Frontline Learning Research
An official journal of EARLI

Frontline Learning Research Vol. 13 No. 4 (2025) 103 -131
ISSN 2295-3159

Exploring the dynamic interplay of cognitive load and emotional arousal by using multimodal measurements: correlation of pupil diameter and emotional arousal in emotionally engaging tasks

Selina Michel*†[1], Christian Kosel†[1], Tina Seidel[1] and Manuel Förster[1]

[1] Department Educational Sciences, TUM School of Social Sciences and Technology, Technical University of Munich, Munich, Germany

† Joint first authors

Article received 21 February 2025 / revised 29 May 2025 / accepted 19 August 2025 / available online 19 June 2026

Abstract

Multimodal data analysis and validation based on streams from state-of-the-art sensor technology, such as eye-tracking or emotion recognition using the Facial Action Coding System (FACS) with deep learning, allows educational researchers to study multifaceted learning and problem-solving processes and to improve educational experiences. This study aims to investigate the correlation between two continuous sensor streams—pupil diameter as an indicator of cognitive workload and FACS with deep learning as an indicator of emotional arousal (RQ1a)—specifically for epochs of high, medium, and low arousal (RQ1b). Furthermore, the time lag between emotional arousal and pupil diameter data will be analyzed (RQ2). A total of 28 participants worked on three cognitively demanding and emotionally engaging everyday moral dilemmas while eye-tracking and emotion recognition data were collected. The data were preprocessed in Phyton (synchronization, blink control, and downsampling) and analyzed using correlation analysis and Granger causality tests. The results show negative and statistically significant correlations between the data streams for emotional arousal and pupil diameter. However, the correlation is negative and significant only for epochs of high arousal, while positive but nonsignificant relationships were found for epochs of medium or low arousal. The average time lag for the relationship between arousal and pupil diameter was 2.8 ms. In contrast to previous findings without a multimodal approach suggesting a positive correlation between the constructs, the results contribute to the state of research by highlighting the importance of multimodal data validation and convergent validity. Future research should consider emotional regulation strategies and emotional valence.

Keywords: Eye-Tracking, Emotion Recognition, Multimodal Measurements, Cognitive Load, Pupil Diameter, Emotional Arousal.

Corresponding author: Selina Michel, Technical University of Munich, selina.michel@tum.de
Doi: http://doi.org/10.14786/flr.v13i4.1663

1. Introduction

The assessment of learners' cognitive and emotional states has long been the focus of educational research, with the primary goal of optimizing the learning experience and fostering performance (Moon et al., 2020). Cognitive and emotional states play a central role in determining learning effectiveness, as they significantly influence a learner's attention span, information processing, and holistic performance (Kärner et al., 2016; MacDougall et al., 2013). Considering learning and problem-solving as multifaceted processes involving interaction between cognitive and emotional states (Noorozi et al., 2019), multimodal measurements using state-of-the-art sensor technology to assess cognitive and affective processes in real time are promising to analyze learning in a more holistic way and improve individualized and adaptive learning activities (Cloude et al., 2020; Mayer et al., 2023).

Traditional methods, particularly self-report measures, have frequently been employed as prevalent tools for evaluating cognitive and emotional states (Schmidt-Weigand & Scheiter, 2011). For example, learners may be asked to rate the perceived difficulty of a task or to describe their emotions after completing a learning activity (Schmidt-Weigand et al., 2010). Sensor technology, such as eye-tracking, emotion recognition, heart rate detection, or pupillometry, now allows researchers to collect real-time data from different data streams at the same time (Cloude et al., 2020). This opens up opportunities for multimodal measurements to analyze more than one indicator and allows for the study of interactions between cognitive and affective states and regulation processes in multifaceted learning processes. Importantly, there is an observable shift from reliance on self-reported measures to incorporating objective or direct measurement techniques. This transition highlights the differentiation between subjective perceptions and externally observable data, enabling a more precise and unbiased analysis of learning behaviors and outcomes. Such objective measures can provide insights into the actual dynamics of learning processes, beyond what learners may report about their own experiences. This methodological evolution underscores the importance of leveraging diverse measurement approaches to capture the complex interplay between cognitive functions and emotional states in education (Noorozi et al., 2019).

These multimodal approaches are emerging in the educational sciences (Noorozi et al., 2019). For example, Dubovi (2022) demonstrated that 51% of pre-post knowledge achievements can be explained by a combination of modalities measured by different psychophysiological sensor streams. However, Noorozi et al. (2020) identified that out of 207 publications focusing multimodal measurements of the learning process from early childhood education to graduate level, only 14 used cognitive and emotional states at the same time, and only 10 used a combination of cognitive, motivational, and emotional states. One reason might be that there are several challenges in multimodal research (e.g., in data integration and analysis across data channels [e.g., different sampling rates], uncertainty about the amount of data that needs to be collected to investigate the underlying processes, and the impact of transformation from raw data to interpretable and actionable data, which is often done by data aggregation or filters) (Azevedo et al., 2018).

In the field of multimodal learning analytics research, several questions concerning data integration have emerged (Mu et al., 2022). Samuelsen et al. (2019) emphasize that besides data storage and data processing requirements, meaningful data integration is a challenge to consider in multimodal studies. For meaningful data integration, it is crucial to ask which type of data is suitable for learning indicators (Noorozi et al., 2020), how multimodal data can be aligned to ensure that all indicators are well reflected, and how to consider complementary correlations from inter- and cross-modality effectively (Mu et al., 2022). These correlations are sometimes hidden and neglected (Mu et al., 2022). Therefore, "many-to-one" approaches to multimodal data integration to improve the accuracy of measurements (e.g. measuring emotions by using indicators from video and audio source as described by Ez-Zaouia and Lavoué, 2018) and "many-to-many" approaches to multimodal data integration for improving information richness (e.g. capturing log files to measure student activity and using emotion recognition for affective states during the learning process as described by Azevedo et al., 2017) might feed into "mutual verification between multimodal data" for empirical evidence on data fusion and integration (please see Mu et al. [2022] for further review). Multimodal data validation and verification also allow for the analysis of relationships and correlations between different data streams and indicators (Mu et al., 2022).

Taking a closer look at these relationships and correlations also raises the question of the concurrent and congruent validity of data streams, which is important for educational practice. Psychophysiological sensor streams, such as heart rate and heart rate variability (Delliaux et al., 2019), electroencephalogram (brain wave levels) (Yoo et al., 2023), electrocardiogram (Chanel et al., 2019), galvanic skin response (Elahi & Islam, 2019), and respiration (Grassmann et al., 2016), share one notable limitation: the need for participants to be physically present in a laboratory setting and to wear specialized hardware, suggesting that these psycho-physiological sensor streams are not easily transferable outside of the laboratory. Recent technological advancements have catalyzed a paradigmatic shift toward digital and remote technology-enhanced learning, thereby significantly elevating the relevance and appeal of psychophysiological sensor streams beyond traditional laboratory settings (Chernikova et al., 2020). One way of assessing learners' cognitive load in remote digital learning environments is to analyze their log file data streams, for example, how much time they spend on tasks, the number of attempts at a problem, or patterns of resource access (Hulshof, 2005). However, emotions are hard to assess using log-file methods (Janning et al., 2016). Researchers have recognized this challenge and have initiated endeavors to incorporate psychophysiological measures, extending their applicability to remote contexts beyond the confines of laboratory settings (Chanel et al., 2019). Therefore, psychophysiological measures that can be used remotely, such as pupillometry or emotion recognition, move into focus. To sum up, it is key to reflect on multimodal data integration by considering intercorrelations between indicators for cognitive and affective states to enhance the data economy and practical usability without losing the benefits of the multimodal measurements described above. In this work, we focus on investigating systematic relationships between two established indicators highly relevant to the educational context: pupil diameter as an indicator of *cognitive load* and an emotion recognition system based on the Facial Action Coding System (FACS) with deep learning for *emotional arousal*.

1.1 Cognitive load

"The level of an individual's measured effort in order to cope with one or more cognitively demanding tasks" can be applied as a definition of cognitive load (Skarmagkas et al., 2021, p. 2). Cognitive load is closely related to constructs such as mental workload, mental effort, or mental demand in ergonomics or human factors literature, which also describes spending cognitive resources and the mental activity required to perform a task (van de Acker et al., 2018; Vanneste et al., 2020; Young et al., 2014). To make a more precise differentiation, cognitive load incorporates demands from the learning environment to perform learning tasks (task-driven mental load), as well as the more human-centered dimension of actively spending cognitive resources while processing the learning task in terms of cognitive engagement (mental effort) (Pass & van Merriënboer, 1994; Schnaubert & Schneider, 2022). The constructs of mental load and cognitive load are often used interchangeably (Schnaubert & Schneider, 2022), whereas the construct of cognitive load is more common in educational sciences. Since we did not aim to differentiate between task-driven mental load and mental effort, we decided to focus on the concept and term of cognitive load in this study.

Cognitive load researchers traditionally aim to assess learners' cognitive functioning for instructional adaptations because optimal learning environments or tasks should neither cause cognitive overload nor underload (Martin et al., 2021; Sweller et al., 1998). Research centered on cognitive load theory emphasizes the finite capacity of working memory, as opposed to the expansive storage potential of long-term memory (Paas, 2014; Sweller et al., 1998). Evidence suggests that flooding working memory with too much information can lead to cognitive overload, impairing the brain's ability to efficiently process and assimilate the information presented (Sörqvist et al., 2016). Based on this assumption, in instructional design theory, a well-known framework is the cognitive load theory by Sweller et al. (1998). This theory differentiates intrinsic load (element interactivity), extraneous load (instructional design/design of the learning environment) and germane load (schema construction in learning) (Sweller et al., 1998). For this publication we are not differentiating different types of loads described by Sweller et al. (1998) but respect their contribution to the overall cognitive load and its measurement.

In line with the seminal work of Löwenstein (1920), one stream of research—pupillometry—is concerned with small changes in pupil diameter attributed to changes in human cognition (Van Der Wel & Van Steenbergen, 2018; Löwenstein, 1920). Cognitive activity during the learning process aligns with sympathic activity and reduced parasympathic activity of the autonomic nervous system (Alshanskaia et al., 2024). As sympathetic activity of the nervous system leads to dilation of the pupil by the dilater muscle and an inhibition of parasympatic activitiy affects the constriction of the pupil by the sphincter muscle (van Der Wel & Sternbergen, 2018; Beatty & Lucero-Wagoner, 2000). Based on this, an increase of pupil diameter is aligned with higher cognitive load. Even though the pupil size is sensitive to changes in light, in light-controlled environments pupillometry is a promising method to measure cognitive effort (Karch et al., 2019) and, following from this, pupil dilation is a well-research indicator for assessing cognitive load. Studies on task-evoked pupillary responses focus on how changes in cognitive load during a task can be measured through changes in pupil size compared to the baseline. According to Mallick et al. (2016), pupil dilation showed a positive relationship with cognitive workload in dynamic and unconstrained tasks such as video games (Tetris®). Pupil size showed the strongest positive correlation with changes in workload compared to other metrics like blink duration, fixation duration and saccade peak velocity (Mallik et al., 2016). Soussou et al. (2012) also reported small to medium positive correlations in measuring pupil dilation by eye-tracking technology and EEG-based gauges as indicators of cognitive workload during a X-Ray screening task. Rodemer et al. (2023) used pupil dilation as an operationalization for cognitive load in instructional videos on complex chemical representations comparing a dynamic, static and control condition and found significant correlations between pupil dilation and reported extraneous cognitive load. The review from Van der Wel and Van Steenbergen (2018) wraps up the discussion by suggesting that pupil diameter rather reflects cognitive effort that can differ between experts and novices than task demand. Based on the studies described above, we assume that higher cognitive load goes hand in hand with an increase of the pupil diameter.

Concurrently, the advent of flexible open-source platforms dedicated to high-resolution pupillometry, stemming from vision research (i.e., Zandi et al., 2021), presents a promising direction for researchers exploring the incorporation of pupillometry within remote technology-enhanced learning modalities to assess cognitive load. While there has been extensive research on cognitive load in isolation (Antonenko et al., 2010), recent findings suggest that a learner's emotional state, especially their level of emotional arousal, can influence learning by altering cognitive load (LeDoux, 2021; Tyng et al., 2017).

1.2 Emotional arousal

Emotional arousal, defined as "a state that describes the level of calmness (i.e., low arousal) or excitation (i.e., high arousal) elicited by a stimulus" (Skaramagkas et al., 2020, p. 2), is a bipolar dimension used to group emotions in multidimensional scaling (Russel, 1980) and, thus, proves to be a higher-level dimension for describing emotional states. Also, for the context of learning, Pekrun (1992, 2006) suggests considering emotional valence and arousal to differentiate between positive-activating (e.g. joy), positive-deactivating (e.g. relief), negative-activating (e.g. anger) and negative-deactivating (e.g. boredom) emotions relevant for learning and performance. Nevertheless, there are findings for a relationship of pupil diameter and emotional valence (e.g. Babiker et al., 2013; Kinner et al., 2017) in this study we focus on the dimension of emotional arousal considering the close relationship of cognitive load and arousal e.g. regarding the activation of the noradrenergic system (van der Wel & van Steenbergen, 2018) described in chapter 1.3.

The rapidly growing methodology of facial recognition offers another promising avenue, providing nuanced real-time analysis for emotion detection (Lewinski et al., 2014). One of these emerging methods is called FACS with deep learning. The FACS categorizes and describes all conceivable facial muscle movements (termed "action units" or AUs) and their visible effects. FACS is often used in psychology to objectively describe facial expressions, which can then be associated with specific emotions or states (Lewinski et al., 2014). It allows for high-frequency and unobtrusive observations of changes in student emotions during learning and problem-solving situations (Dindar et al., 2020; Dubovi, 2022). Combining FACS with deep learning techniques, such as convolutional neural networks, can provide a powerful approach for extracting and interpreting complex facial muscle actions to automatically recognize and classify facial action units from images or video feeds (Bhattacharyya et al., 2021). This automation can allow for real-time emotion detection, which has applications in various fields, such as human–computer interactions (Podder et al., 2023) and psychology (Gao & Ma, 2020; Song, 2021). Current research suggests that detecting emotions with FACS with deep learning is comparable to electromyography data in identifying happy or angry facial expressions (Kulke et al., 2020) and self-report data in the Emotions-Value Questionnaire (Harley et al., 2014). Comparing the method of FACS with deep learning to other measures of emotional experience adds evidence to the validity of FACS. To summarize, in multimodal measurements, emotional arousal can be captured using FACS with deep learning.

1.3 Interplay between cognitive load and emotional arousal

A variety of findings suggests a dynamic interplay between cognitive load and emotional arousal. On the one hand, emotional arousal not only acts as a source of extraneous cognitive load (Mehrotra & Gunalakshmi, 2020) but also plays a role in affecting the human memory system (LeDoux, 2021). When learners are in a state of high emotional arousal, some of their cognitive resources are directed toward processing these emotions, leaving fewer resources available for other tasks and, hence, increasing cognitive load.

According to the Yerkes–Dodson law, there is an optimal level of emotional arousal that supports learning; levels that are too low or too intense can hinder the learning process (Sherwood, 1965; Yerkes & Dodson, 1908). On the other hand, cognitive load can affect emotional arousal (Jerčić et al., 2018). A task that imposes a high cognitive load can induce emotions such as frustration or anxiety, leading to an increase in emotional arousal. For example, if a learner finds a task too difficult or complex, this could induce feelings of stress or anxiety, which could result in heightened emotional arousal (Jerčić et al., 2018).

Current research supports this outlined reciprocal relationship between cognitive load and emotional arousal, with changes in one often leading to changes in the other (Tyng et al., 2017). Neurobiological research indicates that changes in pupil diameter can be attributed to the activation of the locus coeruleus within the noradrenergic system.

This activation, often precipitated by emotional arousal, is postulated to have a consequential role in the process of memory consolidation (Van Der Wel & Van Steenbergen, 2018). Subsequently, several studies have confirmed that pupil dilation serves as a credible biomarker of cognitive load and emotional arousal, highlighting its diagnostic potential in various research contexts (Bradley et al., 2008; Steinhauer, 1983). Although a growing number of studies have used pupil diameter to indicate the level of arousal (Ebitz & Platt, 2015; Kreuzmair et al., 2016; Murphy et al., 2014; Nassar et al., 2012), research that directly examines the relationship between pupil diameter and emotional arousal, both measured with psychophysiological sensor streams, is limited. For example, Kreuzmair et al. (2017) found that participants' pupil diameter was a measure of their emotional arousal in hypothetical medical scenarios. In their experiment, they varied the "risk level" in their scenarios as a proxy for different levels of arousal; however, they aggregated data to mean values for specific times of interest instead of using process data based on real-time psychophysiological sensor streams of emotional arousal (Kreuzmair et al., 2017). This also reflects the challenges of the multimodal research described above. Therefore, there is currently limited evidence about how pupil diameter and emotional arousal correlate in time-series data. To address this research desiderate, different methodological considerations need to be undertaken.

Considering this interplay between cognitive load and emotional arousal, both high-resolution pupillometry and face recognition based on deep learning techniques have the potential to be significant methods for assessing learners' cognitive load and emotional arousal within technology-enhanced remote learning modalities (Azevedo et al., 2018; Cloude et al., 2022). However, it is noteworthy that both data sampling methods use different indicators that do not intersect, and questions of data synchronization and integration arise. While high-resolution pupillometry focuses on variations in pupil diameter, face recognition uses complex facial muscle actions to classify emotional states—ignoring the eyes (Lewinski et al., 2014).

Both psychophysiological data streams aim to capture facets of cognitive load and emotional arousal using different data streams. Consequently, within an individual, these streams exhibit a substantial correlation. From the perspective of multimodal data validation (Mu et al., 2020), the validity of these streams as reliable biomarkers must be questioned if there is no correlation between the two data streams. Therefore, the correlation between the two streams of data remains a significant question. Furthermore, the time lag between emotional arousal and physiological responses, such as pupil diameter, is largely unclear. Generally, emotional responses are described as preceding physiological responses (Mauss & Robinson, 2009). However, there is currently no concrete evidence detailing the time lag between these two physiological indicators, especially in relation to their respective sampling methods. For multimodal data validation and practical use in, for example, adaptive learning systems, we need to understand whether both multimodal data streams for cognitive load and emotional arousal correlate and which stream first shows an indication that might be contextually relevant for feedback or interventions in the future. This is also important considering the analytic bottleneck described by Azevedo and Gasevic (2019) that occurs when merging data streams in educational practice and can lead to latencies in delivering cues and inferences, which might also have negative effects on learning. Therefore, it is key to also reflect on time lags between correlating indicators and consider the first indications as "initial reactions."

1.4 Everyday moral dilemmas as emotionally engaging and cognitively demanding tasks

The dynamic interplay between cognitive load and emotional arousal may occur in different contexts in which learning and problem-solving take place. One problem-solving context that is cognitively demanding and emotionally engaging is everyday moral dilemmas. Dilemmas are short stories or vignettes presenting a moral conflict that confronts a person with two incompatible courses of action, leading to consequences that represent rival moral principles or values (Christensen & Gomila, 2012). These moral conflicts can be between personal interests and accepted moral values, different duties, a set of apparently incommensurable values, or conflicts stemming from one unique moral principle (Christensen & Gomila, 2012). We argue that everyday dilemmas, by their very nature, require individuals to weigh multiple factors, outcomes, and personal values. Typical sacrificial moral dilemmas can be criticized for showing a lack of external validity and realism (e.g., Baumann et al., 2014; Kahane, 2015).

This study uses three everyday dilemma situations from a work context as the emotional dilemma task. Everyday dilemma situations can also be cognitively demanding and stimulate emotions (e.g., Thomson & Berenbaum, 2006). For example, goal or role conflicts in everyday work situations, such as moral conflicts, are described as challenging for employees and can trigger emotions (e.g., Fisher & Ashkanasy, 2000). Research on moral judgment is continuously discussing the role of emotions in the moral evaluation process (e.g., Blum, 2023; Greene et al., 2001, 2009). One focus is on investigating the role of strong emotional reactions, working as a kind of alarm bell stimulated by personal dilemmas, as described in dual process theory (see Greene et al. [2001, 2009] for further review). Accordingly, an analysis of the interplay between cognition and emotions is also particularly interesting for intervals where high arousal is perceived.

To sum up, the complexity of everyday dilemmas can lead to an increase in cognitive load as one's brain processes the information and contemplates potential consequences, which also affect emotional arousal (Bluhm, 2013; Cummings & Cummings, 2012; Greene et al., 2009). Therefore, we argue that everyday moral dilemmas are suitable for use as emotionally engaging problem-solving tasks. Using dilemmas might also have the advantage that previous research already showed—they are suitable for studies using psychophysiological sensors such as eye-tracking (Gaffari & Fiedler, 2018; Garon et al., 2018; Fiedler et al., 2013; Pärnaments et al., 2015) and emotion recognition (Cumming & Cumming, 2012; Gleichgerrcht & Young, 2013; Valdesolo & DeSteno, 2006). For more information on the advantages of text-based dilemmas, please refer to the methodology section.

In summary, based on the evidence described above, we argue that pupil size as an indicator of cognitive load (e.g., Mallick et al., 2016; Soussou et al., 2021; Van Der Wel & Van Steenbergen, 2018) and emotional arousal analyzed by FACS with deep learning (e.g., Bhattacharyya et al., 2021) correlate, particularly in an emotionally engaging task. This correlation, as part of multimodal validation (Mu et al., 2020), can provide starting points to improve learning environments in, for example, adaptive learning settings, using only one of the two indicators to assess cognitive and emotional states as a cue for adaptations (Chanel et al., 2019; Sailer et al., 2023). Therefore, questions of concurrent validity can also be addressed. To understand whether this is a reliable perspective for learning interventions, a process-oriented approach analyzing the correlation between two continuous psychophysiological data streams, pupil size operationalizing cognitive load and FACTS analysis to operationalize emotional arousal, is needed.

1.5 Aim of the study

The aim of this research is to synchronize pupil diameter and emotional arousal during emotionally engaging problem-solving tasks, such as everyday moral dilemmas, and to examine whether and how these two measures (i.e., psychophysiological data streams) correlate. Therefore, we propose a methodological approach for describing the interplay of cognitive and emotional processes. We postulate that synchronizing these data streams would allow us to delineate the dynamic correlation between cognitive and emotional processes, thereby going beyond studies that use either pupillometry or emotion recognition and no multimodal measurement in the context of moral dilemma tasks (e.g., Doerflinger & Gollwitzer, 2020; Fiedler et al., 2013; Garon et al., 2018). Given the research gap, the current study will focus on the research question (RQ) described below:

RQ1a: Can changes in emotional arousal, as captured through facial expressions, be observed concurrently with corresponding variations in cognitive load, as indicated by changes in pupil diameter?

Our aim is to explore the relationship between facially coded emotional arousal and pupil diameter as two psychophysiological data streams. We acknowledge that pupil diameter is not a pure indicator of cognitive load but rather a composite measure influenced by both cognitive and emotional processes. Accordingly, we do not interpret it as a construct-isolated marker of mental effort. Instead, our analysis focuses on the temporal co-fluctuation of these streams to examine their convergence and potential interdependence. This contributes to multimodal data validation and informs future adaptive system design. Furthermore, we analyze correlations during distinct arousal epochs to explore whether the strength of the relationship varies with emotional intensity, especially in cognitively demanding and emotionally engaging tasks such as moral dilemmas.

RQ1b: How does pupil diameter correlate with emotional arousal during epochs of high, medium, and low emotional arousal?

Second, contributing to research on the role of strong emotional reactions in personal moral dilemma situations (e.g., Greene et al., 2001, 2009) and examining changes in pupil diameter during periods of high and low emotional arousal provides distinctive insight into the dynamic relationship between physiological responses. In addition, differentiating episodes of different levels of arousal reflects assumptions from the Yerkes–Dodson law (Yerkes & Dodson, 1908) and can provide evidence on the assumed interaction of emotional arousal and cognition in this model. Acknowledging intra-individual variability in the relationship between emotional arousal and pupil diameter, it is posited that the most significant correlations may emerge specifically at the peaks of emotional intensity. Therefore, differentiating episodes characterized by high, medium, and low emotional arousal is of interest. Moreover, for educational research and practice, the time lag between the two data streams is also crucial.

RQ2: Does a systematic and significant time lag exist between the onset of emotional arousal and the corresponding changes in pupil diameter measurements?

Third, assessing the time lag between emotional arousal based on facial expression analysis and physiological measures, such as pupil diameter, is crucial, especially for future adaptive learning systems. In real-world scenarios, emotional responses might not be immediate, and there could be time lags between emotional arousal and physiological reactions. In general, emotional responses typically occur before physiological responses. When a person experiences an emotion, such as fear, joy, anger, or surprise, the emotional response is reflected in activation in associated brain areas and triggers a series of physiological changes in the body, such as increased heart rate or changes in skin conductance (sweating), or as outlined pupil diameter (Šimić et al., 2021). By assessing the mean time lag between emotional arousal and pupil diameter, adaptive learning systems can provide more timely and contextually appropriate feedback or interventions. This adds to the call that indicators need to be aligned in such a way that all indicators are well reflected (Mu et al., 2022). As described above, when merging data in educational practice, there is a risk of an analytic bottleneck, which can result in latencies in delivering cues and inferences (Azevedo & Gasavi, 2019). For instance, if a learner shows signs of frustration or disengagement based on facial expressions, physiological measures can validate and reinforce these emotional indicators, prompting the system to offer support or encourage the learner appropriately. Therefore, it is important to know the initial reactions and time lags between different psychophysiological responses measured by different data streams.

2. Methodology

2.1 Participants

Requirements and criteria for participants' inclusion in the study were normal or corrected to normal vision and sufficient German language and reading skills to process the emotional dilemma task. Beyond that, no criteria for exclusion were defined. In summary, 36 participants took part in the study. Due to technical issues (e.g., hardware failure when recording videos for emotion recognition and loss of connection between devices due to unstable internet connection), 6 participants were excluded from the analysis. Two additional participants were excluded during the pre-processing of the data streams. This exclusion was due to extreme outliers concerning the time taken for task execution and substantial gaps in the data, which could not be rectified through standard imputation techniques or outlier management strategies. The final sample consisted of 28 participants. All the participants gave their consent for the multimodal data collection after being informed about the conditions of participation.

Among the participants, 13 (46.43%) were male and 15 (53.57%) were female. A total of 8 participants (cumulative percentage: 28.57%) reported a secondary school or university entry degree to be their highest educational degree certificate, 14 (50%) reported a bachelor's degree to be their highest educational degree, and 6 (21.43%) reported a master's degree highest educational degree. Because the study uses dilemmas from work contexts as an emotionally engaging task, the participants were also asked for their work experience using a multiple-choice item: 14 (50%) participants completed vocational training, 6 (21.34%) completed an internship for longer than six months, 9 (32.14%) worked part-time during their studies, and 17 (60.76%) were fully employed for longer than one year. Two (7.14%) participants indicated no work experience. Based on the educational degree and work experience, the sample can be considered suitable for the emotional dilemma task described below.

2.2 Emotional dilemma task

In this study, we used three text-based everyday moral dilemma tasks (Dilemma 1: 336 words, Dilemma 2: 317 words, and Dilemma 3: 248 words). We choose for a text-based format as from literature it is known that the design of the task can increase the extraneous load (e.g. Rodemer et al., 2023) and we aimed to design the three dilemma tasks as comparable as possible without unintentionally manipulating the extraneous load by choosing different visual stimuli. Also, the pupil diameter is sensitive to light (Karch et al., 2019; Mathôt, 2018) we tried to keep the luminescence of the stimulus constant using the same background color and font color and size to differentiate effects of luminescence and cognition and ensured constant artificial light instead of natural light in our Lab. The dilemmas contained a moral conflict, as described by Christensen and Gomila (2012), as an emotionally engaging task. The topic and content of the moral dilemmas were chosen with regard to findings that work events, such as goal or role conflicts, are cognitively challenging and can trigger emotions (e.g., Fisher, 2000) and structured as outlined in the following. The participants in the study were asked to put themselves in the situation described in the dilemmas. The emotional dilemma tasks were structured in an outline of the situation (paragraph 1), a description of a conflict (paragraph 2), and the consequences of different courses of action to solve the dilemma (paragraphs 3 and 4). After gathering the information by reading the text, the participants were asked to decide between two courses of action and justify their decisions in a written format in three to five sentences. In terms of content, the dilemmas present either role, goal, or between-person conflicts, which show the potential to trigger emotions (Fisher, 2000) and can lead to moral conflicts (Christensen & Gomila, 2012). The first dilemma presented a moral conflict in an academic group work setting, and the second and third dilemmas presented work conflicts in an occupational setting.

The dilemmas were validated by asking the participants about their perceptions of the realism of the dilemma, the occurrence of an inner conflict, the seriousness of the conflict for the protagonist, and the difficulty of decision-making (operationalized as the opportunity to make a decision after a short moment of thought) using a 5-point rating scale (1 = not at all; 5 = highly applicable). The validation questions were adapted from the work of Heinrichs and Schadt (2019). The difficulty of the emotional dilemma task is also reflected when taking into account the decision of the participants (Dilemma 1: 35% option 1 and 65% option 2, Dilemma 2: 56% option 1 and 44% option 2, and Dilemma 3: 41% option 1 and 59% option 2). Based on these findings, the emotional dilemma tasks are relatively comparable and assumed to be challenging and realistic. Please see Table 1 in the results section for detailed values.

Participants also rated 14 discrete emotional states (e.g., guilt, shame, pride, determination, insecurity) on 5-point Likert scales. These served both as manipulation checks and as variables to examine the concurrent validity of physiological indicators via correlation analyses.

2.3 Apparatus

In our lab, we used state-of-the-art equipment and software to collect eye-tracking data and analyze emotional reactions. The primary eye-tracking tool was the Tobii Pro Spectrum (Tobii Pro AB, 2014). This device allowed us to track both eyes simultaneously, ensuring comprehensive data collection. The eye tracker was set up as a desktop mount to provide a stable and controlled environment for the participants. To increase the quality of the gaze data, we asked the participants to maintain minimal head movement and a constant viewing distance (~55 cm). For the calibration process, we implemented a standard 9-point procedure to ensure the accuracy of the eye-tracking data. This step was critical in matching the eye tracker to the individual characteristics of the participants, thereby improving the reliability of the data. For this study, we were mainly interested in exporting pupil dilation data for each participant. Eye-tracking data were sampled at 120 Hz. The data on pupil size was exported in millimeters and an average between the right and left eye was calculated. In addition to the eye tracker, we used the Noldus FaceReader (Noldus, 2021) for emotion recognition, which allowed us to correlate eye movement data with emotional responses. FaceReader 9 used video material that was sampled at 5 Hz simultaneously with the eye-tracking data via a Logitech HD 1080p webcam. The value for emotional arousal ranges from 0 to 1 and is calculated based on activation values from twenty FACS action units (1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 24, 25, 26, 27 and inverse 43) (Loijens & Krips, n.d.). While values close 0 zero indices a calm emotional state, values close to 1 indicates high activation (Landmann, 2023).

During validation studies of a previous version of Face Reader (Face Reader 6) recognized 89% of emotional labels for basic emotions in the Amsterdam Dynamic Facial Expression Set (154 images) (van der Schalk, et al., 2011) and 88% in the Warsaw Set of Emotional Facial Expression Pictures (207 images) (Olszanowski et al., 2008) (Lewinski et al., 2014). For the specific action units used to calculate emotional arousal in this study, accuracy was between 0.69 (action unit 7 - Lids tight) and 1.00 (action unit 12 - Lip corner puller) (Lewinski et al., 2014). Following from this, the previous version of the Face Reader falls under the accuracy level of .70 where a human coder would receive FACS certification only for action unit 12 (Lewinski et al., 2014). For Face Reader 7, Skiendziel et al. (2019) compared manual and automated facial coding based on the FACS and the category agreement of the majority of action units were >.70 while only action units 14, 15, 18, 20, 23 and 43 performed worse than <.60 (for more details please see original publication). Beyond that, the study of Skiendziel et al. (2019) did not find remarkable differences between trials with and without prior calibration. Based on the results described above and noticing the improvements in Face Reader 9, we are considering Face Reader 9 as a reliable and valid measurement tool for emotional arousal also without prior calibration.

2.4 Data analysis

Data pre-processing

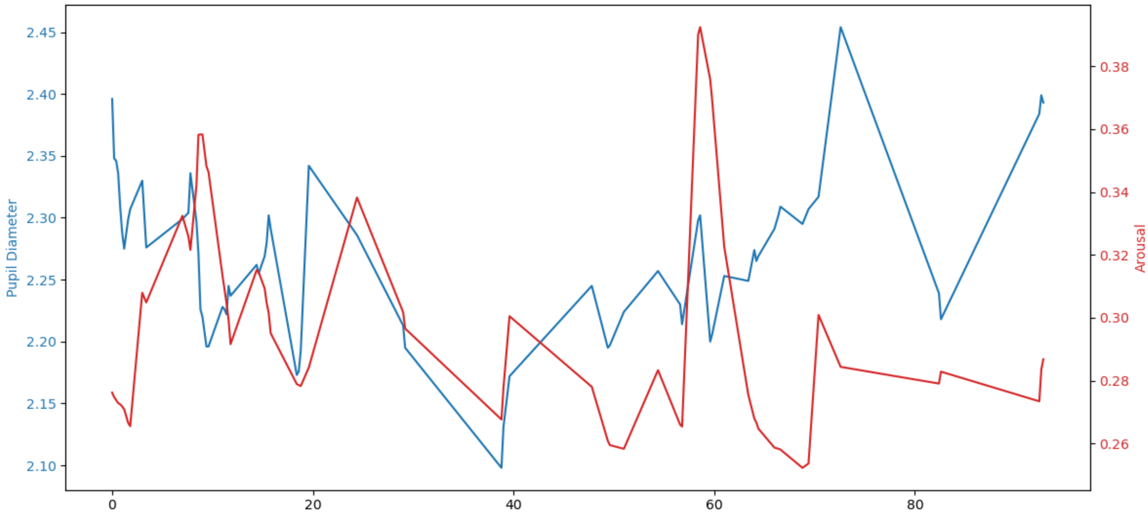
Data pre-processing is an essential phase when aiming to synchronize pupil diameter with emotional arousal (Fink et al., 2023; Kret & Sjak-Shie, 2018). Data processing and analysis were conducted using Python, leveraging its robust ecosystem of libraries to handle a range of tasks: Pandas (McKinney, 2010), SciPy (Virtanen et al., 2020), and NumPy (Harris et al., 2020). We preprocessed all dilemma scenarios and the corresponding data individually. The preprocessing pipeline is shown in more detail below:

1. All participants successfully passed the standard 9-point calibration procedure in Tobii Pro Lab, and calibration accuracy was visually verified prior to each recording session. Eye-tracking data were retained only if both eyes were tracked for at least 85% of the recording time.
2. To address the disruptions caused by blinks in our pupil diameter streams, we implemented a systematic threshold-based correction procedure. As described by Hollander and Huette (2022), there is no consensus on the best way to determine the most appropriate lower blink threshold parameter. We followed Huette's (2016) protocol: blinks were primarily identified based on a minimum blink threshold, which refers to the amount of time that consecutive samples must be missing for an event to be labeled as a blink (here, 100 ms).
3. Once blinks were detected, our next task was to interpolate and correct the affected data points. For this, we employed cubic spline interpolation, a method well suited for ensuring a smooth and natural transition between data points (Dyer & Dyer, 2001). To ensure the reliability of our interpolation, we selected three data points immediately preceding the blink (pre-blink) and three data points following the blink (post-blink). These points served as "anchors," enabling the interpolation to be rooted in actual recorded values and ensuring that the interpolated section seamlessly integrated with the genuine data.
4. Downsampling is a technique used to reduce a signal's sampling rate. One prevalent method involves averaging *N* consecutive samples (Fink et al., 2023; Kret & Sjak-Shie, 2018). This strategy is particularly advantageous for signals with noise, as it offers a smoother representation by counteracting random fluctuations. The value of *N* directly influences the resulting sampling rate. In our case, given the disparity in sampling rates between pupil data (120 Hz) and emotion data (5 Hz), we adjusted the gaze data to match the emotion data's 5 Hz rate. This was achieved by averaging every 24 consecutive samples. The process began by segmenting the signal into non-overlapping chunks, each containing 24 samples. Subsequently, the mean of each segment was computed, yielding a downsampled value for that portion of the data. As a result of the downsampling procedure, we had both pupil diameter and emotional arousal in the dataset, with a sampling rate of 5 Hz.
5. The decision to downsample the eye-tracking data was guided by the need to align both data streams—pupil diameter (120 Hz) and emotional arousal (5 Hz)—for synchronized analysis. As our task involved extended, self-paced reading and decision-making over several seconds, we did not expect short-latency, phasic pupil reactions to specific events. Instead, we were interested in slow, continuous fluctuations in pupil diameter and emotional arousal that reflect sustained cognitive and emotional engagement. Downsampling is therefore not expected to obscure meaningful variance in this context. Following recommendations in multimodal preprocessing literature (Kret & Sjak-Shie, 2018; Fink et al., 2023), we averaged across non-overlapping 24-sample windows to reduce high-frequency noise and produce a smoothed signal suitable for correlation and Granger causality analyses.
6. We did not apply baseline correction relative to a pre-task resting state, as our research focus was not on absolute pupil dilation but rather on the temporal co-variation between emotional arousal and pupil diameter during the task. To reduce inter-individual variability and ensure comparability across participants, pupil diameter values were normalized within each participant and dilemma prior to correlation and causality analyses.
7. Ensuring accurate temporal alignment of the two data streams is vital when studying the relationship between pupil diameter and emotional arousal (see Figure 1). The first step in this process involved an examination of the initial timestamps from both data streams. As they were synchronized using

Observer XT software (Zimmerman et al., 2009), they commenced at identical milliseconds. We then adjusted the timestamps so that each participant's data started at 0 ms. Upon examination of the data, synchronization proved successful, with no discernible offset detected.

Figure 1

Downsampled and synchronized data streams over time (in sec.): pupil diameter plotted against the emotional arousal of a random participant



Data analysis

Analyzing the correlation and exploring the time lag between pupil diameter and emotional arousal data streams involves several steps of analysis:

1. We commenced our analysis by utilizing the full normalized and interpolated data streams for each participant, upon which we calculated Spearman correlation coefficients to evaluate the relationship between emotional arousal and pupil diameter (RQ1a).
2. We then explored epochs of high, medium, and low arousal and the corresponding Spearman correlation coefficients between these different epochs and pupil diameter. First, we defined a high in arousal as any epochs of local maximum that were in the upper third quantile (> 66th percentile), epochs of medium arousal in the middle third quantile (between 33rd and 66th percentiles), and epochs of low arousal in the lower third quantile (< 33rd percentile). Within these epochs, we again calculated the Spearman rank correlation between arousal and pupil diameter, providing a non-parametric measure of association that did not assume a linear relationship (RQ1b).
3. To identify significant time lags between both data streams (RQ2), we applied Granger causality (Granger, 1969). Granger causality is a statistical concept based on predictions. Named after the Nobel laureate Clive Granger, it operates on the principle that if a signal *X* "Granger causes" a signal *Y*, past values of *X* should contain information that helps predict *Y* beyond the information contained in past values of *Y* alone. Granger causality tests are particularly insightful when examining the specific lags at which one time series may forecast another. In essence, the test evaluates whether past values of a time series *X* at various lags contribute unique information that can predict the current value of another time series *Y*. For instance, if we find that *X* Granger causes *Y* at lag 2, it suggests that information from *X* two time units ago provides significant predictive power for the current value of *Y*. This lagged relationship can uncover dynamics that are not immediately apparent, revealing how effects propagate over time in a system. The identification of these specific lags is crucial because it pinpoints the temporal distance over which the predictive relationship holds, which can be vital for understanding

the underlying processes or for developing forecasting models. It is important to clarify that Granger causality should not be interpreted in the same sense as causality in the physical or deterministic sense; it is better understood as "predictive causality."

3. Results

3.1 Task validation and Manipulation checks

Participants' ratings on a scale from 1 to 5 supported the effectiveness of the dilemmas as emotionally and cognitively engaging tasks. Table 1 shows that participants rated the dilemmas as realistic (M = 3.18–3.85), morally serious (M = 3.24–3.65), and emotionally challenging, with high inner conflict (M = 3.41–4.18) and moderate-to-high decision difficulty (M = 3.15–3.94).

Table 1

Descriptive Data of Validation Questions for Dilemmas 1–3

	Dilemma 1		Dilemma 2		Dilemma 3	
	M	*SD*	*M*	*SD*	*M*	*SD*
Realism of dilemma	3.85	1.35	3.18	1.42	3.56	0.79
Seriousness of the conflict	3.24	1.18	3.65	1.28	3.35	1.20
Manifestation of the inner conflict	3.68	1.27	4.18	1.03	3.41	1.21
Difficulty of decision making[a]	3.15	1.26	3.94	1.20	3.26	1.14

Note. [a]Reverse coded item

Emotional ratings from 1 to 5 indicated that participants experienced a differentiated range of affective responses across the three dilemmas, reflecting the intended design of the scenarios. Notably, the emotion "determined" consistently received the highest mean rating across all dilemmas (D1: M = 3.23, SD = 1.10; D2: M = 3.37, SD = 1.42; D3: M = 3.50; SD = 1.42), suggesting a strong motivational engagement with the tasks, regardless of emotional valence.

In Dilemmas 1 and 2, participants reported elevated levels of high arousal and negatively valenced moral emotions, including guilt (D1: M = 1.77, SD = 1.30; D2: M = 2.39, SD = 1.34), shame (D1: M = 1.66, SD = 1.10; D2: M = 1.92, SD = 1.23), sadness (D1: M = 2.09, SD = 1.14; D2: 2.87, SD = 1.39), and anger (D1: M = 2.22, SD = 1.10; D2: M =2.42, SD = 1.32). These ratings align with the moral complexity and emotionally difficult choices presented in the earlier dilemmas.

By contrast, Dilemma 3 elicited relatively higher ratings for positively valenced emotions such as pride (M = 2.71; 1.23), curiosity (M = 2.28; SD = .78), and satisfaction (M = 2.64; SD = 1.30), indicating a shift toward more affirming emotional responses in scenarios with less moral conflict or clearer resolution. In summary, the low to medium means in the self-reported emotions align with low to medium means values from Emotion Recognition. It should be noted that the mean values calculated based on values from continuous data streams with fluctuating values with peaks and lows tend to average out the peaks and lows. Therefore, a low to medium mean value is not surprising and emphasizes the demand for analysis strategies that can cope with dynamic changes as the one reported in this study.

On average, the participants were engaged in the dilemmas M_{d1} = 55.04 sec (SD = 36,24), M_{d2} = 39.35 sec (SD = 17.30), and M_{d3} = 42.09 sec (SD = 31.14). The descriptive data of pupil diameter and emotional arousal are presented in Table 2.

Table 2

Descriptive data of pupil diameter and emotional arousal separated by dilemma (D1, D2, and D3)

	Mean	SD	Min.	Md.	Q3	Max.
Pupil Diameter						
D1	2.54	0.25	1.73	2.53	2.70	3.48
D2	2.57	0.24	1.85	2.57	2.69	4.21
D3	2.59	0.24	2.05	2.55	2.70	3.39
Emotional Arousal						
D1	0.28	0.04	0.01	0.27	0.29	0.77
D2	0.29	0.06	0.02	0.26	0.29	0.74
D3	0.29	0.05	0.05	0.28	0.30	0.68

Note. Q3 indicates mean values for the upper third quartile (>66th percentile); values for pupil diameter are in millimetres; values for emotional arousal ranged from 0 to 1

3.2 Correlation between pupil diameter and emotional arousal (RQ1a)

To assess the distribution characteristics of emotional arousal and pupil diameter across participants, we employed the Shapiro–Wilk test. Our findings revealed significant deviations from a normal distribution for both variables, indicated by p-values below the 0.05 threshold. Consequently, adopting non-parametric methods for subsequent analyses is imperative, aligning with the recommendations of Fink et al. (2023). Consequently, we utilized Spearman rank correlation coefficients to investigate the relationship between emotional arousal and pupil diameter. Our analysis yielded the following insights, categorized by Dilemmas 1–3.

For the first dilemma (D1), the analysis revealed a mean Spearman correlation coefficient of -0.13 across the dataset, indicating a generally weak inverse relationship between the two data streams, which was statistically significant at a p-value of less than 0.001. This correlation coefficient varied significantly among the participants, with the strength of negative correlations reaching as low as -0.65 and positive correlations peaking at 0.38. This wide range underscores the significant individual variability in the relationship between emotional arousal and changes in pupil diameter, suggesting that the connection between these data streams is not only variable but also highly individualized.

For the second dilemma (D2), the Spearman correlation coefficient was found to be -0.09, with the data showing a range from -0.42 to 0.22. Despite being weaker than the correlation observed in the first dilemma (D1), these results still indicate a significant negative correlation, supported by a p-value of less than 0.001, and also highlight a high level of variability among the participants.

Regarding the third dilemma (D3), the analysis yielded a mean Spearman correlation coefficient of -0.11, with correlations ranging from -0.53 to 0.25. This outcome, akin to D1, points to a moderate negative association between the variables, which is statistically significant ($p < 0.001$), underscoring the consistent presence of a significant yet variable inverse relationship across different dilemmas.

Using Spearman rank correlation coefficients, we observed consistent and statistically significant overall negative associations between emotional arousal and pupil diameter across all three dilemmas. Notably, the range of correlation coefficients was quite large, suggesting variability in the strength of the relationship across and within individuals. Therefore, further investigation is needed to elucidate the underlying factors contributing to this variability. One of these further investigations is to examine the relationship between pupil diameter and emotional arousal in different epochs of high, medium, and low emotional arousal.

3.3 Correlation of pupil diameter and emotional arousal in epochs of high, medium, and low emotional arousal (RQ1b)

We then implemented a quantile-based thresholding approach to identify highs in the emotional arousal data stream. Using a quantile-based thresholding approach to classify arousal states into high, medium, and low epochs revealed a balanced distribution among participants across the three dilemmas, with only marginal variations observed between them. This approach ensured that the categorization of arousal levels was directly tied to the data's distribution, making it inherently balanced. This method divides arousal scores into equal-sized groups based on their distribution, which naturally leads to a balanced number of observations across the defined categories (high, medium, and low). The balanced distribution suggests that the dilemmas were effective in eliciting a wide range of emotional responses, with no single arousal epoch dominating across the scenarios.

The Spearman rank correlation coefficient between high emotional epochs (defined as epochs above the 66th percentile) and pupil size is -0.13 ($p < 0.001$) in D1, -0.18 in D2 ($p < 0.001$), and -0.16 in D3 ($p < 0.001$). These results show a significant moderate negative correlation between heightened arousal highs and pupil diameter in all three dilemmas, suggesting that heightened arousal levels are moderately associated with smaller pupil sizes.

Our investigation then turned to epochs characterized by medium emotional arousal. The Spearman rank correlation coefficients for these epochs (defined as falling between the 33rd and 66th percentiles) and the corresponding pupil sizes are approximately D1 = 0.08 ($p < 0.11$), D2 = 0.04 ($p < 0.09$), and D3 = 0.09 ($p < 0.14$). These results suggest a slightly positive but not statistically significant correlation.

We then conducted a parallel investigation focusing on epochs of low emotional arousal. The Spearman rank correlation coefficients for these epochs (defined as falling below the 33rd percentile) and pupil size are approximately D1 = 0.07 ($p < 0.12$), D2 = 0.04 ($p < 0.13$), and D3 = 0.10 ($p < 0.09$). Similar to the moderate arousal epochs, these results indicate a slightly positive but not statistically significant correlation.

These results suggest variability in the relationship between arousal and pupil diameter across different arousal levels. Specifically, in all three dilemmas, a transition from positive correlations in lower arousal states to highly significant negative correlations in higher arousal states was found.

3.4 Time lags between the onset of emotional arousal and corresponding changes in pupil diameter measurements (RQ2)

Granger causality is a statistical hypothesis test used to determine whether one time series can predict another. To calculate the mean time lag between the two time series streams for each participant, we need to determine the specific time lags at which Granger causality is significant for each participant. A significant Granger causality result indicates a predictive relationship between the two time series (see *p*-values in Tables 3–5). If the emotional arousal stream Granger causes the pupil diameter stream, it suggests that past values of emotional arousal contain information that is useful in predicting future values of pupil diameter.

Table 3.

Granger causality for emotional arousal effects pupil diameter at different time lags in Dilemma 1

ID	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6	Lag 7	Lag 8	Lag 9	Lag 10
1	**0.003**	—	—	—	—	—	—	—	—	—
2	0.22	0.62	0.79	0.71	0.81	0.20	0.07	0.10	0.17	0.16
3	0.44	0.99	0.88	0.81	0.70	0.69	0.83	0.74	0.81	0.87
4	**0.05**	—	—	—	—	—	—	—	—	—
5	0.46	0.24	0.19	0.15	**0.05**	—	—	—	—	—
6	**0.05**	—	—	—	—	—	—	—	—	—
7	0.06	**<0.001**	—	—	—	—	—	—	—	—
8	0.11	0.05	**0.05**	—	—	—	—	—	—	—
9	0.28	0.71	0.75	0.40	0.59	0.56	0.75	0.24	0.26	0.28
10	0.12	0.51	0.83	0.88	0.61	0.73	0.91	0.94	0.93	0.91
11	0.63	**0.01**	—	—	—	—	—	—	—	—
12	0.26	0.34	0.41	0.06	0.06	**0.05**	—	—	—	—
13	**<0.001**	—	—	—	—	—	—	—	—	—
14	0.44	0.74	0.88	0.95	0.97	0.98	0.99	0.98	0.99	0.99
15	0.33	0.49	0.62	0.69	0.81	0.87	0.92	0.92	0.93	0.96
16	0.**002**	—	—	—	—	—	—	—	—	—
17	0.84	0.98	0.90	0.79	0.60	**0.04**	—	—	—	—
18	0.13	0.39	0.44	0.41	0.58	0.40	0.48	0.49	0.37	0.43
19	**0.05**	—	—	—	—	—	—	—	0.21	0.29
20	0.19	0.75	0.89	0.83	0.89	0.93	0.87	0.91	0.33	**0.02**
21	**0.05**	—	—	—	—	—	—	—	—	—
22	**0.05**	—	—	—	—	—	—	—	—	—
23	**0.05**	—	—	—	—	—	—	—	—	—
24	0.51	0.57	0.75	0.86	0.95	0.91	0.94	0.97	0.99	0.97
25	0.09	**<0.001**	—	—	—	—	—	—	—	—
26	0.42	0.36	0.06	0.07	0.11	0.21	0.40	0.45	0.29	0.34
27	0.96	0.07	0.14	0.58	0.55	0.30	0.35	0.29	0.31	0.34
28	0.07	0.93	0.84	0.47	0.15	0.22	0.24	0.22	0.15	0.20

Note. *p*-values of Granger causality tests for every participant. Cells are replaced with "—" after a lower time lag is significant for the corresponding participant.

The mean time lag for the first dilemma across all the participants who showed significant Granger causality is approximately 2.81 (or 2.81 ms). This value represents the average lag at which the relationship between arousal and pupil diameter becomes significant, indicating a typical delay in the effect across participants.

Table 4 w

Granger Causality for emotional arousal effects pupil diameter at different time lags in Dilemma 2

ID	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6	Lag 7	Lag 8	Lag 9	Lag 10
1	0.03	—	—	—	—	—	—	—	—	—
2	0.76	0.96	0.48	0.59	0.42	0.23	0.33	0.44	0.53	0.62
3	0.26	0.50	0.58	0.75	0.94	0.96	0.96	0.82	0.67	0.67
4	0.46	0.37	0.58	0.50	0.89	0.95	0.87	**<0.001**	—	—
5	**0.05**	—	—	—	—	—	—	—	—	—
6	0.28	0.36	0.62	0.76	0.80	0.73	0.89	0.90	0.67	0.67
7	**<0.001**	—	—	—	—	—	—	—	—	—
8	0.98	0.20	0.73	0.65	0.65	0.70	**0.03**	—	—	—
9	0.22	**0.01**	—	—	—	—	—	—	—	—
10	0.35	0.65	0.57	0.86	0.93	0.92	0.93	0.85	0.60	0.69
11	**0.03**	—	—	—	—	—	—	—	—	—
12	**0.01**	—	—	—	—	—	—	—	—	—
13	0.56	0.72	0.77	0.46	0.54	0.66	0.62	0.63	0.76	0.54
14	0.26	0.30	0.42	0.28	0.33	0.29	0.14	**0.03**	—	—
15	**<0.001**	—	—	—	—	—	—	—	—	—
16	**0.01**	—	—	—	—	—	—	—	—	—
17	0.60	**<0.001**	—	—	—	—	—	—	—	—
18	0.62	0.69	0.75	0.49	0.22	0.06	0.12	0.13	0.14	0.16
19	0.55	0.15	**0.03**	—	—	—	—	—	—	—
20	**0.02**	—	—	—	—	—	—	—	—	—
21	0.15	0.31	0.34	0.26	0.18	0.52	**0.03**	—	—	—
22	0.88	0.61	0.71	0.47	0.21	0.26	0.11	0.08	0.12	0.17
23	**<0.001**	—	—	—	—	—	—	—	—	—
24	**0.02**	—	—	—	—	—	—	—	—	—
25	0.76	0.96	0.48	0.59	0.42	0.23	0.33	0.44	0.53	0.62
26	0.26	0.50	0.58	0.75	0.94	0.96	0.96	0.82	0.67	0.67
27	0.46	0.37	0.58	0.50	0.89	0.95	0.87	**<0.001**	—	—
28	0.08	0.71	0.88	0.93	0.98	0.98	0.99	0.89	0.88	0.78

Note. *p*-values of Granger causality tests for every participant. Cells are replaced with "—" after a lower time lag is significant for the corresponding participant.

The mean time lag for the second dilemma across all the participants who showed significant Granger causality is approximately 3.44 (or 3.44 ms).

Table 5

Granger Causality for emotional arousal effects pupil diameter at different time lags in Dilemma 3

ID	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6	Lag 7	Lag 8	Lag 9	Lag 10
1	0.60	0.11	0.20	0.12	**0.02**	—	—	—	—	—
2	0.77	0.68	0.85	0.58	0.36	0.24	0.29	0.30	0.33	0.41
3	**<0.001**	—	—	—	—	—	—	—	—	—
4	0.62	0.79	0.91	0.90	0.49	0.15	0.19	0.21	0.09	0.09
5	0.28	0.21	0.24	0.34	0.47	0.48	0.28	0.37	0.32	0.33
6	0.46	0.54	0.78	0.90	0.88	0.63	0.71	0.77	0.85	0.84
7	0.12	0.91	0.79	0.64	0.36	0.43	0.09	0.10	0.16	0.21
8	0.30	0.63	0.55	0.17	0.07	0.09	0.12	0.16	0.22	0.20
9	0.72	0.33	0.47	0.58	0.27	0.13	0.19	0.26	0.32	0.31
10	0.80	0.89	0.19	0.08	0.12	0.19	0.22	0.28	0.16	0.14
11	0.98	0.61	0.71	0.90	0.89	0.75	0.85	0.28	0.06	0.05
12	0.61	0.67	0.76	**0.05**	—	—	—	—	—	—
13	0.75	0.44	0.30	0.51	0.56	0.29	0.37	0.45	0.55	0.64
14	0.47	**<0.001**	—	—	—	—	—	—	—	—
15	0.07	0.52	0.65	0.80	0.88	0.60	0.67	0.40	0.37	0.45
16	**<0.001**	—	—	—	—	—	—	—	—	—
17	**0.05**	—	—	—	—	—	—	—	—	—
18	**<0.001**	—	—	—	—	—	—	—	—	—
19	0.89	0.20	0.38	0.51	0.60	0.60	0.44	0.51	0.70	0.76
20	**0.02**	—	—	—	—	—	—	—	—	—
21	**0.03**	—	—	—	—	—	—	—	—	—
22	0.30	0.85	0.60	0.30	0.26	0.35	0.50	0.59	0.63	0.71
23	0.86	0.65	0.66	0.80	0.90	0.83	0.89	0.89	0.78	0.20
24	0.18	0.41	0.64	0.79	0.73	0.81	0.85	0.91	0.93	0.96
25	0.14	0.65	0.83	0.94	0.93	0.99	0.96	0.96	0.76	0.85
26	0.39	**0.04**	—	—	—	—	—	—	—	—
27	0.33	0.97	1.00	0.90	0.84	0.91	0.81	0.66	0.21	0.21
28	0.60	0.11	0.20	0.12	**0.02**	—	—	—	—	—

Note. *p*-values of Granger causality tests for every participant. Cells are replaced with "—" after a lower time

The mean time lag for the third dilemma across all the participants who showed significant Granger causality is approximately 2.18 (or 2.18 ms). In summary, we observed time lags between 2.81ms for dilemma 1 and 3.44ms for dilemma 2.

4. Discussion

This study investigates the synchronization and correlation of pupil diameter, measured using a stationary eye tracker, and emotional arousal, measured through facial expression recognition, during emotional dilemma tasks. We utilized three everyday moral dilemma scenarios characterized by moral conflicts, chosen based on previous literature indicating their capacity to elicit emotions and cognitive demands, as observed in similar studies (e.g., Fisher, 2000). Our aim was to unveil potential correlations and explore the time lags between these psychophysiological data streams. The analysis revealed moderate negative associations between emotional arousal and pupil diameter across all three emotional dilemmas. Additionally, heightened emotional arousal was notably correlated with smaller pupil size, while moderate and low arousal epochs showed slightly positive but nonsignificant correlations with pupil diameter. These findings underline the nuanced relationship between arousal and pupil diameter across varying arousal levels. We proceeded to investigate time lags, assessing whether one data stream reliably predicts the other while considering (i.e., simulate) different time lags, using Granger causality analysis. Our findings revealed an average lag of approximately 2.81 ms, at which point the relationship between arousal and pupil diameter became significant. This indicates a typical delay in the effects observed among the participants.

In addition to these physiological findings, the subjective validation data provide important converging evidence that the dilemmas successfully engaged both emotional and cognitive processes. Participants rated the dilemmas as realistic, morally serious, and emotionally challenging, with high levels of inner conflict and decision difficulty. Emotional self-report data revealed differentiated emotional experiences, with guilt, shame, and sadness more prominent in Dilemmas 1 and 2, and pride and curiosity more frequent in Dilemma 3. Determination was consistently rated highest across all dilemmas, suggesting sustained motivational engagement. Overall, the mean values indicate low to medium values for self-reported emotions across dilemmas aligning with low to moderate values from emotion recognition. This triangulation strengthens the internal validity of our findings and provides a richer basis for interpreting the observed physiological dynamics.

4.1 Pupil diameter and emotional arousal correlate during epochs of high emotional arousal

Our research delves into the realm of multimodal approaches that integrate cognitive and emotional indicators, thus filling the gap identified by Noorozi et al. (2020). Inspired by the work of Mu et al. (2022), we explored the benefits of multimodal data validation by analyzing correlations between data streams and challenging their concurrent and convergent validity. We found a significant negative correlation systematically between pupil dilation and emotional arousal across all three moral dilemmas. This finding seems to contradict the reported phenomenon on an aggregated level that increased pupil size (as an indicator of cognitive load) goes along with states of higher emotional arousal (Babiker et al., 2013; Bradley et al., 2008; Partala & Surraka, 2003). Nevertheless, it needs to be emphasized that these discrepancies can also be derived from different methodological differences, as a variety of previous studies have chosen approaches such as measuring one indicator during experimental manipulations and emotion induction instead of multimodal measurements.

In our study, we focused on analyzing the correlation between two data streams *over time* as indicators of emotional arousal and cognitive load, which is an essential part of multimodal data validation (Mu et al., 2020). However, while the majority of the above-listed research suggests a positive correlation between pupil diameter and emotional arousal, indicating that pupil diameter tends to increase with increasing emotional arousal, there are instances where a negative correlation or no correlation has been observed (Bebko et al., 2011; Hess & Polt, 1960).

One reason for this discrepancy could be attributed to the everyday moral dilemma embedded within our experimental design. In our study, we assessed both pupil diameter and emotional arousal while participants were actively engaged in moral dilemmas, requiring them to employ strategies to regulate their emotions. This complex interplay between moral judgment and emotional regulation likely influences the observed patterns in pupil responses. For example, Bebko et al. (2011) observed declining pupil diameter during the process of emotion regulation. They observed that when individuals engage in strategies to regulate their emotions, such as suppressing negative emotions, pupil size tends to decrease. On the other hand, other studies have found that when individuals use reappraisal to regulate their emotions (such as reframing a negative situation in a more positive light), both increasing and decreasing negative emotions can lead to pupil dilation.

This dilation is likely a result of the increased cognitive effort required to regulate emotional responses through reappraisal. So, why the apparent discrepancy? One possible explanation is that different emotion regulation strategies may have distinct effects on pupil size. For instance, reappraisal (in comparison to suppression) involves cognitive restructuring and may require more cognitive effort, leading to larger pupil sizes (Bebko et al., 2011). Moreover, we posit that the variance (between participants) in the correlation between emotional arousal and pupil diameter could also be attributed to the utilization of different emotion regulation strategies. Our findings indicate that we also observed positive correlations between arousal and pupil diameter for some participants. This suggests that distinct regulation strategies, such as reappraisal versus suppression, may influence the direction and strength of the correlation. For instance, while suppression of emotions may lead to decreased pupil size due to the inhibition of emotional expressions, reappraisal may entail increased cognitive effort, resulting in larger pupil sizes. Therefore, the participants' emotion regulation strategies could contribute to the observed variability in the relationship between arousal and pupil diameter.

This line of reasoning is also supported by Kinner et al. (2017), who suggest that pupil diameter is modulated by emotional arousal and that it is initially related to the amount of mental effort required to regulate automatic emotional responses.

Beyond that, early research conducted by Stanners et al. (1979), as well as more recent findings by Chen and Epps (2013), demonstrated that the pupil dilates in response to emotional arousal, primarily when cognitive task demands are minimized. These studies suggest that cognitive processes may exert greater control over pupillary responses in tasks that involve both cognitive and emotional components. These previous studies further support our results, as we have shown the absence of any correlation between data streams during epochs characterized by low or moderate emotional arousal states. This also aligns closely with the principles of the Yerkes–Dodson law, which posits that excessively high levels of emotional arousal might impede cognitive processes crucial for learning (Sherwood, 1965; Yerkes & Dodson, 1908). Considering that pupil diameter is an indicator of cognitive load (Mallick et al., 2016; Rodemer et al., 2023; Soussou et al., 2012), high arousal might lead to reduced pupil diameters because it affects working memory and reduces the capability to spend effort on cognitively demanding tasks. This is also in line with findings showing that stress can hinder learning and that stressed individuals show attenuated pupillary responses (de Berker et al., 2016). In addition, research on moral judgment indicates that strong emotional triggers lead to more automatic emotional processes than deliberately controlled processes typically associated with cognitive load (Greene et al., 2001, 2009). In summary, it can be concluded that a closer analysis of the correlation of the psychophysiological data streams in intervals with different levels of arousal is a useful addition to the current state of research.

In addition to the influence of emotion regulation strategies on pupil size and cognitive information processing, it is essential to consider the impact of emotional valence on this relationship (Alsheri & Alghowinem, 2013; Babiker et al., 2013; Kinner et al., 2017; Partala & Surraka, 2003). Emotional valence, or the positivity or negativity of an emotion, can significantly influence pupillary responses. For instance, Hess and Polt's (1960) early discovery of a bidirectional effect remains influential in the study of pupil responses to emotion. Both Hess and Polt (1960) and Kinner et al. (2017) demonstrated that emotional valence can modulate the relationship between pupil diameter and emotional arousal, indicating that specific emotional content plays a crucial role in shaping pupillary responses. This suggests that the relationship between pupil diameter and emotional arousal may be complex and context-dependent, with different emotional states (valence and arousal) eliciting different patterns of pupil responses.

4.2 Emotional arousal precedes and triggers changes in pupil diameter

The methodological framework of our study marks a notable advance by integrating complex Granger causality analysis with psychophysiological data streams. This innovative approach not only represents a technical advance in psychophysiological data analysis (Shojaie & Fox, 2022) but also enriches our understanding of the temporal relationships between emotional arousal and changes in pupil diameter. The Granger causality test, a key component of our methodology, is unique in its ability to statistically determine how one time series may predict future changes in another. Therefore, this feature facilitates the identification of directional relationships between time series, providing a solid foundation for revealing the influence of one variable on another (Barnett et al., 2009).

Through this analytical lens, we are able to dive deeper into the dynamics at play, providing a clearer picture of the pathways underlying the physiological manifestations of emotional processes and, therefore, extending the literature systematically (Oliva & Anikin, 2018). We can confirm that research utilizing a diverse array of psychophysiological sensor streams has consistently shown that within the broad spectrum of physiological reactions, emotional arousal often precedes alterations in pupil diameter (Bradley et al., 2008, 2017; Oliva & Anikin, 2018).

This body of research highlights that emotional arousal stimulates the sympathetic nervous system, leading to a range of physiological effects, including increased heart rate, elevated perspiration levels, and variations in pupil size, with our study concentrating specifically on the latter. Notably, pupil dilation is directly linked to increased sympathetic nervous system activity, whether in response to external stimuli or spontaneously. This is supported by the finding that working with arousing materials leads to an activation of the amygdala, which is important for recall and learning (McGaugh, 2004). Consequently, our results lend further support to the hypothesis that emotional arousal initiates changes in pupil diameter, which aligns with the findings of prior research, such as Bradley et al. (2008).

Uniquely, to our knowledge, our study is the first to apply an emotion recognition system that captures facial expressions to produce a continuous stream of emotional arousal data for the purpose of correlating this stream with physiological data streams, such as pupil diameter. Previous methodologies required participants to manually indicate their arousal or valence levels using button presses (Child et al., 2020; Kinner et al., 2017; Oliva & Anikin, 2018). Our approach, offering millisecond temporal resolution of the emotional arousal stream, not only allows for a precise correlation between emotional arousal and pupil diameter changes but also confirms that emotional arousal precedes these changes. Moreover, our analysis did not uncover any significant Granger causality in the reverse direction, suggesting that pupil diameter changes follow rather than lead to emotional arousal. This absence of reverse causality strongly supports the notion that emotional arousal is a primary driver of pupil diameter changes rather than a converse.

For educational practice and research, it is important to consider possible analytic bottlenecks that can occur when a wide variety of multimodal data streams are merged. As described by Azevedo and Gasevic (2019), latencies in delivering inferences and cues caused by this analytic bottleneck might negatively affect the learning process. Therefore, it is crucial to be aware of the congruent validity of measures to select the right data streams, combine them in a beneficial way (Mu et al., 2022), and use the knowledge of time lags to identify and use “initial reactions” for cues in adaptive and remote learning environments.

4.3 Limitations and future research

While this investigation has shed light on the complex relationship between emotional arousal and changes in pupil diameter, it is important to acknowledge several limitations that may inform future research directions. First, our study did not consider the potential influence of individual differences in emotion regulation strategies, which are known to significantly influence physiological responses to emotional stimuli. It is critical that future studies assess participants’ emotion regulation strategies to disentangle how these strategies may play a role in mediating or moderating physiological outcomes in response to emotional arousal.

In addition, our methodology omitted the assessment of the participants’ stress levels, which may have missed a crucial element that affects both emotional arousal and physiological responses. Implementing measures to assess stress levels both before and during the experimental procedure could provide critical insight into how stress modulates emotional arousal and its associated physiological responses.

Furthermore, the experimental design did not adequately account for variability in the intensity of the dilemmas presented to the participants, which could have led to discrepancies in task load and, by extension, cognitive workload. Cognitive workload, which encompasses both task load and mental effort, likely influenced the participants’ emotional and physiological responses, but this variable was not methodologically controlled or quantified in our study. Future research should strive for a standardized approach to dilemmas or task presentation to ensure consistent cognitive load across different scenarios. In addition, examining how cognitive workload interacts with emotional arousal may shed light on the mechanisms by which these factors collectively influence physiological responses.

It is important to note that our results indicate that Granger causality analysis in RQ2 did not yield statistically significant results for each participant when considering time lags of up to 10 milliseconds. This observation suggests a complex interplay between emotional arousal and pupil diameter changes that may not be captured consistently across individuals within such a short time window. This lack of consistency across participants underscores the inherent variability in psychophysiological responses to emotional stimuli, reflecting individual differences in the rate and manner of processing such stimuli. This suggests the potential need for a more detailed analysis, or possibly the inclusion of additional data points or variables, that could help to better understand these dynamics. Furthermore, this finding encourages further investigation into the temporal resolution and sensitivity of our methods when exploring intricate causal relationships within psychophysiological data streams. Beyond that, we observed a difference in the time lag between dilemmas (dilemma 1: 2.81ms, dilemma 2: 3.44ms, dilemma 3: 2.18ms). When interpreting differences in the time lag between dilemmas, also differences in the material might contribute. For example, dilemma 2 was, on average, rated as a little less realistic than the other dilemmas, but decision-making was harder as the inner conflict and the seriousness of the conflict was described as the highest (please see table 2 for detailed values). Nevertheless, we did not test whether experiencing a more intensive inner conflict and rating the seriousness of the conflict higher in a less authentic dilemma task directly leads to more emotional arousal on an individual level or not and, therefore, we cannot directly measure its impact on the time lag. Future research should take this into consideration.

In light of our findings (especially RQ1a & RQ1b), future research should seek to enrich our understanding of the nuanced relationship between emotional arousal and physiological indicators, such as changes in pupil diameter, paying particular attention to the variability observed among individual participants. The discovery that the correlation between emotional arousal and pupil diameter tends to be slightly below zero, albeit with considerable variability across participants, highlights the need to explore individual differences more deeply. This exploration is crucial for disentangling the layers of complexity that characterize the interplay between emotional processes and physiological responses.

4.4 Conclusion

This study explores the complex relationship between emotional arousal and pupil diameter changes using moral dilemmas to elicit emotional and cognitive responses. By integrating stationary eye-tracking and facial expression recognition technologies, we aimed to identify correlations and time lags between these psychophysiological data streams. Our results show a moderate negative correlation between emotional arousal and pupil diameter, suggesting a nuanced interaction that varies with arousal intensity. Specifically, higher levels of emotional arousal are associated with smaller pupil sizes, whereas lower levels of arousal have a less pronounced, nonsignificant effect on pupil diameter. Further exploration using Granger causality analysis revealed a typical delay of approximately 2.8 ms in the significant relationship between arousal and pupil changes, suggesting a short latency in the physiological response to emotional stimuli across individuals. These results contribute to our understanding of the dynamic interplay between emotional states and physiological responses and highlight the complexity of the impact of emotional arousal on pupil diameter in cognitive–emotional contexts. The results emphasize the relevance of research using multimodal approaches and considering the convergent validity of measures for cognitive workload and emotional arousal to improve adaptive learning environments.

Keypoints

- This research introduces a novel multimodal approach, combining eye-tracking and deep-learning-based emotion recognition.
- We explore how cognitive load and emotional arousal interact during emotional and cognitive engaging problem-solving tasks.
- Findings reveal distinct, intensity-dependent relationships between pupil diameter and emotional arousal and an average time lag of 2,8ms.
- The findings are challenging earlier assumptions and underscoring the importance of nuanced, real-time data analysis in education.

Acknowledgments

Data availability statement: The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

Informed consent: Written informed consent was obtained from all participants prior to participation.

Funding: No funding was received for conducting this study.

Financial interest: The authors have no relevant financial or non-financial interests to disclose.

References

Abdallah, T. B., Elleuch, I., & Guermazi, R. (2021). Student behavior recognition in classroom using deep transfer learning with VGG-16. *Procedia Computer Science, 192*, 951–960. https://doi.org/10.1016/j.procs.2021.08.098

Alshanskaia, E. I., Portnova, G. V., Liaukovich, K. & Martynova, O. V. (2024). Pupillometry and autonomic nervous system responses to cognitive load and false feedback: an unsupervised machine learning approach. *Frontiers in Neuroscience, 18*. https://doi.org/10.3389/fnins.2024.1445697

Al-Elq, A. H. (2010). Simulation-based medical teaching and learning. Journal of Family and *Community Medicine, 17*(1), 35. https://doi.org/10.4103/1319-1683.68787

Antonenko, P. D., Paas, F., Grabner, R. H. & Van Gog, T. (2010). Using electro-encephalography to measure cognitive load. *Educational Psychology Review, 22*(4), 425–438. https://doi.org/10.1007/s10648-010-9130-y

Alshehri, M. & Alghowinem, S. (2013). An exploratory study of detecting emotion states using eye-tracking technology. *Proceedings of 2013 Science and Information Conference*.

Azevedo, R. & Gašević, D. (2019). Analyzing Multimodal Multichannel Data about Self-Regulated Learning with Advanced Learning Technologies: Issues and Challenges. *Computers in Human Behavior*, *96*, 207–210. https://doi.org/10.1016/j.chb.2019.03.025

Azevedo, R., Taub, M., & Mudrick, N. V. (2018). Understanding and reasoning about real-time cognitive, affective, and metacognitive processes to foster self-regulation with advanced learning technologies. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (2nd ed., pp. 254–270). Routledge/Taylor & Francis Group. https://doi.org/10.4324/9781315697048-17

Babiker, A., Faye, I., & Malik, A. (2013). Pupillary behavior in positive and negative emotions. *2013 IEEE International Conference on Signal and Image Processing Applications*. https://doi.org/10.1109/icsipa.2013.6708037

Barnett, L., Barrett, A. B., & Seth, A. K. (2009). Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables. *Physical Review Letters*, *103*(23). https://doi.org/10.1103/physrevlett.103.238701

Beatty, J., & Lucero-Wagoner, B. (2000). The pupillar system. In C. JT, T. LG, & B. GG (Eds.), Handbook of Psychophysiology (pp. 142– 162). Cambridge, UK: Cambridge University Press

Bebko, G. M., Franconeri, S. L., Ochsner, K. N., & Chiao, J. Y. (2011). Look before you regulate: Differential perceptual strategies underlying expressive suppression and cognitive reappraisal. *Emotion*, *11*(4), 732–742. https://doi.org/10.1037/a0024009

Bhattacharyya, A., Chatterjee, S., Sen, S., Sinitca, A. M., Kaplun, D. & Sarkar, R. (2021). A deep learning model for classifying human facial expressions from infrared thermal images. *Scientific Reports, 11*(1). https://doi.org/10.1038/s41598-021-99998-z

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, *45*(4), 602–607. https://doi.org/10.1111/j.1469-8986.2008.00654.x

Bauman, C. W., McGraw, A. P., Bartels, D. M., & Warren, C. (2014). Revisiting External Validity: Concerns about Trolley Problems and Other Sacrificial Dilemmas in Moral Psychology. *Social and Personality Psychology Compass*, *8*(9), 536–554. https://doi.org/10.1111/spc3.12131

Chanel, C., Wilson, M. D., & Scannella, S. (2019). Online ECG-based Features for Cognitive Load Assessment. *Open Archive Toulouse Archive Ouverte (University of Toulouse)*. https://doi.org/10.1109/smc.2019.8914002
Chernikova, O., Heitzmann, N., Stadler, M., Holzberger, D., Seidel, T. & Fischer, F. (2020). Simulation-Based Learning in Higher Education: A Meta-Analysis. *Review of Educational Research, 90*(4), 499–541. https://doi.org/10.3102/0034654320933544
Chen, S. and Epps, J. (2013) Automatic Classification of Eye Activity for Cognitive Load Measurement with Emotion Interference. *Computer Methods and Programs in Biomedicine*, *110*, 111-124. https://doi.org/10.1016/j.cmpb.2012.10.021
Child, S., Oakhill, J., & Garnham, A. (2020). Tracking your emotions: An eye-tracking study on reader's engagement with perspective during text comprehension. *Quarterly Journal of Experimental Psychology*, https://doi.org/10.1177/1747021820905561
Christensen, J. F. & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: A principled review. *Neuroscience & Biobehavioral Reviews*, *36*(4), 1249–1264. https://doi.org/10.1016/j.neubiorev.2012.02.008
Cloude, E. B., Azevedo, R., Winne, P. H., Biswas, G., & Jang, E. E. (2022). System design for using multimodal trace data in modeling self-regulated learning. *Frontiers in Education*, *7*. https://doi.org/10.3389/feduc.2022.928632
de Berker, A. O., Tirole, M., Rutledge, R. B., Cross, G. F., Dolan, R. J., & Bestmann, S. (2016). Acute stress selectively impairs learning to act. *Scientific Reports*, *6*(1). https://doi.org/10.1038/srep29816
Delliaux, S., Delaforge, A., Deharo, J. & Chaumet, G. (2019). Mental workload alters heart rate variability, lowering non-linear dynamics. *Frontiers in Physiology, 10*. https://doi.org/10.3389/fphys.2019.00565
Cloude, E. B., Azevedo, R., Winne, P. H., Biswas, G. & Jang, E. E. (2022). System design for using multimodal trace data in modeling self-regulated learning. *Frontiers in Education*, *7*. https://doi.org/10.3389/feduc.2022.928632
Dindar, M., Järvelä, S., Ahola, S., Huang, X. & Zhao, G. (2022). Leaders and Followers Identified by Emotional Mimicry During Collaborative Learning: A Facial Expression Recognition Study on Emotional Valence. *IEEE Transactions On Affective Computing*, *13*(3), 1390–1400. https://doi.org/10.1109/taffc.2020.3003243
Dubovi, I. (2022). Cognitive and emotional engagement while learning with VR: The perspective of multimodal methodology. *Computers & Education*, *183*, 104495. https://doi.org/10.1016/j.compedu.2022.104495
Dyer, S. A. & Dyer, J. S. (2001). Cubic-spline interpolation. *IEEE Instrumentation & Measurement Magazine*, *4*(1), 44–46. https://doi.org/10.1109/5289.911175
Elahi, E., & Islam, D. (2019). *Galvanic Skin Response signal based Cognitive Load classification using Machine Learning classifier*. https://doi.org/10.1109/icecte48615.2019.9303564
Ez-zaouia, M. & Lavoué, E. (2018). EMODA: a Tutor Oriented Multimodal and Contextual Emotional Dashboard. LAK '17: Proceedings of the Seventh International Learning Analytics & Knowledge Conference. https://doi.org/10.1145/3027385.3027434
Fiedler, S., Glöckner, A., Dickert, S. & Nicklisch, A. (2013). Social Value Orientation and information search in social dilemmas: An eye-tracking analysis. *Organizational Behavior And Human Decision Processes, 120*(2), 272–284. https://doi.org/10.1016/j.obhdp.2012.07.002
Fink, L., Simola, J., Tavano, A., Lange, E. B., Wallot, S. & Laeng, B. (2023). From pre-processing to advanced dynamic modeling of pupil data. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02098-1
Fisher, C.D. & Ashkanasy, N.M. (2000). The emerging role of emotions in work life: an introduction. *Journal of Organizational Behavior*, *21* (2), 123-129.
Frei-Landau, R. & Levin, O. (2023). Simulation-based learning in teacher Education: Using Maslow's hierarchy of needs to conceptualize instructors' needs. *Frontiers in Psychology, 14*. https://doi.org/10.3389/fpsyg.2023.1149576
Fritz, T., Begel, A., Müller, S., Yiğit-Elliott, S. & Züger, M. (2014). Using psycho-physiological measures to assess task difficulty in software development. *Proceedings of the 36th International Conference on Software Engineering.* https://doi.org/10.1145/2568225.2568266

Ghaffari, M. & Fiedler, S. (2018). The Power of Attention: Using Eye Gaze to Predict Other-Regarding and Moral Choices. *Psychological Science, 29*(11), 1878–1889. https://doi.org/10.1177/0956797618799301
Gagné, R. M., Wager, W. W., Golas, K. & Keller, J. M. (2005). Principles of instructional design. Cengage Learning.
Gao, H. & Ma, B. (2020). A robust improved network for facial expression recognition. *Frontiers in signal processing, 4*(4). https://doi.org/10.22606/fsp.2020.44001
Garon, M., Lavallée, M. M., Estay, E. V. & Beauchamp, M. H. (2018). Visual encoding of social cues predicts sociomoral reasoning. *PLOS ONE, 13*(7). https://doi.org/10.1371/journal.pone.0201099
Gleichgerrcht, E. & Young, L. (2013). Low Levels of Empathic Concern Predict Utilitarian Moral Judgment. PLOS ONE, 8(4). https://doi.org/10.1371/journal.pone.0060418
Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica, 37*(3), 424. https://doi.org/10.2307/1912791
Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: the interaction between personal force and intention in moral judgment. *Cognition*, *111*(3), 364–371. https://doi.org/10.1016/j.cognition.2009.02.001
Grassmann, M., Vlemincx, E., Von Leupoldt, A., Mittelstädt, J. & Van Den Bergh, O. (2016). Respiratory Changes in Response to Cognitive Load: A Systematic review. *Neural Plasticity, 2016*, 1–16. https://doi.org/10.1155/2016/8146809
Harley, J. M., Bouchet, F., Hussain, S., Azevedo, R., & Calvo, R. (2014). A multi-componential analysis of emotions during complex learning with an intelligent multi-agent system. *Paper to be presented at a symposium on Interdisciplinary Approaches for Analyzing Data from Multiple Affective Channels with Computer-Based Learning Environments at the 2014 annual meeting of the American Educational Research Association*, Philadelphia, PA.
Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Van Kerkwijk, M. H., Brett, M., Haldane, A., Del Río, J. F., Wiebe, M., Peterson, P., . . . Oliphant, T. E. (2020). Array programming with NumPy. *Nature, 585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2
Hess, E. H., & Polt, J. M. (1960). Pupil size as related to interest value of visual stimuli. *Science*, *132*, 349–350. https://doi.org/10.1126/science.132.3423.34
Hollander, J. & Huette, S. (2022). Extracting blinks from continuous eye-tracking data in a mind wandering paradigm. *Consciousness and Cognition, 100*, 103303. https://doi.org/10.1016/j.concog.2022.103303
Horvers, A., Tombeng, N., Bosse, T., Lazonder, A. W. & Molenaar, I. (2021). Detecting Emotions through electrodermal Activity in learning Contexts: A Systematic review. *Sensors, 21*(23), 7869. https://doi.org/10.3390/s21237869
Huette, S. (2016). Blink durations reflect mind wandering during reading. https://pdfs.semanticscholar.org/fba2/314dea3944723ddc0348dc41a6e823dd5410.pdf
Hulshof, C. D. (2005). Log File Analysis. *Encyclopedia of Social Measurement*, 577–583. https://doi.org/10.1016/b0-12-369398-5/00509-0
Janning, R., Schatten, C. & Schmidt-Thieme, L. (2016). Perceived Task-Difficulty recognition from log-file information for the use in adaptive intelligent tutoring systems. *International Journal of Artificial Intelligence in Education*, *26*(3), 855–876. https://doi.org/10.1007/s40593-016-0097-9
Jerčić, P., Sennersten, C. & Lindley, C. A. (2018). Modeling cognitive load and physiological arousal through pupil diameter and heart rate. *Multimedia Tools and Applications*, *79*(5–6), 3145–3159. https://doi.org/10.1007/s11042-018-6518-z
Kahane, G. (2015). Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. *Social Neuroscience*, *10*(5), 551–560. https://doi.org/10.1080/17470919.2015.1023400
Karch, J. M., Valles, J. C. G. & Sevian, H. (2019). Looking into the Black Box: Using Gaze and Pupillometric Data to Probe How Cognitive Load Changes with Mental Tasks. *Journal Of Chemical Education*, *96*(5), 830–840. https://doi.org/10.1021/acs.jchemed.9b00014
Kinner, V. L., Kuchinke, L., Dierolf, A. M., Merz, C. J., Otto, T. & Wolf, O. T. (2017). What our eyes tell us about feelings: Tracking pupillary responses during emotion regulation processes. *Psychophysiology, 54*(4), 508–518. https://doi.org/10.1111/psyp.12816

Kulke, L., Feyerabend, D. & Schacht, A. (2020). A Comparison of the Affectiva iMotions Facial Expression Analysis Software With EMG for Identifying Facial Expressions of Emotion. *Frontiers in Psychology*, 11. https://doi.org/10.3389/fpsyg.2020.00329

Kret, M. E. & Sjak-Shie, E. E. (2018). Preprocessing pupil size data: guidelines and code. *Behavior Research Methods, 51*(3), 1336–1342. https://doi.org/10.3758/s13428-018-1075-y

Landmann, E. (2023). I can see how you feel—Methodological considerations and handling of Noldus's FaceReader software for emotion measurement. *Technological Forecasting And Social Change*, 197, 122889. https://doi.org/10.1016/j.techfore.2023.122889

LeDoux, J. E. (2021). As soon as there was life, there was danger: the deep history of survival behaviours and the shallower history of consciousness. *Philosophical Transactions of the Royal Society B*, *377*(1844). https://doi.org/10.1098/rstb.2021.0292

Liu, Z. (2004). Measuring the degree of synchronization from time series data. *Europhysics Letters (EPL)*, *68*(1), 19–25. https://doi.org/10.1209/epl/i2004-10173-x

Löwenstein, O. (1920). Experimentelle Beiträge zur Lehre von den katatonischen Pupillenveränderungen. European *Neurology, 47*(4), 194–215. https://doi.org/10.1159/000190690

Loijens & Krips (n.d.), Face reader – Methodology Note. Wageningen, The Netherlands: Noldus Information Technology.

Lu, J., Hallinger, P. & Showanasai, P. (2014). Simulation-based learning in management education. *Journal of Management Development, 33*(3), 218–244. https://doi.org/10.1108/jmd-11-2011-0115

Mallick, R., Slayback, D., Touryan, J., Ries, A. J. & Lance, B. J. (2016). The use of eye metrics to index cognitive workload in video games. *2016 IEEE Conference On Eye Tracking And Visualization (ETVIS)*, 60–64. https://doi.org/10.1109/etvis.2016.7851168

Malmberg, J., Haataja, E., Seppänen, T. & Järvelä, S. (2019). Are we together or not? The temporal interplay of monitoring, physiological arousal and physiological synchrony during a collaborative exam. *International Journal of Computer-Supported Collaborative Learning, 14*(4), 467–490. https://doi.org/10.1007/s11412-019-09311-4

Martin, A. J., Ginns, P., Burns, E., Kennett, R., Munro-Smith, V., Collie, R. J. & Pearson, J. (2021). Assessing instructional cognitive load in the context of students' psychological challenge and threat orientations: A Multi-Level Latent Profile Analysis of students and Classrooms. *Frontiers in Psychology, 12*. https://doi.org/10.3389/fpsyg.2021.656994

Mathôt, S. (2018). Pupillometry: Psychology, Physiology, and Function. Journal Of Cognition, 1(1). https://doi.org/10.5334/joc.18

Mauss, I. B. & Robinson, M. D. (2009). Measures of Emotion: A review. *Cognition & Emotion, 23*(2), 209–237. https://doi.org/10.1080/02699930802204677

Mayer, C., Rausch, A. & Seifried, J. (2023). Analysing domain-specific problem-solving processes within authentic computer-based learning and training environments by using eye-tracking: a scoping review. *Empirical Research in Vocational Education And Training*, *15*(1). https://doi.org/10.1186/s40461-023-00140-2

McGaugh, J. L. (2004). The amygdala modulates the consulidation of memories of emotionally arousing experiences. *Annual Review Of Neuroscience*, 27(1), 1–28. https://doi.org/10.1146/annurev.neuro.27.070203.144157

McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the Python in Science Conferences*. https://doi.org/10.25080/majora-92bf1922-00a

Mehrotra, S. & Gunalakshmi, K. (2020). Impact of intrinsic cognitive load and extraneous cognitive load over emotions. *International journal of scientific and research publications, 10*(06), 318–328. https://doi.org/10.29322/ijsrp.10.06.2020.p10237

Mu, S., Cui, M. & Huang, X. (2020). Multimodal Data Fusion in Learning Analytics: A Systematic review. *Sensors*, *20*(23), 6856. https://doi.org/10.3390/s20236856

Noldus (2021). FaceReader™ 8: Tool for automatic analysis of facial expressions. Wageningen, The Netherlands: Noldus Information Technology.

Noroozi, O., Alikhani, I., Järvelä, S., Kirschner, P. A., Juuso, I. & Seppänen, T. (2019). Multimodal data to design visual learning analytics for understanding regulation of learning. *Computers in Human Behavior*, *100*, 298–304. https://doi.org/10.1016/j.chb.2018.12.019

Noroozi, O., Pijeira-Díaz, H. J., Sobocinski, M., Dindar, M., Järvelä, S. & Kirschner, P. A. (2020). Multimodal data indicators for capturing cognitive, motivational, and emotional learning processes: A systematic literature review. *Education And Information Technologies*, *25*(6), 5499–5547. https://doi.org/10.1007/s10639-020-10229-w

Oliva, M., & Anikin, A. (2018). Pupil dilation reflects the time course of emotion recognition in human vocalizations. *Scientific Reports*, *8*(1), 1–10. https://doi.org/10.1038/s41598-018-23265-x

Olszanowski, M., Pochwatko, G., Kuklinski, K., Scibor-Rylski, M., & Ohme, R. (2008, June). *Warsaw set of emotional facial expression pictures validation study*. Opatija, Croatia: EAESP General Meeting

Paas, F. & Ayres, P. (2014). Cognitive Load Theory: A Broader view on the role of memory in learning and education. *Educational Psychology Review, 26*(2), 191–195. https://doi.org/10.1007/s10648-014-9263-5

Paas, F., Van Merriënboer, J. J. G. & Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Perceptual and Motor Skills, 79*(1), 419–430. https://doi.org/10.2466/pms.1994.79.1.419

Pärnamets, P., Johansson, P., Häll, L., Balkenius, C., Spivey, M. J. & Richardson, D. C. (2015). Biasing moral decisions by exploiting the dynamics of eye gaze. *Proceedings Of The National Academy Of Sciences Of The United States Of America, 112*(13), 4170–4175. https://doi.org/10.1073/pnas.1415250112

Partala, T. & Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International Journal Of Human-Computer Studies, 59*(1–2), 185–198. https://doi.org/10.1016/s1071-5819(03)00017-x

Paulus, P. C., Dabas, A., Felber, A. & Benoit, R. G. (2022). Simulation-based learning influences real-life attitudes. *Cognition, 227*, 105202. https://doi.org/10.1016/j.cognition.2022.105202

Pekrun, R. (1992). The impact of emotions on learning and achievement: Towards a theory of cognitive/motivational mediators. *Applied Psychology: An International Review*, 41(4), 359–376.

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, 18(4), 315–341.

Podder, T., Bhattacharya, D., Majumder, P. & Balas, V. E. (2023). A feature boosted deep learning method for automatic facial expression recognition. *PeerJ, 9*, e1216. https://doi.org/10.7717/peerj-cs.1216

Rodemer, M., Karch, J. M. & Bernholt, S. (2023). Pupil dilation as cognitive load measure in instructional videos on complex chemical representations. *Frontiers in Education, 8*. https://doi.org/10.3389/feduc.2023.1062053

Sailer, M., Bauer, E., Hofmann, R., Kiesewetter, J., Glas, J., Gurevych, I. & Fischer, F. (2023). Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning. *Learning and Instruction, 83*, 101620. https://doi.org/10.1016/j.learninstruc.2022.101620

Schmidt-Weigand, F., Kohnert, A. & Glowalla, U. (2010). A closer look at split visual attention in system- and self-paced instruction in multimedia learning. *Learning and Instruction, 20*(2), 100–110. https://doi.org/10.1016/j.learninstruc.2009.02.011

Skiendziel, T., Rösch, A. G. & Schultheiss, O. C. (2019). Assessing the convergent validity between the automated emotion recognition software Noldus FaceReader 7 and Facial Action Coding System Scoring. PLoS ONE, 14(10), e0223905. https://doi.org/10.1371/journal.pone.0223905

Schmidt-Weigand, F. & Scheiter, K. (2011). Cognitive Load Questionnaire. *PsycTESTS Dataset*. https://doi.org/10.1037/t12856-000

Samuelsen, J., Chen, E. & Wasson, B. (2019). Integrating multiple data sources for learning analytics – review of literature. Research and Practice in Technology Enhanced Learning, 14(11). https://doi.org/10.1186/s41039-019-0105-4

Schnaubert, L. & Schneider, S. (2022). Analysing the relationship between mental load or mental effort and metacomprehension under different conditions of multimedia design. *Frontiers in Education, 6*. https://doi.org/10.3389/feduc.2021.648319

Sherwood, J. J. (1965). A relation between arousal and performance. American Journal of *Psychology, 78*(3), 461. https://doi.org/10.2307/1420581

Shojaie, A. & Fox, E. B. (2022). Granger Causality: A Review and Recent Advances. *Annual Review Of Statistics And Its Application, 9*(1), 289–319. https://doi.org/10.1146/annurev-statistics-040120-010930

Šimić, G., Tkalčić, M., Vukić, V., Mulc, D., Španić, E., Šagud, M., Olucha-Bordonau, F. E., Vukšić, M. & Hof, P. R. (2021). *Understanding Emotions: Origins and Roles of the Amygdala. Biomolecules, 11*(6), 823. https://doi.org/10.3390/biom11060823

Song, Z. (2021). Facial Expression Emotion Recognition model integrating philosophy and machine learning theory. *Frontiers in Psychology, 12*. https://doi.org/10.3389/fpsyg.2021.759485

Sörqvist, P., Dahlström, Ö., Karlsson, T. & Rönnberg, J. (2016). Concentration: the neural underpinnings of how cognitive load shields against distraction. *Frontiers in Human Neuroscience, 10*. https://doi.org/10.3389/fnhum.2016.00221

Su, Y., Lin, Y. & Liu, T. (2022). Applying machine learning technologies to explore students' learning features and performance prediction. *Frontiers in Neuroscience, 16*. https://doi.org/10.3389/fnins.2022.1018005

Stanners, R. F., Coulter, I. M., Sweet, A. W. & Murphy, P. (1979). The pupillary response as an indicator of arousal and cognition. *Motivation And Emotion, 3*(4), 319–340. https://doi.org/10.1007/bf00994048

Sweller, J., Van Merriënboer, J. J. G. & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251–296. https://doi.org/10.1023/a:1022193728205

Thompson, R. J. & Berenbaum, H. (2006). Shame Reactions to Everyday Dilemmas are Associated with Depressive Disorder. *Cognitive Therapy And Research*, *30*(4), 415–425. https://doi.org/10.1007/s10608-006-9056-3

Tobii Pro AB (2014). Tobii Pro Lab User Manual [Apparatus and software]. Tobii Pro AB, Danderyd, Sweden.

Voigt, P., & von dem Bussche, A. (2017). *The EU General Data Protection Regulation (GDPR)*. Springer International Publishing. https://doi.org/10.1007/978-3-319-57959-7

Tyng, C. M., Amin, H. U., Saad, M. N. M. & Malik, A. S. (2017). The influences of emotion on learning and memory. *Frontiers in Psychology, 8*. https://doi.org/10.3389/fpsyg.2017.01454

Valdesolo, P. & DeSteno, D. (2006). Manipulations of Emotional Context Shape Moral Judgment. *Psychological Science, 17*(6), 476–477. https://doi.org/10.1111/j.1467-9280.2006.01731.x

Van Acker, B. B., Parmentier, D., Vlerick, P. & Saldien, J. (2018). Understanding mental workload: from a clarifying concept analysis toward an implementable framework. *Cognition, Technology & Work, 20*(3), 351–365. https://doi.org/10.1007/s10111-018-0481-3

van der Schalk, J., Hawk, S. T., Fischer, A. H., & Doosje, B. (2011). Moving faces, looking places: Validation of the Amsterdam Dynamic Facial Expression Set (ADFES). *Emotion*, 11, 907–920. http://dx.doi.org/10.1037/a0023853

Van Der Wel, P. & Van Steenbergen, H. (2018). Pupil dilation as an index of effort in Cognitive control Tasks: a review. *Psychonomic Bulletin & Review, 25*(6), 2005–2015. https://doi.org/10.3758/s13423-018-1432-y

Vanneste, P., Raes, A., Morton, J., Bombeke, K., Van Acker, B. B., Larmuseau, C., Depaepe, F. & Van Den Noortgate, W. (2020). Towards measuring cognitive load through multimodal physiological data. *Cognition, Technology & Work, 23*(3), 567–585. https://doi.org/10.1007/s10111-020-00641-0

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., Van Der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A., Jones, E. D., Kern, R., Larson, E. B., . . . Vázquez-Baeza, Y. (2020). SCIPY 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods, 17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Yerkes, R. M. & Dodson, J. (1908). The relation of strength of stimulus to rapidity of habit-formation. *The Journal of Comparative Neurology, 18*(5), 459–482. https://doi.org/10.1002/cne.920180503

Yoo, G., Kim, H. & Hong, S. (2023). Prediction of cognitive load from electroencephalography signals using Long Short-Term Memory Network. *Bioengineering, 10*(3), 361. https://doi.org/10.3390/bioengineering10030361

Young, J. Q., Van Merriënboer, J., Durning, S. & Cate, O. T. (2014). Cognitive Load Theory: Implications for Medical Education: AMEE Guide No. 86. *Medical Teacher, 36*(5), 371–384. https://doi.org/10.3109/0142159x.2014.889290

Zandi, B., Lode, M., Herzog, A. G., Sakas, G. & Khanh, T. Q. (2021). PupilEXT: flexible Open-Source platform for High-Resolution Pupillometry in vision Research. *Frontiers in Neuroscience, 15*. https://doi.org/10.3389/fnins.2021.676220

Zimmerman, P., Bolhuis, J., Willemsen, A., Meyer, E. S. & Noldus, L. (2009). The Observer XT: a tool for the integration and synchronization of multimodal signals. *Behavior Research Methods, 41*(3), 731–735. https://doi.org/10.3758/brm.41.3.731