MSP-GEO Corpus: A Multimodal Database for Understanding Video-Learning Experience

Ali N. Salman
ali.salman@utdallas.edu
Department of Electrical and Computer Engineering
The University of Texas at Dallas
Richardson, Texas, USA

Ning Wang
Ning.Wang@utdallas.edu
Department of Sustainable Earth Systems Science
The University of Texas at Dallas
Richardson, Texas, USA

Luz Martinez-Lucas
Luz.Martinez-Lucas@utdallas.com
Department of Electrical and Computer Engineering
The University of Texas at Dallas
Richardson, Texas, USA

Andrea Vidal
AXV170003@utdallas.edu
Department of Electrical and Computer Engineering
The University of Texas at Dallas
Richardson, Texas, USA

Carlos Busso
busso@utdallas.edu
Department of Electrical and Computer Engineering
The University of Texas at Dallas
Richardson, Texas, USA

ABSTRACT

Video-based learning has become a popular, scalable, and effective approach for students to learn new skills. Many of the challenges for video-based learning can be addressed with machine learning models. However, the available datasets often lack the rich source of data that is needed to accurately predict students' learning experiences and outcomes. To address this limitation, we introduce the MSP-GEO corpus, a new multimodal database that contains detailed demographic and educational data, recordings of the students and their screens, and meta-data about the lecture during the learning experience. The MSP-GEO corpus was collected using a quasi-experimental pre-test/post-test design. It consists of more than 39,600 seconds (11 hours) of continuous facial footage from 76 participants watching one of three experimental videos on the topic of fossil formation, resulting in over one million facial images. The data collected includes 21 gaze synchronization points, webcam and monitor recordings, and metadata for pauses, plays, and timeline navigation. Additionally, we annotated the recordings for engagement, boredom, and confusion using human evaluators. The MSP-GEO corpus has the potential to improve the accuracy of video-based learning outcomes and experience predictions, facilitate research on the psychological processes of video-based learning, inform the design of instructional videos, and advance the development of learning analytics methods.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; • **Applied computing** → Learning management systems; Distance learning; **E-learning**.

ICMI '24, November 04–08, 2024, San José, Costa Rica

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/10.1145/3678957.3685737

KEYWORDS

engagement, boredom, confusion, learning experiences, facial analysis, gaze, affective computing, e-learning, online learning.

ACM Reference Format:
Ali N. Salman, Ning Wang, Luz Martinez-Lucas, Andrea Vidal, and Carlos Busso. 2024. MSP-GEO Corpus: A Multimodal Database for Understanding Video-Learning Experience . In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '24), November 4–8, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3678957.3685737

1 INTRODUCTION

Collecting and analyzing large-scale video-learning data of a diverse set of students and interactions is valuable for designing educational materials, strategies, and tools. Such holistic data is also critical for developing individualized *intelligent tutoring systems* (ITSs) supported by *machine learning* (ML) models to predict learning experiences and outcomes [34]. Video-based learning initiatives have resulted in important advances in different digital learning environments[1, 2, 6, 21, 40]. However, most of the resources used to understand cognitive and emotional states are not tailored toward education or a learning environment (e.g., the AffectNET corpus [28]). Moreover, the existing databases used to train models in a learning environment mainly deal with perceived emotions or attention (e.g., DaiSEE [10]). These databases are helpful in training models to predict the learners' visible reactions, such as sleepiness or tiredness, but not enough to understand or predict *deeper* learning experiences from the learners, such as cognitive and engagement changes, satisfaction, or real learning outcomes. As mentioned by some researchers [42], the combination of traditional and novel methods should consider both pedagogical and educational designs. However, currently available databases often lack such a design, so they can rarely provide full and reliable information about the learning input and the learners. Some important learner information is not well reported, including backgrounds, learning styles, experiences, and outcomes.

This study presents the new MSP-GEO dataset, which was collected to provide the ML community a resource to answer profound

learning questions and advance AI solutions to video-learning experiences. The MSP-GEO dataset contains recordings of 76 students watching educational geoscience videos. Collectively, the MSP-GEO dataset includes the learners' data, assessment data of the educational videos, learners' recordings, and annotation data for training ML models. Following a standard procedure, participants were asked to finish pre-video questions, watch an assigned geoscience educational video about how fossils form, and finally answer some post-video questionnaires. This whole video-watching process was conducted with software that we created and installed on their computers to collect facial data and matching learning logs. The facial data of the learners was collected using their webcams as they watched the videos. The corpus also has rich educational data about the learners. Traditional methods incorporated in the data collection protocol include concept inventory questions and surveys. While these traditional methods cannot provide continuous data in daily learning activities and are hard to scale [29, 32], we use these methods to collect reliable and detailed data about the background of the learners, their learning outcomes, and their overall learning experience. The interface designed for the data collection extracts useful information from the digital learning environments, providing rich multimodal data in a continuous fashion. The experiment was conducted in each participant's preferred learning environment(e.g., home, library, coffee shop, or labs) and preferred time. Therefore, the videos show video learning in the learners' natural learning environments on their personal laptops, without device requirements.

The MSP-GEO dataset considers three specially designed geoscience educational videos, where each learner was randomly assigned to one of them. The MSP-GEO database includes more than 39,600 secs (11 hours) of continuous facial footage of 76 learners watching the three educational videos (over 1 million facial images). The corpus includes annotations for the perceived level of engagement, boredom, and confusion. It also contains learners' demographic information, educational information, and learning style preferences. The pre-test and post-test contain information about video learning satisfaction, personal interest, prior knowledge, engagement, and learning outcomes.

The recordings were collected using a rigorous educational experimental design. We carefully processed and organized the information so it can be directly used to explore problems in learning analytics. We envision the MSP-GEO database as a key resource to develop ML algorithms not only to recognize the perceived learning states of the students, but also to study deeper learning questions about the relationship between students' behaviors and video learning experience outcomes. Moreover, the database can be used to improve educational video designs, understand video learning experiences, and predict video learning outcomes.

2 BACKGROUND

Researchers and institutions often collect educational datasets. However, they traditionally focus on collecting and analyzing structured data (e.g., course enrollment and test grades). The area of *educational data science* (EDS) has been actively explored in recent years, due to the development of learning analytics and educational data mining technology. The higher potential of computational models to analyze big data effectively and efficiently have added new dimensions to static and structured data, focusing on more dynamic and unstructured data (e.g., logs during interactions and multimedia recordings) [26]. Fischer et al. [8] provided a framework to classify unstructured data into three categories: microlevel data, mesolevel data, and macrolevel data. Microlevel data has a temporal component associated with fine-grained interaction data with closely tied learner interactions that can capture individual data (such as recordings and activity logs). Mesolevel data contains information on the assessment of learners and learning environments (such as discussions showing cognitive ability or processes, or social relationships in learning). Lastly, macrolevel datasets for EDS can include both static information (e.g., student demographics) and dynamic data (e.g., pre- and post-video personal interest changes). Most of the data collected by researchers contains only one or at most two types of data. For example, learning analytics datasets often contain clickstream, such as play, pause, skip, and rewind, log datasets (microdata), or a combination of logs and assessment scores (micro and macro data). Even the databases containing all three levels of data [14] rarely have fine-grained and complete data about the learners, learning input, assessment results, learning experience data, and learning environment.

Most current studies focus on the learning management system data. The MOOC learning platform has used these types of data to predict and understand course level features, such as learning engagement, and drop-out rates [30]. Focusing solely on learning analytics can be problematic, as it oversimplifies the complex nature of learning. Metrics-based systems, while useful in pattern analysis, often lack direct, detailed information about the learners' experiences and outcomes, hindering a deeper understanding of the learning process [12]. Furthermore, despite the importance of having the complete fine-grained learning experience outcome data, educational datasets rarely include all three levels of meaningful information about learners and their learning processes [31, 37].

The concept of engagement has been studied in the context of learning analytics, where several databases have been published. Table 1 provides an overview of some of these databases. The *dataset for affective states in e-environments* (DAiSEE) [11] is a dataset that captures videos of learners while they watch educational and recreational videos. During the task, the faces of the learners were recorded using a webcam. The corpus includes annotations at the video level for boredom, confusion, engagement, and frustration. Following a similar approach, Kaur et al. [15] created a database to measure the learner's intensity of engagement while watching educational videos. The learners were asked to provide feedback on the tasks. In contrast, Whitehill et al. [41] collected a database in a controlled environment where the learners played a cognitive skill game. The participants' faces and engagement levels were recorded during the game. Delgado et al. [5] introduced the *student engagement* database in a controlled environment where students' faces were recorded while they solved math problems. This database measures the level of engagement by annotating whether the learner is watching the screen or is distracted in some other way.

Due to the limitations of current learning analytics and traditional educational methods, researchers have gradually realized the importance of using microlevel, mesolevel, and macrolevel data to get insights into the students' learning experiences [30]. The MSP-GEO database has all three categories of unstructured and

Table 1: Databases to study engagement during the learning process. We mark with the symbol "-" when a specific description was not provided by the authors.

Database	Video data	Environment	Size [hrs]	Number of subjects
DAiSEE [10]	Participant face	Wild	≈ 25.18	112
Kaur [15]	Participant face	Wild	≈ 16.25	78
Student Engagement [5]	Participant face	Controlled	≈ 5.2	19
Whitehill [41]	Participant face	Controlled	-	34
proposed corpus	Participant face Task recording Eyes calibration	Wild	≈ 11	76

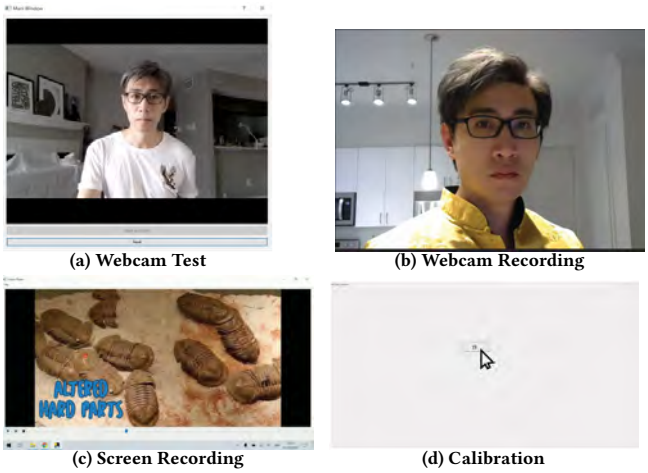
structured data directly related to the a learning topic, used as a case study.

3 THE MSP-GEO VIDEO LEARNING CORPUS

It is important to consider macrolevel, mesolevel, and microlevel data to obtain a more comprehensive understanding of the experience of video-based learning. We propose the MSP-GEO video learning corpus, which addresses this issue by providing a more complete dataset. Over a period of three years, we recruited 76 participants from The University of Texas at Dallas (UTD) campus using flyers, online advertisements, and class advertisements. Some participants completed the experiment online, while others were able to complete the experiment in-person on the UTD campus. As a case study, we consider video lectures on fossil formation. During the experiment, participants were first asked to complete a pre-survey, which provided demographic information, major, personal interests, and a test on the topic of fossil formation. Afterward, we provide a *graphical user interface* (GUI) (Sec. 3.1), which played a pre-selected video lecture. Prior to the start of the video lecture, participants are asked to adjust their webcams, and provide data to calibrate appearance-based gaze models. The gaze model is used for later attention analysis. The GUI records the face of the learners, the monitor with the educational video, and the viewing actions used by the learner (e.g., play, pause, and rewind). Once the lecture was completed, the participants were able to exit the application and complete a post-video survey that measure their satisfaction, learning outcome, and engagement. The macrolevel data includes demographic information and surveys. The microlevel data includes video recordings, activity logs, and annotations of engagement, boredom, and confusion. The mesolevel data includes pre-test and post-test questions to assess the understanding of the learners on fossil formation after watching the videos. We describe the key components of the data collection. The study is approved by UTD's institutional review board (IRB-21-59), and all participants were informed and signed consent forms. The dataset will be available for the interested party or research community if requested.

3.1 Graphical User Interface

We are interested in collecting as much information during the video-learning experience. For this purpose, we create a *graphical user interface* (GUI) with powerful capabilities. The program was built in Python. The GUI was provided to the participants. Upon initiation, the application displays the webcam feed of the participants, allowing them to adjust their seating position and environment



(a) Webcam Test (b) Webcam Recording (c) Screen Recording (d) Calibration

Figure 1: Illustrations of the videos collected in the corpus. (a) The webcam test set, where participants can adjust their webcam and the lighting condition, (b) an example of the webcam recordings during a lecture, (c) an example of the screen recording captured during a session, and (d) an illustration of a randomly placed calibration point.

lighting condition to ensure optimal visibility of their faces. Figure 1a illustrates this process. Once the participants completed this preliminary step, the application begins recording their webcam feed at a target frame rate of 30 *frames per second* (FPS) and their monitor at a target rate of 2 FPS. Figures 1b and 1c show examples of the webcam recordings and screen recordings, respectively. The frame rate of webcam recordings can vary significantly across devices. In our corpus, we observe an average of 29 FPS with only one video with a FPS below 24FPS (19FPS). Our GUI also records the time associated with each frame of the webcam and monitor, facilitating the synchronization between the recordings.

The data collection application also includes a gaze calibration feature, which serves as reference data to train appearance-based gaze approaches [18, 19]. This feature involves the random placement of 21 buttons around the screen in sequential order, asking the participants to click on each point as they appear. Figure 1d shows an example of this step. When a button is clicked, the application records the button and mouse location and the corresponding time to facilitate synchronization between screen location, and the frame displaying the eye-gaze. Additionally, once a button is clicked, it remains on the screen for additional 500 milliseconds before it disappears and the next button appears. This delay before the next point appears has been found to encourage participants to maintain their gaze directed to the target location, resulting in a more accurate synchronization between the time the button is clicked and the frame with the eye appearance captured by the webcam recording. To provide a more diverse distribution of the button locations, the application window is divided into 7 horizontal and 3 vertical regions, resulting in a total of 21 rectangles. Each button appears in each of these 21 rectangles in random order. The placement of the button within the rectangle is also random.

After the completion of the gaze calibration process, the lecture starts playing using the VLC *application programming interface*

(API) for Python. This API provides the participants the ability to control the lecture playback, including features such as play, pause, and rewind. Importantly, each time a participant engages with these playback controls, the application records the corresponding time and location within the lecture. This information is valuable in analyzing participant behavior, particularly in instances where a student revisits particular sections of the lecture. Furthermore, recording this microlevel information allows us to synchronize the lecture playback location with the webcam recordings, facilitating future analysis of the participants' engagement with the lecture material.

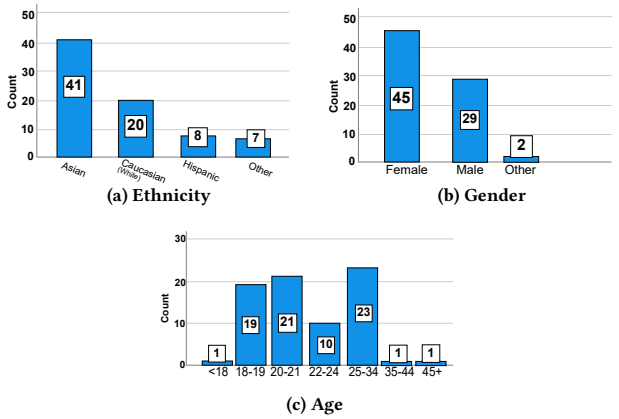
3.2 Educational Videos

We use three different educational geoscience videos, which were designed and created by UTD Geoscience Studio. The videos were created under the supervision of paleontologists and geology professors to ensure their scientific accuracy. The three videos describe how fossils form. Video #1 is 4 mins 33 secs long, video #2 is 6 mins 58 secs long, and video #3 is 6 min 27 secs long. All these videos were designed with animations, static photos, and real-world footage. Videos #2 and #3 were created based on video #1 so that they have the same video style and the same educational content on how fossils form. The only differences between these videos are the number of real examples of fossils discussed in the lecture, where videos #2 and #3 have more real examples of fossils than video #1. Video #2 gives these extra examples in a section at the beginning of the video. In contrast, video #3 separates the same examples into segments and embedded them into different sub-sections within the video.

3.3 Participants

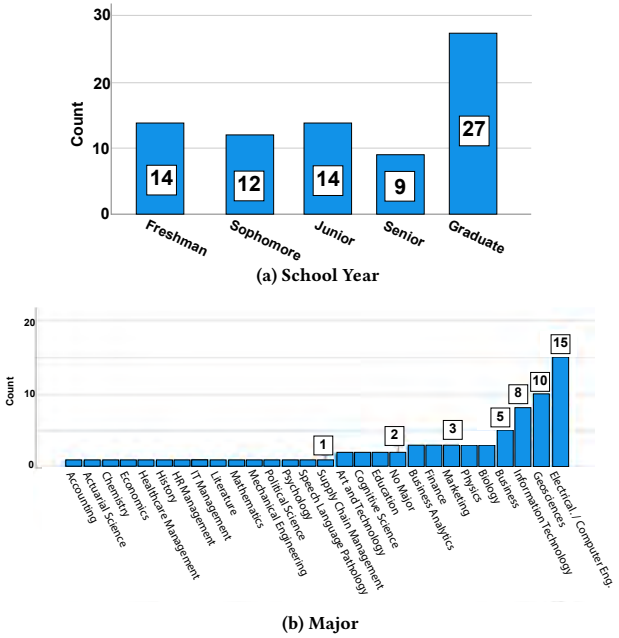
We recruited 101 participants during 3 years (2019-2022) using posters and online advertisements at the UT Dallas. The first part of the dataset was collected using computers provided by the MSP Laboratory, where everything was carefully prepared for the participants. Participants schedule a time slot in which they can use the provided hardware and software to collect the data. In the middle of our effort, the COVID-19 pandemic started, which changed our strategy. Our experiment was forced to be remote using the laptops of the learners. About 40% of the data in the MSP-GEO database were remotely collected. Since this part of the data collection was conducted in the wild, not all videos were good enough for further analysis. For example, some learners wore masks while watching the videos or used dual monitors. Our software was not designed for dual monitors, resulting in recordings of the wrong screen. In total, we encounter problems in the data of 25 subjects. Therefore, the corpus only includes high-quality data from 76 learners that can be used for ML research purposes. Out of the 76 learners, 21 participants watched video #1, 44 participants watched video #2, and 11 participants watched video #3.

We obtain consent forms for all the participants. Video research can easily generate ethical issues, which is important since many studies lack a thoughtful consent process with an appropriate data protection plan [17]. For the collection of this corpus, we created a relatively complex consent process with the help of the UTD's IRB staff (IRB-21-59). The process included in-person and remote



(a) Ethnicity (b) Gender (c) Age

Figure 2: Demographic information of the participants, including ethnicity, gender, and age.



(a) School Year (b) Major

Figure 3: Educational information of the participants, including school year, and major.

consent forms, a description of the protocol, confirmation questions to the participants to assess their understanding, and the opportunity for the participants to ask questions. We used REDCap [13] to collect the data to reduce the risk of leaking the identity of the participants. After the data collection, we encrypted and anonymized the raw data.

4 EDUCATIONAL INFORMATION OF PARTICIPANTS

One of the important strengths of the MSP-GEO Video Learning Dataset is the inclusion of educational information from the participants. We refer to this information as *Ed-Info* data, which includes

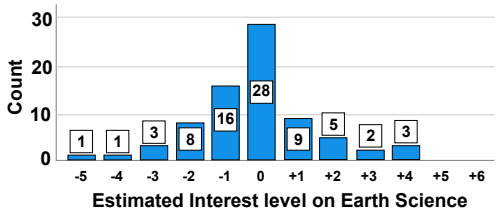


Figure 4: Distribution of the estimated interest level of learners on Earth science. The metric is obtained according to the answers of learners to reasons for taking geoscience-related classes.

the demographic and learning background of the learners. The Ed-Info data was collected via traditional educational experiments under a quasi-experiment design. The Ed-Info data has a total of 18 fields, with the first two including the participant's ID and their assigned lecture. The remaining 16 fields are educational information, which can be further categorized into six categories. The first category is *demographic information*, which includes age, ethnicity, hometown, and gender (Fig. 2). The second category is *school information*, which includes school year, major, and learning style (Fig. 3). The third category is *personal interest* and contains two fields that capture the participant's interest in the general video subject (i.e., Earth Science) and the specific video subject (i.e., fossil formation). The fourth category is *prior knowledge* and consists of two additional fields that capture prior training in the general subject and knowledge of the specific video subject. The fifth category is *learning outcome*, which is a score obtained after completing open-ended questions. The sixth category is *learning experience*, which contains the last four fields. These fields comprise self-report metrics from the participants, including satisfaction, engagement, virtual learning output, defined as the effectiveness of the provided video compared to in-person lecture, and the overall learning outcome, which is defined as the effectiveness of the provided video lecture to convey the target concept. The Ed-info data, which covers the mesolevel to macrolevel information of learners and their video learning experiences was collected from each learner.

4.1 Personal Interest

The participants were asked about their reasons for taking past Earth science courses. We provided the 12 most common general reasons that students take a course. Half of the answer reasons suggest that the student may be personally interested in learning Earth science (e.g., "content seems interesting"). The rest of the answers do not indicate a particular interest in Earth science (e.g., "It is an easy elective"). We coded the positive answers with a '+1' and negative answers with '-1.' We use the sum of the coded answers for individuals to indicate their relative personal interest level compared to other participants in the study.

Figure 4 shows the distribution of the interest level of the participants. The mean of the personal interest metric is -0.17. On average, the participants do not show particular interest in learning about Earth science. The standard deviation is 1.66, with a range of 9, a maximum of 4, and a minimum of -5. The skewness is 0.15. In the post-video survey, students also reported if they liked the subject matter of the assigned video. The results show that most students like the subject matter (70 of 76). The pretest result about personal interests in Earth sciences shows that the interest level of participants is randomly distributed. This almost normal distribution of the interest levels in the general domain knowledge demonstrates good diversity, allowing us to observe how interest levels of students change while watching the video.

4.2 Prior Knowledge

We estimated participants' prior knowledge on fossil forming and general geoscience. We used the *geoscience concept inventory* (GCI) to evaluate their knowledge on fossil formation. GCI is a multiple-choice assessment tool [20]. The GCI v.3.0 includes 73 questions covering topics related to general physical geology concepts, and fundamental concepts in physics and chemistry that are integral to understanding the conceptual Earth (e.g., gravity and radioactivity). We only selected the three questions related to fossil formation, since our study only includes videos describing this topic (questions #38, #40, and #53 in version 3 of GCI). Each question has a difficulty level on the Rasch scale. For these three questions, the scale is zero, indicating that the level of difficulty is average. These questions can identify participants with prior knowledge on fossil formation.

Table 2 shows the result of the participants for the three CGI questions. Few participants obtained either zero correct answers (7.9%) or all correct answers (15.8%). 27 learners obtained one correct question (35.5%), and 31 learners obtained two correct questions. These results show that around 43.4% of the learners had limited prior knowledge about fossil formation.

The prior knowledge on the learning domain, in our case, Earth science, is estimated by collecting information about the number of Earth science classes taken in the past. Figure 5 shows the statistics of how many earth science courses participants have previously taken before participating in this data collection. The majority of the participants (61 of 76, 80%) had taken no more than two geoscience classes in the past. Taking one or two Earth science classes gives a basic intro-level knowledge about this subject. However, it is expected that they will know more than students without any Earth science class. The participants with more than two classes may have learned more advanced concepts about geology and fossils, especially the people who have taken more than five courses.

Table 2: Summary of the pre-test evaluation using the three GCI questions on fossil formation. This evaluation assesses the prior knowledge of the learners.

	Frequency	Percent [%]	Cumulative Percent [%]
0 Correct	6	7.9	7.9
1 of 3 Correct	27	35.5	43.4
2 of 3 Correct	31	40.8	84.2
3 of 3 Correct	12	15.8	100.0
Total	76	100.0	

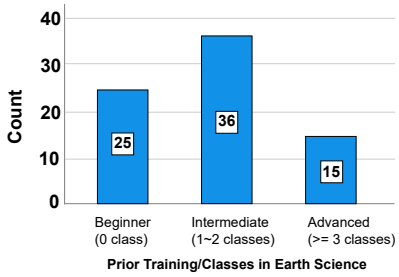


Figure 5: Prior knowledge in Earth science estimated by the number of geoscience classes taken by students before participating in the data collection.

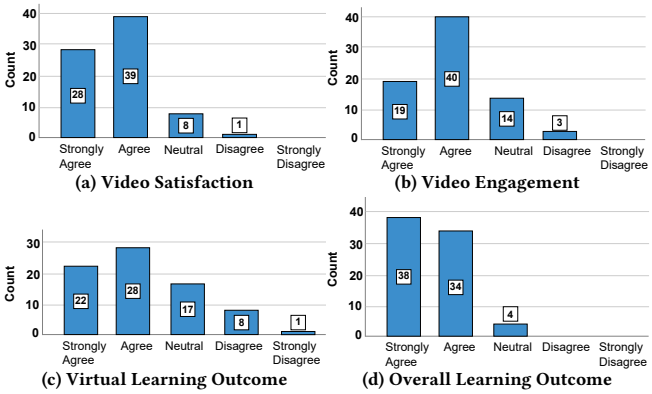


Figure 6: Post-survey for satisfaction, engagement, learning outcome, and effectiveness of video lecture. (a) Overall, I was satisfied with this instructional video, (b) I think the video is engaging, (c), overall, I feel I was able to learn the information from this instructional video as I would have had in an in-person face-to-face class presentation, and (d) this video helped me to understand the topic discussed in the video.

Therefore, we classify the 76 participants into three groups: beginner (taken 0 courses, 25 people), intermediate (taken one or two courses, 36 people), and advanced (taken more than three courses, 15 people).

4.3 Learning Experience

We assess the learning experience of the participants with self-report surveys at the end of the evaluation. We consider video satisfaction (Fig. 6a), video engagement (Fig. 6b), learning outcome compared to in-person lectures (Fig. 6c), and educational effectiveness of the video (Fig. 6d). The learners answered questions in the post-video survey using a 5-item Likert scale where the extreme values were 1 *strongly agree* and 5 *strongly disagree*. These questions are extremely useful as they can be used as target labels to train ML algorithms to improve learning analytics.

The statement in the survey for satisfaction is *Overall, I was satisfied with this instructional video*. Figure 6a shows that 67 out of 76 students were satisfied with the video, with 28 learners answering *strongly agree* and 39 learners answering *agree*. The was only one learner that answered this question with a *disagree*. The results show that the learning experience of the students was mostly positive, in spite of the diversity in their demographic and educational backgrounds.

The statement in the survey for engagement is *I think the video is engaging*. The results for the self-report level of engagement, Figure 6b, show that 59 out of 76 students found the video learning experience engaging, with 19 learners answering *strongly agree* and 40 learners answering *agree*. Three learners answered this engagement question with *disagree*. Most of the students felt engaged during the video learning experience.

One approach that we use to determine the learning outcome is to ask the participants to self-report their video learning experience compared to the in-person learning experience. The statement for the survey for the virtual learning outcome is *Overall, I feel I was able to learn the information from this instructional video as I would have had in an in-person face-to-face class presentation*. Figure 6c shows the results, which show that 50 out of 76 students think the video learning experience can equal an in-person learning session on the same topic. 22 learners answered *strongly agree* and 28 learners answered *agree*. Nine learners answered either *disagree* or *strongly disagree*. These results confirm the potential for online video learning.

The statement for the survey for the educational effectiveness of the video is *this video helped me to understand the topic discussed in the video*. Figure 6 shows the results. We found that the videos were perceived as very effective with 72 out of 76 learners answering this question as either *strongly agree* or *agree*.

4.4 Learning Outcome

The self-report questions are useful to understand the learning outcome. However, they are subjective. After watching the video, we conduct a post-video test to objectively assess their understanding of the fossil-forming process. We use the open question *please explain the fossils formation process*. We created a rubric with three criteria: completeness of the fossil-forming process mentioned in the video, correctness of the statements based on the video content, and depth of the answer (the number of video details mentioned on each topic). For each criterion, the evaluation score range of a participant's learning outcome is from level 0 to 3. A score '0' or '1' indicates a relatively poor performance for that criterion. A score '2' indicates that the learner's performance is acceptable. A score '3' shows a very good (almost perfect) answer for the corresponding criterion. The completeness of the video topics is graded based on the completeness of the potential fossil-forming method considering any missing important topic. The three score levels are complete (level 3), mostly complete (level 2), and missed most of the topics (level 1). The second criterion, 'Accuracy', indicates the correctness and relevance of statements about fossil forming, which is evaluated based on the number of wrong statements: level 3 indicates there are no wrong statements; level 2 indicates a few wrong statements; and level 1 indicates that there are many wrong statements in the answer. Lastly, the third criterion, 'depth', indicates that the participants give enough details or clear explanations about the video topics. Level 3 of this criterion means the answer is very detailed and gives enough explanations; level 2 means that

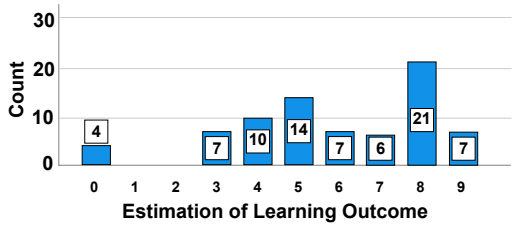


Figure 7: Post-test evaluation of learning outcomes. The objective scores were obtained by rating the open question *please explain the fossils formation process.* The range of the final score is from '0' to '9'.

participants give a general framework with few details, and level 1 indicates a very blurred and general answer. The final score is the addition of the scores assigned to the three criteria (e.g., a number ranging from '0' to '9'). With this number, we estimate and quantify the learning outcome of each participant. The detailed method is in the footnote of the supplementary educational factors table.

Figure 7 shows the distribution of the learning outcomes. The learners obtained different results, where 41 out of 76 achieved a score above 5. This distribution with good and bad scores is ideal to study the relationship between learning states, such as engagement, boredom, and confusion, and learning outcomes.

5 LEARNING STATE

5.1 Annotations of Learners' State

After the webcam and screen recordings are collected, they are synchronized and presented to a team of annotators. We are interested in cognitive and emotional states related to learning. We select engagement, boredom, and confusion. Seven annotators participated in the evaluation. An important difference between the annotations of this corpus with other related corpora is the use of time-continuous annotations. Each annotator used a joystick to annotate each of the target learning states using the CARMA software [9]. The participants watch a synchronized video of the webcam showing the facial expression of the learners, and a video of the monitor recordings to better annotate the data.

The graphical interface has a bar where the extremes represent the extremes of the target learning state (e.g., low engagement versus high engagement). As the evaluator watches the video, she/he moves the joystick to represent the level of the attribute perceived for that particular segment in the video. The interface records traces of the learning state, where the values range from -100 to 100, with -100 indicating a lack of engagement, boredom, or confusion, and 100 indicating a high degree of engagement, boredom, or confusion.

While this annotation method has been used for emotional databases [24, 27, 33], this is the first time this approach has been used to assess learning states to the best of our knowledge. Previous studies have pre-segmented the recordings, providing a global score for each video [10, 15, 41], or provided annotations for some frames [5, 41]. With time-continuous traces, the researcher can determine the temporal unit in the analysis. This approach facilitates the analysis and recognition of learning states at different resolutions. The traces can also provide information to identify the learning states associated with specific segments in the educational video. Figure 8 shows an example of the traces for one recording.

Table 3: Cronbach's Alpha to assess inter-evaluator agreement.

	Engagement		Boredom		Confusion	
All	0.44		0.46		0.35	
Ann	**Incl**	**Excl**	**Incl**	**Excl**	**Incl**	**Excl**
1	0.44	0.34	0.46	0.41	0.35	0.42
2	0.46	0.47	0.59	0.51	0.39	0.29
3	0.44	0.56	0.46	0.45	0.35	0.29
4	0.44	0.34	0.46	0.38	0.35	0.30
6	0.44	0.33	0.45	0.30	0.34	0.21
7	0.44	0.30	0.46	0.36	0.35	0.19

We evaluate the inter-evaluator agreement of the annotations in the corpus by taking the Cronbach's alpha coefficient [3]. We estimate the agreement over the annotations of each video, reporting the average results over the whole corpus. The mean Cronbach's alpha coefficients for the annotations are 0.435 for engagement, 0.457 for boredom, and 0.348 for confusion. We evaluate our current annotators using the inter-evaluator agreement. We take the average Cronbach's alpha coefficient for all the videos a particular annotator has rated, and then we took the same calculations but excluded the ratings of that particular annotator. We expect that for a good annotator, the agreement scores will decrease when their ratings are excluded. Table 3 shows the agreement results for each annotator. In the analysis, we exclude Annotator 5 for a low number of annotations. In most cases, we observe that the agreement drops when an evaluator is removed.

5.2 Learners' State Baselines

We evaluate the corpus as a dataset for dynamic emotion recognition. We build a model to predict the traces, assigning a value for each "frame." We build models for engagement, boredom, and confusion. To facilitate the reproducibility of these results, we rely on a simple model consisting of two *bidirectional long short-term memory* (BiLSTM) layers with a dropout layer between them and an output linear layer. As input to our model, we extract the EMOCA [4] facial features for each frame in the videos. Then, we average the EMOCA features for the 30 video frames centered at each annotation value. The BiLSTM layers have a hidden size of 32, we use a drop rate of 0.5, and the linear layer has a node size of 1. We predict one attribute at a time. Additionally, we use subject independent splits for train, development, and test sets.

The raw annotations we collected using CARMA have inconsistent sample rates, so we first smooth all the annotations using a moving mean filter with a window size of 250ms and a stride of 200ms. Then, we average the annotations for each video. To account for any annotator reaction lags, we shift the mean annotations by 3 seconds into the past, as proposed in Mariooryad and Busso [22, 23]. For training, we use the stochastic gradient descent optimizer with a learning rate of 0.00001 and a momentum of 0.9. Our loss function is the *concordance correlation coefficient* (CCC) loss, i.e. $1-CCC$, and our testing metric is the CCC. We train our engagement model for 100 epochs, and our boredom and confusion models for 150 epochs. We test the model at the last epoch. We obtain a test CCC of 0.552 for engagement, 0.181 for boredom, and 0.119 for confusion.

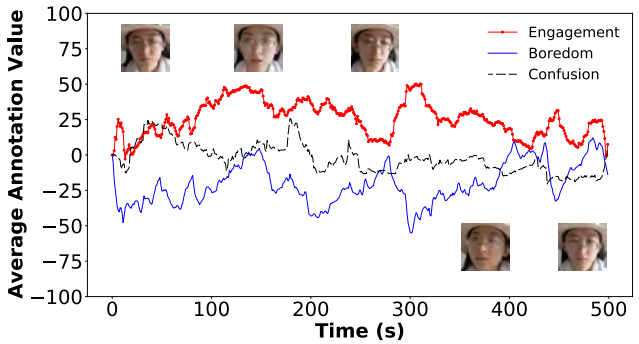
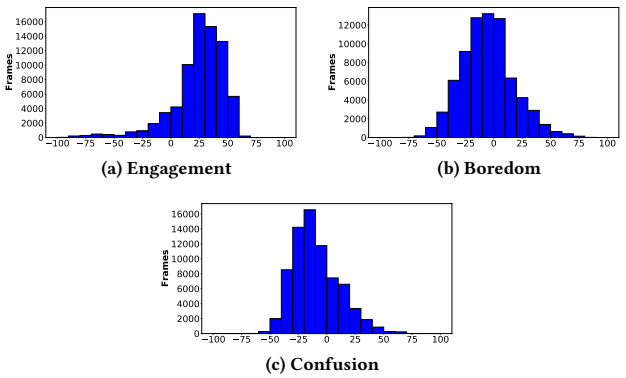


Figure 8: Example of time-continuous annotations for engagement, confusion, and boredom (subject 17). The figure includes frames with facial expressions extracted using the Dlib library [16]. The x-axis represents the time of the recording in seconds, and the y-axis shows the average annotation for each frame.



(a) Engagement (b) Boredom

(c) Confusion

Figure 9: Distribution of learning states in the corpus (engagement, boredom, and confusion). The y-axis is the number of frames, and the x-axis is the value of the attribute.

5.3 Distribution of Learners' States

This section analyzes the distribution of the learning states annotated by the evaluators. We first averaged the time-continuous traces provided by different evaluators to a recording. Then, we plot the distribution of the frame-based values for each learning state. Figure 9 depicts the distribution for each learning state. For engagement, Figure 9a shows that the distribution is predominantly positive, indicating that most subjects were engaged with the lectures. Figures 9b and 9c show the distributions for boredom and confusion. They have higher variance, with many frames above and below 0. These results indicate that while the majority of subjects were not bored or confused, a significant portion of the data contains instances of perceived boredom and confusion. These results match the overall distribution of the survey results of video satisfaction and engagement (Figs. 6a and 6b). However, as reported in previous studies, self-reported results commonly mismatch with test results or other objective data-based analysis metrics [7, 35]. Therefore, we expect that there could be a mismatch in the engagement results between self-reported data, test data, and physiological data. We will explore this direction in our future research directions.

6 CONCLUSIONS

The MSP-GEO corpus is a unique and comprehensive dataset that provides detailed dimensions and rich learning experience data. With more than 39,600 seconds (11 hours) of continuous facial footage from 76 participants, the dataset contains over 1 million facial images, providing a significant amount of data for evaluating how users interact with instructional videos. Additionally, the dataset includes annotated facial videos for engagement, boredom, and confusion, as well as robust learning evaluation data. Moreover, the MSP-GEO corpus has detailed learner information and learning experience data, including demographic information and a summary of self-reported engagement and learning experience results. With the microlevel, mesolevel, and macrolevel data provided, this corpus is ideal for exploring learning analytics solutions using ML.

In the study of cognitive science and multimedia learning, the multi-level features provided in the MSP-GEO corpus are well-suited for understanding the ontological elements and structures within complex knowledge models. These models can be analyzed through ontology design patterns [36] to represent explanations in specific domains, such as psychology and learning sciences. Consequently, the MSP-GEO corpus holds the potential for developing a quantitative understanding of the multimedia learning process. This capability could significantly contribute to advancing theoretical frameworks, such as the Cognitive Theory of Multimedia Learning [25], and support the analysis of video design frameworks, such as the geoscience educational video model [38] and complex system interactive designs [39].

This dataset can be used to develop more accurate video-based learning outcomes or learning experience predictions. The corpus can be valuable for discovering the psychological processes of video learning, optimizing instructional video design, and developing video-based intelligent tutoring systems. Furthermore, domain-based educational researchers could test how place-based examples affect learners' experience and performance. Using this data, learning analytics researchers can explore the correlation between facial analysis and self-report data (using test data as a baseline), evaluate users' preferences about different design elements (e.g., real footage and animation), and test the multimedia design cognitive load-related principles. The MSP-GEO corpus is a valuable resource for researchers in video-based learning.

Data collection in uncontrolled environments has certain limitations. The experiment design has a weak control on the data quality, which results in discarding 1/4 of the subjects' data (25 of 101). The estimation of gaze and video frame is not as accurate as the one obtained in controlled experiments with a given computer and with subjects in the same position and pose. However, the flexibility offered by the corpus due to its rich naturalistic data can open opportunities to advance the area of video-based learning.

In this corpus, we decided to focus on a single educational topic (fossil formation), collecting data from several learners with diverse backgrounds, interests, and expertise. The protocol and GUI used to collect this corpus are flexible. They can be used to collect similar recordings in other domains, broadening the scope of our effort. Likewise, a future research direction is explore ways to use this infrastructure to collect learning data during longitudinal video learning experiences.

REFERENCES

[1] Eric Araka, Elizaphan Maina, Rhoda Gitonga, and Robert Oboko. 2020. Research trends in measurement and intervention tools for self-regulated learning for e-learning environments—systematic review (2008–2018). *Research and Practice in Technology Enhanced Learning* 15 (12 2020). https://doi.org/10.1186/s41039-020-00129-5

[2] Rachel Baker, Di Xu, Jihyun Park, Renzhe Yu, Qiujie Li, Bianca Cung, Christian Fischer, Fernando Rodriguez, Mark Warschauer, and Padhraic Smyth. 2020. The benefits and caveats of using clickstream data to understand student self-regulatory behaviors: opening the black box of learning processes. *International Journal of Educational Technology in Higher Education volume* 17 (04 2020). https://doi.org/10.1186/s41239-020-00187-1

[3] L.J. Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 3 (September 1951), 297–334.

[4] Radek Danecek, Michael J. Black, and Timo Bolkart. 2022. EMOCA: Emotion Driven Monocular Face Capture and Animation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[5] Kevin Delgado, Juan Manuel Origgi, Tania Hasanpoor, Hao Yu, Danielle Allessio, Ivon Arroyo, William Lee, Margrit Betke, Beverly Woolf, and Sarah Adel Bargal. 2021. Student Engagement Dataset. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. 3621–3629. https://doi.org/10.1109/ICCVW54120.2021.00405

[6] Blaženka Divjak, Bart Rienties, Francisco Iniesto, Petra Vondra, and Mirza Žižak. 2022;19. Flipped classrooms in higher education during the COVID-19 pandemic: findings and future research recommendations. *Research and Practice in Technology Enhanced Learning* (2022;19). https://doi.org/10.1186/s41239-021-00316-4

[7] Nancy Falchikov and David Boud. 1989. Student self-assessment in higher education: A meta-analysis. *Review of educational research* 59, 4 (1989), 395–430.

[8] Christian Fischer, Zachary A. Pardos, Ryan Shaun Baker, Joseph Jay Williams, Padhraic Smyth, Renzhe Yu, Stefan Slater, Rachel Baker, and Mark Warschauer. 2020. Mining Big Data in Education: Affordances and Challenges. *Review of Research in Education* 44, 1 (2020), 130–160. https://doi.org/10.3102/0091732X20903304 arXiv:https://doi.org/10.3102/0091732X20903304

[9] Jeffrey M Girard. 2014. CARMA: Software for continuous affect rating and media annotation. *Journal of Open Research Software* 2, 1 (2014), e5. https://doi.org/10.5334/jors.ar

[10] Abhay Gupta, Richik Jaiswal, Sagar Adhikari, and Vineeth Balasubramanian. 2016. DAISEE: Dataset for Affective States in E-Learning Environments. *CoRR* abs/1609.01885 (2016). arXiv:1609.01885 http://arxiv.org/abs/1609.01885

[11] R. Gupta, K. Audhkhasi, Z. Jacokes, A. Rozga, and S.S. Narayanan. 2018. Modeling multiple time series annotations based on ground truth inference and distortion. *IEEE Transactions on Affective Computing* 9, 1 (January-March 2018), 76–89. https://doi.org/10.1109/TAFFC.2016.2592918

[12] Carolina Guzmán-Valenzuela, Carolina Gómez-González, Andrés Rojas-Murphy Tagle, and Alejandro Lorca-Vyhmeister. 2021. Learning analytics in higher education: a preponderance of analytics but very little learning? *International Journal of Educational Technology in Higher Education* 18 (2021).

[13] Paul A. Harris, Robert Taylor, Brenda L. Minor, Veida Elliott, Michelle Fernandez, Lindsay O'Neal, Laura McLeod, Giovanni Delacqua, Francesco Delacqua, Jacqueline Kirby, and Stephany N. Duda. 2019. The REDCap consortium: Building an international community of software platform partners. *Journal of Biomedical Informatics* 95 (2019), 103208. https://doi.org/10.1016/j.jbi.2019.103208

[14] Kintu Justice and Chang Zhu. 2017. Blended learning effectiveness: the relationship between student characteristics, design features and outcomes. *International Journal of Educational Technology in Higher Education* 14 (02 2017). https://doi.org/10.1186/s41239-017-0043-4

[15] Amanjot Kaur, Aamir Mustafa, Love Mehta, and Abhinav Dhall. 2018. Prediction and Localization of Student Engagement in the Wild. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*. 1–8. https://doi.org/10.1109/DICTA.2018.8615851

[16] D.E. King. 2009. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research* 10 (July 2009), 1755–1758.

[17] Nicolas Legewie and Anne Nassauer. 2018. YouTube, Google, Facebook: 21st Century Online Video Research and Research Ethics. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* Vol 19 (2018), No 3 (2018): Research Ethics in Qualitative Research. https://doi.org/10.17169/FQS-19.3.3130

[18] N. Li and C. Busso. 2014. User-Independent Gaze Estimation by Exploiting Similarity Measures in the Eye Pair Appearance Eigenspace. In *International conference on multimodal interaction (ICMI 2014)*. Istanbul, Turkey, 335–338. https://doi.org/10.1145/2663204.2663250

[19] N. Li and C. Busso. 2018. Calibration Free, User Independent Gaze Estimation with Tensor Analysis. *Image and Vision Computing* 74 (June 2018), 10–20. https://doi.org/10.1016/j.imavis.2018.04.001

[20] Julie C Libarkin and Steven W Anderson. 2005. Assessment of learning in entry-level geoscience courses: Results from the Geoscience Concept Inventory. *Journal of Geoscience Education* 53, 4 (2005), 394–401.

[21] Tiecheng Liu and John R Kender. 2004. Lecture videos for e-learning: Current research and challenges. In *IEEE Sixth International Symposium on Multimedia Software Engineering*. IEEE, 574–578.

[22] S. Mariooryad and C. Busso. 2013. Analysis and Compensation of the Reaction Lag of Evaluators in Continuous Emotional Annotations. In *Affective Computing and Intelligent Interaction (ACII 2013)*. Geneva, Switzerland, 85–90. https://doi.org/10.1109/ACII.2013.21

[23] S. Mariooryad and C. Busso. 2015. Correcting Time-Continuous Emotional Labels by Modeling the Reaction Lag of Evaluators. *IEEE Transactions on Affective Computing* 6, 2 (April-June 2015), 97–108. https://doi.org/10.1109/TAFFC.2014.2334294 Special Issue Best of ACII.

[24] L. Martinez-Lucas, Mohammed Abdelwahab, and Carlos Busso. 2020. The MSP-Conversation Corpus. In *Interspeech 2020*. Shanghai, China, 1823–1827. https://doi.org/10.21437/Interspeech.2020-2444

[25] Richard E Mayer. 2005. Cognitive theory of multimedia learning. *The Cambridge handbook of multimedia learning* 41, 1 (2005), 31–48.

[26] Daniel A. McFarland, Saurabh Khanna, Benjamin W. Domingue, and Zachary A. Pardos. 2021. Education Data Science: Past, Present, Future. *AERA Open* 7 (2021), 23328584211052055. https://doi.org/10.1177/23328584211052055 arXiv:https://doi.org/10.1177/23328584211052055

[27] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder. 2012. The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent. *IEEE Transactions on Affective Computing* 3, 1 (January-March 2012), 5–17. https://doi.org/10.1109/T-AFFC.2011.20

[28] A. Mollahosseini, B. Hasani, and M. H. Mahoor. 2019. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing* 10, 1 (January-March 2019), 18–31. https://doi.org/10.1109/TAFFC.2017.2740923

[29] Christian Mühl, Brendan Allison, Anton Nijholt, and Guillaume Chanel. 2014. A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges. *Brain-Computer Interfaces* 1, 2 (2014), 66–84. https://doi.org/10.1080/2326263X.2014.912881 arXiv:https://doi.org/10.1080/2326263X.2014.912881

[30] Larian M. Nkomo, Ben Kei Daniel, and Russell Butson. 2021. Synthesis of student engagement with digital technologies: a systematic review of the literature. *International Journal of Educational Technology in Higher Education* 18 (2021).

[31] Ekene Francis Okagbue, Ujunwa Perpetua Ezeachikulo, Esther Onyinye Nwigwe, and Amina Abedi Juma. 2022. Machine Learning and Artificial Intelligence in Education Research: A Comprehensive Overview of 22 Years of Research indexed in the Scopus Database. *Social Sciences & Humanities Open* (07 2022).

[32] Md. Mustafizur Rahman, Ajay Krishno Sarkar, Md. Amzad Hossain, Md. Selim Hossain, Md. Rabiul Islam, Md. Biplob Hossain, Julian M.W. Quinn, and Mohammad Ali Moni. 2021. Recognition of human emotions using EEG signals: A review. *Computers in Biology and Medicine* 136 (2021), 104696. https://doi.org/10.1016/j.compbiomed.2021.104696

[33] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. 2013. Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions. In *2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE 2013)*. Shanghai, China, 1–8. https://doi.org/10.1109/FG.2013.6553805

[34] David Sharek and Eric Wiebe. 2014. Measuring Video Game Engagement Through the Cognitive and Affective Dimensions. *Simulation & Gaming* 45, 4-5 (2014), 569–592. https://doi.org/10.1177/1046878114554176 arXiv:https://doi.org/10.1177/1046878114554176

[35] Jieun Shin. 2020. How do partisans consume news on social media? A comparison of self-reports with digital trace measures among Twitter users. *Social Media+ Society* 6, 4 (2020), 2056305120981039.

[36] Ilaria Tiddi, Mathieu d'Aquin, and Enrico Motta. 2015. An ontology design pattern to define explanations. In *Proceedings of the 8th International Conference on Knowledge Capture*. 1–8.

[37] Mark Urban-Lurain, Diane Ebert-May, Jennifer Momsen, Ryan McFall, Matthew B Jones, Ben Leinfelder, and Jon Sticklen. 2009. An assessment database for supporting educational research. In *2009 39th IEEE Frontiers in Education Conference*. IEEE, 1–6.

[38] Ning Wang, Zachary Clowdus, Alessandra Sealander, and Robert Stern. 2022. Geonews: Timely geoscience educational YouTube videos about recent geologic events. *Geoscience Communication* 5, 2 (2022), 125–142.

[39] Ning Wang, Robert J Stern, Mary L Urquhart, and Katherine M Seals. 2022. Google earth geoscience video library (GEGVL): Organizing geoscience videos in a google earth environment to support fieldwork teaching methodology in earth science. *Geosciences* 12, 6 (2022), 250.

[40] Ning Wang, Robert J. Stern, and Lowell Waite. 2023. Workflow for designing instructional videos to support place-based geoscience education for geoscience majors. *Journal of Geoscience Education* 71, 1 (2023), 107–125. https://doi.org/10.1080/10899995.2022.2093543 arXiv:https://doi.org/10.1080/10899995.2022.2093543

[41] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R. Movellan. 2014. The Faces of Engagement: Automatic Recognition of Student Engagement from Facial Expressions. *IEEE Transactions on Affective Computing*

5, 1 (2014), 86–98. https://doi.org/10.1109/TAFFC.2014.2316163
[42] Olaf Zawacki-Richter, Victoria Marín, Melissa Bond, and Franziska Gouverneur. 2019. Systematic review of research on artificial intelligence applications in higher education -where are the educators? *International Journal of Educational Technology in Higher Education* 16 (10 2019), 1–27. https://doi.org/10.1186/s41239-019-0171-0