What Do Your Eyes Reveal About Learning from Videos? A Dataset Demonstrating the Link Between Gaze Measures, Engagement, and Learning

Tobias Wagner
Institute of Media Informatics
Ulm University
Ulm, Germany
tobias.wagner@uni-ulm.de

Enrico Rukzio
Ulm University
Ulm, Germany
enrico.rukzio@uni-ulm.de

Teresa Hirzle
Department of Computer Science
University of Copenhagen
Copenhagen, Denmark
tehi@di.ku.dk

Abstract

Sensor-based methods such as eye tracking are promising for user modeling in adaptive educational systems, as they enable continuous, fine-grained assessment of learners' learning processes. However, implementing gaze-based user models requires selecting from a large set of available gaze measures, posing a challenge for adaptive-system designers. To inform this design process, we conduct an exploratory analysis of the associations between gaze measures, learner engagement, and learning outcomes in a video-based learning context. An analysis of data collected from 18 learners indicates that fixation dispersion and fixation rate on active areas of interest are indicators of engagement, while pupil diameter is an indicator of learning outcomes. To support further research on gaze-based user modeling and adaptation, we introduce EnGazement, a dataset comprising over 1.3 million processed eye-tracking data points, enriched with subjective and objective measures of engagement and learning.

CCS Concepts

• **Human-centered computing → Human computer interaction (HCI)**; • **Applied computing → Education**.

Keywords

Engagement, Eye Tracking, Video-based Learning

ACM Reference Format:
Tobias Wagner, Enrico Rukzio, and Teresa Hirzle. 2026. What Do Your Eyes Reveal About Learning from Videos? A Dataset Demonstrating the Link Between Gaze Measures, Engagement, and Learning. In *34th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '26), June 08–11, 2026, Gothenburg, Sweden.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3774935.3806155

1 Introduction

Video-based learning (VBL) has become a popular method for self-directed learning due to the accessibility and availability of learning videos on video-sharing platforms like YouTube[1] and massive open online courses (MOOCs) such as Coursera[2]. VBL allows learners to individualize their learning experience by offering a wide range of content, from lifelong learning to university education and professional skill development. Additionally, it enables learners to define their learning environment, time, and pace individually [67]. In the absence of a teacher's presence, responsibility for planning, monitoring, and evaluating understanding shifts to learners, who often fail to regulate their learning effectively [6, 27]. To address this challenge, prior research has designed and evaluated systems that scaffold self-regulatory processes during learning activities [27]. For such systems to be effective, they require timely information about learners' ongoing processes in order to adapt their support accordingly [18, 40]. While subjective self-reports or post hoc assessments such as quizzes can support evaluation and reflection after video viewing, they fail to capture dynamic factors that directly influence learning performance, including fluctuations in attention or effort [40]. In contrast, log data, video recordings, and physiological measures enable continuous, fine-grained, and more direct assessment of learners' momentary states during learning.

One higher-level learner state that has received attention is engagement, as it is closely associated with learning performance [23]. Engagement, defined as learners' investment or involvement in learning [34], is commonly conceptualized as a multidimensional construct comprising cognitive, behavioral, and emotional engagement [23]. In adaptive learning systems, engagement is a particularly relevant target state, as fluctuations in engagement signal when learners may benefit from interventions. Accordingly, prior work on engagement assessment has explored sensor-based approaches in both classroom settings [1, 20, 25, 50, 62] and VBL environments [31, 89]. These approaches commonly rely on facial expression analysis from video recordings [31, 43, 70] or physiological signals captured by wearable sensors, such as electrodermal activity (EDA) [20].

Another promising sensor-based method for assessing learner engagement is eye tracking, which provides continuous access to learners' visual attention and cognitive processing during learning. While it does not provide a direct measure of engagement, gaze can be interpreted as an indirect indicator and is therefore well-suited for use in adaptive learning systems. In educational research, eye tracking has been used to study processes of selecting, organizing, and integrating information in multimedia learning environments [68]. This builds on the eye–mind assumption [41, 42], which posits that eye movements reflect learners' cognitive processing of

[1]https://www.youtube.com/education, accessed January 26, 2026

[2]https://www.coursera.org/, January 26, 2026

ACM ISBN 979-8-4007-2311-7/26/06
https://doi.org/10.1145/3774935.3806155

visual information [74]. From an adaptive-system perspective, this assumption suggests that eye movement data may serve as a basis for inferring learners' momentary engagement with learning content, as gaze behavior is hypothesized to reflect both the quantity and quality of cognitive investment in comprehension [21, 55, 69].

Existing research linking eye movements to engagement has predominantly focused on detecting mind wandering [36, 38, 89], a specific form of disengagement [72]. These studies typically target isolated processes or single gaze indicators, such as blink rate [14] or pupil dilation as a proxy for cognitive load [88]. For adaptive systems, however, design decisions must be made about which gaze measures to select from the wide range of available ones [12, 35].

This paper addresses this challenge by presenting two contributions [87]. First, an **empirical contribution**: A dataset analysis examining associations between gaze measures, learner engagement, and learning outcomes during applied VBL, informing gaze measure selection for adaptive learning systems. Second, a **dataset contribution**: EnGazement, comprising six measures based on Areas of Interest (AOI) and seven AOI-independent measures, based on over 1.3 million eye-tracking data points from 18 students watching two 10-minute learning videos. As ground truth, the dataset contains subjective ratings of engagement, answers to multiple-choice questions on retention and understanding, ratings of perceived understanding and difficulty, and ratings of mental demand.

By investigating the associations between gaze measures, engagement, and learning outcomes using Bayesian analysis, we found that fixation dispersion and fixation rate on active AOIs are associated with learners' engagement, and that pupil diameter is associated with their learning outcomes when watching learning videos. We discuss the implications of these findings and provide suggestions for future adaptive learning systems that may benefit from gaze-based learning analytics.

2 Related Work

Our work builds on prior research and datasets on assessing engagement during video-based learning using self-report, observational, video-based, physiological, and gaze-based approaches. It addresses the gap in studying how gaze measures relate to learner engagement in a VBL task, as well as the lack of openly available eye-tracking datasets in this area.

2.1 Engagement Assessment

To capture fluctuations in learners' engagement, common approaches are experience sampling [21], stimulated recall [63], clickstream data [21], or observation [61].

Experience sampling methods (ESM) use so-called thought probes that prompt learners within the learning task to rate their current engagement. While this method allows for high temporal resolution in the measurement, upcoming probes interrupt the ongoing task and may influence learners' engagement [63]. As a remedy, stimulated recall can retrospectively assess learners' engagement during a learning task using mnemonic cues such as images, videos, and audio recordings of moments during the task [63]. Pereira et al. [63] found that learners' engagement did not differ between ESM probes and ratings obtained through stimulated recall. However, these methods rely, on the one hand, on learners' ability to self-assess their engagement, and, on the other hand, on the questionnaires used. An alternative to self-assessment methods offers observations. Using protocols like the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP), various coders observe learners during a learning task and code their behavior (e.g., on-task, off-task) and affect (e.g., bored, concentrated, frustrated) [61].

The aforementioned methods are used to collect ground-truth data for training classifiers to automatically detect student engagement using physiological signals (e.g., electrodermal activity, heart rate, skin temperature, and accelerometry [20, 25], and electroencephalography [4]), video-based features such as facial and upper-body movements (e.g., head pose, facial expressions, and gaze vector and angle [8, 31, 43, 56, 86], or eye-tracking data [36, 38, 89].

2.2 Engagement Dimension and Granularity

Previous work differs in the dimensions and granularity of engagement assessed. Student engagement is often assessed overall [86], either with binary classes (engaged or not engaged) [20] or with multiple ordered classes ranging from very low to very high engagement [31], typically using a 3- to 5-point rating scale. The individual dimensions of cognitive, behavioral, and emotional engagement in applied learning tasks were considered in only a few data collections, primarily with video-based approaches. In contrast, gaze-based approaches focus on mind wandering [36, 38, 89], which can be considered a form of disengagement [84] where learners' attention and thoughts shift away from the learning task [72]. Here, the granularity is limited to a binary classification of mind wandering or not. Compared to video-based approaches, eye-tracking data are typically collected in controlled settings with controlled tasks, which may limit their comparability to applied learning contexts, likely due to the elaborate setups required to obtain high-precision eye movement data. Furthermore, prior gaze-based data collections did not consider the dimension-wise assessment of student engagement. However, it has been shown that eye movements, blink rate, and pupil size can provide information about learners' cognitive processing [41], interest [49], and focus of visual attention [35].

2.3 Engagement Datasets

The potential of eye-tracking data for gaze-based learning analytics and adaptive systems is evident not only from prior research and its wide use in cognitive load assessment within HCI [14, 88], but also from their value in educational research as a tool for investigating learning processes by linking gaze measures to cognitive processes [2, 68, 77]. These highlight the potential of eye-tracking data to assess learners' cognitive, behavioral, and emotional engagement, as well as learning-related outcomes. However, available datasets use video-based approaches to measure engagement (e.g., [86]). Previous gaze-based approaches have primarily focused on detecting disengagement with dichotomous classes, particularly for identifying mind-wandering, and here, datasets are not openly available [37, 38]. We address this limitation with the EnGazement dataset.

3 Data Collection

We conducted a data collection with university students who were instructed to watch and understand learning videos and complete

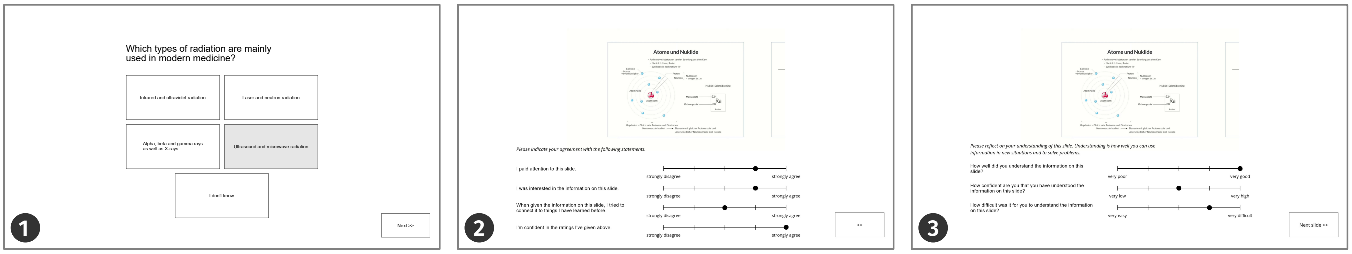

Figure 1: ❶ Answering multiple choice question. ❷ Rating engagement. ❸ Rating self-perceived understanding.

a multiple-choice test afterwards, to investigate the exploratory **research question**: *How are gaze measures associated with learners' engagement and learning outcomes during video-based learning?*

3.1 Design & Measures

The data collection followed a within-participant design with the learning video's topic as a two-level independent variable. They watched a chemistry video about acid-base balance (9:50 mins)[3] and a physics video on ionizing radiation (10:22 mins)[4] in a counter-balanced order. Both videos present information on slides with German narration, without subtitles, a common production style for learning videos [30]. The videos are publicly available on YouTube, produced by professional educators, and are similar in design (i.e., colors, typography, transitions between scenes, and visual aesthetic).

While learners watched these videos, we recorded *eye-tracking data*, and after each video, we collected ground truth data by asking them to answer multiple-choice questions assessing their *learning outcomes* and to rate their *engagement*. These variables were treated as dependent variables. To control for individual differences, we additionally assessed learners' *motivation*, *prior knowledge*, *perceived understanding and difficulty*, and *mental demand* towards the learning video.

We used the stimulated recall method to retrospectively assess learning outcomes, engagement, perceived understanding, and difficulty without interrupting learners' flow or affecting their ongoing attentional state [63]. As a mnemonic cue, learners were shown a single video frame representative of the visual information within an approximately 1-minute window of the learning video (window duration $M = 58s$, $SD = 32s$, $min = 17s$, $max = 118s$).

3.1.1 Eye-Tracking Data. We collected learners' eye movements using the Tobii Pro Spark[5], a research-grade eye tracker commonly used in experimental settings that captures gaze and pupil data using video-based pupil and corneal reflection eye-tracking technique at 60 Hz.

3.1.2 Learning Outcomes. To measure participants' learning outcomes regarding factual knowledge, specifically the cognitive process dimensions of *remembering* and *understanding* [48], we used multiple-choice questions (MCQs) to assess learners' retention and comprehension of learning video parts.

[3]Chemistry video, accessed January 26, 2026
[4]Physics video, accessed January 26, 2026
[5]Tobii Pro Spark: Reported precision of 0.26° root-mean-square (RMS) (using built-in filtering) and an accuracy of 0.45° in optimal conditions, accessed January 26, 2026

3.1.3 Engagement. Consistent with prior work treating engagement as multidimensional construct [19, 25, 75], we operationalized behavioral, emotional, and cognitive engagement using representative single items adapted from the questionnaire by Reeve and Tseng [65], the University Student Engagement Inventory [52], the Engagement versus Disaffection questionnaire [71], and the School Engagement questionnaire [22]. Items were answered on a 5-point Likert scale (1 = strongly disagree to 5 = strongly agree).

3.1.4 Perceived Understanding and Difficulty. Since the sensitivity of objective measures of learning outcomes, like MCQs, is highly dependent on the quality and design of questions and their answer items, we included learners' perceived understanding, which measures what learners think they have understood of the information presented in the learning videos [64]. We used the items from Pilegard and Mayer [64] that assess learners' perceived understanding, confidence, and level of difficulty on a 5-point Likert scale (1 = very poor to 5 = very good). Similar to the engagement questionnaire, those items were assessed using the stimulated recall approach after participants watched the learning video.

3.1.5 Mental Demand. Since the overall difficulty of the videos may affect learners' engagement, we assessed their mental demand after they watched the learning video. We used a translation of the NASA task load index (NASA-TLX) mental demand item (*1 = low* to *20 = high*)[6] [32]. The NASA-TLX is a widely used and valid method in HCI [46] and educational research for measuring subjective intrinsic cognitive load [5].

3.1.6 Prior Knowledge. Based on the cognitive-affective theory of learning with media (CATLM), differences in learners' prior knowledge affect how much is learned [58], which is why we include it as a control variable. We created a questionnaire based on Mayer [53]. We asked participants to self-rate their prior knowledge with a single item on a 5-point Likert scale (e.g., *"How much do you already know about the acid-base balance?"* from *1 = very little* to *5 = very much*) and checking items that apply to them (e.g., for the chemistry video *"I attended a course in chemistry or biochemistry at a university."* and *"I can explain a chemical reaction equation."*). Prior knowledge was determined by summing the value of the single item (min = 1, max = 5) and the number of checked items (min = 0, max = 7). That resulted in a minimum of 1 and a maximum of 12.

3.1.7 Motivation. The CATLM assumes that learners' motivation mediates learning by affecting cognitive engagement [58]; thus,

[6]NASA-TLX translation, accessed January 26, 2026

we controlled for this variable by assessing learners' current motivation using the German version of the questionnaire on current motivation (QCM). It consists of items measuring the factors interest, probability of success, challenge, and anxiety on a 7-point scale (1 = disagree to 7 = agree). [66, 82]. As recommended by Vollmeyer and Rheinberg [82], we adapted the interest items 1 and 4 to our task.

3.2 Participants

We recruited 23 learners via university mailing lists and notices; 5 were excluded because their data exceeded the 0.9th percentile in both videos. The final dataset comprised 18 learners (9 female, 8 male, 1 non-binary), with 16 aged 20–29 and 2 aged 30–39. All participants were fluent in German and proficient in English, which were the inclusion criteria for recruitment. Regarding education, 11 participants held a high school degree (undergraduates), 4 a bachelor's degree (graduates), and 3 a master's degree (postgraduates). Ten participants were studying or had studied Information and Communication Technologies, six Health and Welfare, and two Natural Sciences, Mathematics, and Statistics [76]. Ten participants wore no visual aids, seven wore glasses, and one wore contact lenses. Participants reported using learning videos from sometimes to very often (*1 = Never, 5 = Always*; $M = 3.28$, $SD = 1.23$).

3.3 Apparatus and Environment

Data collection took place in a 3.45 × 4.65 m room with lowered blinds and fluorescent lighting ($M = 808.92$ lux, $SD = 44.81$, measured with a Uni-T Mini Light Meter UT383). Two laptops were placed on a central table: one for pre- and post-questionnaires administered via LimeSurvey, and one[7] equipped with a screen-based eye tracker and webcam for the data collection running PsychoPy and OBS.

3.4 Procedure

Upon arrival, participants provided informed consent[8], received task instructions, and reported motivation and prior knowledge. They then moved to a second laptop running the data collection software to watch one of the two learning videos. We set the initial chair position using a folding rule to ensure a 60 cm eye–screen distance; head movement was unconstrained, and participants were instructed to behave naturally. Each step was briefly explained on-screen and started with a key press.

The procedure began with a 9-point eye tracker calibration and validation, repeated if necessary, followed by a 30 s central fixation to collect baseline data. Participants then watched the video without pausing, rewinding, or taking notes. After each video, they rated mental demand, completed multiple-choice questions, then rated their engagement, followed by perceived understanding and difficulty (see Figure 1). The procedure was then repeated for the second video, with order counterbalanced and prior knowledge assessed beforehand.

Finally, participants completed questionnaires on VBL experience, video similarity, platform use, and demographics, and could provide feedback. The session lasted approximately 60 minutes, and participants received a 10€ allowance.

3.5 Data Preprocessing

We preprocessed collected data with Python (v3.11.4) and Jupyter Notebook, primarily using Pandas (v2.2.2). The code and the resulting dataset are openly accessible on OSF [9]. Table 1 provides an overview of the eye-tracking data included in the dataset, categorized in temporal, spatial, and spatio-temporal measures. The final dataset consists of preprocessed eye-tracking data, identified eye movement events, and calculated gaze measures from over 1.3 million processed eye-tracking data points. These are mapped on windows in the video, together with subjective learner-rated engagement, their perceived understanding and difficulty, and their learning outcomes.

3.5.1 Ground Truth. Ratings on engagement, perceived understanding and difficulty, mental demand, and answers to the multiple-choice questions were extracted from a comma-separated file that results from each PsychoPy session.

3.5.2 Eye-Tracking Data. For each learner, gaze x- and y-positions and pupil diameter measured by the eye tracker, along with a timestamp, were extracted from a hierarchical data file, along with events (i.e., fixation task and watching video). The extracted data showed an accuracy with a mean error of 0.63° ($SD = 0.26°$, $min = 0.24°$, $max = 1.27°$) and a data loss across both videos of 0.12% ($SD = 0.06\%$, $min = 0.03\%$, $max = 0.34\%$). Extracted gaze positions were transformed to pixel values, and REMoDNaV (v1.1.2), a velocity-based eye movement event classification algorithm, was applied to get saccades and fixations [17]. Additionally, we classified blinks by identifying signal-loss windows in the data with a minimum duration of 0.02 seconds.

After identifying eye movement events, the data were segmented according to the duration of each window in the video to calculate AOI-based and AOI-independent measures. We calculated measures that has been shown to be associated with cognitive processing during reading [41, 42], cognitive load in HCI [14, 26, 88], sustained attention [13], cognitive processes of selecting, organizing, and integrating of information in learning material [16], and mind wandering [37].

In general, duration measures are calculated from the start and end of an eye movement event, and are given in seconds. Rates are calculated by dividing the number of eye movement events by the exact duration of the associated window in the videos in seconds, resulting in occurrences per second, and distances are given in pixels or degrees of visual angle (°).

Areas of Interest. Required Areas of Interest (AOIs) to calculate AOI-based measures were identified using the Computer Vision Annotation Tool (CVAT v2.21.1). Text and images appearing in the learning video were annotated with rectangular shapes and categorized as image or text and as *active* if they were currently being displayed or discussed by the speaker. The active AOIs were used to calculate AOI-based gaze measures.

[7] MSI GP66 Leopard 11UG with a 15.6-inch FullHD display (141 ppi)

[8] Ethics: The study followed local regulations and ACM ethical guidelines; formal ethics approval was not required under the university's policies.

[9] Dataset and Code available on OSF

Fixations. We calculated AOI-independent measures average *fixation rate*, *fixation duration* and *fixation dispersion* (i.e., the distance of individual fixation points from a mean fixation position), as well as the AOI-based measures *time to first fixation (TTFF)*, *first fixation duration (FF Duration)*, *first pass time (FPT)*, *average fixation duration*, and *fixation rate* for active AOIs on each slide. Fixations outside the screen were excluded for AOI-independent measures, and those outside defined AOIs were excluded for AOI-based measures. Fixations with durations 1.5 times the interquartile range above the 75th percentile and below the 25th percentile were considered outliers and therefore removed in the case of the AOI-independent measures.

Saccades. Before calculating the AOI-independent measures *saccade duration*, *saccade amplitude*, and *saccade velocity* over windows per learner, we removed saccades that start or end outside the screen and saccades with a duration or amplitude 1.5 times the interquartile range above the 0.75 percentile and below the 0.25 percentile.

Pupil Diameter. Since pupil diameter values vary between individual learners and recordings, we normalized these values within participant and learning video using robust scaling based on the median and interquartile range.

Blink Rate. We calculated *blink rate* by counting blinks per window and dividing this number by the window duration.

	Temporal	Spatial	Spatio-temporal
Processed Eye Tracking Data			
Gaze	time (s)	x- and y-position (px)	
Pupil	time (s)	diameter (mm)	
Eye Movement Events			
Fixation	start, end, duration (s)	mean and SD of position (px), amplitude (°)	mean and peak velocity ($^\circ\,s^{-1}$)
Saccade	start, end, duration (s)	start and end position (px), amplitude (°)	mean and peak velocity ($^\circ\,s^{-1}$)
Blink	start, end, duration (s)		
AOI-independent Gaze Measures			
Fixation	rate (s^{-1}) duration (s)		
Saccade	duration (s)	amplitude (°)	velocity ($^\circ\,s^{-1}$)
Blink	rate (s^{-1})		
Pupil		diameter (mm)	
AOI-based Gaze Measures			
Fixation	rate (s^{-1}) duration (s) FF duration (s) FPT (s) TTFF (s)		

Table 1: Eye-tracking data comprised in the dataset categorized in temporal, spatial, and spatio-temporal measures: Processed gaze signals, eye movement events identified using REMoDNaV [17], AOI-independent and AOI-based gaze measures (*FF: first fixation, FPT: first-pass time, TTFF: time to first fixation*). Rates are reported in s^{-1} (occurences per second) and velocities in $^\circ\,s^{-1}$ (degrees of visual angle per second).

4 Dataset Analysis

We first report the results of the analysis on the video-based CVs mental demand and perceived difficulty. Then, we investigate the link between engagement, learning outcomes, and gaze measures using a Bayesian analysis. The data of the $N = 18$ learners were analyzed using Jasp 0.19.3. and R (v4.4.1).

4.1 Differences between Videos

Differences between the means of mental demand, perceived difficulty, engagement, and learning outcomes were analyzed with a Wilcoxon signed-rank test (for variables measured on an ordinal rating scale) or a paired Student t-test. Assumptions on normality were tested with the Shapiro-Wilk test and visually with a Q-Q plot of the residuals.

The results are reported in detail below. In summary, they revealed no significant differences between the two video conditions in terms of learners' mental demand, perceived difficulty, engagement, and learning outcomes. Given the absence of a significant main effect of video condition and no difference in the tested variables between the two videos, we refrain from further investigating differences between the video conditions and take the whole dataset for subsequent analyses.

Mental Demand. A Wilcoxon signed-rank test was conducted to compare mental demand between the two videos. The results show no significant difference ($W = 62.0$, $z = 0.11$, $p = p = .93$), indicating that learners perceived the chemistry ($Mdn = 13$, $IQR = 5.50$) and physics ($Mdn = 14$, $IQR = 7.75$) video as similarly mentally demanding.

Perceived Difficulty. A Wilcoxon signed-rank test was conducted to assess differences in perceived difficulty between the two videos. Learners perceived the chemistry video slightly more difficult ($Mdn = 2.84$, $IQR = 1.41$) than the physics video ($Mdn = 2.55$, $IQR = 0.93$), but the difference was not significant ($W = 111.50$, $z = 1.13$, $p = p = .27$).

Engagement. Wilcoxon signed-rank tests were conducted to compare cognitive, behavioral, and emotional engagement between the chemistry and physics videos. No significant differences were found for any engagement dimension. Cognitive engagement was slightly higher for the physics video ($Mdn = 4.18$, $IQR = 0.74$) than the chemistry video ($Mdn = 4.00$, $IQR = 0.82$), but the difference was not significant ($W = 48.00$, $z = -1.63$, $p = p = .11$). Behavioral engagement did not differ between the chemistry ($Mdn = 3.84$, $IQR = 0.33$) and physics video ($Mdn = 4.09$, $IQR = 0.61$) ($W = 76.00$, $z = -0.41$, $p = p = .70$). Similarly, emotional engagement was comparable between the chemistry (Mdn = 3.72, IQR = 0.47) and physics video (Mdn = 3.91, IQR = 0.98) (W = 76.00, $z = -0.41$, $p = p$ = .69).

Learning Outcomes. A paired-samples t-test was conducted to compare the average learning outcome (i.e., the average of retention and understanding question results, with uncertain responses coded as 0) between the two videos. The results show no significant difference (t(17) = -1.28, p = .22) between the chemistry (M = 0.62, SD = 0.30) and physics video (M = 0.67, SD = 0.20), indicating that learners achieved similar learning outcomes for both videos.

4.2 Link between Engagement, Learning Outcomes, and Gaze Measures

We analyzed the link between engagement, learning outcomes, and AOI-dependent and independent gaze measures using Bayesian generalized mixed models fitted using *brms*[9], a wrapper for defining Bayesian multilevel models in the STAN programming language in R[11]. To fit models and report the Bayesian analysis, we drew on previous publications in HCI [45, 81, 83] and published Bayesian analyses of eye-tracking data [28, 29, 54].

For each outcome variable, engagement and learning outcome, we fitted two models: with and without video as a fixed effect. All models included gaze measures as fixed effects and learner ID as a random intercept to account for repeated measures and individual differences. Each model was computed with 4 Hamilton-Monte Carlo chains, each with 40000 iterations and 20% warm-up samples, and checked via diagnostic tests (divergent transitions and autocorrelation in trace plots, R-hat < 1.01, Effective Sample Size > 400, and posterior predictive checks) [80]. After model fitting, we compared the models using leave-one-out cross-validation (LOO) to select the best-fitting model for subsequent posterior analysis and to evaluate how adding or removing effects influences predictive performance [79]. Model comparison is based on differences in the Expected Log Predictive Density (ELPD), with higher ELPD values indicating better predictive performance.

For the best-fitting model, we calculated the 95% High-Density Interval (HDI) for all predictors, considering an effect on the outcome variable to be supported if the interval did not include zero. We report the probability of direction (pd), which quantifies the certainty that an effect differs from zero in a specific direction (i.e., posterior distribution proportion either strictly positive or negative) [51]. We utilized δ_t as an effect size estimate, comparable to Cohen's d [33, 54].

4.2.1 Link between Engagement and Gaze Measures. Because the behavioral, emotional, and cognitive engagement items showed strong convergence (Cronbach's $\alpha = 0.89$, 95% CI [0.86, 0.92]; average inter-item correlation = 0.73), we aggregated the three facet ratings into a single ordinal engagement index using the median for analysis. We fit a cumulative model with a logit link and weakly informative priors (normal(0, 0.5)), with engagement as the ordered outcome (1 = low, 5 = high) [9].

Analyzing the posterior distributions of the model without the learning video as fixed effect (ELPD = -428.37) indicates that lower *Fixation Dispersion* ($\tilde{b} = -0.34$, 95% HDI [-0.61, -0.08], $pd = 99.50\%^{**}$, $\delta_t = -0.19$, 95% HDI [-0.38, -0.03]) and a reduced *Fixation Rate* within the active AOI ($\tilde{b} = -0.51$, 95% HDI [-0.93, -0.08], $pd = 99.04\%^{*}$, $\delta_t = -0.27$, 95% HDI [-0.58, -0.03]) were associated with higher levels of engagement. The remaining gaze measures showed no clear evidence of association with *overall engagement*, as their posterior distributions included zero and their credible intervals were wide (see Figure 2).

A subsequent model comparison shows that the full model ($ELPD = -428.37$) achieves the highest predictive performance among the models considered. The model with *Fixation Dispersion* removed has a lower predictive performance ($ELPD_{diff}$ = -1.53, SE_{diff} 2.59), followed by the model with *Fixation Rate on Active AOIs* removed ($ELPD_{diff}$ = -2.77, SE_{diff} = 2.54), and the fully reduced model with both predictors removed ($ELPD_{diff}$ = -4.18, SE_{diff} = 3.45). The ELPD differences for the fully reduced model are larger than four ELPD units, supporting the conclusion that both gaze measures *Fixation Dispersion* and *Fixation Rate on Active AOIs* improve predictive performance [78].

4.2.2 Link between Learning Outcomes and Gaze Measures. For analyzing which of the gaze measures serve as a predictor of learners' *learning outcome*, we fit two Bernoulli models using logit link function with weak informative prior ($b = normal(0, 0.5)$, $Intercept = normal(0, 2)$) as responses are either *0 = wrong* or *1 = correct* [10]. Adding the learning video as a fixed effect did not improve prediction accuracy for *retention*($ELPD_{diff} = -1.95$, SE_{diff} = 0.49), but for *understanding* ($ELPD_{diff} = -0.35$, SE_{diff} = 2.23).

The posterior distribution for the *retention* model ($ELPD$ = -212.63) indicates that a larger *Pupil Diameter* ($\tilde{b}$ = 0.52, 95% HDI [0.05, 1.03], $pd = 98.57\%^{*}$, δ_t = 0.44, 95% HDI [0.03, 0.89]) is associated with a higher probability of correctly answering questions testing retention (see bottom panel in Figure 2). Similarly, the posterior distributions of the *understanding* model ($ELPD = -183.81$) indicates that a larger *Pupil Diameter* ($\tilde{b}$ = 0.85, 95% HDI [0.26, 1.48], $pd = 99.76\%^{**}$, δ_t = 0.42, 95% HDI [0.09, 0.83]) is associated with a higher probability of correctly answering questions testing understanding (see bottom panel in Figure 2). The remaining gaze measures showed no clear association with retention or understanding. Removing *Pupil Diameter* as a predictor from the models results in a lower predictive performance for the retention ($ELPD_{diff}$ = -2.62, SE_{diff} = 1.29) and understanding ($ELPD_{diff}$ = -3.75, SE_{diff} = 1.44).

5 Discussion

The following sections discuss the findings and lessons learned from the EnGazement data collection and the results of our analyses of the dataset, with implications for the design of gaze-based adaptive learning systems.

5.1 Implications for Adaptive Systems

Our results suggest that future adaptive learning systems for VBL that plan to use eye tracking to assess learners' engagement or to predict learning outcomes based on data recorded during learning should prioritize fixation dispersion and pupil diameter as measures, where eye-tracking data from multiple learners are available. Systems without access to others' eye-tracking data must rely on fixation rates within active AOIs and, therefore, implement solutions for defining AOIs. These indicators can be used to detect changes in

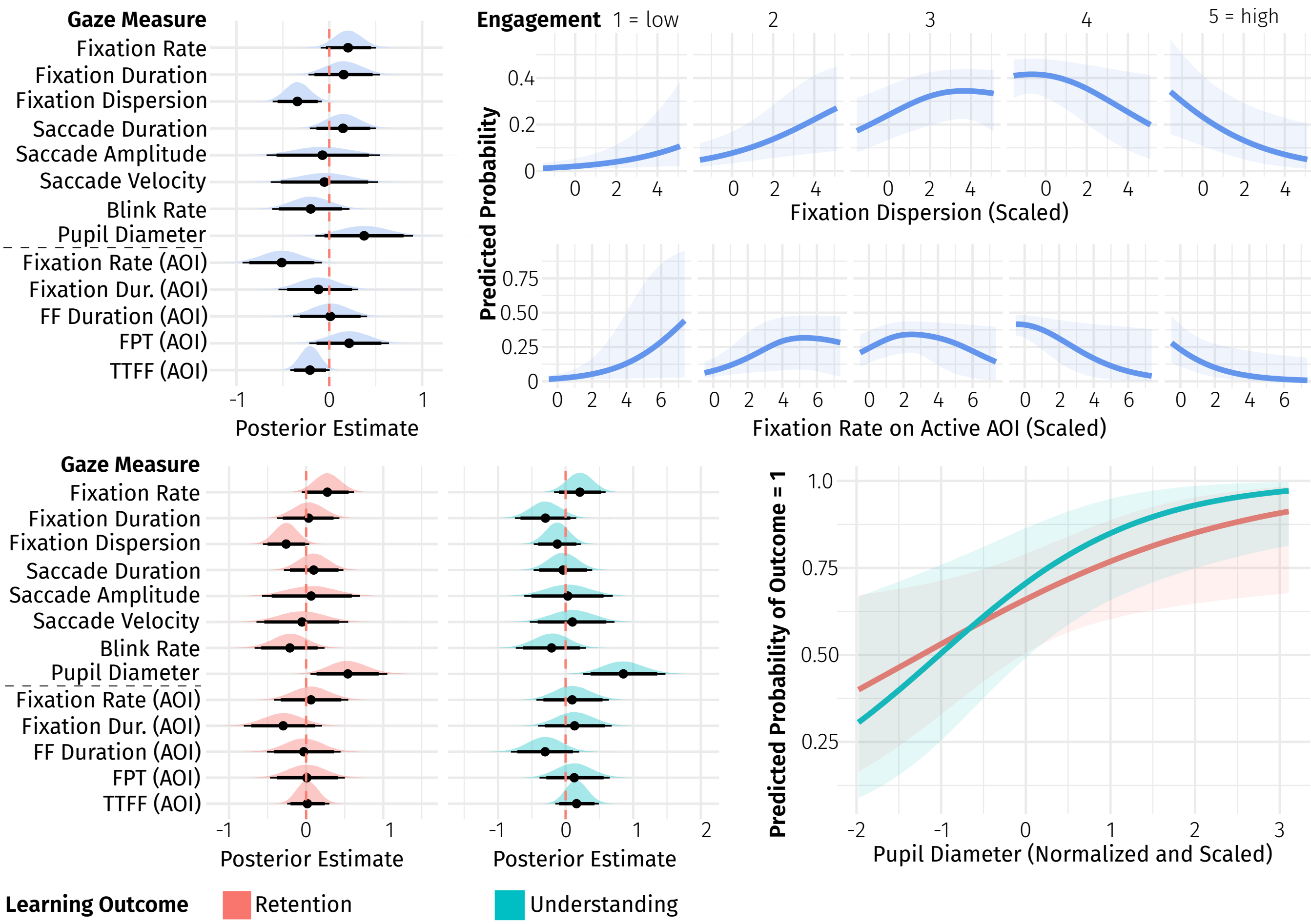

Figure 2: Posterior estimates of gaze measures as predictors of engagement (top) and learning outcomes (bottom), plotted with tidybayes [44]. Dots represent the mean posterior estimates, and horizontal lines indicate 89% and 95% highest density interval (HDI). The AOI-based measures are labeled with (AOI) (*FF: first fixation, FPT: first-pass time, TTFF: time to first fixation*). The predicted probabilities are provided for engagement levels and for answering the question correctly. The shaded area represents the standard error.

engagement during learning and trigger adaptive responses, such as adjusting pacing [15] or pause the video when engagement is low and offer learners to reflect [7, 60].

Gaze Measures as Indicators for Engagement. The analysis of collected data identifies fixation dispersion and fixation rate on active AOIs as promising gaze measures as indicators for engagement. Learners who participated in our data collection showed smaller distances from the distribution of fixations and fewer fixations on active AOIs when more engaged, in contrast to larger distances and more fixations on active AOIs when less engaged. These findings contradict earlier assumptions suggesting that fixation counts on relevant areas (active AOIs) increase with greater cognitive engagement, reflecting deeper information processing [16]. A potential explanation relates to the definition of AOIs. We defined active AOIs as areas that appeared recently or were verbally referenced by the speaker. We hypothesize that highly engaged learners noticed these active AOIs rapidly but quickly processed the information and redirected their attention back to the overall information presented in the video.

Contrary to expectations, we did not find that, besides the two, other gaze measures indicate engagement or learning outcomes, even though previous educational research has shown that measures like first pass time, first fixation duration, and time to first fixation reflect learners' processes of selecting and organizing information in multimedia learning [16]. For example, we expected that learners with higher engagement would have shorter times to first fixation on active AOIs, since we assumed that appearing AOIs might guide learners' attention. Reasons for that may include the fact that many internal processes, such as engagement, reflected in eye movements in controlled setups, might be overshadowed by external influences, such as the visual appearance of the learning content, which are inevitably present in actual VBL.

Gaze Measures as Indicators for Learning Outcomes. Regarding learning outcomes, our analysis indicates that learners in our data collection exhibit larger pupil sizes when achieving higher learning outcomes, as measured by both retention and understanding questions. This result is consistent with prior research demonstrating the utility of pupillary activity as a marker of cognitive load [14, 46]. In particular, larger pupil diameters have been linked to increased cognitive demand. Recent work on video-based online lectures further supports this finding, showing that pupil measures are valuable for estimating learners' subjective comprehension levels [85].

Eye-Tracking Challenges & Privacy. While advancements in eye-tracking hardware and software facilitate the collection and real-time processing of high-quality eye-tracking data, it remains a technically complex task that depends on the eye tracker and data pipeline used, as well as on parameter choices. For instance, the identification of eye movement events, the basis for calculating gaze measures, depends on data quality, the identification algorithm, and its parameters used [3]. Depending on the availability of eye-openness data from the eye tracker used, blinks have to be classified from data loss, which increases the risk of false negatives like partial blinks [59]. Further parameters in the data processing pipeline, such as window size, can affect classifier performance [39], and external factors, such as the visual appearance of the learning content, which are inevitably present in actual VBL, can also affect eye-tracking data (e.g., pupil size).

Beyond these technical challenges, gaze data can reveal sensitive information about learners, and they may not be fully aware of these risks. Therefore, gaze-based adaptive learning systems should consider privacy-aware processing of gaze data [73].

Ground Truth Challenge. For supervised classifier implementation, reliable ground-truth labels across the full range of the desired variable are required. However, in the case of engagement, subjectively rated engagement may not accurately reflect learners' actual engagement [24], and since learning materials are designed to engage learners, data points indicating low engagement are less common than those indicating immediate and high engagement (see distribution of engagement labels in the DAiSEE dataset [31]). Approaches focusing on assessing disengagement (e.g., mind wandering [38]) may intervene when learners disengage, but are limited in detecting moments of high engagement, which may be reflected to learners as an achievement to motivate them. While stimulated recall has been shown to provide reliable ground-truth labels [63], we propose a multimodal approach that complements it with observer-based engagement annotations for recorded video data, as commonly employed in prior datasets [43, 57].

5.2 Opportunities of EnGazement

The EnGazement dataset provides a valuable resource for researching gaze-based measures for adaptive learning systems and learning analytics. It enables reproducibility by allowing comparisons of eye-tracking data collected across different learning materials and video production styles. Educational researchers can use it to investigate how learners' prior knowledge, verbal references, and visual design influence attention allocation during learning. For instance, by comparing scan paths of learners with high against low prior knowledge. The dataset can be used to train classifiers for adaptive learning systems, for example, to predict learners' engagement and learning outcomes or to simulate gaze behavior in VBL tasks. Moreover, EnGazement can serve as an exploratory playground for investigating novel gaze measures that may more accurately indicate learners' engagement and learning outcomes. Finally, the dataset can support multimodal modeling by aligning gaze data with visual or audio features from learning videos, thus helping isolate external content-driven effects from internal cognitive states.

5.3 Limitations

Our study is not without limitations We collected data from a sample of N = 18 learners, all university students in STEM fields, which may limit the generalizability of our findings. Although we used authentic learning videos selected from a popular video-sharing platform, they were similar in production style and primarily targeted medical students, restricting the diversity in video production style [30]. Furthermore, due to practical constraints, our study did not include truly in-the-wild learning environments. Future work could address these limitations by collecting webcam-based eye-tracking data to reach a more diverse set of content and learners, and by replicating our findings with noisier eye-tracking data [47]. Our dataset is also limited by the above-mentioned ground-truth challenge. Low-engagement labels in the dataset are underrepresented, which may limit the generalizability of our findings to disengaged or less attentive learners. Our dataset includes only self-reported engagement measures, which may limit the reliability of our analyses and should be considered when interpreting the results or using the dataset to train classifiers.

6 Conclusion

This paper presented an exploratory analysis of the associations between commonly used gaze measures, learner engagement, and learning outcomes in a VBL context. The analysis was conducted on a dataset comprising over 1.3 million processed eye-tracking data points, derived gaze measures, and ground-truth labels for engagement and learning outcomes collected from 18 learners watching two learning videos. The results indicate that both an AOI-based measure, fixation rate on active AOIs, and an AOI-independent measure, fixation dispersion, are associated with learner engagement, whereas pupil diameter is associated with learning outcomes. From a design perspective, these findings suggest that adaptive learning systems may rely on pupil diameter as a parameter for modeling learning-related outcomes, while modeling learners' engagement can be supported either by fixation dispersion when gaze data across learners are available or by fixation rate on active AOIs when such data are not available. The latter approach, however, requires reliable methods for defining areas of interest in VBL materials. To support reproducibility and further investigation, the EnGazement dataset is openly available. By providing a comprehensive set of processed gaze measures alongside engagement and learning labels, the dataset enables future work on gaze-based user modeling for adaptive learning systems to support learners' learning processes.

Acknowledgments

We would like to thank the reviewers for their insightful and valuable feedback, Clara Schramm for assistance with the data collection, Lukas Steinwender for his initial contributions to defining areas of interest, and all study participants. This research was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through the project "Playing Eyes: Understanding the Communicative Power of Gaze Expressions when Bluffing and Being Confident in Winning" (project number: 563076722). T.H. was supported by the Pioneer Centre for AI Denmark, DNRF grant number P1. We acknowledge the use of Grammarly and LLM-based AI assistants (ChatGPT, Claude) to identify and improve spelling, grammar, punctuation, and clarity, and to suggest paraphrasing. All AI-generated suggestions were reviewed and adapted by the authors; no AI-generated text was used verbatim, and the original content and meaning were preserved throughout.

References

[1] Karan Ahuja, Dohyun Kim, Franceska Xhakaj, Virag Varga, Anne Xie, Stanley Zhang, Jay Eric Townsend, Chris Harrison, Amy Ogan, and Yuvraj Agarwal. 2019. EduSense: Practical Classroom Sensing at Scale. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (Sept. 2019), 71:1–71:26. doi:10.1145/3351229

[2] Ecenaz Alemdag and Kursat Cagiltay. 2018. A Systematic Review of Eye Tracking Research on Multimedia Learning. *Computers & Education* 125 (Oct. 2018), 413–428. doi:10.1016/j.compedu.2018.06.023

[3] Richard Andersson, Linnea Larsson, Kenneth Holmqvist, Martin Stridh, and Marcus Nyström. 2017. One Algorithm to Rule Them All? An Evaluation and Discussion of Ten Eye Movement Event-Detection Algorithms. *Behavior Research Methods* 49, 2 (April 2017), 616–637. doi:10.3758/s13428-016-0738-9

[4] Andrea Apicella, Pasquale Arpaia, Mirco Frosolone, Giovanni Improta, Nicola Moccaldi, and Andrea Pollastro. 2022. EEG-based measurement system for monitoring student engagement in learning 4.0. *Scientific Reports* 12, 1 (April 2022), 5857. doi:10.1038/s41598-022-09578-y

[5] Paul Ayres, Joy Yeonjoo Lee, Fred Paas, and Jeroen J. G. van Merriënboer. 2021. The Validity of Physiological Measures to Identify Differences in Intrinsic Cognitive Load. *Frontiers in Psychology* 12 (Sept. 2021). doi:10.3389/fpsyg.2021.702538

[6] Roger Azevedo and Allyson F. Hadwin. 2005. Scaffolding self-regulated learning and metacognition – Implications for the design of computer-based scaffolds. *Instructional Science* 33, 5/6 (2005), 367–379. https://www.jstor.org/stable/41953688

[7] Rachna Bachhel and Richa Ghay Thaman. 2014. Effective Use of Pause Procedure to Enhance Student Engagement and Learning. *Journal of Clinical and Diagnostic Research : JCDR* 8, 8 (Aug. 2014), XM01–XM03. doi:10.7860/JCDR/2014/8260.4691

[8] Nigel Bosch, Sidney D'Mello, Ryan Baker, Jaclyn Ocumpaugh, Valerie Shute, Matthew Ventura, Lubin Wang, and Weinan Zhao. 2015. Automatic Detection of Learning-Centered Affective States in the Wild. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15)*. Association for Computing Machinery, New York, NY, USA, 379–388. doi:10.1145/2678025.2701397

[9] Paul-Christian Bürkner. 2017. Brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software* 80 (Aug. 2017), 1–28. doi:10.18637/jss.v080.i01

[10] Paul-Christian Bürkner. 2021. Bayesian Item Response Modeling in R with Brms and Stan. *Journal of Statistical Software* 100 (Nov. 2021), 1–54. doi:10.18637/jss.v100.i05

[11] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. Stan: A Probabilistic Programming Language. *Journal of Statistical Software* 76 (Jan. 2017), 1–32. doi:10.18637/jss.v076.i01

[12] Benjamin T. Carter and Steven G. Luke. 2020. Best practices in eye tracking research. *International Journal of Psychophysiology* 155 (Sept. 2020), 49–62. doi:10.1016/j.ijpsycho.2020.05.010

[13] Siyuan Chen and Julien Epps. 2014. Using Task-Induced Pupil Diameter and Blink Rate to Infer Cognitive Load. *Human–Computer Interaction* 29, 4 (July 2014), 390–413. doi:10.1080/07370024.2014.892428

[14] Siyuan Chen, Julien Epps, Natalie Ruiz, and Fang Chen. 2011. Eye Activity as a Measure of Human Mental Effort in HCI. In *Proceedings of the 16th International Conference on Intelligent User Interfaces (IUI '11)*. Association for Computing Machinery, New York, NY, USA, 315–318. doi:10.1145/1943403.1943454

[15] Yu-Jung Chung, Chen-Wei Hsu, Meng-Hsun Chan, and Fu-Yin Cherng. 2024. Enhancing ESL Learners' Experience and Performance through Gradual Adjustment of Video Speed during Extensive Viewing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3613904.3642366

[16] Atakan Coskun and Kursat Cagiltay. 2022. A Systematic Review of Eye-Tracking-Based Research on Animated Multimedia Learning. *Journal of Computer Assisted Learning* 38, 2 (2022), 581–598. doi:10.1111/jcal.12629

[17] Asim H. Dar, Adina S. Wagner, and Michael Hanke. 2021. REMoDNaV: Robust Eye-Movement Classification for Dynamic Stimulation. *Behavior Research Methods* 53, 1 (Feb. 2021), 399–414. doi:10.3758/s13428-020-01428-x

[18] Snigdha Das, Sandip Chakraborty, and Bivas Mitra. 2022. I Cannot See Students Focusing on My Presentation; Are They Following Me? Continuous Monitoring of Student Engagement through "Stungage". In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '22)*. Association for Computing Machinery, New York, NY, USA, 243–253. doi:10.1145/3503252.3531307

[19] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. 2018. Unobtrusive Assessment of Students' Emotional Engagement during Lectures Using Electrodermal Activity Sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (Sept. 2018), 103:1–103:21. doi:10.1145/3264913

[20] Betsy DiSalvo, Dheeraj Bandaru, Qiaosi Wang, Hong Li, and Thomas Plötz. 2022. Reading the Room: Automated, Momentary Assessment of Student Engagement in the Classroom: Are We There Yet? *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (Sept. 2022), 112:1–112:26. doi:10.1145/3550328

[21] Sidney D'Mello, Ed Dieterle, and Angela Duckworth. 2017. Advanced, Analytic, Automated (AAA) Measurement of Engagement During Learning. *Educational psychologist* 52, 2 (2017), 104–123. doi:10.1080/00461520.2017.1281747

[22] Jennifer A. Fredricks, Phyllis Blumenfeld, Jeanne Friedel, and Alison Paris. 2005. School Engagement. In *What Do Children Need to Flourish? Conceptualizing and Measuring Indicators of Positive Development*, Kristin Anderson Moore and Laura H. Lippman (Eds.). Springer US, Boston, MA, 305–321. doi:10.1007/0-387-23823-9_19

[23] Jennifer A Fredricks, Phyllis C Blumenfeld, and Alison H Paris. 2004. School Engagement: Potential of the Concept, State of the Evidence. *Review of Educational Research* 74, 1 (March 2004), 59–109. doi:10.3102/00346543074001059

[24] Nan Gao, Mohammad Saiedur Rahaman, Wei Shao, and Flora D Salim. 2021. Investigating the Reliability of Self-report Data in the Wild: The Quest for Ground Truth. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers (UbiComp/ISWC '21 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 237–242. doi:10.1145/3460418.3479338

[25] Nan Gao, Wei Shao, Mohammad Saiedur Rahaman, and Flora D. Salim. 2020. N-Gage: Predicting in-Class Emotional, Behavioural and Cognitive Engagement in the Wild. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (Sept. 2020), 79:1–79:26. doi:10.1145/3411813

[26] Eric Granholm, Robert F. Asarnow, Andrew J. Sarkin, and Karen L. Dykes. 1996. Pupillary Responses Index Cognitive Resource Limitations. *Psychophysiology* 33, 4 (1996), 457–461. doi:10.1111/j.1469-8986.1996.tb01071.x

[27] Jeffrey A. Greene, Matthew L. Bernacki, and Allyson F. Hadwin. 2023. Self-Regulation. In *Handbook of Educational Psychology* (4 ed.). Routledge.

[28] Jesse W. Grootjen, Philipp Thalhammer, and Thomas Kosch. 2024. Your Eyes on Speed: Using Pupil Dilation to Adaptively Select Speed-Reading Parameters in Virtual Reality. *Proc. ACM Hum.-Comput. Interact.* 8, MHCI (Sept. 2024), 284:1–284:17. doi:10.1145/3676531

[29] Jesse W. Grootjen, Henrike Weingärtner, and Sven Mayer. 2024. Uncovering and Addressing Blink-Related Challenges in Using Eye Tracking for Interactive Systems. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–23. doi:10.1145/3613904.3642086

[30] Philip J. Guo, Juho Kim, and Rob Rubin. 2014. How Video Production Affects Student Engagement: An Empirical Study of MOOC Videos. In *Proceedings of the First ACM Conference on Learning @ Scale Conference (L@S '14)*. Association for Computing Machinery, New York, NY, USA, 41–50. doi:10.1145/2556325.2566239

[31] Abhay Gupta, Arjun D'Cunha, Kamal Awasthi, and Vineeth Balasubramanian. 2022. DAiSEE: Towards User Engagement Recognition in the Wild. arXiv:1609.01885 [cs] doi:10.48550/arXiv.1609.01885

[32] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*, Peter A. Hancock and Najmedin Meshkati (Eds.). Human Mental Workload, Vol. 52. North-Holland, 139–183. doi:10.1016/S0166-4115(08)62386-9

[33] Larry V. Hedges. 2007. Effect Sizes in Cluster-Randomized Designs. *Journal of Educational and Behavioral Statistics* 32, 4 (Dec. 2007), 341–370. doi:10.3102/1076998606298043

[34] Curtis R. Henrie, Lisa R. Halverson, and Charles R. Graham. 2015. Measuring Student Engagement in Technology-Mediated Learning: A Review. *Computers & Education* 90 (Dec. 2015), 36–53. doi:10.1016/j.compedu.2015.09.005

[35] Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost van de Weijer. 2011. *Eye Tracking: A Comprehensive*

Guide to Methods and Measures. OUP Oxford.
[36] Stephen Hutt, Jessica Hardey, Robert Bixler, Angela Stewart, Evan Risko, and Sidney K. D'Mello. 2017. *Gaze-Based Detection of Mind Wandering during Lecture Viewing*. Technical Report. International Educational Data Mining Society. https://eric.ed.gov/?id=ED596576
[37] Stephen Hutt, Kristina Krasich, Caitlin Mills, Nigel Bosch, Shelby White, James R. Brockmole, and Sidney K. D'Mello. 2019. Automated Gaze-Based Mind Wandering Detection during Computerized Learning in Classrooms. *User Modeling and User-Adapted Interaction* 29, 4 (Sept. 2019), 821–867. doi:10.1007/s11257-019-09228-5
[38] Stephen Hutt, Caitlin Mills, Nigel Bosch, Kristina Krasich, James Brockmole, and Sidney D'Mello. 2017. "Out of the Fr-Eye-ing Pan": Towards Gaze-Based Models of Attention during Learning with Technology in the Classroom. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP '17)*. Association for Computing Machinery, New York, NY, USA, 94–103. doi:10.1145/3079628.3079669
[39] Stephen Hutt, Caitlin Mills, Shelby White, Patrick J. Donnelly, and Sidney K. D'Mello. 2016. *The Eyes Have It: Gaze-Based Detection of Mind Wandering during Learning with an Intelligent Tutoring System*. Technical Report. International Educational Data Mining Society. https://eric.ed.gov/?id=ED592729
[40] Sanna Järvelä, Jonna Malmberg, Eetu Haataja, Marta Sobocinski, and Paul A. Kirschner. 2021. What Multimodal Data Can Tell Us about the Students' Regulation of Their Learning Process? *Learning and Instruction* 72 (April 2021), 101203. doi:10.1016/j.learninstruc.2019.04.004
[41] Marcel Adam Just and Patricia A Carpenter. 1976. Eye Fixations and Cognitive Processes. *Cognitive Psychology* 8, 4 (Oct. 1976), 441–480. doi:10.1016/0010-0285(76)90015-3
[42] Marcel A. Just and Patricia A. Carpenter. 1980. A Theory of Reading: From Eye Fixations to Comprehension. *Psychological Review* 87 (1980), 329–354. doi:10.1037/0033-295X.87.4.329
[43] Amanjot Kaur, Aamir Mustafa, Love Mehta, and Abhinav Dhall. 2018. Prediction and Localization of Student Engagement in the Wild. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*. 1–8. doi:10.1109/DICTA.2018.8615851
[44] Matthew Kay. 2024. Tidybayes: Tidy Data and Geoms for Bayesian Models. Zenodo. doi:10.5281/zenodo.13770114
[45] Agnes Mercedes Kloft, Robin Welsch, Thomas Kosch, and Steeven Villa. 2024. "AI Enhances Our Performance, I Have No Doubt This One Will Do the Same": The Placebo Effect Is Robust to Negative Descriptions of AI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–24. doi:10.1145/3613904.3642633
[46] Thomas Kosch, Jakob Karolus, Johannes Zagermann, Harald Reiterer, Albrecht Schmidt, and Paweł W. Woźniak. 2023. A Survey on Measuring Cognitive Workload in Human-Computer Interaction. *Comput. Surveys* 55, 13s (July 2023), 283:1–283:39. doi:10.1145/3582272
[47] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. 2016. Eye Tracking for Everyone. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2176–2184. https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Krafka_Eye_Tracking_for_CVPR_2016_paper.html
[48] David R. Krathwohl. 2002. A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice* 41, 4 (Nov. 2002), 212–218. doi:10.1207/s15430421tip4104_2
[49] Jacob Leon Kröger, Otto Hans-Martin Lutz, and Florian Müller. 2020. What Does Your Gaze Reveal About You? On the Privacy Implications of Eye Tracking. In *Privacy and Identity Management. Data for Better Living: AI and Privacy: 14th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2.2 International Summer School, Windisch, Switzerland, August 19–23, 2019, Revised Selected Papers*, Michael Friedewald, Melek Önen, Eva Lievens, Stephan Krenn, and Samuel Fricker (Eds.). Springer International Publishing, Cham, 226–241. doi:10.1007/978-3-030-42504-3_15
[50] Matias Laporte, Martin Gjoreski, and Marc Langheinrich. 2023. LAUREATE: A Dataset for Supporting Research in Affective Computing and Human Memory Augmentation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (Sept. 2023), 106:1–106:41. doi:10.1145/3610892
[51] Dominique Makowski, Mattan S. Ben-Shachar, and Daniel Lüdecke. 2019. bayestestR: Describing Effects and Their Uncertainty, Existence and Significance within the Bayesian Framework. *Journal of Open Source Software* 4, 40 (Aug. 2019), 1541. doi:10.21105/joss.01541
[52] João Maroco, Ana Lúcia Maroco, Juliana Alvares Duarte Bonini Campos, and Jennifer A. Fredricks. 2016. University Student's Engagement: Development of the University Student Engagement Inventory (USEI). *Psicologia: Reflexão e Crítica* 29 (June 2016). doi:10.1186/s41155-016-0042-8
[53] Richard E. Mayer. 2020. Science of Assessment: Determining What Was Learned in Multimedia Learning. In *Multimedia Learning* (third ed.). Cambridge University Press, 95–116. https://www.cambridge.org/highereducation/books/multimedia-learning/FB7E79A165D24D47CEACEB4D2C426ECD/science-of-assessment-determining-what-was-learned-in-multimedia-learning/88AD50F758EAA8DDCEE3DF9BFB71F393
[54] Beatriz Mello, Robin Welsch, Marissa Christien Verbokkem, Pascal Knierim, and Martin Johannes Dechant. 2024. Navigating the Virtual Gaze: Social Anxiety's Role in VR Proxemics. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3613904.3642359
[55] Brian W. Miller. 2015. Using Reading Times and Eye-Movements to Measure Cognitive Engagement. *Educational Psychologist* 50, 1 (Jan. 2015), 31–42. doi:10.1080/00461520.2015.1004068
[56] Omid Mohamad Nezami, Mark Dras, Len Hamey, Deborah Richards, Stephen Wan, and Cécile Paris. 2020. Automatic Recognition of Student Engagement Using Deep Learning and Facial Expression. In *Machine Learning and Knowledge Discovery in Databases*, Ulf Brefeld, Elisa Fromont, Andreas Hotho, Arno Knobbe, Marloes Maathuis, and Céline Robardet (Eds.). Springer International Publishing, Cham, 273–289. doi:10.1007/978-3-030-46133-1_17
[57] Hamed Monkaresi, Nigel Bosch, Rafael A. Calvo, and Sidney K. D'Mello. 2017. Automated Detection of Engagement Using Video-Based Estimation of Facial Expressions and Heart Rate. *IEEE Transactions on Affective Computing* 8, 1 (Jan. 2017), 15–28. doi:10.1109/TAFFC.2016.2515084
[58] Roxana Moreno and Richard Mayer. 2007. Interactive Multimodal Learning Environments. *Educational Psychology Review* 19, 3 (Sept. 2007), 309–326. doi:10.1007/s10648-007-9047-2
[59] Marcus Nyström, Richard Andersson, Diederick C. Niehorster, Roy S. Hessels, and Ignace T. C. Hooge. 2024. What Is a Blink? Classifying and Characterizing Blinks in Eye Openness Signals. *Behavior Research Methods* (Feb. 2024). doi:10.3758/s13428-023-02333-9
[60] Heather L. O'Brien, Ido Roll, Andrea Kampen, and Nilou Davoudi. 2022. Rethinking (Dis)engagement in human-computer interaction. *Computers in Human Behavior* 128 (March 2022), 107109. doi:10.1016/j.chb.2021.107109
[61] Jaclyn Ocumpaugh, Ryan S Baker, and Ma Mercedes T Rodrigo. 2015. Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual. (2015).
[62] Prasoon Patidar, Tricia J. Ngoon, John Zimmerman, Amy Ogan, and Yuvraj Agarwal. 2024. ClassID: Enabling Student Behavior Attribution from Ambient Classroom Sensing Systems. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2 (May 2024), 55:1–55:28. doi:10.1145/3659586
[63] Effie J. Pereira, Samantha Ayers-Glassey, Jeffrey D. Wammes, and Daniel Smilek. 2023. Attention in Hindsight: Using Stimulated Recall to Capture Dynamic Fluctuations in Attentional Engagement. *Behavior Research Methods* (Nov. 2023). doi:10.3758/s13428-023-02273-4
[64] Celeste Pilegard and Richard E. Mayer. 2015. Adding Judgments of Understanding to the Metacognitive Toolbox. *Learning and Individual Differences* 41 (July 2015), 62–72. doi:10.1016/j.lindif.2015.07.002
[65] Johnmarshall Reeve and Ching-Mei Tseng. 2011. Agency as a Fourth Aspect of Students' Engagement during Learning Activities. *Contemporary Educational Psychology* 36, 4 (Oct. 2011), 257–267. doi:10.1016/j.cedpsych.2011.05.002
[66] Falko Rheinberg, Regina Vollmeyer, and Bruce D. Burns. 2001. FAM: Ein Fragebogen Zur Erfassung Aktuller Motivation in Lern- Und Leistungssituationen. *Diagnostica* 47, 2 (April 2001), 57–66. doi:10.1026//0012-1924.47.2.57
[67] Daniel L. Schacter and Karl K. Szpunar. 2015. Enhancing Attention and Memory during Video-Recorded Lectures. *Scholarship of Teaching and Learning in Psychology* 1, 1 (2015), 60–71. doi:10.1037/stl0000011
[68] Katharina Scheiter and Alexander Eitel. 2017. The Use of Eye Tracking as a Research and Instructional Tool in Multimedia Learning. In *Eye-Tracking Technology Applications in Educational Research*. IGI Global, 143–164. doi:10.4018/978-1-5225-1005-5.ch008
[69] Gale M. Sinatra, Benjamin C. Heddy, and Doug Lombardi. 2015. The Challenges of Defining and Measuring Student Engagement in Science. *Educational Psychologist* 50, 1 (Jan. 2015), 1–13. doi:10.1080/00461520.2014.1002924
[70] Monisha Singh, Ximi Hoque, Donghuo Zeng, Yanan Wang, Kazushi Ikeda, and Abhinav Dhall. 2023. Do I Have Your Attention: A Large Scale Engagement Prediction Dataset and Baselines. In *Proceedings of the 25th International Conference on Multimodal Interaction (ICMI '23)*. Association for Computing Machinery, New York, NY, USA, 174–182. doi:10.1145/3577190.3614164
[71] Ellen A. Skinner, Thomas A. Kindermann, and Carrie J. Furrer. 2009. A Motivational Perspective on Engagement and Disaffection: Conceptualization and Assessment of Children's Behavioral and Emotional Participation in Academic Activities in the Classroom. *Educational and Psychological Measurement* 69, 3 (June 2009), 493–525. doi:10.1177/0013164408323233
[72] Jonathan Smallwood, Daniel J. Fishman, and Jonathan W. Schooler. 2007. Counting the Cost of an Absent Mind: Mind Wandering as an Underrecognized Influence on Educational Performance. *Psychonomic Bulletin & Review* 14, 2 (April 2007), 230–236. doi:10.3758/BF03194057
[73] Julian Steil, Inken Hagestedt, Michael Xuelin Huang, and Andreas Bulling. 2019. Privacy-aware eye tracking using differential privacy. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications (ETRA '19)*. Association for Computing Machinery, New York, NY, USA, 1–9. doi:10.1145/3314111.3319915
[74] Anselm R. Strohmaier, Kelsey J. MacKay, Andreas Obersteiner, and Kristina M. Reiss. 2020. Eye-Tracking Methodology in Mathematics Education Research: A Systematic Literature Review. *Educational Studies in Mathematics* 104, 2 (June 2020), 147–200. doi:10.1007/s10649-020-09948-1

[75] Thitaree Tanprasert, Sidney S Fels, Luanne Sinnamon, and Dongwook Yoon. 2023. Scripted Vicarious Dialogues: Educational Video Augmentation Method for Increasing Isolated Students' Engagement. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–25. doi:10.1145/3544548.3581153

[76] UNESCO Institute for Statistics. 2015. *International Standard Classification of Education: Fields of Education and Training 2013 (ISCED-F 2013) Detailed Field Descriptions*. UNESCO Institute for Statistics. doi:10.15220/978-92-9189-179-5-en

[77] Tamara van Gog and Katharina Scheiter. 2010. Eye Tracking as a Tool to Study and Enhance Multimedia Learning. *Learning and Instruction* 20, 2 (April 2010), 95–99. doi:10.1016/j.learninstruc.2009.02.009

[78] Aki Vehtari. 2025. *Cross-validation FAQ*. Retrieved April 29, 2025 from https://users.aalto.fi/%7Eave/CV-FAQ.html

[79] Aki Vehtari, Andrew Gelman, and Jonah Gabry. 2017. Practical Bayesian Model Evaluation Using Leave-One-out Cross-Validation and WAIC. *Statistics and Computing* 27, 5 (Sept. 2017), 1413–1432. doi:10.1007/s11222-016-9696-4

[80] Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. 2021. Rank-Normalization, Folding, and Localization: An Improved R^ for Assessing Convergence of MCMC (with Discussion). *Bayesian Analysis* 16, 2 (June 2021), 667–718. doi:10.1214/20-BA1221

[81] Steeven Villa, Thomas Kosch, Felix Grelka, Albrecht Schmidt, and Robin Welsch. 2023. The Placebo Effect of Human Augmentation: Anticipating Cognitive Augmentation Increases Risk-Taking Behavior. *Computers in Human Behavior* 146 (Sept. 2023), 107787. doi:10.1016/j.chb.2023.107787

[82] Regina Vollmeyer and Falko Rheinberg. 2006. Motivational Effects on Self-Regulated Learning with Different Tasks. *Educational Psychology Review* 18, 3 (Sept. 2006), 239–253. doi:10.1007/s10648-006-9017-0

[83] Samangi Wadinambiarachchi, Ryan M. Kelly, Saumya Pareek, Qiushi Zhou, and Eduardo Velloso. 2024. The Effects of Generative AI on Design Fixation and Divergent Thinking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3613904.3642919

[84] Jeffrey D. Wammes, Brandon C. W. Ralph, Caitlin Mills, Nigel Bosch, Tracy L. Duncan, and Daniel Smilek. 2019. Disengagement during Lectures: Media Multitasking and Mind Wandering in University Classrooms. *Computers & Education* 132 (April 2019), 76–89. doi:10.1016/j.compedu.2018.12.007

[85] Ko Watanabe, Gitesh Gund, Jayasankar Santhosh, Haruka Sakagami, Yuki Matsuda, Andreas Dengel, and Shoya Ishimaru. 2025. EyeUnderstand: Dashboard for Gaze and Deep-Learning Driven Comprehension Estimation in Online Lectures. *IEEE Access* 13 (2025), 102220–102233. doi:10.1109/ACCESS.2025.3578185

[86] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R. Movellan. 2014. The Faces of Engagement: Automatic Recognition of Student Engagementfrom Facial Expressions. *IEEE Transactions on Affective Computing* 5, 1 (Jan. 2014), 86–98. doi:10.1109/TAFFC.2014.2316163

[87] Jacob O. Wobbrock and Julie A. Kientz. 2016. Research Contributions in Human-Computer Interaction. *Interactions* 23, 3 (April 2016), 38–44. doi:10.1145/2907069

[88] Johannes Zagermann, Ulrike Pfeil, and Harald Reiterer. 2016. Measuring Cognitive Load Using Eye Tracking Technology in Visual Computing. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization (BELIV '16)*. Association for Computing Machinery, New York, NY, USA, 78–85. doi:10.1145/2993901.2993908

[89] Han Zhang, Kevin F. Miller, Xin Sun, and Kai S. Cortina. 2020. Wandering Eyes: Eye Movements during Mind Wandering in Video Lectures. *Applied Cognitive Psychology* 34, 2 (2020), 449–464. doi:10.1002/acp.3632