arousal	Heart Rate
	Heart Rate Variability	Autonomic Flexibility Parasympathetic Activity Stress Monitoring	SDNN PNN50 RMSSD

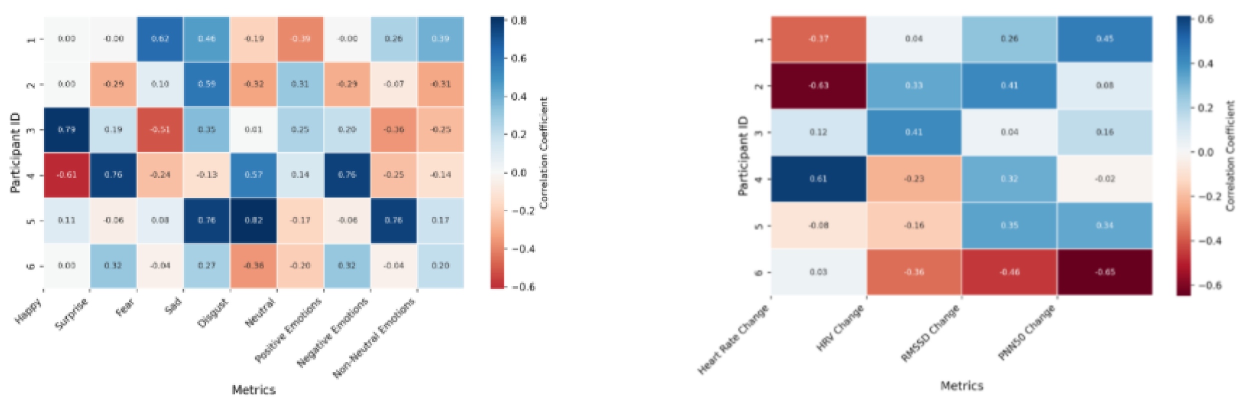
Table 3: Comprehensive Physiological Responses and Broad Representations

Table 3 was compiled and informed by the measurement categories and response indicators described in [14, 15, 16, 17, 18, 19, 20, 21, 22, 23]. As a concrete example, consider Participant 1 watching a video on Genes. This participant answered three baseline questions correctly before the video and five correctly afterward, yielding a Δ (delta) of 2 (i.e., two more correct answers than baseline), suggesting meaningful learning. During the video, the participant's emotional data indicated 638 frames classified as Fear, 166 frames as Neutral, and just 12 frames as other emotions. Similarly, other parameters, such as fixations (grouped by duration and dispersion) and pupil diameter (small, medium, large), were logged and converted into percentages. Heart activity metrics (e.g., heart rate) were collected and averaged over the same time window. Next, we carried out the same measurements for the 30-second intervals designated as non-learning phases, during which a delta of zero was assigned (indicating no learning gain). To isolate the change in biomarker prevalence for facial expressions and eye tracking, we subtracted the non-learning percentage from the learning percentage. For heart activity, we subtracted the non-learning average from the learning average. We refer to these differences as the "net biomarker" values, effectively removing baseline effects from our analysis.

To complement effect-size screening ($|r| \geq 0.4$), we computed nominal p-values for Pearson correlations and controlled for multiple comparisons using Benjamini–Hochberg FDR. Because each participant contributed repeated measurements across videos, pooled analyses were additionally evaluated with linear mixed-effects models using participant as a random intercept. We also explored pairwise biomarker interactions to test whether cross-modal combinations predict delta beyond single-feature effects, reporting both $q < 0.05$ (primary) and $q < 0.10$ (exploratory) thresholds.

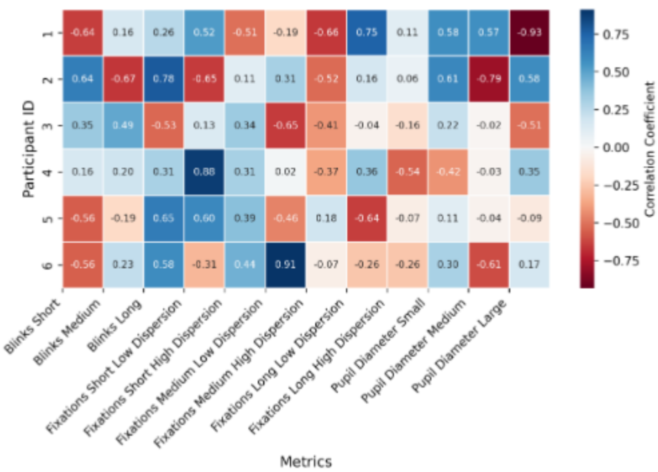
4. Results

To examine how the changes in various physiological responses relate to learning outcomes, we computed Pearson correlation coefficients between each net biomarker percentage and the participants' Δ across all videos. The resulting correlation values summarized in Figure 3 provide an at-a-glance view of how physiological responses shift from different sensing modalities like ECG, eye-tracking, and webcam correlate with performance gains.



(a) Facial Expression vs Delta

(b) ECG vs Delta



(c) Eye Tracking vs Delta

Figure 3: Correlation of Physiological Responses with Delta

4.1. Physiological Responses Correlates of Learning: Eye-Tracking as the Strongest Predictor of Knowledge Gains

In our study, we investigated three primary data sources — (1) facial expressions (captured via a webcam), (2) eye-tracking metrics, and (3) heart activity measurements—to understand how biometric signals correlate with learning outcomes. We used $|r| \geq 0.4$ as an effect-size screening threshold to highlight potentially meaningful associations. Statistical inference was evaluated separately using p-values and FDR correction; under FDR control, no single biomarker reached $q < 0.10$ in the pooled analysis.

The preliminary insight from our analysis (Figure 3) is that many physiological markers are correlated with learning across participants, even though not all may be correlated in all participants. This is because of the diversity in human physiology and the different learning curves of participants due to different baselines.

Facial Expressions: This category included ten features, such as happy, surprise, fear, sad, disgust, anger, neutral, positive, negative, and non-neutral (Figure 3a). Observing these variables across six participants revealed that only one or two participants exhibited a correlation above +0.4 (or below −0.4) for any single emotion metric. Consequently, while facial expressions provide some information, their relationship to learning gains appears relatively weak within our sample.

Heart Activity Measurements: Heart activity features showed fewer large single-feature correlations; however, interaction analysis suggested that HRV (RMSSD) may moderate the relationship between focused gaze and learning gain.

Eye-Tracking Metrics: This data encompassed blinks (short, medium, and long), fixations (low and high dispersion across short, medium, and long durations), and pupil diameter (small, medium, and large) (Figure 3c). Among these 12 distinct eye-tracking features multiplied by 6 participants (72 total "feature-participant" pairs), nearly half of the correlations exceeded +0.4 or fell below −0.4. Across participants, eye-tracking features produced a higher proportion of large-magnitude correlations ($|r| \geq 0.4$) than facial-expression or ECG features, suggesting that gaze measures may be *promising candidates* for capturing learning-related changes in this pilot. However, these counts reflect effect-size screening only; after correcting for multiple comparisons, no *single* biomarker remained significant at $q < 0.10$ in pooled analyses. Notably, exploratory interaction modeling suggests that eye-tracking may be most informative when combined with cardiac measures (e.g., a fixation × RMSSD interaction).

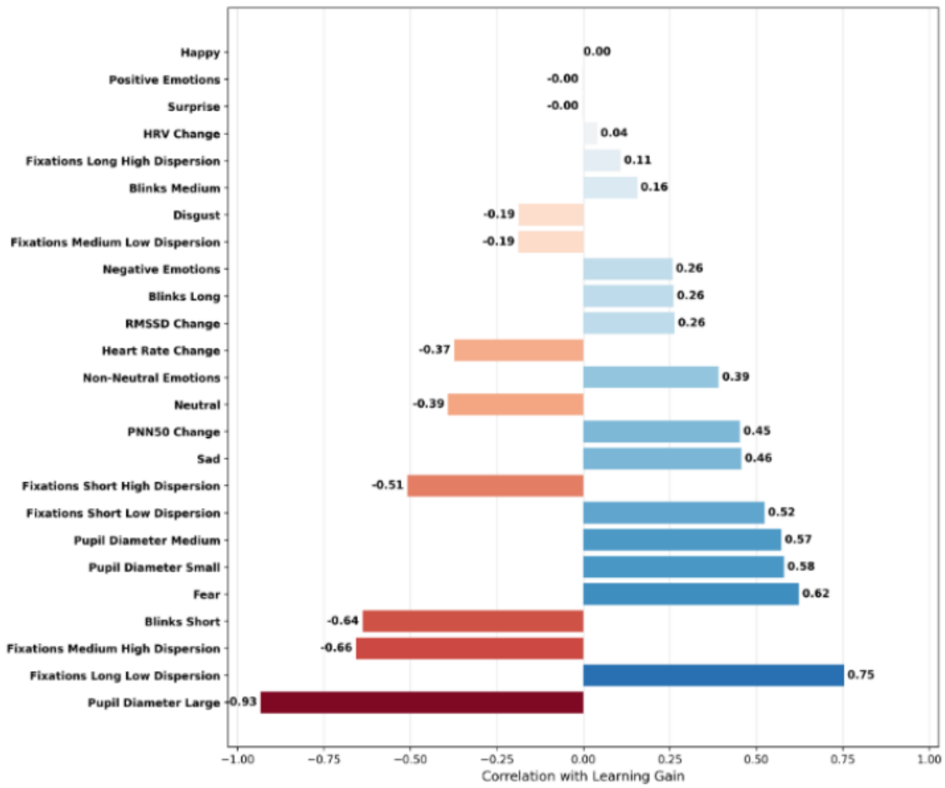


Figure 4: Correlation of Physiological Response with Delta for Participant 1

Figure 4 illustrates the correlation coefficients between the change percentage of various biomarkers and the post-video performance delta for Participant 1. The x-axis represents the correlation coefficient, while the y-axis lists the features (e.g., facial expressions, eye-tracking metrics). In alignment with our threshold of ±0.4, correlations with absolute values exceeding 0.4 are interpreted as strong relationships (positive or negative), whereas those lying between −0.4 and +0.4 are treated as weak or moderate.

Emotion-Related (Webcam) Features: Among the emotion-based features (Fear, Happy, Sad, Disgust, Anger, Neutral, Positive, Negative, Non-neutral), only a few exceed the ±0.4 threshold for Participant 1. Specifically, Happy shows a strong negative correlation (−0.61), indicating that increases in "happy" expressions are associated with lower learning gains for this participant. By contrast, Disgust, Positive, and Surprise each exhibit strong positive correlations (all above +0.4). Referring to Table 3, these outcomes may reflect varying emotional engagement levels. For instance, "positive engagement" or "unexpected moments" might heighten cognitive processing and benefit learning (hence the positive correlation), while the negative association with "happy" could suggest that too much ease or enjoyment might reduce the focused attention required for higher achievement. Other emotional markers, such as Fear, Sad, Anger, and Neutral, remain well within the −0.4 to +0.4 band, pointing to minimal predictive strength for this participant.

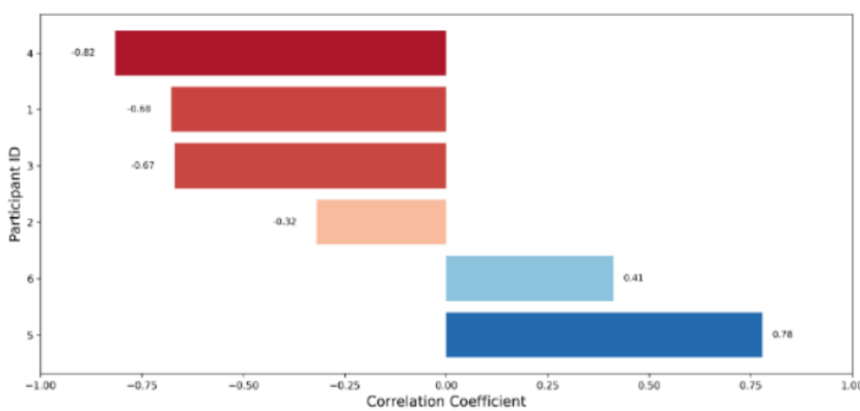
Eye-Tracking Features: Eye-tracking measures display the largest number of strong correlations, consistent with the broader findings in Section 5.1:

a. Strong Negative Correlations Several features, such as fixations_medium _high_dispersion (−0.85), pupil_diameter_small (−0.74), and fixations_short_high_dispersion (−0.68), are associated with a decline in learning gains. According to Table 3, medium or short fixations with high dispersion may indicate scattered viewing or rapid scanning, suggesting insufficient focused attention. Similarly, a small pupil diameter implies low arousal or reduced cognitive load, possibly leading to weaker learning outcomes.

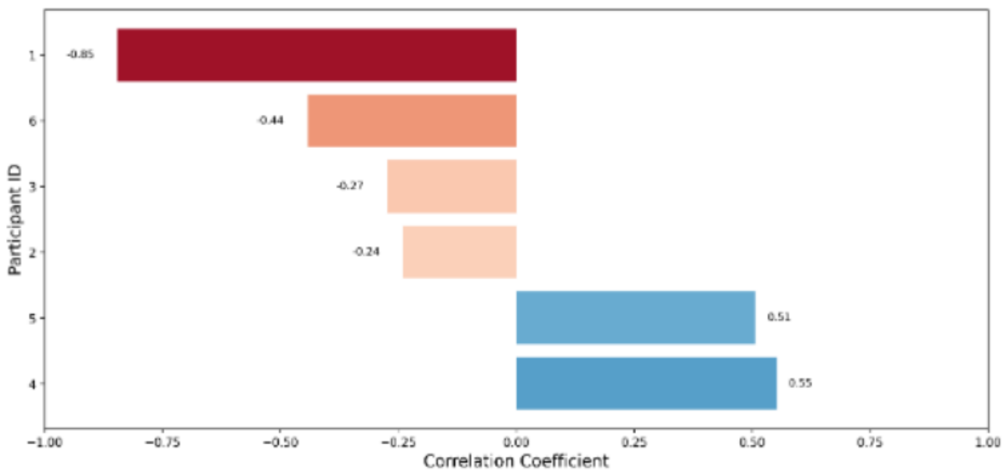
b. Strong Positive Correlations In contrast, fixations_short_low_ dispersion (+0.83), fixations_medium_low_dispersion (+0.82), and pupil_diameter _large (+0.66) exhibit high positive correlations with delta. These metrics point to more targeted attention (low dispersion) and heightened cognitive engagement (large pupil diameter), both of which appear to facilitate better learning outcomes. Furthermore, blinks_medium at +0.40 is near the threshold, suggesting moderately beneficial blinking patterns that may align with typical cognitive processing and help avoid eye fatigue.

Interpretation: For Participant 1, the strongest predictors of learning gains are closely tied to eye-tracking metrics. A higher degree of focused attention (short or medium fixations with low dispersion) and elevated cognitive effort (large pupils) correlate with improved performance, while fragmented viewing habits (high-dispersion fixations) or low arousal (small pupil diameter) correlate negatively with delta. Emotional responses like disgust and surprise also show positive relationships, but the magnitude and direction of these emotional correlations can be highly individual and context-dependent.

4.2. Individual Variability in Physiological Correlations: Divergent Patterns Across Participants



(a) Fixations short high dispersion



(b) Fixations medium high dispersion

Figure 5: Comparison plots for two features across all six participants

We now shift our focus to a broader comparison across all six participants. Specifically, we examine two eye-tracking parameters that often show strong correlations with learning outcomes: fixations_short_high_dispersion and fixations_medium_high_dispersion. Figure 5 demonstrates that these features yield starkly varying correlation coefficients among participants. For example, fixations_short_high_dispersion may exhibit a highly negative correlation (below −0.4) for one participant, suggesting that a high degree of rapid, scattered fixation adversely affects learning outcomes, while for another participant, this same feature might be moderately positive (above +0.4). A similar inconsistency is observed in "fixations medium high dispersion", where participant 1 shows a negative correlation while participant 4 shows a positive one, reinforcing the absence of a universal response profile.

In essence, these findings suggest that each participant exhibits a unique pattern of correlations between learning outcomes and specific eye-tracking features. Finally, the concluding insight from our work is that before applying this approach in real-world scenarios, one should first establish a baseline for each individual to understand which parameters positively or negatively influence their learning. Only then can these insights be effectively leveraged to optimize and personalize instruction.

4.3 Statistical significance and cross-modal interactions

To complement the effect-size screening based on large-magnitude correlations ($|r| \geq 0.4$), we evaluated statistical significance using pooled models and controlled for multiple comparisons using the Benjamini–Hochberg false discovery rate (FDR). Under an FDR threshold of $q < 0.10$, no individual biomarker (eye-tracking, facial expression, or ECG feature considered alone) remained significant in the pooled analysis.

We then explored whether cross-modal combinations provide stronger predictive signals by testing pairwise interaction terms (feature A × feature B) in the pooled model. Two interaction terms reached significance after FDR correction:

Interaction term	**Interaction coefficient (β)**	**p (raw)**	**q (FDR)**
fixations_medium_low_dispersion × RMSSD (diff_ecg_rmssd)	0.625	0.0012	0.027
RMSSD (diff_ecg_rmssd) × heart rate change (diff_ecg_heart_rate)	-0.133	0.0065	0.071

These findings suggest that learning gain (delta) may be better captured by multi-sensor interactions—particularly a gaze × HRV (RMSSD) relationship—than by any single biomarker alone. However, this study has limited statistical power due to the small sample size and repeated-measures structure (multiple videos per participant), so these interaction effects should be interpreted as exploratory and require validation in a larger cohort.

5. Discussions

5.1. Addressing the Gap in Traditional Assessments

The primary research question examines the quantifiable link between student physiology and knowledge acquisition. Our findings suggest a qualified affirmative response: multi-modal biomarker data can correlate with learning improvements from baseline to post-video assessments. Critically, this study directly addresses the limitation identified in traditional assessments—the inability to measure knowledge change during learning. By operationalizing learning as the delta between pre- and post-assessment scores and correlating this with physiological signals captured during the learning event itself, we demonstrate a methodological approach that moves beyond endpoint assessment toward process-oriented evaluation.

Eye-tracking features showed the largest effect sizes in this pilot; however, no individual biomarker remained significant after FDR correction. The strongest evidence was a cross-modal interaction between focused fixations and HRV (RMSSD), suggesting that multi-sensor combinations may be more informative than single channels. These findings suggest that physiological indicators captured in real-time can serve as proxies for the cognitive processes underlying learning, providing insights that traditional post-hoc assessments cannot offer. While traditional methods reveal only whether a student knows something after instruction, our approach offers preliminary evidence that biomarkers may indicate whether and how effectively learning is occurring during instruction.

5.2. New Directions for Educational Evaluation

These preliminary findings point toward several potential directions for educational evaluation and learning system design. First, the strong associations between eye-tracking metrics and learning outcomes suggest that gaze-based measures could serve as non-intrusive indicators of learning effectiveness. Unlike traditional assessments that interrupt the learning process, eye-tracking can be captured continuously and unobtrusively, enabling evaluation that is embedded within the learning experience itself.

Second, the individual variability observed across participants highlights the potential for personalized assessment approaches. Rather than applying uniform evaluation criteria, biomarker-based systems could establish individual baseline profiles and track deviations that signal effective or ineffective learning states for each student. This represents a departure from standardized testing toward individualized evaluation metrics.

Third, the multimodal nature of our approach—integrating eye-tracking, facial expression analysis, and cardiac measurements—demonstrates that different physiological channels capture different aspects of the learning experience. While eye-tracking emerged as the strongest correlate in our study, the framework allows for future refinement as understanding of physiological-learning relationships deepens.

5.3. Practical Applications in Instructional Contexts

Our findings have implications for three distinct instructional applications: formative assessment, adaptive feedback systems, and learning analytics for instructors.

Regarding formative assessment, the real-time nature of physiological monitoring suggests potential applications for ongoing, low-stakes assessment during learning activities. Rather than relying solely on periodic quizzes or assignments, instructors could receive aggregated indicators of class-wide engagement and comprehension. For instance, a sudden increase in high-dispersion fixations across multiple students might signal confusion with particular content, prompting the instructor to pause and clarify.

For adaptive feedback systems, particularly in digital learning environments, physiological signals could trigger automated interventions. If a learning management system detects biomarker patterns associated with low learning gains (such as small pupil diameter or scattered fixations), it could automatically adjust content pacing, offer additional explanations, or prompt the learner to take a break. This closed-loop approach would enable learning systems to respond to individual cognitive states in real time, a capability that current systems—which rely on explicit learner inputs or post-hoc performance data—cannot provide.

For learning analytics, instructors and learning designers could use aggregated physiological data to evaluate the effectiveness of different instructional materials or approaches. By comparing biomarker patterns and associated learning gains across different videos or modules, educators could identify which content formats best support cognitive engagement for their student populations. This evidence-based approach to instructional design moves beyond self-reported student preferences or end-of-course evaluations toward objective measures of learning process quality.

It is important to note that our current findings are preliminary and based on a small sample, and thus these applications remain speculative until validated through larger-scale studies. However, they illustrate the practical potential that motivates continued research in this area.

5.4. Individual Variability and Personalized Learning

Learners exhibit unique biomarker profiles, with eye-tracking metrics affecting learning gains differently for each individual. For one participant, larger pupil diameters correlated inversely with learning, while another showed a mild positive association. This variability highlights learning's personalized nature—emotional triggers, cognitive loads, and attention patterns may either impede or enhance comprehension depending on the student. By identifying these personal biomarker patterns, educators and adaptive learning systems could potentially deliver more effective targeted interventions than traditional one-size-fits-all approaches.

This finding reinforces the central argument of our work: that measuring learning requires attending to individual differences rather than applying universal metrics. Traditional assessments assume that the same test measures the same construct for all students, but physiological evidence suggests that the pathways to learning may differ substantially across individuals. A biomarker-informed approach to assessment could account for these differences, evaluating each student against their own baseline rather than a normative standard.

6. Limitations

There are three key limitations of our work. First, our small sample size (n=6) limits generalizability, requiring larger, more diverse participant pools to validate patterns across demographics, learning styles, and subjects. The correlational patterns observed may not replicate in broader populations, and the individual variability we documented underscores the need for larger samples to distinguish genuine individual differences from noise. Second, contextual factors (stress, time of day, personal interest) can influence biometric responses, necessitating better controls in future studies to ensure observed patterns reflect actual learning rather than external variables. Additionally, while we have demonstrated correlations between biomarkers and learning gains, establishing causal relationships will require experimental designs that manipulate physiological states or instructional conditions.

7. Future Directions

Future work should focus on scaling data collection across larger, diverse populations to enhance reliability and enable subgroup analyses. Methodologically, integrating real-time biometric feedback loops into learning platforms could enable experimental tests of whether biomarker-triggered interventions improve learning outcomes compared to control conditions. Such studies would move beyond correlational evidence toward causal validation of biomarker-based assessment approaches.

From a practical standpoint, developing user-friendly interfaces that translate physiological data into actionable insights for instructors represents an important implementation challenge. The gap between raw biomarker data and pedagogically meaningful information must be bridged through thoughtful dashboard design and clear interpretation guidelines. Longitudinal studies tracking how individual biomarker profiles evolve over a semester or academic year could also reveal whether physiological correlates of learning remain stable or shift as students develop expertise.

These advances would work toward establishing biomarker-based assessment as a credible, individualized tool for measuring authentic learning—addressing the fundamental gap in traditional assessments identified at the outset of this work.

8. References

[1] B. E. Seely, "The other re-engineering of engineering education, 1900-1965," Journal of Engineering Education, vol. 88, no. 3, pp. 285-294, 1999.

[2] M. Calvert, The Mechanical Engineer in America: Professional cultures in conflict., Baltimore: The John Hopkins University Press, 1967.

[3] Christopher J. H., Krajcik S. J., Pellegrino J. W., DeBarger A. H., "Designing Knowledge-In-Use Assessments to Promote Deeper Learning," Educational Measurement: Issues and Practice, vol. 38, no. 2, pp. 53-67, 2019.

[4] A. Johri, International Handbook of Engineering Education Research, New York: Routledge, 2023.

[5] N. K. Segool, J. S. Carlson, A. N. Goforth, N. V. derEmbse, "Heightened test anxiety among young children: Elementary school students' anxious responses to high-stakes testing," Psychology in the Schools, vol. 50, no. 5, pp. 489-499, 2013.

[6] T. W. Joseph A. Rios, "Essay Items," in The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation, Thousand Oaks, SAGE Publications, Inc., 2018, pp. 602-605.

[7] J. W. Pellegrino, L. V. DiBello, S. P. Brophy, "The Science and Design of Assessment in Engineering Education," in Cambridge Handbook of Engineering Education Research, New York, Cambridge University Press, 2014, pp. 694-728.

[8] R. M. Luecht, "Computerized Adaptive Testing," in The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation, Thousand Oaks, SAGE Publications, Inc., 2018, pp. 344-350.

[9] B. Lund, et al., "Student Perceptions of AI-Assisted Writing and Academic Integrity: Ethical Concerns, Academic Misconduct, and Use of Generative AI in Higher Education," AI in Education, vol. 1, no. 1, 2025.

[10] D. K. Darnell and P. A. Krieg, "Student engagement, assessed using heart rate, shows no reset following active learning sessions in lectures," PLoS ONE, vol. 14, no. 2, 2019.

[11] T. Piquado, D. Isaacowitz, A. Wingfield, "Pupillometry as a measure of cognitive effort in younger and older adults," Psychophysiology, vol. 47, no. 3, pp. 560-569, 2010.

[12] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster and J. R. Movellan, "The Faces of Engagement: Automatic Recognition of Student Engagement from Facial Expressions," IEEE Transactions on Affective Computing, vol. 5, no. 1, pp. 86-98, 2014.

[13] M. I. Posner, "Cognitive Neuroscience," in The SAGE encyclopedia of educational research, measurement and evaluation, Thousand Oaks, SAGE Publications, Inc, 2018.

[14] J. Beatty, "Task-evoked pupillary responses, processing load, and the structure of processing resources.," Psychological Bulletin, vol. 91, no. 2, pp. 276-292, 1982.

[15] A. T. Duchowski, Eye Tracking Methodology, Springer, 2017.

[16] C. Fogarty and J. A. Stern, "Eye movements and blinks: their relationship to higher cognitive processes," . International Journal of Psychophysiology, vol. 8, no. 1, pp. 35-42, 1989.

[17] E. Granholm and S. R. Steinhauer, "Pupillometric measures of cognitive and emotional," International Journal of Psychophysiology, vol. 52, no. 1, pp. 1-6, 2004.

[18] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, "Eye tracking: A comprehensive guide to methods and measures," Oxford: Oxford University Press, 2011.

[19] M. A. Just and P. A. Carpenter, "Eye fixations and cognitive processes," Cognitive Psychology, vol. 8, no. 4, pp. 441-480, 1976.

[20] D. Kahneman and J. Beatty, "Pupil diameter and load on memory," Science, vol. 154, no. 3759, pp. 1583-1585, 1966.

[21] J. Wang et al., "Facex-zoo: A pytorch toolbox for face recognition," in Proceedings of the ACM International Conference on Multimedia , 2021.

[22] A. Poole and L. J. Ball, "Eye tracking in human-computer interaction and usability research: Current status and future prospects," in Encyclopedia of Human Computer Interaction, Idea Group Reference, 2006, pp. 211-219.

[23] K. Rayner, "ye movements in reading and information processing," Psychological Bulletin, vol. 124, no. 3, pp. 372-422, 1998.

[24] L. R. Lattuca, P. T. Terenzini, J. F. Volkwein, "Engineering Change," ABET , Baltimore, USA, 2006.

[25] A. M. Eren Bilen, "Online cheating amid Covid-19".

[26] J. Stern, D. Boyer, D. Schroeder, "Blink rate: a possible measure of fatigue," Human Factors, vol. 36, no. 2, pp. 285-297, 1994.

[27] F. V. Orden, W. Limbert, S. Makeig, T. Jung, "Eye activity correlates of workload during a visuospatial memory task," Human Factors, vol. 43, no. 1, pp. 111-121, 2001.