Assessing Medical Training Skills via Eye and Head Movements

Kayhan Latifzadeh*
University of Luxembourg
Luxembourg, Luxembourg
kayhan.latifzadeh@uni.lu

Luis A. Leiva*
University of Luxembourg
Luxembourg, Luxembourg
luis.leiva@uni.lu

Klen Čopič Pucihar*
University of Primorska
Koper, Slovenia
Stellenbosch University
Stellenbosch, South Africa
klen.copic@famnit.upr.si

Matjaž Kljun
University of Primorska
Koper, Slovenia
Stellenbosch University
Stellenbosch, South Africa
matjaz.kljun@upr.si

Iztok Devetak
University of Ljubljana
Ljubljana, Slovenia
iztok.devetak@pef.uni-lj.si

Lili Steblovnik*
University Medical Centre
Ljubljana
Ljubljana, Slovenia
lili.steblovnik@mf.uni-lj.si

Abstract

We examined eye and head movements to gain insights into skill development in clinical settings. A total of 24 practitioners participated in simulated baby delivery training sessions. We calculated key metrics, including pupillary response rate, fixation duration, or angular velocity. Our findings indicate that eye and head tracking can effectively differentiate between trained and untrained practitioners, particularly during labor tasks. For example, head-related features achieved an F1 score of 0.85 and AUC of 0.86, whereas pupil-related features achieved F1 score of 0.77 and AUC of 0.85. The results lay the groundwork for computational models that support implicit skill assessment and training in clinical settings by using commodity eye-tracking glasses as a complementary device to more traditional evaluation methods such as subjective scores.

Keywords

eye movements, head movements, simulation training

ACM Reference Format:
Kayhan Latifzadeh, Luis A. Leiva, Klen Čopič Pucihar, Matjaž Kljun, Iztok Devetak, and Lili Steblovnik. 2025. Assessing Medical Training Skills via Eye and Head Movements. In *33rd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '25), June 16–19, 2025, New York City, NY, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3699682.3728330

1 Introduction

Simulation based training plays a crucial role in preparing specialized medical professionals by providing a safe environment for skill development and hands-on experience [11, 34, 61]. This includes both technical and non-technical skills, such as communication and teamwork [8, 44]. Practicing these skills in a controlled setting reinforces muscle memory and builds confidence, which can lead to improved performance in real-life situations.

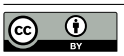
In addition, simulated medical environments enable instructors to deliver immediate feedback, supporting reflective learning. This process encourages practitioners to critically examine their experiences, thoughts, and responses, allowing them to gain deeper insights and improve future performance. This method facilitates more effective learning compared to or in combination with traditional classroom-based learning. However, it requires continuous observation, which can be time-intensive and prone to human biases [10, 17, 45].

Nowadays, integrating biosignals into simulation training offers several benefits that can enhance learning outcomes and overall training effectiveness [3, 18]. These signals offer an implicit and objective means of skill assessment, providing faster and less biased evaluation. Among the various biosignals, eye and head movement tracking are commonly used, as they can be collected unobtrusively using lightweight wearable devices such as eye and head tracking glasses [16, 23, 26, 37, 39]. Despite numerous studies exploring the use of eye tracking and head movement in simulation training, the impact of different features from eye (e.g., pupil size, fixation, and saccades) and head movements (velocity and rotation) is still poorly understood. Since these factors may influence different tasks in varying ways, further asessment is necessary [52].

This study investigates the role of eye and head movements in assessing skill acquisition during simulation based training for medical professionals in the context of breech delivery–a childbirth scenario that has not yet been explored. By addressing this gap, we contribute insights into how eye and head tracking technologies can enhance both the assessment and training of healthcare professionals, ultimately leading to improved patient care during childbirth. Our key contributions include:

- Exploration of the potential of eye and head movements as indicators of skill acquisition in medical training.
- Identification and analysis of key metrics from eye and head movements for this purpose.
- Introduction of an eye/head movements dataset [1] of 48 breech delivery procedures, worth of over 8 h of movement data, featuring annotated time segments and post-session skill scores. We also release python scripts, and preprocessed data [2].

*Authors contributed equally to this research.



UMAP '25, New York City, NY, USA

ACM ISBN 979-8-4007-1313-2/2025/06
https://doi.org/10.1145/3699682.3728330

[1] https://zenodo.org/records/15163456
[2] https://github.com/kayhan-latifzadeh/LaborTrack

2 Background and related work

Eye-tracking glasses have become a valuable tool in medical research [6, 19, 28, 59], providing insights into the performance of healthcare professionals during task execution. These devices enable researchers to evaluate key metrics such as fixations, saccades, or pupil size, which have been, for example, linked to decision-making [2] and attention [25]. These metrics can be used for modeling skill development and adapting content in medical training. For example, studies have demonstrated that healthcare professionals can use eye-tracking metrics to identify areas for improvement and optimize simulation-based learning procedures for both novice and expert practitioners [24, 46, 47]. In this section we review related work on eye and head tracking in clinical settings, and outline the research hypotheses we aim to validate through a user study.

2.1 Eye tracking

Traditionally, eye tracking research has used Areas of Interest (AOI) analysis to evaluate how individuals visually engage with specific regions within a given environment [41, 42, 57]. Metrics such as Time to First Fixation (TFF), fixation counts, and dwell time have been effective in assessing attention [36], perception [51], and decision-making [43, 54]. TFF, in particular, has proved important in medical settings for understanding decision-making and situational awareness. For instance, experienced emergency medicine residents in a simulated environment had a shorter TFF for the "ECG monitor" AOI (22 vs. 30 seconds) and focused more quickly on critical equipment like the pacing unit, improving emergency decision-making [51]. In neonatal care, a study of visual attention during positive pressure ventilation showed that the exhaled tidal volume waveform received the highest total gaze duration and visit count, compared to other respiratory function monitor parameters, indicating its perceived importance during the procedure [30] .

In a simulated echocardiography study [23], experts fixated earlier and spent longer dwell times on key AOIs, completing ultrasound exams faster than non-experts. Research on fourth-year medical students' non-technical skills in emergency care simulations [5] revealed that prolonged visual attention on the patient correlated negatively with leadership and communication, whereas a focused gaze on specific elements, like intravenous access, was linked to poorer decision-making and situational awareness.

In summary, AOI analysis plays an important role in understanding medical training, but annotating AOI is challenging, especially with eye-tracking glasses. For one, frequent head movements cause rapid shifts in the video stream's view orientation, complicating automated AOI annotation. Additionally, AOIs are environment-specific and cannot be applied across different simulations. Due to these limitations, our study does not focus on AOI-related metrics and instead focuses on more general eye and head tracking measures discussed hereafter.

2.1.1 Fixation-related metrics. Fixations are among the most widely used metrics in eye-tracking research, as they capture moments when the gaze remains steady on a single point. Both fixation duration (time spent fixating) and fixation count (number of fixations) have been linked to cognitive load [12, 21, 56]. Chen et al. [14] found that surgical residents who spent more time fixating on a feedback screen during needle insertion tasks performed better in later sessions. Similarly, Capogna et al. [13] observed that expert anesthesiologists performing epidural blocks had fewer but more precise fixations, completing procedures more efficiently, compared to novices. Another study [12] showed that hands-on training helped novice anesthesia trainees develop improved focus, leading to fewer but longer fixation duration, indicating enhanced precision.

Building on these findings, we propose the following hypothesis:

H1: Trained practitioners develop different fixation counts (**H1a**) and fixation durations (**H1b**) than untrained practitioners.

2.1.2 Cognition-related metrics. A metric for analyzing skill development called Task-Evoked Pupillary Response (TEPR) measures pupil dilation (increase in size) as an indicators of cognitive load. TERP has been used as a metric in studies involving medical professionals with different experience levels [50]. and has also been used for clinical performance assessment [35], highlighting its potential for user modeling. Another relevant metric is Eye Blink Rate (EBR), which has been shown to be a good proxy of cognitive flexibility [29], a key factor for problem-solving, creativity, and learning. Furthermore, variations in EBR have been shown to provide valuable information for assessing cognitive abilities [40]. Building on these findings, we propose the following hypothesis:

H2: Trained practitioners develop different TEPR (**H2a**) and EBR (**H2b**) than untrained practitioners.

2.1.3 Saccade-related metrics. Saccadic movements are rapid eye shift that allow redirecting focus from one point to another. During high precision tasks, saccade amplitude, velocity, and acceleration tend to decrease as the brain prioritizes accuracy and control over speed [7]. Kessler et al. [31] explored saccade-related metrics in a simulated neonatal intubation, tracking the visual focus of pediatric and neonatal practitioners. Their findings showed that more experienced practitioners demonstrated better visual attention and situational awareness, though training did not significantly enhance performance. Ahmadi et al. [1] monitored ICU nurses using Tobii Pro Glasses 2 and Empatica E4 devices during 12-hour shifts. They found that stress increased both gaze entropy and eye fixations, but reduced saccade duration and pupil diameter, particularly during high-stress periods like initial handoffs. Building on these findings, we propose the following hypothesis:

H3: Trained practitioners develop different saccade amplitude (**H3a**), saccade velocity (**H3b**), and saccade acceleration (**H3c**) than untrained practitioners.

2.2 Head tracking

Head movements are commonly characterized by acceleration. Viriyasiripong et al. [55] measured head movements during simulated laparoscopic suturing surgery and found that novices exhibited significantly higher acceleration than experts along both vertical and horizontal axes, which proved to be a useful metric to evaluate skill development. Another approach to describe head movement is angular velocity [60], which reflects the speed of head rotation. Additionally, cumulative rotation (the total amount of head movement during a task) may provide useful insights for user modeling [58]. Building on these findings, we propose the following hypothesis:

H4: Trained practitioners develop different angular velocity (**H4a**) and cumulative rotation (**H4b**) than untrained practitioners.

3 Method

Our aim was to assess skill acquisition during simulation training for breech deliveries, a childbirth scenario where a baby is born bottom-first instead of head-first. This training is part of the Training in Obstetric Emergencies (TUPS) program which is organized up to four times a year at Medical Simulation Centre at the University Medical Centre Ljubljana.

TUPS is aimed at specialists in gynecology, obstetrics, and anesthesiology, as well as qualified midwives and nurse anesthetists. The main goal is to teach professional skills for managing obstetric emergencies, emphasizing adherence to professional guidelines and technical executions of standardized procedures. The training lasts ten hours and is divided in thematic modules, where participants take part in various simulated scenarios. The breach delivery training follows the Simulation Education Model in Obstetrics–Pelvic Insertion Parturition (SIP-MV), introduced by Steblovnik et al. [48, 49], which promotes active participation in a breach delivery scenario in a simulated delivery room.

3.1 Participants

We recruited 24 practitioners from the TUPS training program. All participants already finished a 6-year medical degree, had prior experience of working in an obstetric room, and are currently at various stages of their 5-year specialization training program for gynecology ($M = 2.72$, $SD = 1.32$). All participants had normal or corrected-to-normal vision.

3.2 Apparatus

The breech delivery training was conducted in a simulated delivery room (Figure 1) which comprised a high-tech NOELLE® S550 manikin that breathes, delivers, speaks, and changes clinical parameters under guidance. The manikin was operated by an expert medical doctor, who was also an actor playing the role of a patient, voicing their concerns and hardships. The room was equipped with a CTG monitor, an infusion pump, a rotating chair on wheels, a trolley, and sterile equipment that is usually required for the procedure. Besides the expert doctor and the participant, a midwife was also present during simulation training (see Figure 1 right). The participant was wearing Tobii Pro Glasses 2, equipped with a Full HD scene camera.[3] Training sessions were also recorded by two cameras positioned on the ceiling, recording a 360 panoramic view of the room and a view overlooking the patient.

3.3 Procedure

The procedure began with a welcome and briefing on the training plan. Participants were then asked to sign a consent form and fill in a demographic questionnaire collecting information about specialization status, prior simulation training experience, and experience with breech delivery. Next, participants attended a lecture covering various aspects of breech delivery, providing theoretical knowledge and context before proceeding on to the simulation training.

Each participant completed two breech delivery simulation training sessions, both conducted in the same delivery room, with each session lasting about seven minutes. All of them, including the expert, wore masks as a Covid-19 prevention method as well as to preserve their anonymity during the recordings. Between the sessions, participants engaged in other TUPS-related activities that lasted approximately 60 minutes. Following each session, participants received approximately five minutes of feedback from an expert doctor, focusing on reflective learning.

Before each training session, the eye-tracking glasses were calibrated using the manufacturer's calibrating software. If a participant wore corrective glasses, we replaced them with specialized lenses that fit directly on the eye-tracking glasses.

3.4 Collected data

The eye-tracking glasses recorded data at 100 Hz, including pupil size, fixations, saccades, and blinks. They also captured head movements at 100 Hz using a built-in gyroscope and accelerometer, along with a video stream that enabled visualization of fixations throughout the training sessions. All recorded data had synchronized timestamps.

3.5 Task description

During the training sessions, participants took the role of a doctor overseeing a natural breech delivery that gets complicated. Throughout the procedure, the trainees must demonstrate a set of 14 skills:

- Introduce themselves to the patient (1).
- Gather patient's medical history (2).
- Vaginal examination (3).
- Prepare for delivery by explaining the procedure to the patient and inserting an intravenous line (4), cleaning the vaginal area (5), and placing a catheter (6).
- Determine the appropriate timing and dosage of Sintocinon to address the obstetric arrest (7).
- Administer analgesia (8) and perform an episiotomy if needed (9).
- Call in additional team members at the appropriate moment to assist with the labor (10).
- During labor, apply gentle pressure to the baby's buttocks to facilitate delivery (11). Identify the active delivery stage when baby's scapulas become visible (12), at which point the doctor must free the baby's hands (13) to enable the Bracht maneuver assisting birth, without pulling on the baby (14).

3.6 Analysis

Our goal was to assess the predictive power of eye and head tracking data in assessing medical skills during simulation training. To achieve this, we first needed to establish a benchmark with the help of an expert medical doctor. The doctor viewed all recordings of the training sessions, in no particular order, and graded each participant on a 1–5 scale for each of the 14 skills (see Section 3.5).

For data analysis, we needed a meaningful way to segment the training sessions into smaller time segments to assess changes in

[3] https://go.tobii.com/Glasses2UM

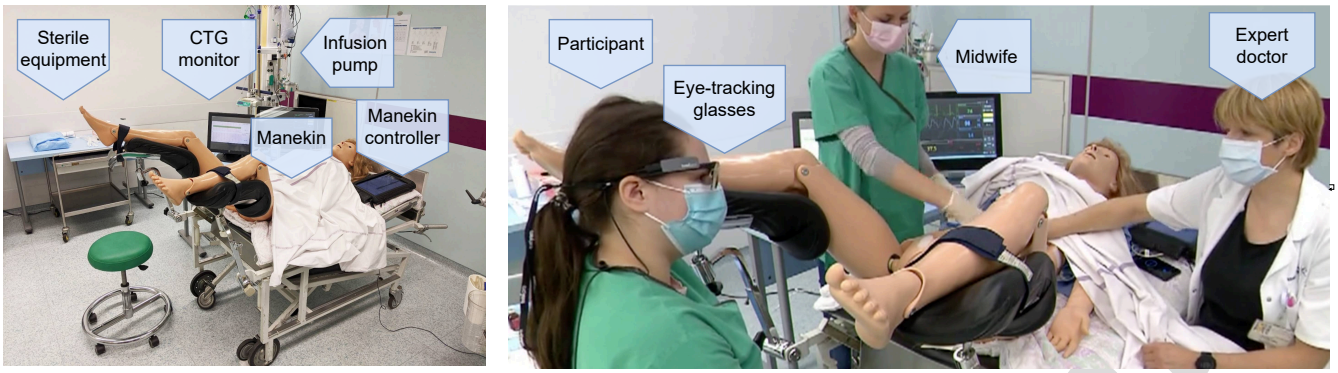


Figure 1: Apparatus and delivery room setup. Left: Setup with manikin, controller, CTG monitor, infusion pump, and sterile equipment on a trolley. Right: Expert doctor providing feedback to participant after simulation training.

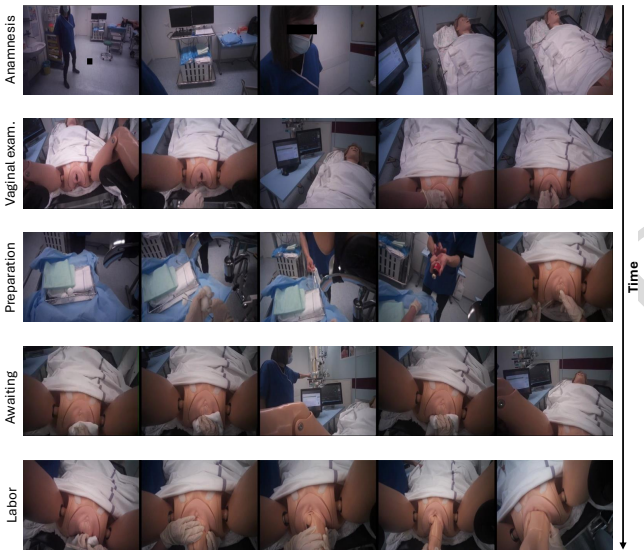


Figure 2: Example screenshots of a recorded video from different time segments in breech delivery, ordered chronologically from top to bottom.

the metrics defined in Background and related work. One approach is to create a new segment for each skill demonstrated. However, as some skills are demonstrated sequentially and others in parallel, sometimes within a very short time frame, we decided to group them under the guidance of the expert doctor. This process led to the identification of 5 key segments of breech delivery simulation training (see Figure 2), namely:

(1) *Anamnesis*: Checking patient's medical history (Skills: 1, 2).
(2) *Vaginal examination*: Assessing the readiness of the cervix and the baby's position (Skill: 3).
(3) *Preparation*: Setting up the delivery area, ensuring all necessary medical equipment is sterilized and the patient is ready for labor (Skills: 5, 6, 8, 9, 11).
(4) *Awaiting*: The time before going into labor, when instructions are given, intravenous line is inserted, analgesia is admitted, and episiotomy is done (Skills: 4, 8, 10).
(5) *Labor*: The actual process of delivering the baby (Skills: 7, 12, 13, 14)

4 Results

We first checked whether training led to any improvements in skill acquisition. We compared the scores assigned by the expert medical doctor to all participants in both sessions (Figure 3). Participants performed better in the second session ($M = 4.67$, $SD = 0.25$, $Mdn = 4.73$) compared to the first session ($M = 4.13$, $SD = 0.55$, $Mdn = 4.2$). A paired t-test (two-tails) revealed statistically significant differences: $t(23) = -5.08, p < .001$.

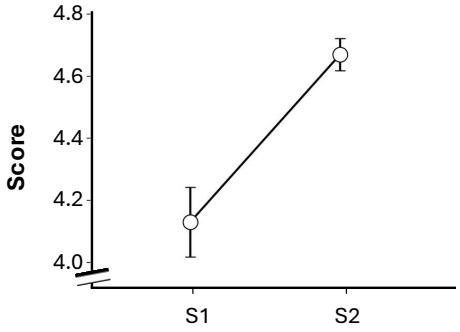


Figure 3: Skill scores between sessions (S1: first session, S2: second). Error bars denote standard error of the mean.

Next, we compared time spent by participants for each task. The results are summarized in Table 1. For Anamnesis, Vaginal examination, and Preparation segments, participants performed faster during the second session. A paired t-test (two-tails) was significant [Anamnesis: $t(23) = 3.80, p < .001$; Vaginal examination: $t(23) = 3.01, p < .01$; Preparation: $t(23) = 4.84, p < .001$]. No differences were found in the Awaiting segment: $t(23) = 0.39, p = .698$. Participants took significantly more time during Labor in the second session: $t(23) = -5.01,\ p < .001$.

Table 1: Duration (in seconds) of each segment for both training sessions (S1: first session, S2: second session).

Segment	Mean		Mdn		SD	
	S1	S2	S1	S2	S1	S2
Anamnesis	56.79	41.67	54.5	42	14.14	12.24
Vaginal exam.	44.08	31.08	40.5	31.5	15.28	13.72
Preparation	81.12	58.04	79.5	56.5	19.11	12.48
Awaiting	105.5	101.12	108.5	106.5	33.29	39.79
Labor	39.04	67.92	30	64.5	16.50	17.89

4.1 Fixation-related metrics

To ensure the same signal length per segment per participant and to facilitate within-segment comparisons, we normalized the time in the range of 0 (start of segment) to 1 (end of segment).

Fixation count. Although there were more fixations in the first session, the paired t-test (two-tails) revealed no significant differences within any of the segments ($p > .05$). Figure 4 summarizes the results, while Figure 5 presents heatmaps of fixation points aggregated from all users across time segments. We can observe more concentrated fixations during the Vaginal examination in the second session compared to the first.

Fixation duration. Although fixations lasted longer in the first session, the paired t-test (two-tails) revealed no significant differences within any of the segments ($p > .05$). Figure 6 summarizes the results.

4.2 Cognition-related metrics

Task-Evoked Pupillary Response. We took the average of the left and right pupil sizes, and normalized them using Min-Max normalization, considering the entire signal from both sessions. While TEPR was higher in the second session, the paired t-test (two-tails) was significant only during Labor: $t(23) = -2.41, p = .017$. Figure 7 summarizes the results.

Eye Blink Rate. We excluded any blinks with a duration of less than 100 ms as this threshold is considered the minimum duration for a valid blink [20, 27]. While EBR was higher in the first session, especially during Anamnesis and Labor, no significant differences were observed within any of the segments ($p > .05$). Figure 8 summarizes the results.

4.3 Saccade-related metrics

Saccade amplitude. While saccade amplitude was higher for the first session, the paired t-test (two-tails) was significant only during Awaiting: $t(23) = 2.9465, p < .01$. Figure 9 summarizes the results.

Saccade velocity. While saccade velocity was faster in the first session, the paired t-test (two-tails) revealed no significant differences within any of the segments ($p > .05$). Figure 10 summarizes the results.

Saccade acceleration. While saccade acceleration was higher for the first session, the paired t-test (two-tails) was significant only during Awaiting: $t(23) = 2.78, p < .01$. Figure 11 summarizes the results.

4.4 Head-related metrics

Angular velocity. While angular velocity was higher for the first session, the paired t-test (two-tails) was significant only during Anamnesis: $t(23) = -2.897, p < .01$. Figure 12 summarizes the results.

Cumulative rotation. Cumulative rotation was significantly higher for the first session, as corroborated by the paired t-test (two-tails), during Anamnesis ($t(23) = 10.30, p < .001$), Vaginal examination ($t(23) = 7.40, p < .001$), and Preparation ($t(23) = 10.94, p < .001$). No statistically significant differences were found during Awaiting ($p > .05$). Finally, cumulative rotation was significantly lower during Labor ($t(23) = -12{,}08, p < .001$). Figure 13 summarizes the results.

5 Machine Learning models

We observed significant improvements in skill acquisition but most eye/head related metrics revealed no statistically significant differences between training sessions. Therefore, aimed at further investigating the role of eye and head movements for user modeling, we trained Machine Learning classifiers to tell trained and untrained practitioners apart (binary classification task).

We trained Support Vector Machine (SVM) classifiers, since they have been widely used for eye-tracking classification tasks [22, 32] given their efficiency and adequacy in handling small sample sizes. We employed AutoML with Bayesian Optimization to tune the following SVM hyperparameters: kernel type[4] $\in$ {Linear, RBF, Polynomial}, regularization parameter C within $0.1 \leq C \leq 100$, and decay for non-linear kernels within $0.01 \leq \gamma \leq 10$.

We divided each segment into smaller non-overlapping sampling windows. For each window, we engineered a feature vector specific to each modality (Table 2) consisting of: Min, Max, Mean, Mdn, and SD of the values in each window. As in the previous section, time was normalized between 0 and 1, relative to the start and end of each segment. We also concatenated a numerical code to the feature vectors, to inform the model about the segment from which each sample originates: {1: Anamnesis, 2: Vaginal examination, 3: Preparation, 4: Awaiting, 5: Labor}.

The target label, to predict as model output, was either '0', representing the first session (untrained practitioners), or '1', representing the second session (trained practitioners). We first trained each model on individual segments, to assess how effectively each of them contributes to discriminating between practitioners. We then considered the concatenation of all segments at once. In any case, we used 80% of the data as the training set, 10% for validation, and 10% as the test set. Note that each modality has a different number of feature vectors (Table 2) since the sampling windows are different.

[4] All polynomial kernels had a degree of 3.

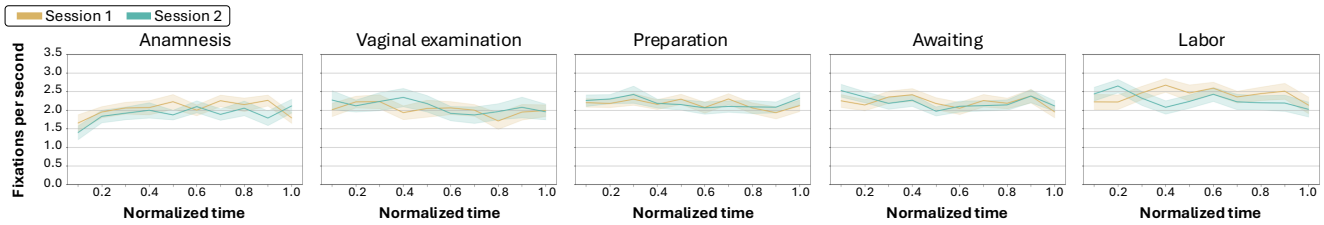


Figure 4: Fixation count across different segments. Shaded areas represent the standard error of the mean.

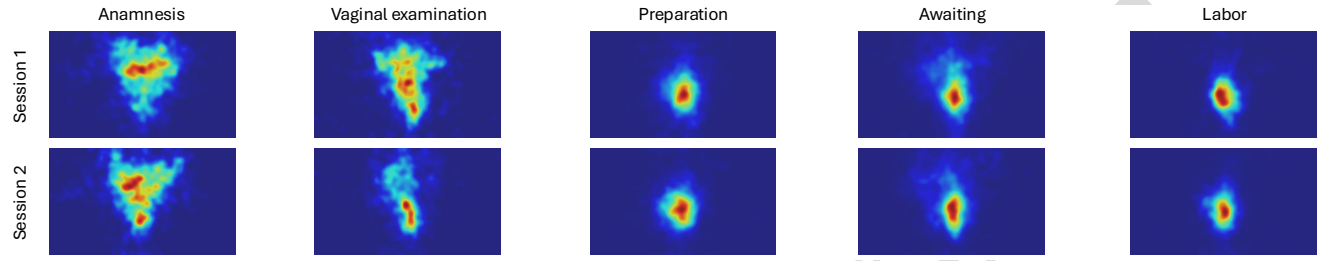


Figure 5: Heatmaps of eye fixations.

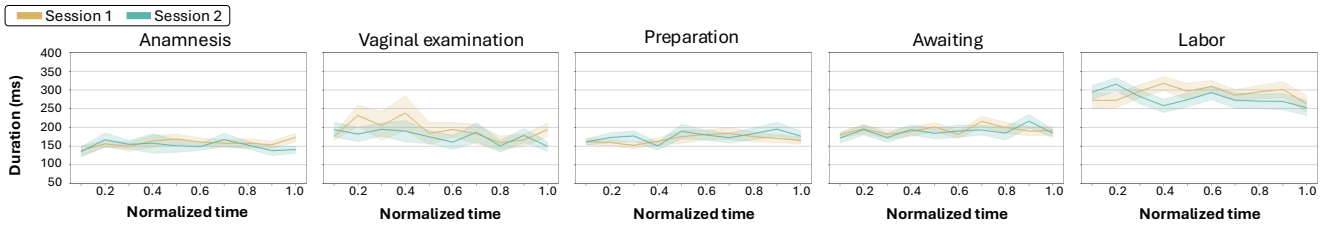


Figure 6: Fixation duration across different segments. Shaded areas represent the standard error of the mean.

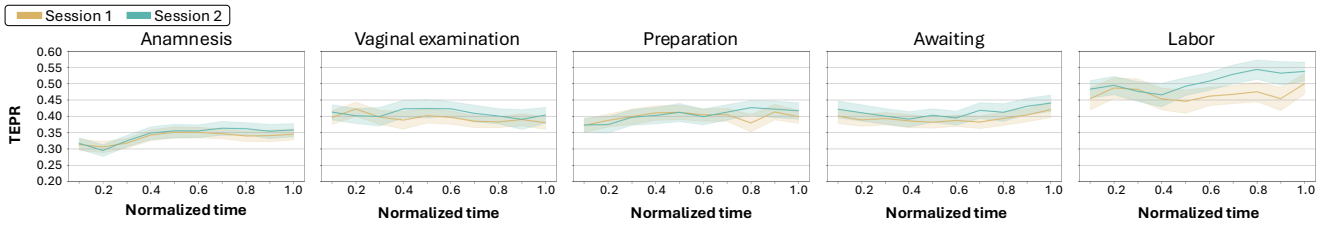


Figure 7: Pupil size across different segments. Shaded areas represent the standard error of the mean.

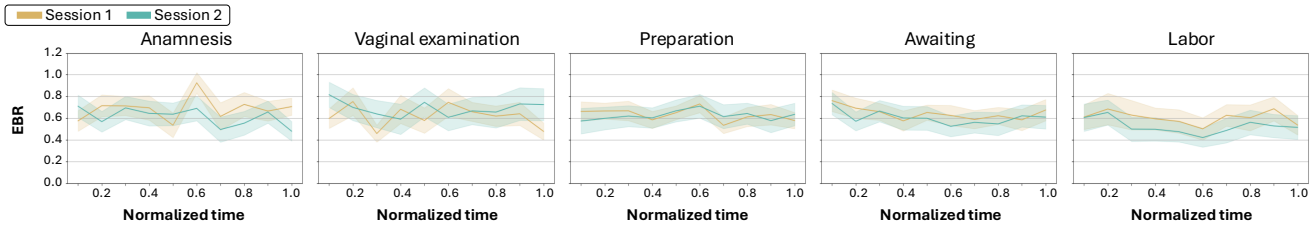


Figure 8: Blink rate across different segments. Shaded areas represent the standard error of the mean.

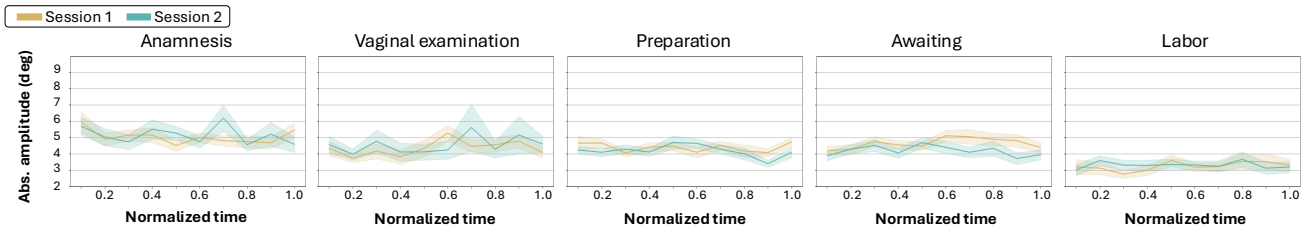


Figure 9: Saccade amplitude across different segments. Shaded areas represent the standard error of the mean.

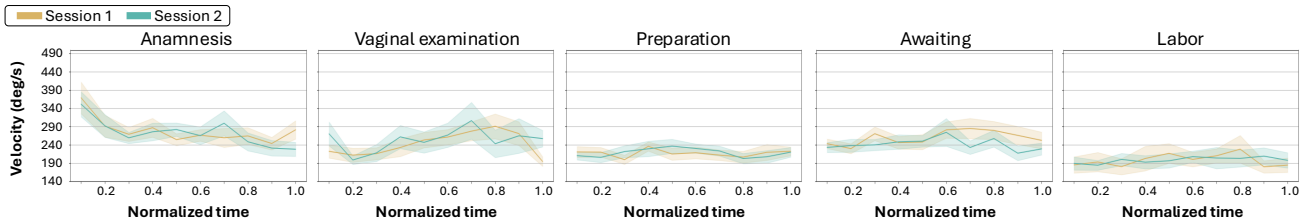


Figure 10: Saccade velocity across different segments. Shaded areas represent the standard error of the mean.

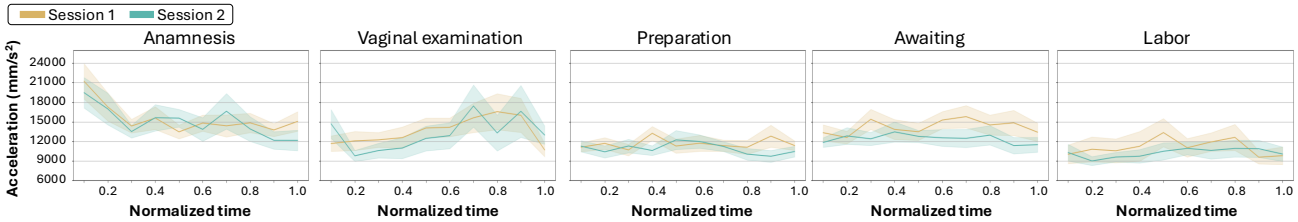


Figure 11: Saccade acceleration across different segments. Shaded areas represent the standard error of the mean.

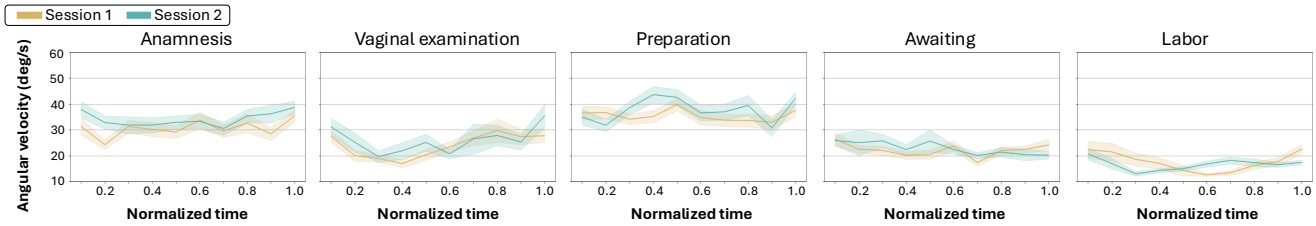


Figure 12: Angular velocity of head movements. Shaded areas represent the standard error of the mean.

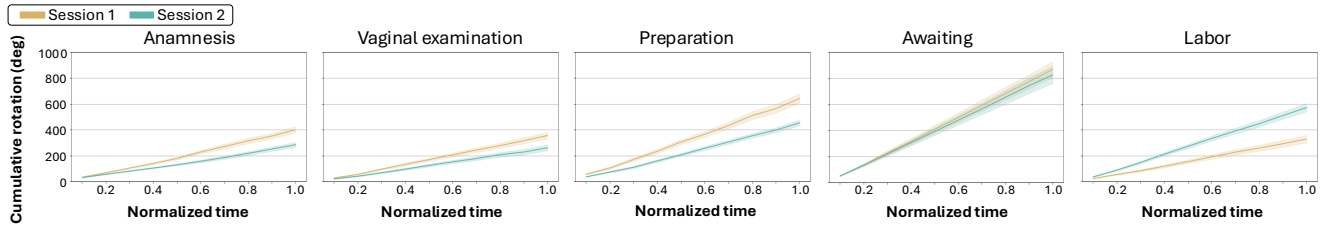


Figure 13: Cumulative rotation of head movements. Shaded areas represent the standard error of the mean.

Table 2: List of handcrafted features for SVM classifiers.

Modality	Features	Sampling window
Pupil	Timestamp X and Y coordinates Normalized pupil size Segment code	100 data points (1 s) 10751 feat vectors
Fixation	Timestamp X and Y coordinates Duration Segment code	10 successive fixations 3361 feat vectors
Saccade	Timestamp Amplitude Peak velocity Peak acceleration Segment code	10 successive saccades 8729 feat vectors
Blinks	Timestamp Duration Segment code	10 successive blinks 1051 feat vectors
Head	Timestamp Rotational speed in X, Y, Z Segment code	100 data points (1 s) 15107 feat vectors

5.1 Classification Performance

Table 3 reports the weighted F_1 and AUC scores of each classifier, highlighting the discriminating power of each modality in distinguishing between trained and untrained practitioners. Overall, the head-based classifiers performed best. The highest performance was observed for the Labor segment, with 85% F_1 and 86% AUC, followed by 77% F_1 and 84% AUC for the Preparation segment. Among the eye-based classifiers, the highest performance was also observed for the Labor segment, with an F_1 of 77% and AUC of 85% using pupillary responses and 67% F_1 and 80% AUC using fixation data. Blink data also performed similarly. These results are particularly encouraging, as discussed in the next section.

Table 3: Classification performance results. Best result in bold. Second best result underlined.

Modality	Segment	Hyperparameters			Adj. F_1	AUC
		kernel	C	γ		
Pupil	Anamnesis	linear	47.35	0.01	0.68	0.74
	Vaginal exam.	linear	12.58	0.01	0.58	0.69
	Preparation	linear	100.0	10.0	<u>0.70</u>	<u>0.79</u>
	Awaiting	RBF	25.21	0.58	0.67	0.74
	Labor	linear	100.0	0.85	**0.77**	**0.85**
	All segments	RBF	19.46	0.40	0.65	0.70
Fixation	Anamnesis	linear	6.64	0.01	0.62	0.64
	Vaginal exam.	RBF	1.70	6.29	0.44	0.54
	Preparation	linear	54.39	0.01	<u>0.62</u>	<u>0.81</u>
	Awaiting	RBF	32.98	0.01	0.59	0.61
	Labor	linear	73.52	0.80	**0.67**	**0.80**
	All segments	RBF	13.50	0.01	0.56	0.58
Saccade	Anamnesis	linear	44.23	10.0	0.59	0.63
	Vaginal exam.	linear	27.37	0.62	0.61	<u>0.64</u>
	Preparation	linear	56.50	0.01	<u>0.63</u>	**0.69**
	Awaiting	RBF	100.0	0.02	0.54	0.56
	Labor	linear	27.36	0.62	**0.64**	**0.69**
	All segments	RBF	93.19	0.01	0.55	0.57
Blink	Anamnesis	RBF	49.38	0.32	0.42	0.52
	Vaginal exam.	RBF	0.64	9.89	0.42	0.44
	Preparation	RBF	100.0	7.33	<u>0.52</u>	0.48
	Awaiting	polynomial	21.58	2.83	0.39	<u>0.58</u>
	Labor	RBF	15.92	0.03	**0.66**	**0.84**
	All segments	RBF	27.36	0.62	0.46	0.52
Head	Anamnesis	linear	77.60	9.97	0.63	0.69
	Vaginal exam.	linear	81.10	0.01	0.65	0.72
	Preparation	linear	100.0	0.56	<u>0.77</u>	<u>0.84</u>
	Awaiting	RBF	0.1	0.49	0.47	0.48
	Labor	linear	31.76	10.0	**0.85**	**0.86**
	All segments	RBF	100.0	0.15	0.55	0.58

6 Discussion

Participants demonstrated improved expertise after both training sessions, as reflected in their skill assessment scores. While eye and head movements effectively distinguish between trained and untrained practitioners, our findings suggest that certain segments of the training process—particularly Labor—play a more critical role in skill assessment. Furthermore, eye and head movement data provide valuable insights into different aspects of skill development.

During the second session, we observed an increase in TEPR during Labor, indicating both higher cognitive load and engagement [53]. This segment also exhibited longer fixation durations but fewer fixation counts, suggesting a heightened level of concentration as participants became more skilled [4, 33]. We attribute this behavior to trained practitioners being able to consider more options before making a critical decision, as indicated by greater pupil dilation during the second session Figure 7. This also explains the longer duration of Labor after the first session Table 1 while trained participants performed faster in the rest of the segments. Consequently, we reject **H1a** and **H1b**, concluding that trained practitioners exhibit similar fixation counts and fixation durations as untrained practitioners. We partially accept **H2a**, as TEPR showed significant differences during Labor, but we reject **H2b** due to the lack of differences in EBR.

Saccadic movements primarily revealed differences between the two training sessions during Awaiting and Labor. Lower saccade amplitude and increased velocity acceleration in the first session suggests that participants developed quicker yet more focused movements as their skills improved after the first training session. As a result, we reject **H3b**, as trained and untrained practitioners demonstrated similar saccade velocity. However, we partially accept **H3a** and **H3c**, as saccade amplitude and acceleration were significantly different during Awaiting.

Finally, our findings confirm that head movements are a reliable indicator of skill progression. The increase in angular velocity during Anamnesis, Vaginal examination, and Preparation suggests greater fluency in executing these preparatory steps before Labor. Additionally, the decrease in cumulative rotation during the second session (except for Labor) indicates increased focus, resulting in reduced head movements to complete the tasks. Based on this, we partially accept **H4a** and **H4b**, since angular velocity was significantly different in Amnesis, and cumulative rotation was significantly different in all segments except Awaiting.

7 Limitations and future work

We acknowledge a relatively small sample size in our study (24 participants), however this is a common challenge in medical research [38], given the difficulty of recruiting professionals [9]. Another limitation of our study is that we normalized the time of our collected signals, to facilitate the comparisons within each segment, which assumes that each practitioner spent similar time in each segment. While this holds for most cases, the Awaiting segment showed greater variability (see Table 1).

We should point out that accurately tracking saccadic movements requires an eye tracker operating at least at 200 Hz. These high-frequency eye trackers are usually available in stationary form only. Due to the physical constraints of the delivery room simulation, we used eye-tracking glasses, that operated at 100 Hz. Despite this, our classification results indicate that meaningful data can still be extracted at this frequency. Future advances in eye-tracking technology are expected to enhance accuracy further.

Additionally, our findings highlight the potential of SVM models for accurate skill classification. We should note that, in addition to the SVM classifiers, we trained Recurrent Neural Network (RNN) and XGBoost models, also utilizing AutoML with Bayesian Optimization, but they did not perform well. The RNN model had an input layer of N dimensions (where N is the size of the sampling window, see Table 2) followed by a hidden LSTM layer using either hyperbolic tangent or ReLU activation. The embedding size of the hidden layer ranged from 50 to 100, in increments of 10. This was followed by a dropout layer with values between 0.1 and 0.5, in increments of 0.1. Then a fully connected layer with a single neuron and sigmoid activation was added for the final output. The candidate learning rates were selected from the set $\{10^{-n} \mid n = 3, 4\}$. For the XGBoost classifiers, the parameter space included the number of estimators (50–500), maximum depth (3–10), learning rate (0.01–0.3, using a log-uniform distribution), subsample ratio (0.5–1.0), and the fraction of features considered per split (0.5–1.0). The main reason for lower performance using LSTM is that the number of samples (time series) per segment is not large enough for RNN model training. One possibility for future work would be to apply data augmentation strategies. Future work should also consider multimodal fusion, combining features from eye and head movements to improve further model performance. Moreover, given the discriminative power of head movements, further analysis of specific types, such as neck extension and lateral bending [15], could provide deeper insights. Additionally, exploring a range of different clinical tasks could provide valuable insights into the robustness of our methodology and results.

8 Conclusion

We conducted a user study exploring eye and head tracking to understand skill development in clinical settings, specifically during simulated baby delivery. Our results show that eye and head movements can effectively distinguish trained from untrained practitioners with remarkable performance, with some tasks (e.g., Labor) being more informative than others. These findings lay the groundwork for computational models that support implicit and objective skill assessment using commodity eye-tracking glasses, complementing traditional evaluation methods in clinical settings such as explicit and objective questionnaires. Ultimately, our results pave the way to faster and less biased assessment of the skills of medical doctors in training activities.

Acknowledgments

This research is supported by the European Innovation Council Pathfinder program (SYMBIOTIK project, grant 101071147), the Slovenian Research Agency (grants N2-0354, BI-NO/25-27-007, P5-0433, IO-0035, J5-50155, J7-50096), and the CogniCom (grant 0013103) program of the University of Primorska. Authors would like to thank all medical and technical support staff of the Medical Simulation Unit of the University Medical Center Ljubljana.

References

[1] N. Ahmadi, F. Sasangohar, J. Yang, D. Yu, V. Danesh, S. Klahn, and F. Masud. 2024. Quantifying workload and stress in intensive care unit nurses: preliminary evaluation using continuous eye-tracking. *Human factors* 66, 3 (2024), 714–728.

[2] M. O. Al-Moteri, M. Symmons, V. Plummer, and S. Cooper. 2017. Eye tracking to investigate cue processing in medical decision-making: A scoping review. *Computers in Human Behavior* 66 (2017), 52–66.

[3] S. M. Ali, S. Aich, A. Athar, and H.-C. Kim. 2023. Medical education, training and treatment using XR in healthcare. In *2023 25th international conference on advanced communication technology (ICACT)*. IEEE, 388–393.

[4] S. D. Aljehane, B. Sharif, and J. I. Maletic. 2023. Studying developer eye movements to measure cognitive workload and visual effort for expertise assessment. *Proceedings of the ACM on Human-Computer Interaction* 7, ETRA (2023), 1–18.

[5] N. E. Anton, J. S. Cha, E. Hernandez, D. I. Athanasiadis, J. Yang, G. Zhou, D. Stefanidis, and D. Yu. 2023. Utilizing eye tracking to assess medical student non-technical performance during scenario-based simulation: results of a pilot study. *Global Surgical Education-Journal of the Association for Surgical Education* 2, 1 (2023), 49.

[6] T. Bapna, J. Valles, S. Leng, M. Pacilli, and R. M. Nataraja. 2023. Eye-tracking in surgery: a systematic review. *ANZ Journal of Surgery* 93, 11 (2023), 2600–2608.

[7] R. Bogacz, E.-J. Wagenmakers, B. U. Forstmann, and S. Nieuwenhuis. 2010. The neural basis of the speed–accuracy tradeoff. *Trends in neurosciences* 33, 1 (2010), 10–16.

[8] C. F. S. Brandão and D. C. Fernandes. 2018. Importance and challenges of simulation training in healthcare. *Scientia Medica* 28, 1 (2018), 1.

[9] J. Bruneau, D. Moralejo, C. Donovan, and K. Parsons. 2021. Recruitment of healthcare providers into research studies. *Canadian Journal of Nursing Research* 53, 4 (2021), 426–432.

[10] B. Bühler, E. Bozkir, H. Deininger, P. Gerjets, U. Trautwein, and E. Kasneci. 2024. On Task and in Sync: Examining the Relationship between Gaze Synchrony and Self-Reported Attention During Video Lecture Learning. *Proceedings of the ACM on Human-Computer Interaction* 8, ETRA (2024), 1–18.

[11] M. Cai, B. Zheng, and C. Demmans Epp. 2022. Towards supporting adaptive training of injection procedures: Detecting differences in the visual attention of nursing students and experts. In *Proceedings of the 30th ACM conference on user modeling, Adaptation and Personalization*. 286–294.

[12] E. Capogna, F. Salvi, A. Del Vecchio, M. Velardo, G. Capogna, et al. 2020. Changes in gaze behavior during the learning of the epidural technique with a simulator in anesthesia novices. *Open Journal of Anesthesiology* 10, 11 (2020), 361.

[13] E. Capogna, F. Salvi, L. Delvino, A. Di Giacinto, and M. Velardo. 2020. Novice and expert anesthesiologists' eye-tracking metrics during simulated epidural block: a preliminary, brief observational report. *Local and Regional Anesthesia* (2020), 105–109.

[14] H.-E. Chen, R. R. Bhide, D. F. Pepley, C. C. Sonntag, J. Z. Moore, D. C. Han, and S. R. Miller. 2019. Can eye tracking be used to predict performance improvements in simulated medical training? A case study in central venous catheterization. In *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*, Vol. 8. SAGE Publications Sage CA: Los Angeles, CA, 110–114.

[15] S. Chen and J. Epps. 2019. Atomic head movement analysis for wearable four-dimensional task load recognition. *IEEE journal of biomedical and health informatics* 23, 6 (2019), 2464–2474.

[16] M. Cognolato, M. Atzori, and H. Müller. 2018. Head-mounted eye gaze tracking devices: An overview of modern devices and recent advances. *Journal of rehabilitation and assistive technologies engineering* 5 (2018), 2055668318773991.

[17] A. Das, Z. Wu, I. Skrjanec, and A. M. Feit. 2024. Shifting Focus with HCEye: Exploring the Dynamics of Visual Highlighting and Cognitive Load on User Attention and Saliency Prediction. *Proceedings of the ACM on Human-Computer Interaction* 8, ETRA (2024), 1–18.

[18] O. Faust, Y. Hagiwara, T. J. Hong, O. S. Lih, and U. R. Acharya. 2018. Deep learning for healthcare applications based on physiological signals: A review. *Computer methods and programs in biomedicine* 161 (2018), 1–13.

[19] D. F. Ferreira, S. Ferreira, C. Mateus, N. Barbosa-Rocha, L. Coelho, and M. A. Rodrigues. 2024. Advancing the understanding of pupil size variation in occupational safety and health: A systematic review and evaluation of open-source methodologies. *Safety science* 175 (2024), 106490.

[20] D. J. Frank, B. Nara, M. Zavagnin, D. R. Touron, and M. J. Kane. 2015. Validating older adults' reports of less mind-wandering: An examination of eye movements and dispositional influences. *Psychology and Aging* 30, 2 (2015), 266.

[21] B. Fu, A. R. Soriano, K. Chu, P. Gatsby, and N. Guardado. 2024. Modelling Visual Attention for Future Intelligent Flight Deck-A Case Study of Pilot Eye Tracking in Simulated Flight Takeoff. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*. 170–175.

[22] S. Haddad, K. Latifzadeh, S. Duraisamy, J. Vanderdonckt, O. Daassi, S. Belghith, and L. A. Leiva. 2024. Good GUIs, Bad GUIs: Affective Evaluation of Graphical User Interfaces. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*. 232–243.

[23] C. Hafner, V. Scharner, M. Hermann, P. Metelka, B. Hurch, D. A. Klaus, W. Schaubmayr, M. Wagner, A. Gleiss, H. Willschke, et al. 2023. Eye-tracking during simulation-based echocardiography: a feasibility study. *BMC Medical Education* 23, 1 (2023), 490.

[24] J. E. Hoffman. 2016. Visual attention and eye movements. *Attention* (2016), 119–153.

[25] D. A. Hofmaenner, A. Herling, S. Klinzing, S. Wegner, Q. Lohmeyer, R. A. Schuepbach, and P. K. Buehler. 2021. Use of eye tracking in analyzing distribution of visual attention among critical care nurses in daily professional life: an observational study. *Journal of Clinical Monitoring and Computing* 35 (2021), 1511–1518.

[26] Z. Huang, X. Duan, G. Zhu, S. Zhang, R. Wang, and Z. Wang. 2024. Assessing the data quality of AdHawk MindLink eye-tracking glasses. *Behavior Research Methods* (2024), 1–17.

[27] S. Huette, A. Mathis, and A. Graesser. 2016. Blink durations reflect mind wandering during reading.. In *CogSci*.

[28] B. Ibragimov and C. Mello-Thoms. 2024. The Use of Machine Learning in Eye Tracking Studies in Medical Imaging: A Review. *IEEE Journal of Biomedical and Health Informatics* (2024).

[29] B. J. Jongkees and L. S. Colzato. 2016. Spontaneous eye blink rate as predictor of dopamine-related cognitive function—A review. *Neuroscience & Biobehavioral Reviews* 71 (2016), 58–82. https://doi.org/10.1016/j.neubiorev.2016.08.020

[30] T. A. Katz, D. D. Weinberg, C. E. Fishman, V. Nadkarni, P. Tremoulet, A. B. Te Pas, A. Sarcevic, and E. E. Foglia. 2019. Visual attention on a respiratory function monitor during simulated neonatal resuscitation: an eye-tracking study. *Archives of Disease in Childhood-Fetal and Neonatal Edition* 104, 3 (2019), F259–F264.

[31] L. Kessler, P. Gröpel, H. Aichner, G. Aspalter, L. Kuster, G. M. Schmölzer, A. Berger, M. Wagner, and B. Simma. 2023. Eye-tracking during simulated endotracheal newborn intubation: a prospective, observational multi-center study. *Pediatric research* 94, 2 (2023), 443–449.

[32] J. Z. Lim, J. Mountstephens, and J. Teo. 2022. Eye-tracking feature extraction for biometric machine learning. *Frontiers in neurorobotics* 15 (2022), 796895.

[33] B. Mahanama, Y. Jayawardana, S. Rengarajan, G. Jayawardena, L. Chukoskie, J. Snider, and S. Jayarathna. 2022. Eye movement and pupil measures: A review. *frontiers in Computer Science* 3 (2022), 733531.

[34] P. Mannella, R. Antonelli, M. M. Montt-Guevara, M. Caretto, G. Palla, A. Giannini, F. Pancetti, A. Cuttano, and T. Simoncini. 2018. Simulation of childbirth improves clinical management capacity and self-confidence in medical students. *BMJ Simulation & Technology Enhanced Learning* 4, 4 (2018), 184.

[35] E. Mauriz, S. Caloca-Amber, and A. M. Vázquez-Casares. 2023. Using Task-Evoked Pupillary Response to Predict Clinical Performance during a Simulation Training. *Healthcare (Basel)* 11, 4 (2023), 455. https://doi.org/10.3390/healthcare11040455

[36] J. Mercier, O. Ertz, and E. Bocher. 2024. Quantifying Dwell Time With Location-based Augmented Reality: Dynamic AOI Analysis on Mobile Eye Tracking Data With Vision Transformer. *Journal of Eye Movement Research* 17, 3 (2024).

[37] J. Meyer, A. Frank, T. Schlebusch, and E. Kasneci. 2022. U-har: A convolutional approach to human activity recognition combining head and eye movements for context-aware smart glasses. *Proceedings of the ACM on Human-Computer Interaction* 6, ETRA (2022), 1–19.

[38] A. A. Mitani and S. Haneuse. 2020. Small data challenges of studying rare diseases. *JAMA network open* 3, 3 (2020), e201965–e201965.

[39] V. Onkhar, D. Dodou, and J. De Winter. 2024. Evaluating the Tobii Pro Glasses 2 and 3 in static and dynamic conditions. *Behavior Research Methods* 56, 5 (2024), 4221–4238.

[40] R. Paprocki and A. Lenskiy. 2017. What Does Eye-Blink Rate Variability Dynamics Tell Us About Cognitive Performance? *Frontiers in Human Neuroscience* 11 (2017). https://doi.org/10.3389/fnhum.2017.00620

[41] C. M. Privitera and L. W. Stark. 2000. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on pattern analysis and machine intelligence* 22, 9 (2000), 970–982.

[42] N. W. Rim, K. W. Choe, C. Scrivner, and M. G. Berman. 2021. Introducing Point-of-Interest as an alternative to Area-of-Interest for fixation duration analysis. *PLoS One* 16, 5 (2021), e0250170.

[43] A. Rosner, I. Basieva, A. Barque-Duran, A. Glöckner, B. von Helversen, A. Khrennikov, and E. M. Pothos. 2022. Ambivalence in decision making: An eye tracking study. *Cognitive Psychology* 134 (2022), 101464.

[44] M. Saleem and Z. Khan. 2023. Healthcare Simulation: An effective way of learning in health care. *Pakistan Journal of Medical Sciences* 39, 4 (2023), 1185.

[45] J. Shah and A. Darzi. 2001. Simulation and skills assessment. In *Proceedings International Workshop on Medical Imaging and Augmented Reality*. IEEE, 5–9.

[46] V. Skaramagkas, G. Giannakakis, E. Ktistakis, D. Manousos, I. Karatzanis, N. S. Tachos, E. Tripoliti, K. Marias, D. I. Fotiadis, and M. Tsiknakis. 2021. Review of eye tracking metrics involved in emotional and cognitive processes. *IEEE Reviews in Biomedical Engineering* 16 (2021), 260–277.

[47] A. D. Souchet, S. Philippe, D. Lourdeaux, and L. Leroy. 2022. Measuring visual fatigue and cognitive load via eye tracking while learning with virtual reality head-mounted displays: A review. *International Journal of Human–Computer Interaction* 38, 9 (2022), 801–824.

[48] L. Steblovnik. 2014. *TUPS : trening urgentnih porodniških stanj.* Klinični oddelek za perinatologijo Ginekološke klinike UKCLJ: Združenje za perinatalno medicino SZD.

[49] L. Steblovnik, J. Drame, T. Drolc, M. Druškovič, G. Kavšek, S. Kofol, M. Lučovnik, M. Narin, and D. Trošt. 2021. Simulation-based education and training in obstetrics in cooperation with the medical simulation centre. *10th anniversary of the Medical Simulation Centre at the University Medical Centre Ljubljana* (2021), 213–221.

[50] A. Szulewski, N. Roth, and D. Howes. 2015. The Use of Task-Evoked Pupillary Response as an Objective Measure of Cognitive Load in Novices and Trained Physicians: A New Tool for the Assessment of Expertise. *Academic Medicine* 90, 7 (2015). https://doi.org/10.1097/ACM.0000000000000677

[51] I. Tanoubi, M. Tourangeau, K. Sodoké, R. Perron, P. Drolet, M.-È. Bélanger, J. Morris, C. Ranger, M.-R. Paradis, A. Robitaille, et al. 2021. Comparing the visual perception according to the performance using the eye-tracking technology in high-fidelity simulation settings. *Behavioral Sciences* 11, 3 (2021), 31.

[52] N. V. Valtakari, I. T. Hooge, C. Viktorsson, P. Nyström, T. Falck-Ytter, and R. S. Hessels. 2021. Eye tracking in human interaction: Possibilities and limitations. *Behavior Research Methods* (2021), 1–17.

[53] R. L. Van Den Brink, P. R. Murphy, and S. Nieuwenhuis. 2016. Pupil diameter tracks lapses of attention. *PloS one* 11, 10 (2016), e0165274.

[54] L. N. van der Laan, I. T. Hooge, D. T. De Ridder, M. A. Viergever, and P. A. Smeets. 2015. Do you like what you see? The role of first fixation and total fixation duration in consumer choice. *Food Quality and Preference* 39 (2015), 46–55.

[55] S. Viriyasiripong, A. Lopez, S. H. Mandava, W. R. Lai, G. C. Mitchell, A. Boonjindasup, M. K. Powers, J. L. Silberstein, and B. R. Lee. 2016. Accelerometer Measurement of Head Movement During Laparoscopic Surgery as a Tool to Evaluate Skill Development of Surgeons. *Journal of Surgical Education* 73, 4 (2016), 589–594. https://doi.org/10.1016/j.jsurg.2016.01.008

[56] K. Walter and P. Bex. 2021. Cognitive load influences oculomotor behavior in natural scenes. *Scientific Reports* 11, 1 (2021), 12405. https://doi.org/10.1038/s41598-021-91845-5

[57] D. S. Wooding. 2002. Eye movements of large populations: II. Deriving regions of interest, coverage, and similarity using fixation maps. *Behavior Research Methods, Instruments, & Computers* 34 (2002), 518–528.

[58] Z. Zhao, Z. Zhu, X. Zhang, H. Tang, J. Xing, X. Hu, J. Lu, Q. Peng, and X. Qu. 2021. Atypical head movement during face-to-face interaction in children with autism spectrum disorder. *Autism Research* 14, 6 (2021), 1197–1208.

[59] L. Zhu, J. Chen, H. Yang, X. Zhou, Q. Gao, R. Loureiro, S. Gao, and H. Zhao. 2024. Wearable Near-Eye Tracking Technologies for Health: A Review. *Bioengineering* 11, 7 (2024).

[60] O. A. Zobeiri, B. Ostrander, J. Roat, Y. Agrawal, and K. E. Cullen. 2021. Loss of peripheral vestibular input alters the statistics of head movement experienced during natural self-motion. *The Journal of physiology* 599, 8 (2021), 2239–2254.

[61] S. I. Çalim, S. C. Ulaş, H. Demirci, and E. Tayhan. 2020. Effect of simulation training on students' childbirth skills and satisfaction in Turkey. *Nurse Education in Practice* 46 (2020), 102808.