Inconsistencies in the Definition and Annotation of Student Engagement in Virtual Learning Datasets: A Critical Review

Shehroz S. Khan[1], Ali Abedi[1*] and Tracey J.F. Colella[1]

[1]KITE Research Institute, University Health Network, Canada.

*Corresponding author(s). E-mail(s): ali.abedi@uhn.ca;
Contributing authors: shehroz.khan@uhn.ca;
tracey.colella@uhn.ca;

Abstract

Background: Student engagement (SE) in virtual learning can have a major impact on meeting learning objectives and program dropout risks. Developing Artificial Intelligence (AI) models for automatic SE measurement requires annotated datasets. However, existing SE datasets suffer from inconsistent definitions and annotation protocols mostly unaligned with the definition of SE in educational psychology. This issue could be misleading in developing generalizable AI models and make it hard to compare the performance of these models developed on different datasets. The objective of this critical review was to explore the existing SE datasets and highlight inconsistencies in terms of differing engagement definitions and annotation protocols. **Methods:** Several academic databases were searched for publications introducing new SE datasets. The datasets containing students' single- or multi-modal data in online or offline computer-based virtual learning sessions were included. The definition and annotation of SE in the existing datasets were analyzed based on our defined seven dimensions of engagement annotation: sources, data modalities, timing, temporal resolution, level of abstraction, combination, and quantification. **Results:** Thirty SE measurement datasets met the inclusion criteria. The reviewed SE datasets used very diverse and inconsistent definitions and annotation protocols. Unexpectedly, very few of the reviewed datasets used existing psychometrically validated scales in their definition of SE. **Discussion:** The inconsistent definition and annotation of SE are problematic for research on developing

comparable AI models for automatic SE measurement. Some of the existing SE definitions and protocols in settings other than virtual learning that have the potential to be used in virtual learning are introduced.

Keywords: Virtual Learning, Student Engagement, Engagement Measurement, Engagement Definition, Artificial Intelligence, Machine Learning

1 Introduction

As the use of internet services becomes more widespread, virtual learning programs are becoming increasingly common and accepted as a mainstream form of education [1]. In contrast to traditional in-person learning, virtual learning offers several benefits, including increased accessibility, lower costs, and the ability to provide personalized instruction [2]. However, virtual learning also presents its own set of challenges, particularly when it comes to assessing Student Engagement (SE). In a virtual learning environment, it can be difficult for instructors to measure the level of engagement of their students, especially when working with large groups in online virtual learning settings [3]. This is a significant issue because SE has been shown to have a direct impact on the achievement of learning objectives [4]. Therefore, it is important for instructors to be able to assess SE in order to provide real-time feedback and take necessary actions to maximize engagement.

In recent years, advances in Artificial Intelligence (AI) have led to the successful development of algorithms to objectively and automatically measure SE in virtual learning environments, especially in academia and online classrooms [5, 6]. Most of the published results in this area rely on supervised machine-learning approaches [5, 6], requiring annotated ground-truth data to develop models and to provide SE-related outcomes (e.g., Engaged versus Not-engaged or different levels of engagement). A major concern is that most of the datasets used non-standard definitions of engagement; thus, the data samples in many of these datasets are annotated very differently across multiple datasets. Unless a standardized SE definition and measurement scale are in place, annotating data to develop AI algorithms is very challenging. This further constrains the development of AI algorithms to objectively quantify SE and to compare the SE measurement algorithms fairly.

The objective of this critical review was to identify inconsistencies in definitions and annotation protocols used in the existing SE datasets. The research question of this study was as follows: How inconsistent was the definition and annotation of SE in the existing datasets based on seven dimensions of engagement annotation (described in Section 2.1), including

1. Sources: the observers performing the annotation,
2. Data modality: the information that is observed by the observers for annotation,
3. Timing: the time when the annotation takes place,

4. Temporal resolution: the timesteps in which the annotation takes place,
5. Level of abstraction: whether engagement is defined and annotated as a single- or multi-component variable,
6. Combination: the way the components of engagement are combined to create one value for engagement, and
7. Quantification: the way the engagement is represented numerically.

2 Student Engagement Annotation

Researchers have identified SE to encompass three primary components [7, 8]. These components include behavioral engagement, which refers to behaviors such as attendance, involvement, and being On-Task; affective engagement, which refers to emotional reactions such as excitement and desirability; and cognitive engagement, which refers to a student's investment in learning and willingness to embrace challenges. An additional dimension of SE, referred to as "agentic engagement," has also been proposed, which involves a student constructively contributing to the flow of instructions [9].

The concept of engagement may differ based on the perspective from which it is being analyzed and the level of detail at which it is being studied; a concept referred to as "grain size" [10]. When the grain size is considered at a micro level, engagement may refer to an individual's involvement in a specific task or learning activity. At a macro level, on the other hand, engagement may pertain to a group of students within a class or community. The National Survey of Student Engagement [11] is an example of a measure that is suited for evaluating engagement at the institutional level but may not be as effective in identifying correlations between an individual student's engagement and their learning experience.

There are various methods that have been employed to assess student engagement in traditional in-person learning settings, including self-reporting, observational scales, experience sampling, teacher rating, and interviews [12]. Henrie et al. [13] conducted a review of the various self-reporting and observational (both qualitative and quantitative) scales that have been used to measure student engagement in technology-mediated learning environments and identified their strengths and limitations.

Virtual learning platforms often utilize video and audio mediums for both content delivery and communication between instructors and students. A variety of features, such as body pose, valence, arousal, and audio pitch, can be extracted from video and audio data [5, 6, 14]. The extracted features can then be used to build AI models for engagement measurement. These features can also be learned through deep learning approaches [5, 6, 15]. However, the extent of information captured is restricted due to limited modalities. Audio data, for example, cannot be used to learn or extract facial features. In theory, other types of sensors can also be employed, for example, electrocardiogram, electroencephalogram, and wearable devices to collect other physiological information (e.g., electrodermal activity, skin temperature, heart rate) [16].

However, in a real-world scenario, the use of many sensors in the educational environment and on the body of a student is impractical. Therefore, the key question to consider is whether these (extracted or learned) features from a sensing modality correspond or correlate to a measurement scale. Recent advancements in machine learning, especially deep learning, have allowed for the extraction of temporal affective and behavioral information from video and audio datasets for various tasks in the field of affective computing [17, 18]. However, in the context of SE, the existing virtual learning datasets used diverse observational scales for collecting ground-truth annotations (as discussed in Section 5). Sometimes these scales are arbitrarily contrived, invented, or based on general knowledge rather than complete psychometric analysis. Therefore, there appears to be no direct concordance between the information extracted from the video, audio, or other sensing modalities and the measurement scales. In such cases, it is very hard to establish a clear interpretation between What we wanted to train an AI algorithm on and what actually the AI algorithm is trained on. In most of the existing AI-driven engagement measurement approaches [5, 6], the focus was on building sophisticated AI models without as much emphasis on the correctness of annotations upon which they are trained. The outcomes of a successful AI model are as good as the quality of ground-truth annotations assigned to it [15, 19, 20]. This further leads to questions about the validity of performance values reported by the existing AI-driven methods for SE measurement.

2.1 Dimensions of Engagement Annotation

D'Mello [21] identified five dimensions of affect annotation for developing affect detection systems: *sources*, *data modality*, *timing*, *temporal resolution (timescale)*, and *level of abstraction*. There are two key differences between affect detection and engagement measurement. Firstly, affect is only one component of engagement, which also includes behavioral and cognitive components, [7]. Secondly, contrary to the affect "detection", an annotation for engagement "measurement" must not only identify engagement versus disengagement but also determine the "level" of engagement. Considering the differences between affect and engagement, we modified the five dimensions above and added two additional dimensions, *combination* and *quantification (scale)*, as described below. These seven dimensions of engagement annotation were used to analyze the SE definition and annotation used in the existing datasets.

2.1.1 Sources

The first dimension of engagement annotation, *sources*, refers to the types and number of individuals performing the annotation [21]. Observer-based annotation is the most common approach in the reviewed SE datasets (see Section 5), which is categorized into expert (trained) observers and non-expert (or

untrained) observers. An example of the expert observers, which is considerably high cost per observer, would be a group of educational psychology experts who are asked to annotate students' engagement in a dataset. Conversely, non-expert observers would be students without any prior training in psychology who are asked to annotate engagement in a dataset according to their perception of engagement. The annotation by non-expert observers is usually performed in crowdsourcing settings [22]. In crowdsourcing, a task is usually given over the internet to a less-specific and more-public-oriented group of observers. This approach can yield a large number of annotations in a short span of time. However, due to the lack of expertise in observers, it could lead to noisy annotations [23].

Observer-based annotations do not interrupt the learning process of a student. However, they are time, and labor-consuming [24], and can suffer from observation bias (such as *seeing what one is looking for and missing what one is not* [25]). These measures are hard to scale but can "*measure SE as it occurs*" [13], which can be used to annotate various segments or transience of a student's engagement in a learning session. The observer-based measures are often used in conjunction with other measures for additional evidence rather than a stand-alone source of information [26].

Annotations can also be performed by the student themselves, which is called self-report. Collecting concurrent self-report data at regular intervals can be disruptive and disengage individuals during their tasks [26]. Retrospective self-report also requires a student to reconstruct past states of engagement on a post-hoc basis, which may be biased. Different students may also differ in their own sense of what it means to be engaged. On the other hand, since SE also encompasses cognitive and emotional components, it is argued that self-report is the most valid measure to capture aspects of engagement that focus heavily on students' perception of their experience [13].

2.1.2 Data Modality

The *data modality* dimension [21] pertains to the information that is observed by observers for annotation, such as video, audio, computer screen recording, mouse cursor tracking data, or their combination.

2.1.3 Timing

Annotation *timing* [21] refers to the point at which the annotation takes place. This can take place in real-time, as when students are prompted to self-report their engagement through the use of pop-up windows during virtual learning sessions. In observer-based annotation, observers may be asked to watch the recorded (e.g., audio-visual) data of students and perform annotation in an offline manner [27] or in an online manner, e.g., live annotation of students' engagement in a learning session [28].

2.1.4 Temporal resolution

Temporal resolution (timescale) [21] refers to whether the annotation is performed at frame-level (e.g., still images extracted from video frames), segment-level (e.g., pop-up window self-reports shown every 10 minutes in a session), or session-level (e.g., retrospective self-reports at the end of learning session). The majority of the existing engagement annotation datasets are recorded videos of students. Observers have either annotated single frames of videos, video segments of a predetermined length, or videos of the entire learning session. Some datasets have been annotated in an adaptive segment-level manner [27, 28] in which the timescale (e.g., the length of video segments in a video dataset) is determined according to the changes in the engagement states of the students.

2.1.5 Level of Abstraction

Regarding the *level of abstraction*, engagement can be annotated at a high level without considering the components of engagement, e.g., into two classes of engagement and disengagement. In a different setting, the affective, behavioral, and cognitive components of engagement are first separately annotated and then combined to result in a numerical value for engagement. In the field of affect annotation, Pomsta et al. [29] have defined the above two settings as discrete response, and dimensional response, respectively. Each of the affective, behavioral, and cognitive components of engagement can also be annotated at different levels. To illustrate, behavioral engagement can be in two states of Off-Task and On-Task. The On-Task behavior itself can be in different categories of On-Task Conversation, On-Task Giving Answers, and so on [28].

2.1.6 Combination

An annotated dataset suitable for developing AI algorithms requires a numerical value or a class label for each sample in the dataset. The *combination* dimension is concerned with how the annotated affective, behavioral, and cognitive components of engagement are combined to derive a numerical value or a class label for engagement. For instance, in the SE annotation protocol proposed by Aslan et al. [27], the combination of the On-Task behavioral state and Highly-Motivated affective state results in the state of engagement (versus disengagement). Naibert et al. [30] proposed different architectures for the combination of affective, behavioral, and cognitive components of engagement collected through self-report questionnaires. It should be noted that the practice of combining the multiple components precludes examining distinctions among the components, and important information may be lost [31]. In addition, a strategy for combination should take into account the correlation between the affective, behavioral, and cognitive states of students [32]. As described for the level of abstraction, if one value is directly assigned to a specific level of engagement without considering its components, no combination is required.

2.1.7 Quantification

The *quantification (scale)* dimension refers to how engagement level is represented numerically, i.e., the type of the variable used for definition and annotation of engagement. The quantification of engagement level as a psychology term must ensure objectivity, precision, and rigor [33]. The engagement was quantified and annotated as a dichotomous (binary) variable having two states of engagement and disengagement. It was also quantified as a categorical variable in some datasets. Most datasets quantified engagement as an ordinal, or interval variable representing discrete, or continuous levels of engagement, respectively.

3 Related Reviews

A few reviews have been published on automatic SE measurement, most of which focused on the AI methodologies and algorithms used for SE measurement. Dewan et al. [6] divided the existing SE measurement methods into three categories: automatic, semi-automatic, and manual, considering the methods' dependencies on students' participation. They identified the challenges involved in SE methods and briefly explored the available datasets and performance metrics for SE measurement techniques. Karimah and Hasegawa [5, 34] conducted a systematic review and studied available engagement definitions, datasets, and methods at a high level. After explaining the characteristics of the available datasets, they reviewed and explained different steps for engagement measurement as a supervised machine-learning problem. The authors covered pre-processing (including face detection, feature extraction, data augmentation, feature selection, dimensionality reduction, and imbalanced data handling), classic machine learning, deep learning, fine-tuning and transfer learning and performance evaluation for engagement measurement methods. Salam et al. [35] surveyed different aspects of context-driven engagement inference, entailing definition, engagement components and factors, publicly available datasets, ground truth assessment, features, and methods. The above aspects of engagement inference were studied in different settings, including human-human, human-computer, human-agent, and human-robot interaction. In the category of human-computer interaction, they covered a few SE in virtual learning datasets and measurement methods. Researchers in the area of educational psychology reviewed publications on the theory of student engagement, in general, [36], and in technology-mediated learning, [13, 37, 38]. The previous reviews in the area of computer science mainly were focused on machine-learning and deep-learning methodologies for engagement measurement [5, 6, 34, 35]. Karimah and Hasegawa [5, 34] and Salam et al. [35] briefly studied the existing SE datasets. However, the previous reviews did not emphasize the inconsistencies in the annotation of SE datasets and the difficulties associated with developing comparable AI models. We introduced the seven dimensions for engagement annotation (described in Section 2.1)

to critically examine these inconsistencies. Additionally, we provide recommendations for appropriate SE definitions and annotation protocols in virtual learning environments.

4 Methods

IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, and Google Scholar were searched for English journals and conference publications published between 2010 and 2022. Different combinations of the following keywords were adopted: engagement measurement/detection/prediction/recognition/-classification/regression, machine learning, deep learning, artificial intelligence, and dataset. Reviewers screened all studies in order to identify those in which a new dataset was proposed for developing AI models for automatic SE measurement. The studies introducing datasets containing single- or multi-modal data of individual students in online or offline computer-based virtual learning sessions were included. Correspondingly, the exclusion criteria are as follows: (i) the studies introducing SE datasets containing students in groups, (ii) the studies introducing SE datasets containing students in in-person classrooms, (iii) the studies introducing student datasets for general affect detection, such as basic affect state recognition or valence and arousal recognition, and not for engagement measurement, and (iv) the studies on developing AI models for SE measurement without introducing a new dataset. Although this review endeavored to provide a comprehensive overview of the literature, it should be noted that it was not conducted in a systematic review manner. The focus of this review was not on analyzing the use of AI techniques for SE measurement. Other reviews that delve into AI, machine learning, and deep learning [5, 6, 34, 35] are available for interested readers, Section 3.

The datasets were analyzed in terms of the definition and annotation of SE based on the seven dimensions of engagement annotation, described in Sections 2.1. In addition, the following seven baseline characteristics of the datasets were analyzed: computer-based lecture watching, reading activity, writing activity, or working with educational software; interactive or non-interactive activity; data collected in-the-wild or in-the-lab; the number of students; their sex; and age in the dataset; and the distribution of samples in different levels of engagement in the dataset.

5 Results and Discussion

Thirty studies introducing new datasets for SE measurement in virtual learning were included. Tables 1, and 2 show the seven dimensions of engagement annotation in the existing datasets, and the baseline characteristics of the datasets, respectively. The inconsistencies in the seven dimensions of engagement in the dataset are analyzed as follows.

5.1 Sources

The sources of annotation in the reviewed datasets were expert [39–45], non-expert [24, 46–53], crowdsourcers [23, 54–56], or through self-reported questionnaires [57–63]. Some of the reviewed datasets also combined self-reports with observer-based annotation [64–68]. In most studies, non-expert or crowdsourcer observers were untrained students or freelancers. Booth et al. [47] pointed out that the observers did not receive any clarification or guidance regarding how to interpret the term engagement. In the event that the observers were unfamiliar with the concept they were annotating, they may have used their uninformed definition of engagement, which can lead to inaccurate annotations. The AI models built on such an annotation strategy could learn erroneous concepts. A specialized concept of SE needs to be annotated by either experts or people with training in the field. Otherwise, the validity of these labels may be under question. Gupta et al. [23] commented on noisy data after using a crowdsourcing platform for obtaining engagement annotations. In some other studies, noise filters were applied to labels [48]. Expert annotators and a validated SE definition and annotation protocol would prevent the need for such post-processing. The cost and time to annotate the data is a known challenge, but accurate data annotation is paramount to building generalizable AI models. Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) [28] and Human Expert Labeling Process (HELP) [27] (explained in Section 6) are two engagement annotation protocols that are used in some of the studies, [39–42, 44], where training is provided to observers. Most of the questionnaires used in the reviewed datasets contained very few questions, sometimes including only one question [47, 58]. A short questionnaire could indicate a flawed data collection process that could influence the value of reported metrics [26].

In some of the reviewed datasets with multiple annotators, different inter-rater reliability or correlation metrics, e.g., Cronbach's alpha, Cohen's kappa, Fleiss' kappa, Krippendorff's alpha, and Pearson correlation, were used to evaluate the quality of the annotations produced by the annotators.

5.2 Data Modality

A variety of types of information were used by observers for engagement annotation in the existing datasets, such as video [23, 24, 39, 41, 42, 45–47, 49–51, 53–55, 62, 64–68], image [24, 46, 52, 56], audio [40, 42], screen capture [39, 40, 40–42], and mouse cursor tracking [40–42], or self-reports by students [57–63]. The diversity of data modalities used for annotation across datasets may not pose a problem; however, the inconsistencies between the data modalities used for annotation and for developing AI models may cause problems. For instance, Chen et al. [60] and Thomas et al. [63] used retrospective self-report questionnaires for annotation, but videos were used for training AI models. Retrospective self-reports are collected after the occurrence of engagement states. Therefore, as opposed to observer-based annotations with appropriate

time resolutions, self-reports are not reliable reflections of the in-situ engagement states of students. Thus, it will be difficult to develop AI models on such data. It is important to investigate the capacity of different data modalities to represent different components of engagement. That is, to what extent each affective, behavioral, and cognitive component of engagement can be annotated based on which data modalities. For instance, Bosch [67] investigated the feasibility of measuring cognitive engagement using video data modality. According to Alyuz et al. [40], the distribution of affective and behavioral dimensions of engagement are different in students in different high school grades and in diverse ethnicities. The expression of emotions and affect also differs across genders [69]. This demographic information, such as sex, gender, age, ethnicity, students' major, and the relevance of the virtual learning materials to students' majors, should be taken into consideration in engagement data collection and annotation (see Tables 1 and 2).

5.3 Timing

There were inconsistencies in the existing engagement datasets in terms of the timing of self-reports, most of which were annotated retrospectively [58–63], and a few were based on concurrent self-reports [57, 58, 67]. Moreover, Monkaresi et al. [58] used a combination of both. Apart from only one dataset [44], in which in-situ annotation was performed using BROMP protocol [28], all other observer-based annotations used retrospective annotation, which was performed using recorded data [23, 24, 39–47, 49–52, 54–56, 64–67, 67, 68, 70].

5.4 Temporal Resolution

D'Mello and Graesser [32] have differentiated between mood states and affect states. While moods, e.g., depression, have been defined for an entire learning session (several minutes or a few hours), engagement, defined as an affect state, arises and decays at much faster timescales (a few seconds). According to the extensive experiments on the dynamics of affect states during learning [32], [71], and [72], there is an affect state transition approximately every 30 seconds, every 10-40 seconds, and every one minute. The temporal resolution of engagement annotation should be determined based on this affect dynamics (in different populations and in different learning situations). The existing datasets used inconsistent temporal resolutions for engagement annotation, starting from frame-level annotation [24, 46, 49, 51, 52, 54–56, 56], segment-level annotation with segments of 1-second length to 30-minute length [23, 24, 43, 45, 47, 48, 66, 67, 73], and session-level annotation [59, 63]. Temporal resolution in some of the existing datasets is close to the timescales mentioned above, e.g., 10 seconds and 60 seconds in Whitehill et al. [24] and 10 seconds in Gupta et al. [23]. Whitehill et al. [24] have also reported higher inter-rater reliability for annotation with 10-second segments compared to 60-second segments.

None of the annotation protocols in the reviewed datasets with high temporal resolution indicated how to annotate when there is a transition between different levels of engagement. In the datasets with a relatively low temporal resolution, e.g., five minutes, more than one engagement state may occur in each timescale. Considering each data segment being annotated in the corresponding timescale as a multi-set (or bag of words) [74], none of the existing annotation protocols determined how many engagement or disengagement states (words) must occur in the timescale to be annotated as engagement or disengagement. A plausible solution is to have an adaptive timescale as performed in BROMP [28] and HELP [27] annotation protocols (explained in Section 6).

Inconsistencies in temporal resolution make it difficult to develop and evaluate AI models across datasets. For instance, a sequential machine-learning model with a specific architecture [14, 19, 74] is not capable of simultaneously handling video segments of lengths 10 seconds in the dataset presented in [23], 100 seconds in [63], and 5 minutes in [48].

5.5 Level of Abstraction

Even though engagement is a multi-component state [7, 10], in most of the reviewed datasets, it was defined as a single-component state without clarification of its components [39, 43, 45, 47, 50, 51, 54, 56–58, 62, 63, 66]. Some authors annotated engagement based on only one of its components, e.g., affective [23, 52, 60, 61, 65], behavioral [41, 42, 48, 55], or cognitive engagement [67]. Some works annotated engagement as a multi-component state [24, 40, 44, 46, 49]. The existing datasets containing single or multi-modal data did not provide any rationale for considering only one or more components of engagement.

5.6 Combination

In the datasets in which engagement was defined and annotated as a multi-component state, the components of engagement were combined to generate one numerical value for engagement [40, 46], as in HELP protocol [27], in which various combinations of affective and behavioral components resulted in a dichotomous engagement state. In some other works, one component of engagement was taken into consideration as a prerequisite for other components of engagement. For instance, in Alkabbany et al. [49], the presence of behavioral engagement (and the absence of affective engagement) corresponded to lower levels of engagement. Then, the presence of both affective and behavioral engagements corresponded to higher levels of engagement. In some other works, such as the datasets in which BROMP [28] was used for annotation, engagement was annotated as an affective state, and behavioral states were annotated separately [44]. They did not combine the affective and behavioral dimensions. In a totally different engagement annotation method, Bosch [67], the occurrence of mind-wandering was considered as cognitive engagement.

5.7 Quantification

A major issue in the reviewed datasets is the inconsistency in the use of scales to measure SE. A few researchers, e.g., Vanneste et al. [57], Zaletelj and Košir [43], have stressed the fact that there is no "gold standard" for measuring engagement. There are different quantification methods used in different datasets to represent different levels of engagement, and these methods range from the use of two points to the use of six points, as well as the use of continuous values. Correspondingly, engagement was considered a dichotomous (binary) variable [45, 46, 51, 58, 67, 68], an ordinal variable [23, 24, 43, 48–50, 54, 60, 61, 63–65], or an interval variable [47, 52, 57, 62, 66]. Some datasets defined engagement as a categorical variable [39–42, 55, 56], e.g., three categories of Engaged, Not-engaged, and Unknown in [39]. Moving forward, this inconsistency can be a major constraining factor for progress in the field. In a simplified sense, the concept of an object, "X" must be consistently annotated as "X" based on a commonly accepted measurement instrument across multiple data sources. Otherwise, supervised machine-learning models may not be able to learn that concept effectively. Corresponding to the different engagement scales in the existing datasets, different types of supervised machine-learning models were trained to solve binary or multi-class classification or regression problems [5, 6]. Due to this scale inconsistency, it is infeasible to use a machine-learning model trained on one dataset to make engagement inferences on another dataset annotated with a different engagement scale. Moreover, the performance of machine-learning models trained on different datasets with different engagement scales (e.g., a binary classification model with a regression model) cannot be compared.

5.8 Miscellanies

5.8.1 Publicly Available Datasets

A limited number of the reviewed datasets were available publicly, including Dataset for Affective States in E-Environments (DAiSEE) [23], Emotion Recognition in the Wild-Engagement prediction in the Wild (EmotiW-EW) [48], and Affect Transfer Learning for Behavior Prediction (ATL-BP) Student Engagement Dataset [55, 56]. Few researchers have analyzed and pointed out problems with the annotations in the public datasets. Abedi and Khan [15] indicated the annotation issues in the DAiSEE dataset as a misleading factor in training temporal and non-temporal deep-learning models. Additionally, Liao et al. [19] and Mehta et al. [20] discussed the annotation problems in the DAiSEE dataset by providing examples of videos and annotations of one student in different levels of engagement. Specifically, Liao et al. [19] criticized the use of discrete labels when annotating SE levels in videos and proposed the use of continuous values instead.

5.8.2 Characteristics of Virtual Learning Environment

The characteristics of the virtual learning environment in which the engagement annotation definition and protocol are designed to be applied are another important missing factor in the existing datasets. The existing datasets did not provide a rationale or justification for using a particular SE definition with respect to the characteristics of the virtual learning setting in the dataset. For example, it is important to determine whether the virtual learning environment is interactive or non-interactive. An interactive setting involves more interaction between the student and the computer (e.g., mouse cursor movements) than a non-interactive setting in which the student merely watches a recorded or online video. It is important to consider whether it is: (i) an online course with live communication between students and instructor, (ii) a recorded video of the instructor being viewed offline on a computer, (iii) a writing task on the computer screen, or (iv) a writing task on a piece of paper. As the affective, behavioral, and cognitive components of engagement are different in the above exemplary settings, the characteristics of the virtual learning setting must be considered during the design of engagement annotations.

5.8.3 Imbalanced Distribution

The distribution of samples in different levels of engagement in the existing datasets is presented in Table 2. It can be observed that the number of samples in disengagement or low levels of engagement is typically much lower than the number of samples in higher levels of engagement in almost all the existing datasets. The highly imbalanced data distribution in these datasets must be considered when developing AI models.

6 Student Engagement Definitions and Protocols in Other Settings

The definitions of SE presented in Section 2 and research in education and psychology have led to the definition and design of several SE scales used in various settings. In this section, some of the existing SE definitions and protocols in settings other than virtual learning that have the potential to be used in measuring SE in virtual learning are discussed.

BROMP [28] is an observation protocol for in vivo annotation of students' affective and behavioral states. In the BROMP platform, observers (*sources*) are trained and tested on the BROMP annotation protocol and achieve sufficient inter-rater agreement, Cohen's Kappa ¿ 0.6, in their observations and get a BROMP certification before participating in the annotation. Students in an in-person classroom working with educational software on computers are observed by the observer in-person by a side glance to make a holistic judgment of a student's state based on facial expressions, speech, body posture, gestures, and the student's interaction with the educational software (*data modality*). Observation is performed in a round-robin manner, observing and annotating

one student and moving to the next. The frequency of observations per student varied between class periods depending on the number of students in the class (*timing*). Each student is observed for 20 seconds or until a visible state is detected (*temporal resolution*). The annotation is inserted in a mobile application. In the BROMP protocol, the affective and behavioral states of students are annotated separately (*combination*). Various affect states are included in the BROMP protocol; some commonly used are Boredom, Confusion, Delight, Engaged Concentration, Frustration, and Surprise. The main behavioral states are On-Task and Off-Task (*level of abstraction* and *quantification*).

Aslan et al. [27] stated unaddressed challenges in BROMP [28] as follows: (1) limited chance for revision as annotation is performed in vivo, (2) difficulty of making a decision about a student's state in real-time, (3) fragmented annotation and disregarding state change in students due to the round-robin technique, (4) limited labels for model training, and (5) inevitable observer effect due to the presence of the observer in the learning settings. To address these challenges, they developed the HELP annotation process. HELP has a systematic process of training and evaluation for observers (*source*). It contains an annotation software containing the recorded video, audio, screen capture of students' computers and learning material contextual data, and demographic information of students (*data modality*). The *timing* is post-facto as observers watch the recorded data of students retrospectively. *Temporal resolution* is similar to BROMP, after observing the first state change of the student. The annotation of the affect and behavioral states are separate. The discrete affective states are Satisfied, Bored, Confused, and the behavioral states are On-Task and Off-Task (*level of abstraction* and *scale*). Inspired by Woolf et al. [75], different combinations of affect and behavioral states result in the dichotomous state of Engaged versus Not-engaged (*combination*). Some studies have used BROMP [44] and HELP [40–42] for engagement annotation. Aslan et al. [27] failed to demonstrate how addressing the fifth challenge in BROMP regarding the "inevitable observer effect" resulted in a better annotation. It also needs to be investigated which technique is optimal, observing one student continuously in HELP or using the round-robin technique in BROMP.

Altuwairqi et al. [65] presented an affective model (not an annotation protocol) for engagement, in which different areas of the circumplex model of affect [76], corresponding to different values of valence and arousal, are defined as five ordinal levels of engagement: Disengagement, Low, Medium, High, and Strong Engagement. In combination with self-reports, this affective model of engagement was used for video-based engagement annotation by Altuwairqi et al. [77].

Deng et al. [78] developed the Massive Open Online Course (MOOC) Engagement Scale (MES). The scale is a 12-item questionnaire comprised of four components, behavioural engagement, emotional engagement, cognitive engagement, and social engagement. The students are asked to fill out the questionnaire at the end of MOOC using a 6-point Likert scale for each item. Therefore, the timing and temporal resolution of MES are after the course,

and the entire course semester, respectively. The items were also designed based on the timing, e.g., "I set aside a regular time each week to work on the MOOC". The Online Student Engagement scale (OSE) is a 19-item questionnaire developed by Dixson [79]. Students report on a 5-point Likert scale how well each behavior, thought, or feeling was characteristic of them or their behavior. Examples of questions in OSE are "Making sure to study on a regular basis" and "Doing well on the tests/quizzes".

As is the case with the inconsistent SE annotation scales in the existing virtual learning datasets, the annotation protocols discussed in this section were also inconsistent with respect to the seven dimensions of engagement annotation. These annotation protocols were not specifically designed for the purpose of SE annotation in virtual learning settings and to develop AI models. However, with appropriate modifications (considering the seven dimensions of engagement annotation), these protocols have the potential to be utilized for this purpose, [39–42, 44].

7 Conclusions

In this critical review, we examined the existing SE virtual learning datasets and highlighted inconsistencies in terms of engagement definitions and annotations. Our analysis was based on the seven dimensions of engagement annotation: sources (observers), data modality, timing, temporal resolution, level of abstraction, combination, and quantification (scale). We discussed to what extent different dimensions of engagement annotation in the existing datasets are in accordance with the definition of SE in educational psychology. We explained how the inconsistencies are problematic in developing AI models for automatic SE measurement. We discussed some of the existing SE definitions and protocols in settings other than virtual learning that can be used in measuring SE in virtual learning. We appreciate the previous work by researchers who collected these datasets in order to contribute to progress in the field. However, we also raised doubts about the comparability of labels of different datasets and the generalizations of AI models across these datasets. We strongly recommend in future studies that both the observer-based and self-reporting SE annotations should be used in tandem to provide better evidence of SE in virtual learning. Consistent approaches for observational and self-reporting measurement of engagement should be developed to make progress in the field of developing AI models for SE measurement.

Table 1. The seven dimensions of engagement annotation in the existing datasets.

Reference	Sources	Data Modality	Timing	Time scale	Level of Abstraction	Combination	Quantification
Whitehill et al. 2014 [24]	Observers: 9 trained students	Video	Retrospective	10 and 60 seconds	Affective, behavioral, and cognitive components	Rule-based	Ordinal 1–4: not at all engaged–very engaged
Aslan et al. 2014 [39]	Observers: 3 experts	Video and computer screen	Retrospective	20 minutes	Engagement	NA	Categorical: engaged, not engaged, and unknown
Chen et al. 2015 [60]	self-reports	NA	Retrospective	2 minutes	Affective component	NA	Ordinal 1–6: very little engagement–very much engagement
Bosch et al. 2016 [67]	Observers: trained experts	video	Concurrent	Adaptive or 20 seconds	Affective and behavioral components	No combination	Categorical: boredom, confusion, delight, frustration, engaged concentration and off-task, on-task conversation, and on-task
Chen et al. 2016 [68]	-	Video, skin conductance, and log files	Retrospective	-	Engagement	NA	Dichotomous: attention ON and attention OFF
Monkaresi et al. 2016 [58]	self-reports	NA	Concurrent and retrospective	2 minutes	Engagement	NA	Dichotomous: engaged and not-engaged
Gupta et al. 2016 [23]	Observers: 10 untrained crowdsourcers	Video	Retrospective	10 seconds	Affective component	NA	Ordinal 1–4: very low–very high
Kamath et al. 2016 [54]	Observers: 25 untrained crowdsourcers	Video	Retrospective	single-frame	Engagement	NA	Ordinal 1–3: not-engaged–very engaged
Bosch 2016 [67]	self-reports + observers (untrained annotators)	Video	Concurrent + retrospective	12 seconds	Cognitive component	NA	Dichotomous: mind-wandering (not-engaged) and no mind-wandering (engaged)
Booth et al. 2017 [47]	Observers: 9 untrained students	Video	Retrospective	20 minutes	Engagement	NA	Interval 0–1
Okur et al. 2017 [41]	Observers: 3 experts	Video, audio, computer screen, mouse cursor, and URL logs	Retrospective	Adaptive	Behavioral component	NA	Categorical: on-task, off-task, not applicable, cannot decide
Alyuz et al. 2017 [42]	Observers: 3 experts	Video, audio, computer screen, and mouse cursor	Retrospective	Adaptive	Behavioral component	NA	Categorical: on-task, off-task, not applicable, cannot decide

Table 1 (continued)

Zaletelj et al. 2017 [43]	Observers: 5 experts	Video	Retrospective	1 second	Engagement	NA	Ordinal 1-5: 5 levels of engagement
Kaur et al. 2018 [48]	Observers: 5 experts	Video	Retrospective	5 minutes	Behavioral component	No combination	Ordinal 0–3: completely disengaged–highly engaged
De Carolis et al. 2019 [59]	self-reports	NA	Retrospective	9 minutes	-	-	-
Hutt et al. 2019 [61]	self-reports	NA	Retrospective	Random	Affective component	NA	Ordinal 1–5: not at all engaged–very engaged
Alkabbany et al. 2019 [49]	Observers: 4 annotators	Video	Retrospective	Single-frame	Affective and behavioral components	No combination	Ordinal 0–3: no detected face–emotionally engaged
Nezami et al. 2020 [46]	Observers: 6 trained students	Video	Retrospective	Single-frame	Affective and behavioral components	rule-based	Dichotomous: engaged and not-engaged
Vanneste et al. 2021 [57]	self-reports	NA	Concurrent	5-12 minutes	Engagement	NA	Interval 0–2: totally disengaged–totally engaged
Alyuz et al. 2021 [40]	Observers: 3 experts	Audio, video, and screen capture	Retrospective	Adaptive	Affective and behavioral components	No combination	Categorical: satisfied, bored, confused, on-task, off-task, not available, and cannot decide
Bhardwaj et al. 2021 [50]	Observers: 10 annotators	Video	Retrospective	-	Engagement	NA	Ordinal 0–5: 6 level of engagement
Delgado et al. 2021 [55]	Observers: 3 untrained crowdsourcers	Video	Retrospective	Single-frame	Behavioral component	No combination	Categorical: looking at their screen, looking at their paper, and wandering
Zheng et al. 2021 [64]	self-reports + observers	Video	Retrospective	12 minutes	Engagement	NA	Ordinal 1–3: 3 levels of engagement
Altuwairqi et al. 2021 [77]	self-reports + observers (untrained students)	Video	Retrospective	Single-frame	Affective component	NA	Ordinal 1–5: low–strong engagement
Ma et al. 2021 [66]	Observers: 3 untrained annotators	Video	Retrospective	30 minutes	Engagement	NA	Interval 0–1
Gupta et al. 2022 [52]	-	Image	Retrospective	Single-frame	Affective engagement	NA	Interval 0–1: basic facial expressions were converted to an interval variable
Buono et al. 2022 [62]	self-reports	NA	Retrospective	9 minutes	Engagement	NA	Interval 0–1
Jeong et al. 2022 [45]	Observers: 3 experts	Video	Retrospective	5 seconds	Engagement	NA	Dichotomous: engaged and not-engaged
Thomas et al. 2022 [63]	self-reports	NA	Retrospective	100 seconds	Engagement	NA	Ordinal 1–5
Verma et al. 2022 [53]	Observers: 3 trained annotators	Video	Retrospective	Adaptive	Behavioral engagement	NA	Categorical: disengagement, strange eye movements, presence of some kind of facial expression, yawning, face occlusion, body movements, and pressing keyboard

Table 2. The baseline characteristics of the existing student engagement datasets.

Reference	Activity	Interactive	In-the-wild	# of students	# of females	Age (years)	Class distribution of samples
Whitehill et al. 2014 [24]	software	yes	no	34	25	NA	6% not engaged at all, 10% nominally engaged, 46% engaged in task, and 38% very engaged
Aslan et al. 2014 [39]	lecture	no	no	9	NA	NA	NA
Chen et al. 2015 [60]	reading	no	yes	88	NA	NA	NA
Bosch et al. 2016 [67]	software	yes	yes	137	57	13–15	4% boredom, 2% confusion, 2% delight, 14% frustration, and 78% engaged concentration — 5% off-task, 21% on-task conversation, and 74% on-task
Chen et al. 2016 [68]	software	yes	no	30	17	NA	NA
Monkaresi et al. 2016 [58]	writing	no	no	23	9	20–60	80% engaged and 20% not-engaged
Gupta et al. 2016 [23]	lecture	no	yes	112	32	NA	1% very low, 5% low, 41% high, and 45% and very high
Kamath et al. 2016 [54]	lecture	no	yes	23	NA	18-24	9% not-engaged, 51% nominally engaged, and 40% very engaged
Bosch 2016 [67]	reading	no	yes	98	NA	NA	NA
Booth et al. 2017 [47]	lecture	no	no	12	NA	25	NA
Okur et al. 2017 [41]	lecture	yes	yes	28	NA	14-15	71% on-task and 29% off-task
Alyuz et al. 2017 [42]	lecture	yes	yes	17	NA	14-15	68% on-task and 32% off-task
Zaletelj et al. 2017 [43]	lecture	no	no	22	2	NA	NA
Kaur et al. 2018 [48]	lecture	no	yes	78	25	19–27	5% completely disengaged, 27% barely engaged, 42% engaged, and 26% highly engaged
De Carolis et al. 2019 [59]	lecture	yes	no	19	7	21	NA
Hutt et al. 2019 [61]	lecture	yes	yes	69174	NA	NA	NA
Alkabbany et al. 2019 [49]	lecture	yes	yes	14	NA	NA	0% no face detected, 12% behaviorally not engaged, 56% behaviorally engaged, emotionally not engaged, and 32% emotionally engaged
Nezami et al. 2020 [46]	software	yes	no	20	11	14–16	50% engaged and 50% not-engaged

Table 2 (continued)

Vanneste et al. 2021 [57]	lecture	no	yes	14	4	18	NA
Alyuz et al. 2021 [40]	lecture	yes	no	60	30	NA	NA
Bhardwaj et al. 2021 [50]	lecture	no	yes	1000	NA	NA	NA
Delgado et al. 2021 [55]	software	yes	no	19	NA	NA	25% looking at their screen, 72% looking at their paper, and 3% wandering
Zheng et al. 2021 [64]	software	yes	no	19	NA	NA	33% (1), 40% (2), and 27% (3)
Altuwairqi et al. 2021 [77]	lecture	no	yes	110	NA	NA	NA
Ma et al. 2021 [66]	lecture	no	yes	59	NA	20-32	NA
Gupta et al. 2022 [52]	NA	NA	NA	NA	NA	NA	NA
Buono et al. 2022 [62]	lecture	no	yes	31	14	19	NA
Jeong et al. 2022 [45]	lecture	no	yes	92	NA	20-31	NA
Thomas et al. 2022 [63]	lecture	no	yes	6	NA	NA	NA
Verma et al. 2022 [53]	reading	yes	no	26	NA	NA	11% disengagement, 25% strange eye movements, 12% presence of some kind of facial expression, 2% yawning, 8% face occlusion, 27 % body movements, 15% press keyboard

Data availability
No dataset was used.

Conflict of Interest
The authors declare that they have no conflict of interest.

References

[1] Mukhtar, K., Javed, K., Arooj, M., Sethi, A.: Advantages, limitations and recommendations for online learning during covid-19 pandemic era. Pakistan journal of medical sciences **36**(COVID19-S4), 27 (2020)

[2] Dung, D.T.H.: The advantages and disadvantages of virtual learning. IOSR Journal of Research & Method in Education **10**(3), 45–48 (2020)

[3] Sümer, Ö., Goldberg, P., D'Mello, S., Gerjets, P., Trautwein, U., Kasneci, E.: Multimodal engagement analysis from facial videos in the classroom. IEEE Transactions on Affective Computing (2021)

[4] Gray, J.A., DiLoreto, M.: The effects of student engagement, student satisfaction, and perceived learning in online learning environments. International Journal of Educational Leadership Preparation **11**(1), 1 (2016)

[5] Karimah, S.N., Hasegawa, S.: Automatic engagement estimation in smart education/learning settings: a systematic review of engagement definitions, datasets, and methods. Smart Learning Environments **9**(1), 1–48 (2022)

[6] Dewan, M., Murshed, M., Lin, F.: Engagement detection in online learning: a review. Smart Learning Environments **6**(1), 1–20 (2019)

[7] Fredricks, J.A., Blumenfeld, P.C., Paris, A.H.: School engagement: Potential of the concept, state of the evidence. Review of educational research **74**(1), 59–109 (2004)

[8] Trowler, V.: Student engagement literature review. The higher education academy **11**(1), 1–15 (2010)

[9] Reeve, J., Tseng, C.-M.: Agency as a fourth aspect of students' engagement during learning activities. Contemporary Educational Psychology **36**(4), 257–267 (2011)

[10] Sinatra, G.M., Heddy, B.C., Lombardi, D.: The challenges of defining and measuring student engagement in science. Taylor & Francis (2015)

[11] Kuh, G.D.: Assessing what really matters to student learning inside the

national survey of student engagement. Change: The magazine of higher learning **33**(3), 10–17 (2001)

[12] Fredricks, J.A., McColskey, W.: The measurement of student engagement: A comparative analysis of various methods and student self-report instruments. In: Handbook of Research on Student Engagement, pp. 763–782. Springer, ??? (2012)

[13] Henrie, C.R., Halverson, L.R., Graham, C.R.: Measuring student engagement in technology-mediated learning: A review. Computers & Education **90**, 36–53 (2015)

[14] Abedi, A., Khan, S.: Affect-driven engagement measurement from videos. arXiv preprint arXiv:2106.10882 (2021)

[15] Abedi, A., Khan, S.S.: Improving state-of-the-art in detecting student engagement with resnet and tcn hybrid network. In: 2021 18th Conference on Robots and Vision (CRV), pp. 151–157 (2021). IEEE

[16] Bustos-López, M., Cruz-Ramírez, N., Guerra-Hernández, A., Sánchez-Morales, L.N., Cruz-Ramos, N.A., Alor-Hernández, G.: Wearables for engagement detection in learning environments: A review. Biosensors **12**(7), 509 (2022)

[17] Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.-P.: Openface 2.0: Facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 59–66 (2018). IEEE

[18] Cao, Z., Martinez, G.H., Simon, T., Wei, S., Sheikh, Y.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. in ieee transactions on pattern analysis and machine intelligence (2019)

[19] Liao, J., Liang, Y., Pan, J.: Deep facial spatiotemporal network for engagement prediction in online learning. Applied Intelligence **51**(10), 6609–6621 (2021)

[20] Mehta, N.K., Prasad, S.S., Saurav, S., Saini, R., Singh, S.: Three-dimensional densenet self-attention neural network for automatic detection of student's engagement. Applied Intelligence, 1–21 (2022)

[21] D'Mello, S.K.: On the influence of an iterative affect annotation approach on inter-observer and self-observer reliability. IEEE Transactions on Affective Computing **7**(2), 136–149 (2015)

[22] Brabham, D.C.: Crowdsourcing. Mit Press, ??? (2013)

[23] Gupta, A., D'Cunha, A., Awasthi, K., Balasubramanian, V.: Daisee: Towards user engagement recognition in the wild. arXiv preprint arXiv:1609.01885 (2016)

[24] Whitehill, J., Serpell, Z., Lin, Y.-C., Foster, A., Movellan, J.R.: The faces of engagement: Automatic recognition of student engagementfrom facial expressions. IEEE Transactions on Affective Computing **5**(1), 86–98 (2014)

[25] Minner, D.D., Levy, A.J., Century, J.: Inquiry-based science instruction—what is it and does it matter? results from a research synthesis years 1984 to 2002. Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching **47**(4), 474–496 (2010)

[26] O'Brien, H., Cairns, P.: Why Engagement Matters: Cross-disciplinary Perspectives of User Engagement in Digital Media. Springer, ??? (2016)

[27] Aslan, S., Mete, S.E., Okur, E., Oktay, E., Alyuz, N., Genc, U.E., Stanhill, D., Esme, A.A.: Human expert labeling process (help): towards a reliable higher-order user state labeling process and tool to assess student engagement. Educational Technology, 53–59 (2017)

[28] Ocumpaugh, J.: Baker rodrigo ocumpaugh monitoring protocol (bromp) 2.0 technical and training manual. New York, NY and Manila, Philippines: Teachers College, Columbia University and Ateneo Laboratory for the Learning Sciences **60** (2015)

[29] Porayska-Pomsta, K., Mavrikis, M., D'Mello, S., Conati, C., Baker, R.S.: Knowledge elicitation methods for affect modelling in education. International Journal of Artificial Intelligence in Education **22**(3), 107–140 (2013)

[30] Naibert, N., Barbera, J.: Development and evaluation of a survey to measure student engagement at the activity level in general chemistry. Journal of Chemical Education **99**(3), 1410–1419 (2022)

[31] McNeish, D.: Limitations of the sum-and-alpha approach to measurement in behavioral research. Policy Insights from the Behavioral and Brain Sciences **9**(2), 196–203 (2022)

[32] D'Mello, S., Graesser, A.: Dynamics of affective states during complex learning. Learning and Instruction **22**(2), 145–157 (2012)

[33] Tafreshi, D., Slaney, K.L., Neufeld, S.D.: Quantification in psychology: Critical analysis of an unreflective practice. Journal of Theoretical and Philosophical Psychology **36**(4), 233 (2016)

[34] Karimah, S.N., Hasegawa, S.: Automatic engagement recognition for distance learning systems: A literature study of engagement datasets and methods. In: International Conference on Human-Computer Interaction, pp. 264–276 (2021). Springer

[35] Salam, H., Celiktutan, O., Gunes, H., Chetouani, M.: Automatic context-driven inference of engagement in hmi: A survey. arXiv preprint arXiv:2209.15370 (2022)

[36] Wong, Z.Y., Liem, G.A.D.: Student engagement: Current state of the construct, conceptual refinement, and future research directions. Educational Psychology Review, 1–32 (2021)

[37] Schindler, L.A., Burkholder, G.J., Morad, O.A., Marsh, C.: Computer-based technology and student engagement: a critical review of the literature. International journal of educational technology in higher education **14**(1), 1–28 (2017)

[38] Hu, M., Li, H.: Student engagement in online learning: A review. In: 2017 International Symposium on Educational Technology (ISET), pp. 39–43 (2017). IEEE

[39] Aslan, S., Cataltepe, Z., Diner, I., Dundar, O., Esme, A.A., Ferens, R., Kamhi, G., Oktay, E., Soysal, C., Yener, M.: Learner engagement measurement and classification in 1: 1 learning. In: 2014 13th International Conference on Machine Learning and Applications, pp. 545–552 (2014). IEEE

[40] Alyuz, N., Aslan, S., D'Mello, S.K., Nachman, L., Esme, A.A.: Annotating student engagement across grades 1–12: Associations with demographics and expressivity. In: International Conference on Artificial Intelligence in Education, pp. 42–51 (2021). Springer

[41] Okur, E., Alyuz, N., Aslan, S., Genc, U., Tanriover, C., Arslan Esme, A.: Behavioral engagement detection of students in the wild. In: International Conference on Artificial Intelligence in Education, pp. 250–261 (2017). Springer

[42] Alyuz, N., Okur, E., Genc, U., Aslan, S., Tanriover, C., Esme, A.A.: An unobtrusive and multimodal approach for behavioral engagement detection of students. In: Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education, pp. 26–32 (2017)

[43] Zaletelj, J., Košir, A.: Predicting students' attention in the classroom from kinect facial and body features. EURASIP journal on image and video processing **2017**(1), 1–12 (2017)

[44] Bosch, N., D'mello, S.K., Ocumpaugh, J., Baker, R.S., Shute, V.: Using video to automatically detect learner affect in computer-enabled classrooms. ACM Transactions on Interactive Intelligent Systems (TiiS) **6**(2), 1–26 (2016)

[45] Jeong, Y.-S., Cho, N.-W.: Evaluation of e-learners' concentration using recurrent neural networks. The Journal of Supercomputing, 1–18 (2022)

[46] Mohamad Nezami, O., Dras, M., Hamey, L., Richards, D., Wan, S., Paris, C.: Automatic recognition of student engagement using deep learning and facial expression. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 273–289 (2020). Springer

[47] Booth, B.M., Ali, A.M., Narayanan, S.S., Bennett, I., Farag, A.A.: Toward active and unobtrusive engagement assessment of distance learners. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 470–476 (2017). IEEE

[48] Kaur, A., Mustafa, A., Mehta, L., Dhall, A.: Prediction and localization of student engagement in the wild. In: 2018 Digital Image Computing: Techniques and Applications (DICTA), pp. 1–8 (2018). IEEE

[49] Alkabbany, I., Ali, A., Farag, A., Bennett, I., Ghanoum, M., Farag, A.: Measuring student engagement level using facial information. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 3337–3341 (2019). IEEE

[50] Bhardwaj, P., Gupta, P., Panwar, H., Siddiqui, M.K., Morales-Menendez, R., Bhaik, A.: Application of deep learning on student engagement in e-learning environments. Computers & Electrical Engineering **93**, 107277 (2021)

[51] Binh, H.T., Trung, N.Q., Nguyen, H.-A.T., Duy, B.T.: Detecting student engagement in classrooms for intelligent tutoring systems. In: 2019 23rd International Computer Science and Engineering Conference (ICSEC), pp. 145–149 (2019). IEEE

[52] Gupta, S., Kumar, P., Tekchandani, R.K.: Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. Multimedia Tools and Applications, 1–30 (2022)

[53] Verma, M., Nakashima, Y., Takemura, N., Nagahara, H.: Multi-label disengagement and behavior prediction in online learning. In: International Conference on Artificial Intelligence in Education, pp. 633–639 (2022). Springer

[54] Kamath, A., Biswas, A., Balasubramanian, V.: A crowdsourced approach to student engagement recognition in e-learning environments. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–9 (2016). IEEE

[55] Delgado, K., Origgi, J.M., Hasanpoor, T., Yu, H., Allessio, D., Arroyo, I., Lee, W., Betke, M., Woolf, B., Bargal, S.A.: Student engagement dataset. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3628–3636 (2021)

[56] Ruiz, N., Yu, H., Allessio, D.A., Jalal, M., Joshi, A., Murray, T., Magee, J.J., Delgado, K.M., Ablavsky, V., Sclaroff, S., et al.: Atl-bp: a student engagement dataset and model for affect transfer learning for behavior prediction. IEEE Transactions on Biometrics, Behavior, and Identity Science (2022)

[57] Vanneste, P., Oramas, J., Verelst, T., Tuytelaars, T., Raes, A., Depaepe, F., Van den Noortgate, W.: Computer vision and human behaviour, emotion and cognition detection: A use case on student engagement. Mathematics **9**(3), 287 (2021)

[58] Monkaresi, H., Bosch, N., Calvo, R.A., D'Mello, S.K.: Automated detection of engagement using video-based estimation of facial expressions and heart rate. IEEE Transactions on Affective Computing **8**(1), 15–28 (2016)

[59] De Carolis, B., D'Errico, F., Macchiarulo, N., Palestra, G.: "engaged faces": Measuring and monitoring student engagement from face and gaze behavior. In: IEEE/WIC/ACM International Conference on Web Intelligence-Companion Volume, pp. 80–85 (2019)

[60] Chen, Y., Bosch, N., D'Mello, S.: Video-based affect detection in noninteractive learning environments. International Educational Data Mining Society (2015)

[61] Hutt, S., Grafsgaard, J.F., D'Mello, S.K.: Time to scale: Generalizable affect detection for tens of thousands of students across an entire school year. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–14 (2019)

[62] Buono, P., De Carolis, B., D'Errico, F., Macchiarulo, N., Palestra, G.: Assessing student engagement from facial behavior in on-line learning. Multimedia Tools and Applications, 1–19 (2022)

[63] Thomas, C., Sarma, K.P., Gajula, S.S., Jayagopi, D.B.: Automatic prediction of presentation style and student engagement from videos. Computers and Education: Artificial Intelligence, 100079 (2022)

[64] Zheng, X., Hasegawa, S., Tran, M.-T., Ota, K., Unoki, T.: Estimation of learners' engagement using face and body features by transfer learning. In: International Conference on Human-Computer Interaction, pp. 541–552 (2021). Springer

[65] Altuwairqi, K., Jarraya, S.K., Allinjawi, A., Hammami, M.: A new emotion–based affective model to detect student's engagement. Journal of King Saud University-Computer and Information Sciences **33**(1), 99–109 (2021)

[66] Ma, J., Jiang, X., Xu, S., Qin, X.: Hierarchical temporal multi-instance learning for video-based student learning engagement assessment. In: IJCAI, pp. 2782–2789 (2021)

[67] Bosch, N.: Detecting student engagement: human versus machine. In: Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, pp. 317–320 (2016)

[68] Chen, J., Luo, N., Liu, Y., Liu, L., Zhang, K., Kolodziej, J.: A hybrid intelligence-aided approach to affect-sensitive e-learning. Computing **98**(1), 215–233 (2016)

[69] Brody, L.R., Hall, J.A.: Gender and emotion in context. Handbook of emotions **3**, 395–408 (2008)

[70] Khan, S.S., Mishra, P.K., Javed, N., Ye, B., Newman, K., Mihailidis, A., Iaboni, A.: Unsupervised deep learning to detect agitation from videos in people with dementia. IEEE Access **10**, 10349–10358 (2022)

[71] D'Mello, S., Graesser, A., *et al.*: Monitoring affective trajectories during complex learning. In: Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 29 (2007)

[72] d Baker, R.S., Rodrigo, M., Mercedes, T., Xolocotzin, U.E.: The dynamics of affective transitions in simulation problem-solving environments. In: International Conference on Affective Computing and Intelligent Interaction, pp. 666–677 (2007). Springer

[73] Zhang, H., Xiao, X., Huang, T., Liu, S., Xia, Y., Li, J.: An novel end-to-end network for automatic student engagement recognition. In: 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC), pp. 342–345 (2019). IEEE

[74] Abedi, A., Khan, S.S.: Detecting disengagement in virtual learning as an anomaly. arXiv preprint arXiv:2211.06870 (2022)

[75] Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., Picard,

R.: Affect-aware tutors: recognising and responding to student affect. International Journal of Learning Technology **4**(3/4), 129–164 (2009)

[76] Russell, J.A.: A circumplex model of affect. Journal of personality and social psychology **39**(6), 1161 (1980)

[77] Altuwairqi, K., Jarraya, S.K., Allinjawi, A., Hammami, M.: Student behavior analysis to measure engagement levels in online learning environments. Signal, Image and Video Processing **15**(7), 1387–1395 (2021)

[78] Deng, R., Benckendorff, P., Gannaway, D.: Learner engagement in moocs: Scale development and validation. British Journal of Educational Technology **51**(1), 245–262 (2020)

[79] Dixson, M.D.: Measuring student engagement in the online course: The online student engagement scale (ose). Online Learning **19**(4), 4 (2015)