Engagement Detection and its Applications in Learning: A Tutorial & Selective Review

Brandon M. Booth*, Nigel Bosch†, Sidney K. D'Mello*
*Institute of Cognitive Science, University of Colorado Boulder
†School of Information Sciences and Department of Educational Psychology, University of Illinois Urbana–Champaign
Email: brandon.booth@colorado.edu, pnb@illinois.edu, sidney.dmello@colorado.edu

***Abstract*—Engagement is critical to satisfaction and performance in a number of domains but is challenging to measure and sustain. Thus, there is considerable interest in developing affective computing technologies to automatically measure and enhance engagement, especially in the wild and at scale. This paper provides an accessible introduction to affective computing research on engagement detection and enhancement using educational applications as an application domain. We begin with defining engagement as a multi-componential construct (i.e., a conceptual entity) situated within a context and bounded by time and review how the past six years of research has conceptualized it. Next, we examine traditional and affective computing methods for measuring engagement and discuss their relative strengths and limitations. Then we move to a review of proactive and reactive approaches to enhancing engagement towards improving the learning experience and outcomes. We underscore key concerns in engagement measurement and enhancement, especially in digitally enhanced learning contexts, and conclude with several open questions and promising opportunities for future work.**

***Index Terms*—Engagement, learning outcomes, affective computing, procedural justice**

I. Introduction

> "If you are interested in something, you will focus on it, and if you focus attention on anything, it is likely that you will become interested in it. Many of the things we find interesting are not so by nature, but because we took the trouble of paying attention to them." – Mihaly Csikszentmihalyi, *Finding Flow: The Psychology of Engagement with Everyday Life, pp. 128* [39]

As the renowned psychologist Csikszentmihalyi points out, our engagement is both a byproduct of and an essential precursor to our long-term interests. For many of us, it seems natural that we focus our attention on tasks we already have an interest in, but research and theoretical models of interest development suggest that we can nurture an emerging interest in an activity through repeated engagement with it [40, 41].

In the context of education, this observation is one of the primary reasons why triggering and nurturing early phases of interest is so important. Maintaining engagement, however, taxes cognitive and emotional resources, making it difficult to sustain for prolonged periods of time. In learning contexts especially, many hours of engaged attentive focus are needed to successfully master new and challenging content. Thus, maintaining engagement in the moment as well as a sustainable commitment to reengaging periodically over time are essential. Can intelligent technologies help promote and sustain engagement across extended periods of time?

Recent advancements in remote sensing technologies (e.g., portable cameras, microphones, wearable health trackers) and affective computing approaches offer one potential solution by enabling deeper understanding of people and their behaviors and emotions in different contexts (e.g., facial expression tracking [42], physiological stress detection [43]). In learning contexts, these technologies aim to monitor learner engagement and take actions to help promote engagement (e.g., to help improve the pacing of instructional content or reengage students who may have temporarily disengaged). These approaches fall under the broad umbrella of affect detection [44, 45], affect-aware interaction (e.g., [46, 47]; see [48] for a review), and attention-aware interaction [49].

However, because engagement is a complex psychological construct (i.e., a conceptual entity), purely technical approaches risk oversimplification, resulting in measures that are only tangentially related to engagement [50]. A purely technical perspective also risks adopting narrow value and reward structures such as maximizing predictive accuracy without considering robustness, generalizability, interpretability, bias/fairness, and application contexts [51, 52]. Thus, it is critical that any attempts to measure and improve engagement in a given domain adopt interdisciplinary perspectives that blend psychological, computational, and domain-specific know-how. This is precisely what we do in this manuscript by providing a review of engagement from the fields of cognitive, affective, and motivational science, affective computing, attentional computing, wearable sensing, and machine learning. We focus on research and applications involving learning and education to keep the scope manageable.

So what exactly is engagement and how can technologies help to enhance it? Conceptually, engagement is easy to comprehend, but as we will discuss in this article, it is difficult to precisely define. We will begin by examining notions of engagement broadly, which encompass a very large number of behaviors, emotions, and cognitive features that are differentially relevant when studying engagement in different domains (Section II-A). Next, we will adopt two complementary schemes representing the multiple dimensions and perspectives

Manuscript revision received April 14, 2023

of engagement (Section II-B) and use them to illustrate the breadth of engagement research in the past few years (Section II-C). We will provide an overview of traditional and recent approaches to measuring learner engagement in particular, and also discuss how digital technologies and recent advances in human-centered and affective computing are enabling automated measurement of learner engagement (Section III-A). We will examine how these automated technologies perform both in terms of accuracy of learner engagement assessment and also in terms of the biases and errors they make and how those may negatively impact certain groups of learners. Then, with these traditional and automated engagement measurement approaches in place, we will describe how technologies may be used to help enhance engagement to improve learning outcomes, focusing on systems from the past decade that implemented these strategies (Section IV). Finally, we will conclude with a discussion of the strengths and weaknesses of different approaches to automated engagement detection and feedback systems, propose open research questions for enhancing learner engagement, and highlight important next steps for future research (Section V).

II. Conceptualizing Engagement

A. What is Engagement?

In 2013, an interdisciplinary group of researchers spanning computing sciences and psychology convened to discuss and better understand what "engagement" [53] means to different research fields. They discovered that the notions of engagement in different research areas spans a wide range of behaviors, thoughts, perceptions, feelings, and attitudes towards a particular task, as also noted by Christenson et al. [54]. In particular, the committee produced a list of behaviors indicative of engagement relevant in specific contexts, including: attendance, attention, memory, caring, emotion, inhibited actions, an urge to share, understanding/learning, taking action, willingness, active participation, and mental investment. While this list indeed covers many pertinent aspects of engagement, a rigorous definition that is both broad enough to be generalizable and narrow enough to be scientifically measured and tested continues to remain out of reach. As Eccles and Wang observe [55], generalizable notions of engagement may be more intuitive and accessible to the public, but they offer little guidance for scientific inquiry and uncovering cause–effect relationships between the antecedents and consequents of engagement. Thus, it may be more beneficial to narrowly study engagement in particular contexts in terms of the associated behaviors and mental states rather than adopting all encompassing definitions which equate being engaged to "doing something."

Adopting a narrow focus has been the usual approach in a tradition of theoretical and practical scientific inquiry. When considering motivational aspects of engagement, theories including self-determination theory [56, 57] and self-efficacy theory [58–60] emphasize understanding the precursors of engagement like autonomy, self-efficacy, interest in a particular activity, and a balance of challenge and skill. Theories focused on cognition prioritize understanding the ebb and flow of mental demands and how they impact attention and performance [33, 61, 62]. For instance, in learning contexts, the *Interactive-Constructive-Active-Passive* (ICAP) framework [33] and an attention-based extension [63] propose that cognitive engagement and attention are highest for *interactive* (e.g., debating/discussing) and *constructive* (e.g., generating a self-explanation) activities, decreasing progressively in order for *active* (e.g., copying verbatim notes), and *passive* (e.g., watching a prerecorded lecture). Additionally, affective theories focus on the role of mood and emotion on engagement, such as the assimilation–accommodation framework [64], goal appraisal theories emphasizing how physiological arousal and cognitive appraisal influence emotions, or hedonic schema-based theories examining the interplay between the pleasure of immersion (i.e., "flow" [65]) and interactive engagement [66]. Though engagement is often perceived as a positive mental state, it is sometimes linked with addictive behaviors which may occur when dopamine (a "feel good" chemical in the brain) is confused with happiness causing people to engage with certain activities (e.g., video games) to the detriment of other life goals [67]. It may also be associated with affective states such as confusion and frustration, which may have a negative valence, but are part and parcel of complex learning [68].

Given its multitude of manifestations (e.g., visual attention, activity participation, feelings towards the activity) and the numerous contexts in which it is studied, we adopt a broad perspective of engagement. Specifically, *we operationalize engagement in terms of multi-componential affective states, cognitive states, and behaviors that arise from interactions with a person and a task context and unfold across multiple time scales*. In other words, as reviewed below, engagement is is not one unitary entity, but an umbrella term that is operationalized as a function of component(s), task context, and timescale as noted below:

$$\text{engagement} = f(\text{component}, \text{context}, \text{time})$$

B. Components of Engagement and the Engagement Continuum

Researchers generally agree that engagement is a multidimensional construct that encompasses not only how one feels or behaves in the moment but also longer-term patterns of engaging and wanting to engage with a task [53]. To provide a more structured approach to its study in learning contexts, Fredricks, Blumenfeld, and Paris [69] proposed three components (or facets) of engagement: emotional, behavioral, cognitive. *Emotional engagement* refers to one's feelings and attitudes about a specific task or the context in which it is performed, such as feelings of interest towards a particular subject or liking for a teacher in school [70]. *Behavioral engagement* regards aspects related to one's direct involvement in a task and encompasses behavioral features like participation and persistence (e.g., "hard fun" in learning games [71]) practice, and level of effort. *Cognitive engagement* pertains to the allocation of cognitive resources to a task, ranging from maintaining attentional focus to adopting high-level learning strategies, As such, it captures aspects related

to cognitive outcomes, such as memory, recall, learning, and a deep understanding or mastery of knowledge pertaining to the task. These components provide a useful mechanism for approaching engagement as a multi-componential construct, however, they do not account for the influence of context and time, both of which are crucial dimensions.

Complementary to the multi-componential categorization scheme, Sinatra et al. consider how context and time scale influence engagement [72]. They propose engagement along a continuum with one endpoint corresponding to a *person-oriented* perspective of engagement, which focuses on the cognitive, behavioral, and emotional components of engagement within individuals in a singular task context and across a short time span lasting seconds to minutes. Studies at this extreme would focus on tracking one or more components of engagement while people are engaged in some activity over a short time period (e.g., using eye trackers to measure mind wandering while students interact with an educational technology in a computer-enabled classrooms [12]). Most affective computing research can be aligned to the person-oriented perspective of engagement. At the other extreme is a *context-oriented* perspective where group-level or task-level features of engagement are considered as products of context and across extended time frames lasting weeks, months, or years. For example, understanding how broad learning structures (e.g., whether a school adopts a traditional lecture-based, computer-based instruction, or a blended approach) influence engagement at the school-level (i.e., an aggregate of measurements on individual students). Sinatra et al. [72] propose that between these two extremes, is the *person-in-context* perspective where the focus in on how particular contexts influence individuals' engagement (e.g., how a teacher's behavior differentially influences students from different backgrounds [73]) and unfolds over tens of minutes to hours and days. Of course, every event entails an interaction of person, context, and time, but the framework highlights whether the event is characterized from the perspective of the person, the context, or their interaction.

C. Review of Recent Studies on Engagement and Learning

We conducted a review of how recent research within learning contexts has contextualized engagement with respect to components, contexts, and time scales. To facilitate this, we selected papers from the last six years (published 2017 or later) based on Google Scholar results for *engagement "machine learning" students*, search variations including modality-specific keywords (e.g., "EEG"), and by exhaustively searching for "engage" in all paper titles published in *IEEE Transactions on Affective Computing* during or after 2017.

We filtered papers based on whether they considered engagement in learning, then categorized the contextual focus (i.e., person-oriented to context-oriented) based primarily on the research questions or goals in the papers. Research questions varied from highly person-oriented, in which the purpose was to learn about signals of individuals' engagement, to highly context-oriented questions, in which the purpose was to learn something about the engaging properties of the context itself. For example, Chang et al. [4] focused on engagement detection across contexts in a person-oriented way, relying only on features that could be extracted across contexts (e.g., facial expressions) without accounting for the influence of context on those features. Conversely, Soffer & Cohen [25] addressed person-in-context research questions that integrate person-oriented goals (i.e., detecting student engagement) and context-oriented goals (i.e., discovering which aspects of the context predicted engagement). In highly context-oriented research, Seo et al. [24] investigated how an aspect of the context (i.e., a novel type of video presentation) related to student engagement. Most research projects are between the extremes of the continuum, where they take contextual factors into account for predicting individual engagement or consider a mix of both contextual and person-oriented research questions. Table A1 in the supplementary materials contains the tabulated results from this survey, accounting for the focus along the engagement continuum, engagement components, engagement construct, data modalities, measurement approaches, modeling approaches, and also methods to enhance engagement. Figure 1 plots these papers along the engagement components (discrete) and engagement continuum (continuous) dimensions and illustrates the focus of the past six years of research. The color of the boxes in Figure 1 groups studies by the measurement approaches where black boxes use overt (openly displayed; e.g., gaze, face, logs) signals, white boxes use covert (subconscious or less-controlled; e.g., electroencephalography [EEG], electrodermal activity [EDA], ambient sound) signals, and gray boxes indicate a mixture of the two. Additionally, the data modalities used to monitor and measure engagement are listed within each box.

Based on the results from Table A1 and Figure 1, the past six years of learning engagement research span a variety of components, contexts, and time scales. This is quite a departure from the the bulk of research on learner engagement prior to the last decade (see Christenson, Reschly, and Wylie [54] for an excellent summary), which has mainly focused on the classroom as a traditional learning context. In particular, out of the 32 papers we surveyed, only 22% focus on traditional classroom activities or settings while the rest focus on learning from digital technologies and online learning systems (see the *Learning context* column in Table A1). Collectively, these studies focus on different engagement components (26% affective, 26% cognitive, 48% behavioral), more remote/online than classroom contexts (22% classroom, 37% remote/online, 41% laboratory), and they skew more towards short (person-oriented) than long-term (context-oriented) time scales (47% person-oriented, 38% person-in-context, 15% context-oriented).

1) Engagement Components: Sometimes when learners are willing to engage, but have difficulties sustaining engagement (*cognitive* component), long-term learning outcomes are poor. For instance, when learners are bored with learning content or the learning environment (e.g., school classroom), they may have difficulties maintaining attention, leading to diminished learning outcomes [74–76]. A meta-analysis of 29 research studies found a statistically significant (N = 19,052 students total) negative correlation (r = -.24) between boredom (an *affective* component) and academic success [77]. Perhaps

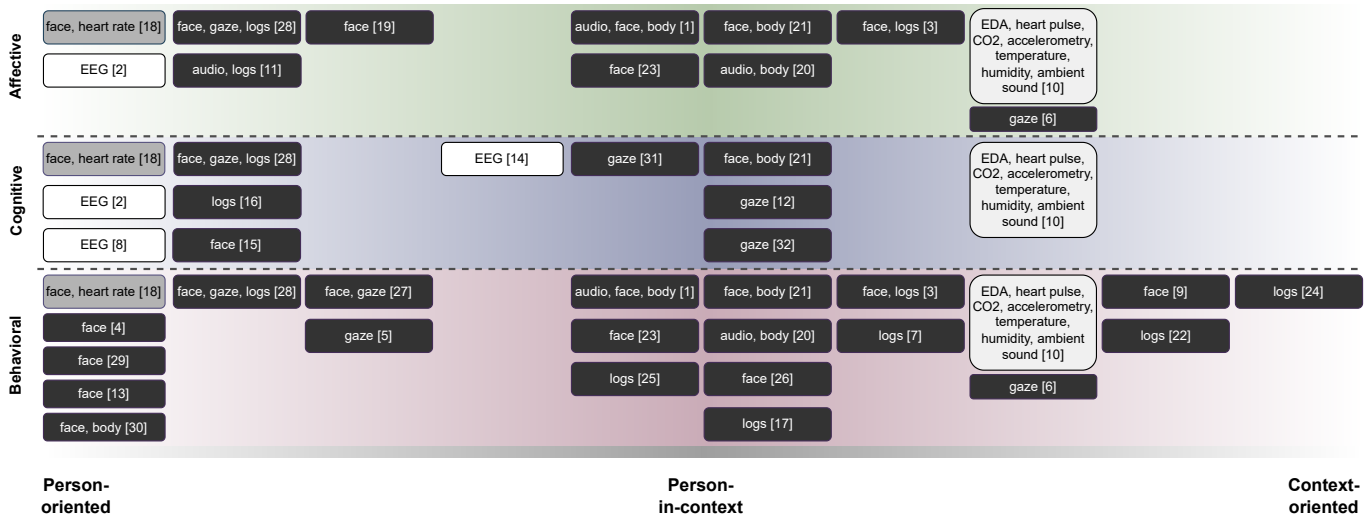

Fig. 1. A plot of our survey of engagement research over the past six years (see Table A1) categorized using the three components (behavioral, cognitive, emotional) proposed by Fredricks et al. [69] and the engagement continuum (ranging from person-oriented to context-oriented) proposed by Sinatra et al. [72]. The signals used to measure engagement within each work are listed in each box. Black boxes use overt (openly displayed) signals, white boxes use covert (subconscious or uncontrolled) signals, and gray boxes indicate a mixture of the two. Additionally, the data modalities used to assess engagement are listed within each box. EEG = electroencephalography, EDA = electrodermal activity, CO2 = carbon dioxide levels.

unsurprisingly, interventions that promote attention and concentration are among the most beneficial of all interventions explored in another recent meta-analysis of training programs for self-regulated learning skills (total N = 5,786; Hedges' g = 0.61) [78]. Stress (an *affective* component) is yet another influential factor, but one which can have both positive and negative effects on learner engagement, depending on its strength, duration, source, and learner personality [79–81].

Certain behavioral manifestations of engagement offer a window into the day-to-day and week-to-week trends in engagement. In the classroom, learner attendance and completed exercises provides some important behavioral indicators linked to positive learning outcomes [82, 83], while in technology-driven and online learning contexts, forum posts and viewed videos are readily apparent behavioral indicators of engagement [84, 85]. Behavioral engagement in online settings has also been used to learn more about how course content benefits students (e.g., [22]).

2) Engagement Contexts: Engagement in learning manifests in different ways across a spectrum of activities ranging from listening and note-taking in traditional classroom learning [33], watching lecture videos and answering questions in remote or digital learning [86], to reading and completing homework [87, 88]. Engagement is relevant for each of these activities over varied time scales as well.

Today, more and more professionally curated learning content is becoming accessible through Massive Open Online Courses (MOOCs; e.g., edX, Coursera) and less traditional presentations of learning content are also available on YouTube and other streaming platforms (e.g., Khan Academy, Veritasium). Though these digital learning platforms offer perhaps the most convenient and approachable access to learning content, with MOOCs providing additional facilities for students and teachers to directly communicate and interact, student disengagement and course dropout rates are problematic and often more troublesome than traditional classroom learning [89].

Engagement manifests in ways that differ substantially between face-to-face and online learning contexts as well. Objective indicators of engagement from methods like eye-gaze tracking and electroencephalography (e.g., [2, 5, 12]) are difficult to implement with high fidelity outside of experimenter-controlled in-person contexts. Consequently, research on engagement in different learning contexts has also adopted varied approaches. For example, Rodriguez et al. [22] clustered students according to their online course behaviors to learn more about how the course context benefited students, whereas Gao et al. [10] explored certain in-person contextual factors directly, such as indoor climate. At the person-oriented end of the spectrum, research questions about the individual may be answered via fine-grained data of engagement-related behaviors during learning with technology [16], though multimodal data collected in face-to-face contexts are perhaps more common (e.g., [2, 11, 29, 30]).

3) Engagement Time Scales: Engagement influences learning in terms of both momentary engagement in a learning task and longer-term engagement with a course or topic [90–92]. Furthermore, prolonged periods of focused attention may occur in a classroom setting during a lecture [30, 93], while it may occur in short bursts when learning in between other responsibilities or distractions within a home [94]. In spite of the complexity of these multi-temporal aspects, students who are engaged for prolonged periods should be better able to learn and retain information compared to students who are disengaged or engaged for shorter and less frequent durations [95, 96].

Since learning requires sustained effort over a period of time, periodic disengagement (e.g., for relaxation, taking breaks) might actually benefit learning outcomes to the extent that it enables learners to reengage in learning content over

time [97]. However, prolonged periods of disengagement are associated with negative learning outcomes, such as a reduced interest in educational activities [98], lower self-efficacy [99], an increase in risk-taking behaviors [100], lower levels of educational achievement [74, 101], or an increased risk of course absenteeism or dropping out from school entirely [102].

4) Summary: In short, engagement in learning is a multi-temporal and multi-componential construct that is positively associated with long-term learning outcomes in a variety of formal (e.g., classroom) and informal (e.g., YouTube learning videos) contexts. Though prolonged periods of disengagement are associated with negative learning outcomes, periodic disengagement may help learners "recharge" in between learning sessions and can benefit long-term learning outcomes. Careful measurement of learner engagement during these focused periods coupled with strategies to promote focused engagement (i.e., to combat fatigue or boredom) are necessary to help improve the effectiveness of these focused sessions and long-term learning outcomes. The next sections focus on these two major issues.

D. Takeaways

Key takeaways from this section on the conceptualization of engagement are:

1) Engagement can be operationalized as a function of component(s), task context, and timescale.
2) Engagement is multicomponential, encompassing affective states, cognitive states, and behaviors.
3) Engagement unfolds as a function of task context and time, where a *person-oriented* perspective focuses on momentary engagement patterns over narrow contexts, a *context-oriented* perspective focuses on the influence of the context on engagement patterns over extended time frames, and a *person-in-context* perspective lies somewhere between these two.

III. Measuring Engagement

Given the multitude of ways in which learners exhibit behaviors, express emotions, and cognitively attend to learning tasks, several methods for measuring engagement have been developed. Since engagement is a latent construct and cannot be directly observed, traditional measures rely on self-reports, observation/annotation, or proxies for inferring engagement, which we will delve into next. Automated approaches then leverage these "ground truth" measures along with sensor data to derive computer estimates of engagement. In both cases, there are several challenges involved in obtaining valid, reliable, and unbiased measures in different learning contexts.

A. Traditional (Manual) Approaches

The traditional approaches to measuring learner engagement can be broadly categorized along two dimensions. The first dimension concerns when the engagement measurement is made, which can either occur in tandem with learning activities (i.e., a *momentary* assessment) or *retrospectively* after the activity. The second dimension concerns the perspective used to make the assessment, which can either come from the learner (i.e., *self-reported* engagement) or from the perspective of an *observer*. We consider the merits and drawbacks of these approaches below.

1) Retrospective Self-report Measures: These measures are among the most commonly used to capture learner engagement in classroom contexts (e.g., [92, 103, 104]). Typically, these are operationalized as Likert-type or yes/no questionnaires where learners reflect on recent learning experiences and respond to a collection of questions. For example, learners may be prompted with, "I like learning new things in class" (an item for emotional engagement) or "I try to match what I already know with what I learn in school" (an item capturing cognitive engagement) (e.g., [105]). In other cases, students may be presented with a list of statements to endorse, like "When I am in class, I listen very carefully" (a behavioral engagement item) or "In school, I do just enough to get by" (a reverse-scored behavioral engagement item). Usually, these items are packaged together into questionnaires to increase their validity and reliability and promote reuse (for example, the *Student Engagement in School Questionnaire* [105]). This is an important step for generalizable and reproducible research. Other non-questionnaire approaches for retrospective and self-reported engagement measurement include day reconstruction [106] and interviews [107] (which perhaps blur the line between self-reported and observer-based measurement).

2) Momentary Self-report Measures: Momentary self-reports are useful measurement approaches for sampling mental states at particular moments when those states may have changed. One approach is called experience sampling [108], which is similar to ecological momentary assessments used in affective computing research [109, 110]). Here, learners may be periodically probed to report their levels of engagement during a learning activity. Though these kinds of assessments are widely used in affective computing research, they are less commonly utilized for measuring engagement in classrooms because of the concern that they disrupt the engagement in the process. However, this concern can be alleviated by careful design of the timings and delivery of the probes [108]. In perhaps the largest study on student engagement during digital learning, Hutt et al. [111] used this method to collect tens of thousands of engagement instances from close to 70,000 students across an entire school year.

3) Retrospective Human Observer-based Measures: These types of engagement measures include video coding and observer-based annotation of recordings of student engagement. For example, several recent works (e.g., [26, 29, 30, 112]) utilized human coders to rate the level of engagement of different students in laboratory and classroom settings based on camera recordings (and eventually used these measures to train machine learning models to automatically infer engagement). This approach gives observers access to some types of information available to teachers in a classroom for assessing student engagement in real-time (e.g., visual and audible information).

Some additional approaches utilize observations about the learning outcomes to retrospectively infer whether learners were engaged in activities or learning content. Examples of

these types of measures are common in today's classroom and digital learning environments: homework grades or completion, absences, test scores, behavior records, time spent watching lecture videos, number of discussion forum views and posts, and more [113, 114]. Traditionally, these metrics are evaluated by humans (e.g., teachers, teaching assistants) to gauge learner engagement, but it is questionable whether they actually reflect engagement (e.g., test scores are designed to measure knowledge rather than engagement). Furthermore, absences conflate legitimate reasons for nonattendance (e.g., illness) with disengagement and homework completion measures do not account for home-life factors (e.g., access to reliable internet). These types of measures need to be taken with a grain of salt and used as a last resort for assessing learner engagement, unless they are validated against accurate alternative measures.

4) Momentary Human Observer-based Measures: Momentary observer-based measures can encompass both live human-based observation (e.g., via a teacher or research assistant in the classroom) and also asynchronous indicators of learner behavior at particular moments (e.g., observation of video recordings of students in classrooms [26], coding of log files during digital learning [115]). One widely used observer protocol is the Baker-Rodrigo-Ocumpaugh Monitoring Protocol (BROMP)—a human-observer coding method where students are observed individually for up to 20 seconds to assess both their affective (e.g., neutral, boredom, confusion) and behavioral states (e.g., on-task, off-task, on-task with conversation) [116]. To avoid ordering effects and the influence of distracting behaviors on the observers, BROMP requires that the sequence of students to be observed is determined *a priori* [117].

B. Recent (Automated) Approaches

Recent approaches to measuring engagement have turned to automated and machine-based methods. The main ideas behind using machines rather than human observers or self-reports are to reduce costs, (ostensibly) reduce biases, improve objectivity in assessment, scale up measurement, and have real-time measures for dynamic interventions. The manner in which these automated systems are built depends on the learning context in which they will be deployed. For example, an automated engagement detection system in the classroom might use video and audio data from camera and microphone recordings of students to infer engagement, similar to how a teacher would use the same visual and audible cues (e.g., [118, 119]). Another example is an automated system for digital learning environments where students' interactions with the system (e.g., time spent watching lectures, number of lectures viewed, quizzes answered) and with each other (e.g., forum views and posts) are used to infer engagement (see [120] for a review). Each of these approaches are forms of machine observation that are trained to infer learner engagement levels at different moments in time using information from an array of features (e.g., video-based gaze or face, audio cues, logs) and prior examples of engagement and disengagement from human-based measurement approaches. Thus, human measurements provide the foundation for automated assessments.

Figure 2 provides an overview of how these automated AI engagement inference systems are built. The end result of this process is a trained machine learning model (i.e., the *ML Model* in the Model/Prediction stage) that is able to infer engagement from (machine-based) observations about learners. However, in order to train this model, additional supporting information needs to be supplied, such as examples of learners' behaviors/cognition/emotions in context, human-provided ground-truth ratings/annotations of engagement using any of the aforementioned measures, and various decisions from stakeholders (e.g., researchers) to determine the ML model algorithm. To summarize, first the ML model is trained to detect and predict learner engagement from examples, then it can be used to infer engagement levels for new groups of learners, as noted below:

ML model training:

$$\text{machine-observed features} + \text{human-provided engagement scores/labels} \rightarrow \text{ML model}$$

ML model deployment:

$$\text{machine-observed features} + \text{trained ML model} \rightarrow \text{machine-provided engagement estimates}$$

In Figure 2, each stage (gray dashed box with bold labels) is comprised of *information*, denoted by the yellow wavy boxes, and processes, people, or systems which act upon and transform that information. We refer to this latter grouping as *agents* and depict them using rounded green boxes. Within and between each stage, *information* is produced by *agents* and passed along to other *agents* that transform it into a new type of information. Thus, the pipeline consists of alternating steps (e.g., agent $\rightarrow$ information $\rightarrow$ agent ...). The red diamond denotes a *stakeholder decision* which influences how information flows through the pipeline and is used to decide when an iteration of the pipeline is complete. In Figure 2, only one decision is depicted at the end where stakeholders control whether to continue training the ML model, but in reality, many stakeholder decisions influence the structure and flow of information throughout the entire pipeline (e.g., which algorithm for the ML model should be used, how many learners should be studied and in what context?). For simplicity here, only one decision is illustrated. We describe each stage in turn next and use italicized font when referring to items within the figure.

1) Engagement Continuum Stage: This stage encompasses the learners and the learning context. As proposed by Sinatra et al. [72], there are a number of influences that learners have on their learning context and vice versa, thus there exists a continuum where engagement can be studied at the level of learners, the learning context, or anywhere in between. This stage captures the details of a particular context, such as learning in a traditional classroom or learning by watching video lectures at or away from home.

2) Human-provided Measurement Stage: This next stage encompasses any of the four aforementioned measurement approaches (e.g., momentary observer-based). This stage is

only necessary for ML model training and would not be used once the ML model is deployed, thus all measurements made in this stage involve human-based assessments of engagement (see [121] for a review). Three paths are depicted in the figure, but typically only one or two are utilized in a particular study, but see D'Mello et al. [122] for a combination of measures including self-reports, peer-judgments, and trained judges. The top path represents both types of observer-based measurements where human observers use their senses (e.g., watching and listening to learners behaviors, expressions, and performance) to make a judgement about their perceived level of learner engagement. Sometimes these judgements are purely subjective and sometimes they may involve questionnaires and assessment items, similar to those previously mentioned. Perceived engagement scores from multiple observers are sometimes gathered to help reduce the influence of any observer's biases, then the scores are fused (e.g., by averaging) to produce an *Engagement Score*. The second path involves self-reported engagement where the learners themselves reflect on their experiences of engagement and report them directly (e.g., at seven-minute intervals [123]). Finally, sometimes engagement is induced rather than measured, which the third path represents. In these types of scenarios, the learning context is carefully controlled and presumed to have some known effect on the engagement levels of learners. For example, Siddiqui et al. suggest that engagement can be induced via peer-to-peer synchronous interactions [124] and Hsu et al. have shown that disengagement can be induced using distractions [125]. Engagement scores are often treated as binary measures (e.g., engaged or disengaged) in these induced engagement settings, offering coarse insights into the effects of (dis)engagement.

3) Automated Sensing/Feature Stage: In this stage, information about the learners and learning context are observed and recorded. Since the aim is to build an ML model that can automatically infer engagement without human intervention (though human oversight is needed at all steps, as we will explain later), a machine-based observer collects observations here rather than humans. Thus, these observations can be generated from any signal that can be digitized. A video is one example of a digitized signal from which features such as facial expressions, gestures, body posture, or eye gaze [126] among others could be automatically extracted using computer vision techniques [42]. Another example for learners in remote or digital contexts involves collecting patterns of interactions with the learning management systems (LMS), including video views, video skips, page view, or click streams, which provides a basis for inferring student engagement (e.g., see [111, 127]). This latter example is called a "sensor-free" approach because it uses overt information obtainable from LMS interactions rather than external sensors. Of the 32 studies surveyed (see Table A1), most relied on overt signals (e.g., face, gaze, logs), whether sensor-based or sensor-free, rather than covert (e.g., heart pulse) signals to infer engagement levels (i.e., 84% overt, 13% covert, 3% mixed).

More recently, advances in deep neural network modeling are enabling automatic feature extraction and the transfer of trained automated sensing technologies across domains. For example, Sümer et al. [26] used pre-trained deep networks called Attention-Net and Affect-Net to learn deep embeddings (i.e., features of engagement) based on facial expressions and head pose in a classroom learning context (a person-oriented approach). Transfer learning (i.e., from outside sources to learning contexts) benefits from large amounts of data to learn how to understand different signals (e.g., gaze, facial expressions, vocalized audio) are proving to be powerful methods for automated construct inference in many domains (see [128] for a review of transfer learning), and it seems likely advances in automated learner engagement inference will follow suit.

4) Model/Prediction Stage: Next is the *Model/Prediction Stage* where the *ML Model* is trained using the *Features of Engagement* and the *Engagement Scores* (i.e., supervised machine learning). The training process, represented by the ML model in Figure 2, entails learning a mathematical function which maps the engagement features to the engagement scores as accurately as possible. This process is typically iterative, as many refinements to the learning model may be needed (e.g., hyperparameter tuning) until a sufficiently optimized mapping is found. Once the ML model is trained, it can be used to make predictions about the level of engagement based solely on the engagement features without using the engagement scores.

Cross-validation is often employed in this stage to improve the generalizability and reliability of the ML model for a particular purpose. This technique entails measuring the ML model's performance on different subsets of data not considered when training the models, and the choice of these subsets influences the model's robustness. For example, if the stakeholders want to train the ML model to make predictions about the engagement levels of new learners in a similar context, the ML model will be trained using a subject-independent cross-validation procedure where examples from a given learner appear only in the training or testing set, but never split among the two. Likewise, ML models meant to assess the effectiveness of different courses at eliciting engaged behaviors (i.e., a context-oriented approach) would use a course-independent cross-validation approach. Details of this procedure are discussed further in Section III-C.

5) Decision/Evaluation Stage: In this final stage, the predicted levels of engagement from the ML model are evaluated to determine whether the model performs well enough for its intended purpose. In most research, this determination is made based on accuracy (i.e., the similarity between the *Engagement Predictions* and the ground truth *Engagement Scores*), reliability (i.e., similarity of *Engagement Predictions* for similar *Features of Engagement*), and generalizability (i.e., how accurately the model predicts engagement for different learners and contexts). Statistical tests are sometimes utilized to assess how well the model's predictions perform compared to suitable baselines (e.g., simple heuristics, educated guessing) both in terms of strength and statistical significance. If the models under-perform compared to expectations, stakeholders may choose to alter portions of the pipeline (e.g., adding or removing engagement features, choosing a different ML model algorithm) and try again. The final output of this pipeline is a trained ML model suitable for automatically predicting the engagement levels of a new set of learners in similar learning

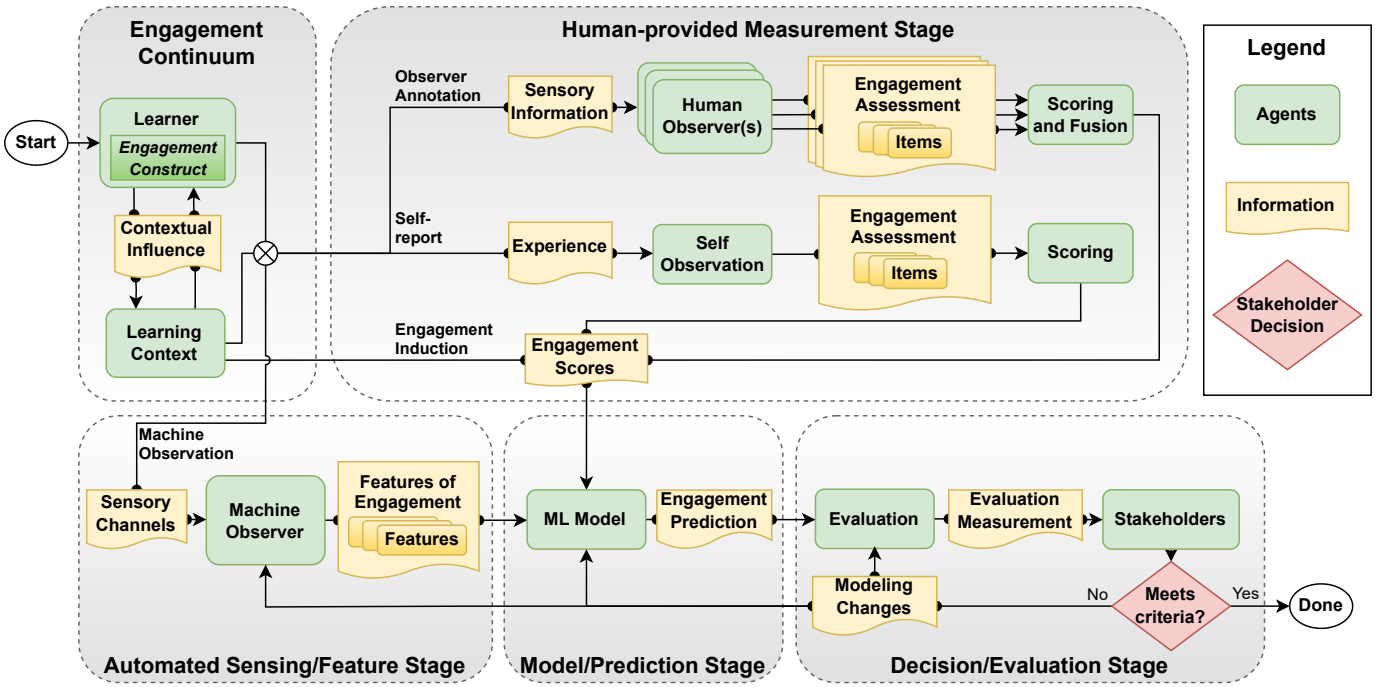

Fig. 2. A pipeline demonstrating how machine learning models for automated engagement assessment are trained. ML = machine learning

contexts using only machine-based observations.

This pipeline for training the ML model is a general-purpose pipeline for human-centered computing tasks where constructs beyond engagement are of interest (e.g., positive or negative affect [129], perceived sleep quality [130], suitability of job candidates for hiring [51]). A person-oriented version of this pipeline tailored for automated learner engagement assessment in digital learning contexts was proposed by D'Mello, Dieterle, and Duckworth [131] called the Advanced, Analytic, Automated (AAA) approach. Many of the studies surveyed (see Table A1) include an automated engagement inference system trained in this fashion. In particular, while 25% of the 32 studies used a more traditional form of statistical inference, the other 75% used machine learning methods (42% classic ML [e.g., k-nearest neighbors, support vector machines], 33% modern ML [e.g., deep and reinforcement learning]). Though research into automated engagement inference is still emerging, these and other studies (e.g., [118, 132–134]) demonstrate the power of this general supervisory approach for developing automated systems for engagement prediction.

C. Challenges in Measurement

Though the specific implementations for each engagement measurement approach aim to be as accurate and reproducible as possible, there are several challenges involved with both human-based and automated machine-based measures of engagement whose predictive accuracies depend on the human-based measures (used as supervisory signals). Here, we discuss some of the challenges to engagement measurement.

1) Validity of Ground Truth Assessments: According to the *Standards for Educational and Psychological Testing* [135], the validity of a measurement refers to "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests." Using the three components of engagement [69] as a guide, this means that a valid measure needs to be accurate and encompass all facets of engagement.

Figure 3 illustrates how different perspectives (self, human observer, machine observer) have varying access to information signals (e.g., introspection, gaze, heart pulse) for determining engagement within each component. Thus, each perspective is restricted to producing engagement measures based on the available channels of information, some of which serve as unreliable proxies for estimating engagement. Proxy measures such as grades, test results, visual cues, and attendance are often limited in what they can reveal about affective and cognitive engagement (e.g., neutral facial expressions are limited proxies for focus, gaze is a limited proxy for visual attention), and some behavioral proxies are non-specific; for example, absences can reflect health conditions and a difficult home life rather than lack of engagement, whereas showing up each day might reflect compliance rather than a genuine desire to learn. Thus, any measure of engagement should strive to incorporate a multitude of perspectives in order to maximize validity.

However the validity within each perspective can be diminished by various forms of bias. Self-report accuracy may be influenced by social desirability bias [136], memory recall limitations [137], cultural contexts [138], and cognitive and recall biases [139]. Though human-based observational measures can mitigate some of the biases from self-reports (e.g., [70, 140–143]), these measures are also subject to the influence of prior experiences, implicit biases, spatial attention [144], and individual differences between learners and observers (see [145] for a catalogued review). Biases are also introduced by the timing of the measure. For example, whereas

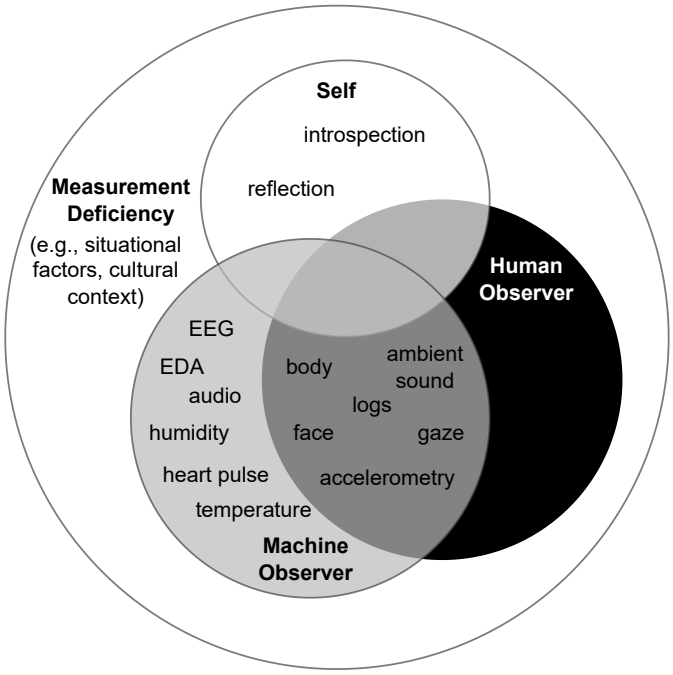

Fig. 3. A Venn diagram illustrating three different perspectives for measuring engagement. Examples of data modalities (i.e., signals) are contained within each circle showing which types of information can be incorporated by each perspective. A complete view of engagement should include a multitude of these perspectives and also momentary and retrospective assessments (not pictured). Biases in the form of contamination or deficiency can arise when irrelevant sources of information (e.g., situational factors, cultural context) are utilized when making an assessment of learner engagement.

the asynchronous nature of retrospective observation allows human observers to more carefully reflect on the behaviors and affect of students, contextual factors not captured in a video cannot be considered. Retrospective self-report measures avoid interrupting the learning activity but suffer from limits of memory reconstruction. As a result of these biases and differences in perspectives, self- and observer- ratings of engagement and affect in general tend to be very weakly correlated [131, 146–151], suggesting each is sensitive to different sources of information. Attempts to mitigate these differences, such as frame-of-reference training that increases observer–observer agreement, does not seem to improve self–observer agreement [152].

Overcoming these challenges is difficult, and at present there is no single best approach. Generally, biases in one measurement procedure can be mitigated by collecting several measures independently (e.g., multiple observers, multiple self-reports), and research should aim to incorporate multiple measures where possible that accurately comprise cognitive, affective, and behavioral engagement. Given that both self- and observer- reports have biases and are privy to different sources of information, perhaps the most defensible approach is to consider a combination of the two as in D'Mello et al. [122]. Thus, a major weakness in current research is that studies typically do not capture or account for these multiple perspectives.

2) Scalability of Assessments: To maximize the validity of engagement measures, multiple approaches to engagement measurement are desirable, but this increases costs and impacts the scalability of research. Human-based observer measures in particular entail considerable human effort, which makes it difficult to replicate studies in similar contexts and across cultures to test the generalizability of findings at scale. These limitations can be partially addressed when using machine-based observation to supplement or replace some human-based observation. For instance audio and video recording systems, such as the Electronically Activated Recorder (sampling audio clips in naturalistic settings) [153] or cameras in a classroom [26], can capture indicators of student engagement passively and easily at scale. These recordings, however, need to be transcribed and annotated by human observers (e.g., [147, 154]) to obtain an engagement measure, which still incurs considerable costs and hinders scalability. This is where automated ML models can help by generating engagement measures from the machine observations without any need for human observers beyond the model training process. This approach saves considerable time, effort, costs, and easily scales to studies involving large populations. There have been several studies utilizing automated AI in this way (e.g., [118, 132–134]), but research along these lines is still relatively new and more work is needed before the ML model's engagement assessment accuracy achieves parity with human-based observation [91, 120].

3) Generalizability of Human-based Assessments: Self-reported measures of engagement are relatively inexpensive and easy to administer, but in addition to the previously mentioned biases, they may not generalize across learners and cultures [136, 148]. For example, on a 5-point questionnaire item asking learners to rate how hard they study each day, one student in a competitive learning environment who studies for at least an hour each day may rate their effort as 4/5 while another student in a more relaxed environment who reviews flash cards for five minutes each day may do the same. Since the learning outcomes and amount of time spent engaged is likely to differ between learners in this instance, the validity of interpretations of these self-reported questionnaires is reduced when comparing them across contexts [148]. This basic argument applies to learners in different contexts (e.g., formal classroom learning with notes vs. informal digital learning with a learning management system) or cultures since the study habits and perceptions of successful studying vary by context. Self-reported measures are more valid when compared over time within a learner since these differences in reference frames are no longer problematic.

4) Validity of ML Models: There is the major issue of how to evaluate the expected validity of automated ML models for engagement prediction on new samples of learner data. Often, the model's accuracy is used (a form of convergent validity), measured as the alignment between automated estimates and an external standard (typically self- or observer- annotations) and quantified, for instance, using recognition rate, kappa, or correlations. Measuring the suitability of an ML model for a purpose in terms of accuracy ultimately requires stakeholders to make a subjective assessment (see the *Stakeholder Decision* in Figure 2). Although it is difficult to specify exact bounds on what constitutes "good" accuracy (as discussed in detail later

on), at a minimum it should exceed random guessing (chance).

As automated machine-based engagement measures have only been the focus of concentrated research effort in recent years, their validity has yet to be thoroughly examined. Efforts to build the best ML model possible (see Figure 2) have focused almost entirely on optimizing accuracy, and thus convergent validity. This provides one form of evidence of the validity (accuracy) of ML-based measures, but as *Standards for Educational and Psychological Testing* [135] clearly states, multiple types of evidence of validity are needed to help establish the suitability of a measure for a particular purpose. To date, there is little to no evidence of the discriminant validity (i.e., ML predictions are uncorrelated with unrelated constructs), predictive validity (i.e., the success at predicting future states of engagement), or external validity (i.e., generalizability). For instance, none of the reactively designed systems we describe later in Section IV-B provide evidence of these additional types of validity. Only a small handful of studies (e.g., [91, 155–157]) link the automated engagement predictions to meaningful outcomes, such as learning gains and college enrollment (i.e., evidence of predictive validity).

5) Generalizability of ML Models: Only another handful of studies (e.g., [91, 158, 159]) consider generalizability beyond predictive accuracy for new learners by measuring the generalizability over time and demographics. The generalizability of a measure is concerned with its validity when applied to data beyond what was used to develop it. In the context of machine learning, this refers to how well the ML model's engagement predictions perform on unseen data from a different set of learners. Generalizability is usually operationalized by dividing the data into two mutually exclusive sets, training the ML model on one set (i.e, the "training data"), and testing its performance on the remaining data (i.e., the "test set"). Cross-validation is a widely employed variant of this procedure where the ML model's performance on the test set is measured over different partitions or test sets within the data. In this type of procedure, each data sample is used as testing data only once and never simultaneously included in the training data, which would be a form of "cheating" since a flexible ML model would be able to recall the exact engagement score when making a prediction. Thus, it is important to ensure that the samples contained within each test set reflect the deployment goals for the ML model. For example, if the model aims to be used to predict engagement levels of previously unseen individuals, then the training and test sets should be constructed so all available samples from one individual are contained entirely in either the test or training sets, but not both (i.e. subject-independent cross-validation folds). This ensures the training data will never contain specific information about a learner in the test data, and hence, measures of the test-set performance will be better indicators of the anticipated performance on new learners. Usually, the average performance across the test sets (e.g., the mean correlation) provides a measure of the expected performance of an automated machine-based engagement measure on new data. Sometimes, the variance is also reported (e.g., the standard deviation of the correlations), which provides some information about the precision of the ML model's engagement predictions.

Models designed to be deployed in heterogeneous settings or in environments where data noise may vary from the training data need to generalize to a range of noise conditions. This is especially true for models trained on sufficiently clean or denoised data (e.g., vocalized audio in a controlled and quiet classroom) which intend to make predictions in naturalistic setting (e.g., groups of students chatting, talking over one another, and making noise while working together) (e.g., [160]). Though modern machine learning techniques have the ability to separate out noise from the relevant data if provided enough samples (e.g., [161, 162]), most research discards noisy samples prior to modeling (e.g., [155]). Systematic modeling of noise processes and how they affect data can be incorporated during the modeling process and improve measurement accuracy [163]. Nonetheless, the ability to handle noisy data and thus generalize to similar contexts with varied noise conditions needs to be a fundamental design constraint rather than an afterthought.

Furthermore, efforts to improve ML model generalizability are only meaningful when models are deployed to predict learner engagement in similar learning contexts. As noted by Sinatra et al. [72], there is a continuum of influence between learners and the learning context, so studies aiming to assess the same type of learner engagement (e.g., via within-lecture quizzes) in different learning contexts (e.g., for remote learners in a controlled laboratory setting vs. in-person learning in a naturalistic classroom setting) should expect different results. For example, studying a construct in its natural context is generally more difficult than studying it in a controlled laboratory setting (e.g., [164]), and research is starting to highlight the gap in performance when an ML model trained in one context (e.g., lab studies on remote learner engagement) is used to make predictions in another context (e.g., classroom engagement) [165–167]. Thus, scientists and practitioners should expect that generalizability measures are only applicable when both the populations and contexts are very similar, and they should take caution when using automated ML systems outside of their intended contexts. Accordingly, the basic learner-independent cross-validation method can be expanded in scope to incorporate groups of learners with particular characteristics [168], temporal changes [169], domain differences [168], amongst others. Even when models fail to generalize, these analyses provide valuable data for further refinement.

6) Robustness of ML Models: We consider the robustness of an ML model as a function of its reliability and handling of missing data. The *Standards for Educational and Psychological Testing* [135] defines the general reliability of a measurement as the "consistency of scores across replications of a testing procedure, regardless of how this consistency is estimated or reported." For ML models inferring engagement, this refers to the precision and consistency of accuracy/errors in engagement predictions for unseen data (i.e., future samples) in the same learning context. Measures of reliability are often obtained during the ML model training process using cross-validation, where out-of-training-sample accuracies provide a measure of a model's reliability across replicated testing procedures. These measures are operationalized using different metrics, such as the standard deviation in prediction

accuracy or test-retest reliability metrics [170]. Reliability is an important measure because it provides insight into the expected performance of the ML model on unseen data in the same learning context, however in practice the performance is further modulated by the degree to which the context surrounding the unseen data is similar to the one used to train the ML model.

Additionally, the robustness of an ML model is affected by its ability to handle missing data. When data is gathered in naturalistic settings (e.g., in the classroom, at home during remote learning), both user- and sensor-based issues often impact the availability and quality of data. For example, learners may forget to turn on or wear sensing devices (e.g., wristband heart rate sensors, cameras), forget to clean them properly, forget to recharge the batteries between uses, or sensors may simply malfunction. Thus, sparse samples or completely missing data are unavoidable. Under ideal conditions, an ML model would only be asked to provide predictions of learner engagement when it has observed a sufficient amount of data to be confident, but in practice, the unpredictable nature of missing data means this is not always possible. In this scenario, a robust ML model should be capable of making a best-guess prediction based on the available information (e.g., [163]), and it should also report low confidence in its assessment(s). In cases where multiple sensors are available capturing an array of multi-modal information, ML models can use information from one channel to compensate for the lack of information from other channels. For examples, Bosch et al. found that multimodal fusion techniques were able to compensate for missing facial data (due to face detector failings of motion, occlusion, poor lighting, etc.) to achieve around 98% coverage when combined with educational game interaction data [171].

7) Privacy, Ethics, and Bias/Fairness: Most of the data used to assess learner engagement, whether based on in-person observation or retrospectively, is sensitive to privacy concerns. Information about grades, test scores, behavioral performance, and more can be legally protected (e.g., by the Family Educational Rights and Privacy Act in the United States), so researchers needs to take extra precautions to ensure that adequate permissions are obtained from learners and that protective measures are in place (e.g., anonymizing data, binning performance scores) to prevent a breach of privacy. The 2010 "WebcamGate" scandal [172], where thousands of compromising video camera and desktop background images were captured from students' computers without their knowledge or permission, illustrates the massive potential for these technologies to cause harm. One effective strategy for ensuring the privacy of engagement features is to obtain and record non-identifiable features from the signals and then immediately discard the signals themselves. This is the approach taken by certain machine-based observation tools, such as the TILES (Tracking IndividuaL pErformance using Sensors) Audio Recorder [173] that randomly samples the environment listening for vocalized audio clips and transforms and records anonymized versions of them (e.g., prosodic information rather than the words uttered). Bosch et al. [112] demonstrate a similar approach to anonymizing facial expressions in classrooms.

Even with these protective measures in place, bias/fairness concerns apply to the *Features of Engagement* (see Figure 2) as well, since these may contain additional information beyond engagement (e.g., about race from skin tone or gender from vocal pitch). Not only does this threaten individual rights to privacy (e.g., via re-identification), it also can manifest as a type of *measurement bias* where some aspect(s) of the input features which are irrelevant to the construct (e.g., race) is treated as relevant (i.e., a contamination of the relevance; see [51, 174] for a full discussion of contamination and deficiency biases). Automated engagement measurement systems can help to prevent these bias/fairness concerns by discarding all portions of the captured signal irrelevant to engagement assessment and keeping only non-identifiable versions of the relevant information.

Furthermore, the manner in which automated engagement tools are trained, evaluated, and used also presents major ethical and fairness concerns. Recent research has uncovered a plethora of examples of sensing technologies that collect high-quality and more representative features for certain groups of people than others. To give a few examples, facial feature recognition software (a foundation for emotional expression recognition) captures Black and female faces less well than lighter colored or male faces [175]; vocalized audio transcription accuracy may be diminished for non-native language speakers due to articulatory differences from native speech [176], longer pause durations [177], or non-normative pause locations [178]; and measures for engagement in digital learning may not account for differences in eye gaze or interaction patterns for learners with attention-deficit disorders [179, 180]. Because these technologies may perform less well for certain people, researchers and practitioners must pay extra attention to potential disparities in the resulting trained ML model's performance across groups, especially groups protected by legal statutes (e.g., race, ethnicity, sex, gender).

In addition to outlining the ML model training process, Figure 2 provides a theoretical framework for systematic investigation into differences in ML model accuracy across protected groups (i.e., bias and fairness concerns). As Booth et al. [51] discuss, versions of this figure tailored to an application domain (learner engagement in this article) can serve as a guide for identifying potential sources of bias leading to differences in accuracy across groups. Each piece of information (yellow wavy boxes) can in principle be inspected for evidence of unnecessary group differences. If these differences exist, for example if the *Evaluation Measurement* indicates engagement detection accuracy is better for one group vs. another when no such differences are to be expected, then the potential sources of bias causing this disparity occur anywhere upstream (i.e., at previous stages). The goal during this investigative process is to identify the agent(s) (i.e., an information "transformer") that causes this developmental difference to appear. For example, possible sources may include: differences in the presence of *Features of Engagement* for different groups, differences in how self-reported measures are interpreted among learners from different groups, or differences in how human observers notice and assess engagement across groups. Once possible sources have been identified, additional steps can be taken to mitigate the influence of these biases, such as in-learning ML

model debiasing strategies [181]. There is yet no prescriptive method for performing this search, but it is important to reduce the potential sources of bias as much as possible when designing and building these automated ML systems for engagement assessment.

Lastly, the contexts of use of trained ML models for engagement prediction and the decisions made by stakeholders presents other ethical concerns. Many ML models are not yet "self-aware" to the extent that they can recognize when the features used to make engagement predictions are coming from a different context. An ML model may thus produce engagement scores to the best of its ability without making its confidence in assessment or confusions known to stakeholders. Noting that this situation is likely a failing of the stakeholders to recognize that the ML model should not be used in this setting, any decisions made by the stakeholders may result in ethical concerns. For example, using a face-based engagement prediction system to measure learner engagement in classrooms consisting of a majority of female and Black faces may fail to adequately capture student engagement due to well-known deficiencies in current facial recognition techniques [175].

Even if all of these sources of biases were mitigated, there is still the question of whether a model should be used for a particular purpose. Specifically, there is a massive concern of these automated models being used to surveil students and for purposes of disciplining and evaluating them. Using models of student engagement to evaluate teachers is similarly alarming and distressing. Thus, even when an ML model seems to function in a particular context, stakeholders' decisions to utilize it without considering its fitness-for-purpose can result in ethical concerns. Therefore, we recommend that these models be used for research purposes, formative feedback (i.e., feedback for improvement not evaluation), or dynamic intervention, and ideally in low-stakes settings. Users should have agency over the measures including the ability to turn them off.

D. Takeaways

The key takeaways pertaining to engagement measurement covered in this section are:

1) Measures of engagement should utilize multiple approaches to measuring engagement because each accesses different sources of information (see Figure 3).
2) Low-levels of agreement between engagement scores from different perspectives should be expected due to differences in information and biases that uniquely affect measurements from each perspective. However, these unique scores provide a more nuanced view of the different ways in which affective, cognitive, and behavioral indicators of engagement manifest.
3) In addition to accuracy, researchers should analyze generalizability, bias/fairness, and robustness when evaluating automated measures of engagement (see Figure 2)
4) The use-for-purpose of automated measures of engagement should be scrutinized for ethical concerns. ML models should only be deployed to measure learner engagement in contexts very similar to how they were trained and never used for evaluation purposes or in high-stakes scenarios.

IV. Enhancing Engagement

Early learning technologies in the 1980s focused primarily on optimizing knowledge and skill acquisition (e.g., [182–184]), in line with learning theories at that time emphasizing knowledge as the predominant learning outcome. This perspective has shifted over the past four decades as newer learning theories have come to realize the role of engagement in deep conceptual learning. Consequently, we focus on the promotion of learner engagement in the context of learning technologies. Methods to enhance engagement via curriculum design, presentation style, and teacher intervention have been a major focus of good pedagogical practice for many decades [185, 186], and more recently in the context of learning technologies [50].

Deep conceptual learning is difficult because it requires sustained effort, rehearsal, practice, and struggle [187]. Short-term distractions and gratifications providing affective rewards (e.g., social connection, "fun") may need to be temporarily deferred [188]. A commitment to genuine and persistent focus on learning needs to be established and become routine to sustain long-term interest [40]. Even when these factors are accounted for, mental lapses in attention are normal occurrences (e.g., students experience "zone outs" around 30% of the time while learning from technology; see [189, 190] for a review) and might need to be regulated to optimize learning and avoid diminished learning outcomes [191, 192].

There is also a fundamental tension between liking and learning. Whereas, "edutainment" games can be highly engaging, it is not clear if they encourage deep comprehension [193, 194]. On the other hand, intelligent tutoring systems (ITSs) designed to mimic one-on-one human tutoring effectively promote deeper learning (e.g., [195–197]), but they also result in increased levels of boredom [150, 198, 199], which may hinder the development of sustained engagement in the long-term.

How do we design learning technologies and contexts that facilitate situational engagement and promote sustained engagement to improve learning outcomes? This question has received concentrated research effort over the past 30 years [200], and researchers have identified two main strategies: proactive and reactive design. *Proactive* design is more of a top-down approach focusing on optimizing the learning context and materials to facilitate engagement (i.e., person-in-context). *Reactive* design is more of a bottom-up approach that monitors and encourages learner engagement either in real-time (intervening with learner engagement) or over a short time-scale (e.g., hours or days) by offering feedback to learners and instructors. Both of these strategies can be utilized in tandem, and Figure 4 provides some examples of both proactive and reactive design for improving learner engagement across the three engagement components (affective, cognitive, behavioral).

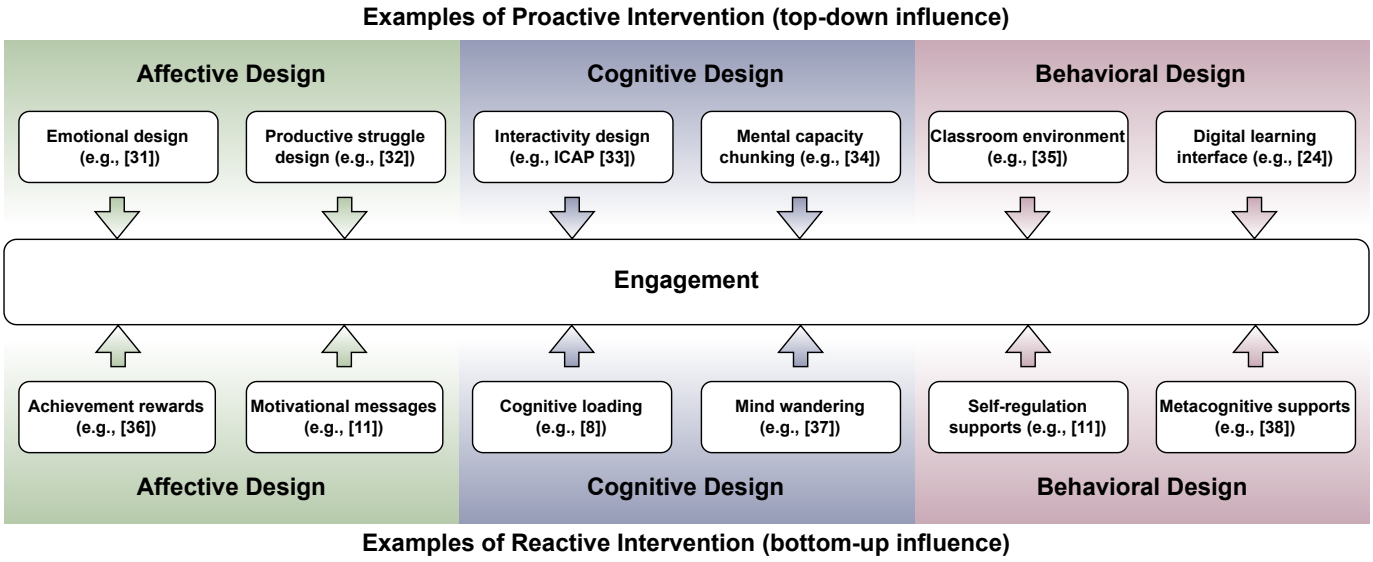

Fig. 4. Examples of proactive (top-down) and reactive (bottom-up) design for improving learner engagement across the affective, cognitive, and behavioral engagement components.

A. Proactive Design

Proactively designed learning experiences are carefully crafted to enhance engagement and successful learning outcomes. This approach is mainly person-in-context focused (see Figure 1) in that the design aims to optimize the likelihood of promoting engagement across learners or learner groups in a particular learning context. These experiences can take on many forms, such as carefully crafted lectures with several interesting asides to break up content, interactive digital technologies where learners are given the freedom to explore the learning content according to their interests [201], well-designed curricula [202], or personalization of math problems based on students' interests [203]. The main aim of the proactive approach is to craft experiences that produce cognitive and affective states associated with engagement (e.g., interest, curiosity, challenge, critical thinking and reflection, surprise, productive struggle), while minimizing events and mental states that reduce engagement and trigger boredom and mind wandering.

Within the digital learning space, early attempts to proactively optimize content for engagement attempted to incorporate elements of games, puzzles, and comics (e.g., point systems, badges, achievements, leaderboards, and more [204, 205]) to increase motivation and engagement. The results were generally unfavorable—students appreciate the ease-of-use that tends to accompany the "gamification" of educational content, but they do not favor the game-enhanced experience [206]. This phenomenon has been described as "chocolate-covered broccoli" [207] and has been a barrier to the uptake of entertainment elements within education. Since then, recent approaches to proactive design have targeted learner affect, cognition, and behaviors (following the three components of engagement; see Figure 4) more directly.

1) Affective Design: Proactive affective design broadly aims to enhance engagement by appealing to affective elements. One approach called *emotional design* is simple: alter the content so the learning materials induce mild positive affect and delight. Some specific implementations of this method include adding anthropomorphic facial features to non-human graphical elements and adding colors to embellish drab imagery [208, 209]. A recent meta-analysis [31] found that emotional design was effective both in increasing learning and improving learner engagement as measured by intrinsic motivation, liking/enjoyment, positive affect, and reductions in perceptions of difficulty. Conversely, another approach named *productive struggle* aims to control the emotional arc when a learner first encounters cognitive disequilibrium—a state of confusion when confronting content that does not match expectations [32]. At the moment when learners would be expected to experience this disequilibrium, additional supports can appear, encouraging the learner to persist in achieving understanding (e.g., motivational language, alternative perspectives) until a productive resolution is reached, resulting in conceptual change [68, 210]

2) Cognitive Design: Proactive cognitive design aims to promote the onset and maintenance of cognitive engagement, for instance through the content design, content ordering, and interactivity within a learning session. In one example, Chi and Wylie [33] organize learner engagement during activities demanding different amounts of attention and interaction in their Interactive-Constructive-Active-Passive (ICAP) framework. The framework is named after four tiers of types of activities with decreasing likelihoods of promoting engagement ($I > C > A > P$). The research suggests that *passive* activities, such as simply watching a lecture video, are least likely to generate engagement. *Active* verbatim note-taking in class produces slightly more engagement, while a *constructive* version where notes are summarized via self-explanations would be much more engaging. *Interactive* tasks, like debating or discussing learning content with peers, is the most likely

to elicit high levels of engagement. Thus, the ICAP approach encourages the incorporation of interactive and constructive tasks in the design of learning experiences to promote high levels of cognitive engagement.

Another approach called *chunking* seeks to improve the ability of learners to store and retain information [211] while also helping teachers manage the anticipated cognitive load of learning content. The main theme in different styles of chunking is to break down complex ideas into smaller portions that are easier for learners to hold in short-term memory, and the general rule of thumb (from Miller [211]) is that about seven "bits" of information is right for most people (e.g., see [34] for an approach using chunking to teach communication skills). While Fiden [212] has observed that microlearning (using bite-sized chunks) reduces cognitive loading on learners in a flipped classroom context (and also improves behavioral and affective engagement), Gao and Kuang [213] have observed that increasing cognitive load improves cognitive engagement in an educational art design setting. Thus, proactive designs using chunking to promote cognitive engagement must balance the chunk size of content to ensure adequate cognitive loading (larger chunks) while using chunks which are small enough to help learners store and retain information.

3) Behavioral Design: Proactive behavioral design attempts to influence context in which learners engage in activities to improve behavioral engagement for instance by reducing environmental distractions or minimizing the amount of effort required to participate in learning activities. For example, Cheryan et al [35] highlight how inadequate environmental conditions (i.e., lighting, noise, air quality, heating) within a classroom can impact behavioral engagement and lead to significantly lower student achievement. Thus, one effective proactive strategy for improving behavioral engagement is ensuring the environmental conditions are conducive to learning. Another example for distance learning students regards how behavioral engagement is impacted by the effectiveness of the digital interface. For instance, Seo et al. [24] studied remote learners who watched lecture videos week-to-week and observed that students spent more time selectively searching for specific content within videos to prepare for exams. The authors suggest that a proactive design where the video player would adapt to students each week as needed, to help them easily locate important information, may improve engagement, perhaps by reducing the tedious task of seeking out information.

4) Multi-component Design: Highly successful proactive interventions will incorporate all three components of engagement in design. More modern attempts at well-designed educational games, for example, that carefully incorporate game-design principles of problem solving, adaptive challenges, and ongoing feedback can trigger and sustain interest and motivation, in turn supporting engagement and learning [214–217]. Good educational game design in this context addresses affective design (e.g., well-timed and meaningful rewards), cognitive design (e.g., activity structure, a balance of challenging content), and behavioral design (e.g., easy-to-use interface). This is very different from designing games simply for entertainment and "fun", which may enhance engagement but not necessarily learning.

B. Reactive Design

The following excerpt borrowed from D'Mello [50] illustrates the potential of reactive designs to enhance engagement in learning:

> "Imagine you are helping your niece prepare for an upcoming examination in evolutionary biology. Things started off quite well, but after a while, you realize that her mind is a million miles away. Although the plan is for the two of you to collaboratively model genetic frequency shifts in populations, you notice that her attention has drifted to unrelated thoughts of lunch, the football game, or an upcoming vacation. You might try to momentarily reorient her attention by asking a probing question. However, if her attentional focus continues to wane, you realise that you must adapt your instruction to better engage her by altering the course of the learning session. You shift the initiative from a collaborative discussion to a student-centered perspective by asking her to develop a strategy for tracking genetic changes in populations. This works and she appears to tackle this task with a renewed gusto and the session progresses quite smoothly. However, sometime later, you notice that she actually appears to be nodding off as you delve into the fundamentals of allele frequencies. So, you suggest switching topics or even taking a break, thereby giving her an opportunity to recharge."

In this example reactive approach, the niece's attentional states are being monitored and responded to in the moment as needed to maintain engagement. This type of momentary intervention nudges learners towards an engaged state when their engagement seems to decline. Reactive approaches influence learners in a person-oriented fashion (see Figure 1) in order to promote affective, cognitive, and behavioral states conducive to successful learning outcomes. The main benefit of this type of design is that it embraces the notion that engagement varies over time as a result of interactions between competing mental and somatic demands (e.g., fatigue, hunger, stress) that result in mind wandering, inattention, and distractions [218]. It also demonstrates that guided and subtle alterations to the learning content, the ordering of the content, and just-in-time motivational feedback can help turn an otherwise mundane learning experience into an engaging one.

Reactive approaches require more awareness and contextual understanding than proactive ones. Dynamic adaptation to the ebb and flow of a learner's engagement requires the ability to measure it and to understand how to intervene to nudge it in the right direction. The measurement needs of such a system have been described in the previous section (e.g., see Figure 2), but the implementation of the intervention mechanism is open-ended as it entails selecting an action among many possibilities. If a learner is engaged, should the intervention mechanism do nothing or provide some motivational reward? If a learner is confused, should it wait while the learner

struggles, provide some supportive message to encourage productive struggling (e.g., [32, 219, 220]), or provide a hint or just-in-time explanation (e.g., [221])? Just as video games must adapt to increases in player skill to maintain engagement, learning experiences must also adapt challenge to abilities or boredom may emerge if learners are underwhelmed or overwhelmed [222]. The best courses of action are not yet well understood, largely due to the variability among individual behaviors and preferences for how and when to engage with learning content.

There have been recent reactive design efforts to optimize learner engagement (see *Reactive Interventions* in Figure 4) following the three components of engagement. We look at some examples of these approaches and delve further into specific reactive intervention systems utilizing affective, cognitive, and behavioral designs.

1) Affective Design: Rewards for demonstrating a learned skill (i.e., achievement rewards [36]) and motivational messaging to help struggling learners [11] are two examples of reactive techniques for improving affective engagement. These examples provide feedback to elicit desired emotional responses while other systems, such as the following *iTalk2Learn* example, vary how feedback is furnished based on learners' affective states pertaining to engagement.

iTalk2Learn: Grawemeyer et al. [11] detected affective components of engagement from voice-based features and automatically transcribed text recorded from students' speech in a computer-based learning environment called *iTalk2Learn*. In iTalk2Learn, students ages 8–12 learn about math fractions by interacting with a graphical interface as well as talking through problems out loud. The system combines speech and interaction log file data in a model to predict affective components of engagement including flow, confusion, frustration, and boredom. Then, a Bayesian network predicts what kind of feedback should be given to students to promote engagement (e.g., encouragement, additional task instructions), and a second Bayesian network predicts how best to display that feedback (i.e., either subtly or more forcefully in a way that will interrupt the learner's activities). The final feedback prediction system produced feedback that aligned well with experts' decisions about feedback (10-fold cross-validated Cohen's κ = .50), and was thus incorporated into iTalk2Learn.

iTalk2Learn researchers evaluated the automatic feedback system in a randomized controlled trial with 77 students in two conditions: one condition using the automatic feedback system, and an active control condition in which the system generated feedback without incorporating affective engagement detection. The engagement detection condition resulted in significantly lower rates of boredom and off-task behavior (both reduced by 50% in the engagement detection condition) as assessed by third-party observers, as well as suggestive (but not significant) evidence of greater learning. In sum, the iTalk2Learn project illustrated the feasibility of increasing real-time engagement based on affective analysis of multimodal speech and interaction data and demonstrated some of the expected resulting benefits for students.

2) Cognitive Design: Examples of reactive designs for cognitive engagement may aim to vary the pacing of content or reengage distracted learners. For instance, a study from Eldenfria and Al-Samarraie [8] aimed to regulate the presentation of learning content (i.e., cognitive loading) based on real-time measures of learner aptitude, while other research has shown that mind wandering can be sensed and used to trigger digital learning interventions to reengage students [37], such as the *Eye-Mind Reader* study.

Eye-Mind Reader: Research shows that adapting interfaces based on cognitive aspects of engagement can benefit learners. In one example, Mills et al. [37] trained a support vector machine to predict instances of *mind wandering*, or "zoning out", from features of learners' eye-gaze patterns including gaze fixation durations, pupil diameters, and other measures. Learners read an instructional scientific text and self-reported when they were mind wandering—that is, a form of cognitive disengagement that occurred when they found themselves thinking about something other than the task at hand. The machine learning approach yielded weighted precision and recall of .722 and .674, respectively. The researchers then incorporated this machine learning model into an adaptive version of the text reading interface, called *Eye-Mind Reader*, which triggered interventions to improve reading comprehension in situations where the model detected mind wandering. When an intervention was triggered, the student would write a short self-explanation of what they had just read in response to a prompt. Their response was then automatically graded via natural language processing. If the summary appeared inaccurate, students would then be prompted to re-read the last few pages and generate a revised summary.

To evaluate Eye-Mind Reader, researchers conducted a randomized controlled trial with experimental and yoked-control conditions. In the experimental condition, 35 learners received interventions triggered by the machine learning model, while in the yoked-control condition a further 35 learners were not asked to self-report mind wandering but still received interventions at the same points in the text as the corresponding yoked-learner in the experimental condition. This careful design ensured equal treatment dosage across conditions, but the dosage was only timed to mind wandering in the intervention condition.

Learners' assessment scores were not significantly different on an assessment directly after the learning experience, but on a follow-up assessment one week later the experimental group significantly outperformed the yoked-control group in terms of both surface-level and deep comprehension questions (Cohen's d = .352 and .307). Thus, Eye-Mind Reader successfully improved longitudinal retention of learned material by adapting to cognitive engagement.

3) Behavioral Design: Research has demonstrated that careful learner feedback can reduce off-task behaviors (i.e., self-regulation supports) [11] and improve metacognitive awareness [38], both of which help improve behavioral engagement. Systems designed to monitor students' behavioral engagement and raise teacher's awareness, such as the *SEAT* [3] system below, can also be effective reactive designs for indirectly improving learner engagement.

Student Engagement Analytics Technology (SEAT): Aslan et al. [3] adopted a multimodal approach to engagement detection that incorporated facial features, interaction log files, and contextual factors in real-world classrooms. The researchers trained two machine learning models with these features, one to predict students' behavioral engagement (specifically, on-task vs. off-task behavior) and one to predict affective facets of engagement including boredom, confusion, and satisfaction. Their multimodal random forest model for behavioral engagement had accuracy well above chance (Cohen's κ = .65) [223], while their affective engagement model had F_1 scores ranging from .558 to .634 depending on the type of learning activity (instructional vs. assessment) [224]. These machine learning models powered an adaptive graphical interface, referred to as *SEAT* (Student Engagement Analytics Technology), which provided teachers with whole-class and student-specific engagement information to enable them to better tailor instructional support based on individual student engagement with technology in the classroom.

The authors conducted a single-user case study with one teacher and two classes (one with SEAT and another without) over several weeks. The goal was not to demonstrate any causal effects but merely to ascertain the usability of SEAT across time. Results stemming from interviews with the teacher showed that SEAT enhanced the teacher's abilities to ascertain engagement and act on it, especially across the whole classroom. The teacher also noted that SEAT enabled more timely interventions to reorient students who were experiencing boredom, confusion, or were off-task, and that without SEAT some disengaged students would have remained unnoticed in some cases. This study helps illustrate the potential of automated engagement inference systems to contribute to improved student engagement in the classroom.

4) Multi-component Design: There is a dearth of recent research incorporating reactive design elements spanning all three components. Table A1 (in the *Enhancement approach* column) lists further examples of studies promoting learner engagement, but only two systems (SEAT [3], Park et al. [20]) demonstrate successful approaches to enhancing engagement by reacting to both affective and behavioral learner cues. We expect that future automated AI systems incorporating reactive designs spanning all three engagement components (examples in Figure 4) will be better equipped to enhance learner engagement.

C. Takeaways

Key takeaways in this section related to the enhancement of learner engagement are:

1) Designs for improving engagement can be proactive (top-down) or reactive (bottom-up), ideally including elements of both
2) Rather than focusing on individual components of engagement (i.e., affective, cognitive, behavioral), the most effective approaches should address multiple components.
3) The promotion of learner engagement in digital learning technologies has only recently become the focus of concentrated research. As such, many strategies have only been tested once and have yet to be tested longitudinally for individual learners, meaning these methods have yet to be robustly validated.
4) Methods to validate efforts to enhance engagement need to be improved. Simple experimental designs that compare systems enhanced with interventions to baseline versions belie dosage and placebo effects and can result in misattributing effects to the intervention itself.

V. Future Research Directions

We have presented an overview of several approaches to measuring learner engagement, including automated, machine-based measurement strategies, and shown examples of how systems can improve learner engagement. So where do we take learner engagement research next? Here we end with a prospective look at promising research directions in this domain.

1) Utilizing Heterogeneous Engagement Measures: Among researchers, practitioners, policy-makers, and learning system designers alike, there is an over-reliance on using a single measurement approach, be it self-reports or observer (informant) reports, to collect ground truth human judgments. These measures have and continue to inform learner engagement theories and intervention strategies, and yet as we discussed in Section III, they have individual biases and only provide one perspective into a complex construct. By utilizing a variety of self-reported, observer-based, and even machine-observed measures, we stand to gain a more comprehensive view of its varied affective, cognitive, and behavioral components and dynamics. Focusing on blending these measures in a valid, fair, and reliable fashion will improve the social and scientific value of research studies and findings.

2) Integrating with Human-sensing Technologies: The market for Consumer-Off-The-Shelf (COTS) devices (i.e., wearable sensors and fitness trackers; e.g., Fitbit, Garmin) has exploded over the past decade. The success of COTS devices means a pervasive network of human-sensing technology is becoming a reality. However, there is still much to learn about how these new-generation technologies for tracking physiological signals (e.g., [225, 226]), eye gaze (e.g., [12, 227]), and vocal audio (e.g., [6]) can inform about learner engagement. Recent research cautions that the effectiveness of different COTS-derived indicators of mental states decreases when clean signals gathered in controlled settings (e.g., digital learning in a lab) are applied to real-world domains (e.g., digital learning from home) [165]. Thus, leveraging these abundant human-sensing options will take considerable effort and testing in naturalistic environments, but will unlock the potential for findings to quickly scale to large populations rather than smaller numbers of learners using specific digital learning platforms or sensing-enabled classrooms.

3) Embracing Multimodal, Multi-componential, and Multi-temporal Complexity: Engagement in any context, not just learning, entails cognitive, behavioral, and affective states (multi-componential) expressed in a variety of manners (multimodal; e.g., focused eye gaze, note-taking, discussions) and over momentary and long-term time scales (multi-temporal).

Most research has focused on one or two of these areas, but none have yet investigated the complexity of interactions among all three. Multimodal measurement (e.g., video, audio, logs) offers more relevant channels through which engagement is expressed and can help to improve ML model robustness. Though, multimodal measures may not always prove unbiased or useful when predicting constructs (e.g., [52]), they at least offer a practical advantage in that the presence of one secondary signal can compensate when a primary signal is unavailable (e.g., muddled speech, blurry camera focus, occluded face). Furthermore, capturing multi-componential information yields a more complete view of engagement, and it may be best achieved through multimodal signal capture. For instance, eye gaze and central physiology are best suited for cognitive engagement [228–232], facial features and peripheral physiology for affective engagement [233–236], and interaction features for behavioral engagement [115, 237–239]. Multimodal measures (capturing multi-componential aspects of engagement) that operate across multiple timescales ranging from milliseconds (physiological signals), milliseconds to seconds (bodily responses), and seconds to minutes (interaction patterns) would likely improve modeling of different components of engagement that manifest across different timescales [240].

4) Incorporating Theories of Engagement-related Experiences: As we mentioned earlier in this section, little research on machine-aided engagement prediction focuses on divergent validity. Thus, when learner engagement is predicted to be low, it is uncertain whether the learner is disengaged or whether the model simply is inaccurate in this case. This can be remedied in part by training the model to accurately predict disengagement as well, but more can be done. For instance, theories of disengagement can be strategically incorporated into the measurement process where indicators of different types of disengagement (e.g., boredom, distraction, disinterest) can inform and improve the accuracy and diagnosticity of the reason(s) for disengagement. Boredom, for example, can stem from multiple factors: understimulation, perception that effort is forced, underchallenge, lack of value, lack of interest, or even a dislike of the teacher [75, 241]. Hence, making an accurate disengagement assessment is helpful, but if it can be correctly attributed to boredom, for example, then subsequent steps aimed at enhancing engagement can directly address the sources of boredom.

5) Blending Person-oriented and Context-oriented Perspectives and Proactive and Reactive Design: Reactive interventions, which aim to address disengagement when it occurs in a person-oriented manner, is a powerful paradigm to promote engagement (see Section IV-B). Proactive designs with context-oriented perspectives have been less studied (see Figure 1), but several examples of successful designs exist. Well-designed video games, for example, lie somewhere between a context-oriented and person-in-context oriented design, and many successfully capture attention and produce hours of engaged interactions [242, 243].

We anticipate that the most successful systems for measuring and enhancing engagement will blend both person-oriented and context-oriented approaches and both proactive and reactive designs. For example, researchers are exploring how to embed both cognitive (focusing on improving learning) and affective (focusing on improving affective/motivation) supports in a video-game design [36, 244], thereby aiming to improve both liking and learning outcomes.

VI. Conclusion

Engagement is one of the most fundamental aspects of the human experience, yet its ubiquity defies its complexity. We presented an accessible overview and selective review to affective computing research on conceptualizing, measuring, and enhancing engagement with an emphasis on educational applications. We conceptualized engagement as a multi-componential construct (i.e., affective, cognitive, behavioral) situated within a context over time where an ebb and flow of influence between a learner and learning context (i.e., the engagement continuum) constantly impacts engagement levels. We examined traditional (manual) and affective computing-based (automated) methods for measuring and enhancing engagement and discussed major challenges to broad adoption of these techniques and technologies across learning contexts and periods of time. Finally, we discussed promising future research directions embracing heterogeneous perspectives, and the multimodal, multi-componential, multi-temporal nature of engagement to get one step closer to generalizable, scalable, and effective technologies for enhancing learner engagement and improving learning outcomes.

Acknowledgment

This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL 2019805 and DRL 1920510. The opinions expressed are those of the authors and do not represent views of the NSF.

References

[1] K. Ahuja, D. Kim, F. Xhakaj, V. Varga, A. Xie, S. Zhang, J. E. Townsend, C. Harrison, A. Ogan, and Y. Agarwal, "EduSense: Practical classroom sensing at scale," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1–26, 2019.

[2] A. Apicella, P. Arpaia, M. Frosolone, G. Improta, N. Moccaldi, and A. Pollastro, "EEG-based measurement system for monitoring student engagement in learning 4.0," *Scientific Reports*, vol. 12, no. 1, Apr. 2022.

[3] S. Aslan, N. Alyuz, C. Tanriover, S. E. Mete, E. Okur, S. K. D'Mello, and A. Arslan Esme, "Investigating the impact of a real-time, multimodal student engagement analytics technology in authentic classrooms," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: Association for Computing Machinery, May 2019, pp. 1–12.

[4] C. Chang, C. Zhang, L. Chen, and Y. Liu, "An ensemble model using face and body tracking for engagement detection," in *Proceedings of the 20th ACM International*

Conference on Multimodal Interaction, 2018, pp. 616–622.

[5] H. Chauhan, A. Prasad, and J. Shukla, "Engagement analysis of ADHD students using visual cues from eye tracker," in *Companion Publication of the 2020 International Conference on Multimodal Interaction*, ser. ICMI '20 Companion. New York, NY, USA: Association for Computing Machinery, Oct. 2020, pp. 27–31.

[6] A. Chorianopoulou, E. Tzinis, E. Iosif, A. Papoulidi, C. Papailiou, and A. Potamianos, "Engagement detection for children with autism spectrum disorder," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5055–5059.

[7] R. W. Crues, N. Bosch, M. Perry, L. Angrave, N. Shaik, and S. Bhat, "Refocusing the lens on engagement in moocs," in *Proceedings of the fifth annual ACM conference on learning at scale*, 2018, pp. 1–10.

[8] A. Eldenfria and H. Al-Samarraie, "Towards an online continuous adaptation mechanism (OCAM) for enhanced engagement: An EEG study," *International Journal of Human–Computer Interaction*, vol. 35, no. 20, pp. 1960–1974, 2019.

[9] K. Fujii, P. Marian, D. Clark, Y. Okamoto, and J. Rekimoto, "Sync class: Visualization system for in-class student synchronization," in *Proceedings of the 9th Augmented Human International Conference*, 2018, pp. 1–8.

[10] N. Gao, W. Shao, M. S. Rahaman, and F. D. Salim, "n-Gage: Predicting in-class emotional, behavioural and cognitive engagement in the wild," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 3, pp. 79:1–79:26, Sep. 2020.

[11] B. Grawemeyer, M. Mavrikis, W. Holmes, S. Gutiérrez-Santos, M. Wiedmann, and N. Rummel, "Affective learning: Improving engagement and enhancing learning with affect-aware feedback," *User Modeling and User-Adapted Interaction*, vol. 27, no. 1, pp. 119–158, Mar. 2017.

[12] S. Hutt, K. Krasich, C. Mills, N. Bosch, S. White, J. R. Brockmole, and S. K. D'Mello, "Automated gaze-based mind wandering detection during computerized learning in classrooms," *User Modeling and User-Adapted Interaction*, vol. 29, no. 4, pp. 821–867, 2019.

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

[14] A. B. Khedher, I. Jraidi, C. Frasson *et al.*, "Tracking students' mental engagement using eeg signals during an interaction with a virtual learning environment," *Journal of Intelligent Learning Systems and Applications*, vol. 11, no. 01, p. 1, 2019.

[15] S. Li, S. P. Lajoie, J. Zheng, H. Wu, and H. Cheng, "Automated detection of cognitive engagement to inform the art of staying engaged in problem-solving," *Computers & Education*, vol. 163, p. 104114, 2021.

[16] S. Li, J. Zheng, and S. P. Lajoie, "The relationship between cognitive engagement and students' performance in a simulation-based training environment: An information-processing perspective," *Interactive Learning Environments*, pp. 1–14, Nov. 2020.

[17] O. H. Lu, J. C. Huang, A. Y. Huang, and S. J. Yang, "Applying learning analytics for improving students engagement and learning outcomes in an moocs enabled collaborative programming course," *Interactive Learning Environments*, vol. 25, no. 2, pp. 220–234, 2017.

[18] H. Monkaresi, N. Bosch, R. A. Calvo, and S. K. D'Mello, "Automated detection of engagement using video-based estimation of facial expressions and heart rate," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 15–28, 2016.

[19] M. Ninaus, S. Greipl, K. Kiili, A. Lindstedt, S. Huber, E. Klein, H.-O. Karnath, and K. Moeller, "Increased emotional engagement in game-based learning–a machine learning approach on facial emotion detection data," *Computers & Education*, vol. 142, p. 103641, 2019.

[20] H. W. Park, I. Grover, S. Spaulding, L. Gomez, and C. Breazeal, "A model-free affective reinforcement learning approach to personalization of an autonomous social robot companion for early literacy education," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 687–694.

[21] A. Psaltis, K. C. Apostolakis, K. Dimitropoulos, and P. Daras, "Multimodal student engagement recognition in prosocial games," *IEEE Transactions on Games*, vol. 10, no. 3, pp. 292–303, 2017.

[22] F. Rodriguez, H. R. Lee, T. Rutherford, C. Fischer, E. Potma, and M. Warschauer, "Using clickstream data mining techniques to understand and support first-generation college students in an online chemistry course," in *LAK21: 11th International Learning Analytics and Knowledge Conference*, ser. LAK21. New York, NY, USA: Association for Computing Machinery, Apr. 2021, pp. 313–322.

[23] A. V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network," *IEEE Transactions on Affective Computing*, 2022.

[24] K. Seo, S. Dodson, N. M. Harandi, N. Roberson, S. Fels, and I. Roll, "Active learning with online video: The impact of learning context on engagement," *Computers & Education*, vol. 165, p. 104132, 2021.

[25] T. Soffer and A. Cohen, "Students' engagement characteristics predict success and completion of online courses," *Journal of Computer Assisted Learning*, vol. 35, no. 3, pp. 378–389, 2019.

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

[27] C. Thomas and D. B. Jayagopi, "Predicting student engagement in classrooms using facial behavioral cues,"

in *Proceedings of the 1st ACM SIGCHI international workshop on multimodal interaction for education*, 2017, pp. 33–40.

[28] J. Yue, F. Tian, K.-M. Chao, N. Shah, L. Li, Y. Chen, and Q. Zheng, "Recognizing multidimensional engagement of e-learners based on multi-channel data in e-learning environment," *IEEE Access*, vol. 7, pp. 149 554–149 567, 2019.

[29] W.-H. Yun, D. Lee, C. Park, J. Kim, and J. Kim, "Automatic recognition of children engagement from facial video using convolutional neural networks," *IEEE Transactions on Affective Computing*, vol. 11, no. 4, pp. 696–707, Oct. 2020.

[30] J. Zaletelj and A. Košir, "Predicting students' attention in the classroom from Kinect facial and body features," *EURASIP Journal on Image and Video Processing*, vol. 2017, no. 1, p. 80, Dec. 2017.

[31] C. Brom, T. Starkova, and S. K. D'Mello, "How effective is emotional design? a meta-analysis on facial anthropomorphisms and pleasant colors during multimedia learning," *Educational Research Review*, vol. 25, pp. 100–119, 2018.

[32] S. May, K. Todd, S. G. Daley, and G. Rappolt-Schlichtmann, "Measurement of science museum visitors' emotional experiences at exhibits designed to encourage productive struggle," *Curator: The Museum Journal*, vol. 65, no. 1, pp. 161–185, 2022.

[33] M. T. Chi and R. Wylie, "The ICAP framework: Linking cognitive engagement to active learning outcomes," *Educational psychologist*, vol. 49, no. 4, pp. 219–243, 2014.

[34] G. D. Bodie, W. G. Powers, and M. Fitch-Hauser, "Chunking, priming and active learning: Toward an innovative and blended approach to teaching communication-related skills," *Interactive learning environments*, vol. 14, no. 2, pp. 119–135, 2006.

[35] S. Cheryan, S. A. Ziegler, V. C. Plaut, and A. N. Meltzoff, "Designing classrooms to maximize student achievement," *Policy Insights from the Behavioral and Brain Sciences*, vol. 1, no. 1, pp. 4–12, 2014.

[36] K. Bainbridge, G. L. Smith, V. J. Shute, and S. D'Mello, "Designing and testing affective supports in an educational game," *International Journal of Game-Based Learning (IJGBL)*, vol. 12, no. 1, pp. 1–32, 2022.

[37] C. Mills, J. Gregg, R. Bixler, and S. K. D'Mello, "Eye-mind reader: An intelligent reading interface that promotes long-term comprehension by detecting and responding to mind wandering," *Human–Computer Interaction*, vol. 36, no. 4, pp. 306–332, 2021.

[38] F. G. Karaoglan Yilmaz, "The effect of learning analytics assisted recommendations and guidance feedback on students' metacognitive awareness and academic achievements," *Journal of Computing in Higher Education*, pp. 1–20, 2022.

[39] M. Csikszentmihalhi, *Finding Flow: The Psychology of Engagement with Everyday Life*. Basic Books, 1997.

[40] S. Hidi and K. A. Renninger, "The four-phase model of interest development," *Educational Psychologist*, vol. 41, no. 2, pp. 111–127, 2006.

[41] S. Hidi, "Interest, psychology of," in *International Encyclopedia of the Social & Behavioral Sciences*, N. J. Smelser and P. B. Baltes, Eds. Pergamon, pp. 7712–7715.

[42] N. Samadiani, G. Huang, B. Cai, W. Luo, C.-H. Chi, Y. Xiang, and J. He, "A review on automatic facial expression recognition systems assisted by multimodal sensor data," *Sensors*, vol. 19, no. 8, p. 1863, 2019.

[43] S. Gedam and S. Paul, "A review on mental stress detection using wearable sensors and machine learning techniques," *IEEE Access*, vol. 9, pp. 84 045–84 066, 2021.

[44] R. A. Calvo and S. K. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on affective computing*, vol. 1, no. 1, pp. 18–37, 2010.

[45] S. K. D'Mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Computing Surveys (CSUR)*, vol. 47, no. 3, pp. 1–36, 2015.

[46] N. Thakur and C. Y. Han, "A complex activity based emotion recognition algorithm for affect aware systems," in *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2018, pp. 748–753.

[47] S. K. D'Mello and A. C. Graesser, "Feeling, thinking, and computing with affect-aware learning technologies," *The Oxford Handbook of Affective Computing*, pp. 419–434, 2015.

[48] S. K. D'Mello, N. Blanchard, R. S. Baker, J. Ocumpaugh, and K. Brawner, "I feel your pain: A selective review of affect-sensitive instructional strategies," *Design Recommendations for Intelligent Tutoring Systems: Volume 2-Instructional Management*, vol. 2, pp. 35–48, 2014.

[49] S. K. D'Mello, "Giving eyesight to the blind: Towards attention-aware aied," *International Journal of Artificial Intelligence in Education*, vol. 26, no. 2, pp. 645–659, 2016.

[50] ——, "Improving student engagement in and with digital learning technologies," *OECD Digital Education Outlook 2021 Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots: Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots*, p. 79, 2021.

[51] B. M. Booth, L. Hickman, S. K. Subburaj, L. Tay, S. E. Woo, and S. K. D'Mello, "Integrating psychometrics and computing perspectives on bias and fairness in affective computing: A case study of automated video interviews," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 84–95, 2021.

[52] B. M. Booth, L. Hickman, S. K. Subburaj, L. Tay, S. E. Woo, and S. K. D'Mello, "Bias and fairness in multimodal machine learning: A case study of automated video interviews," in *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 268–277.

[53] H. J. Witchel, "Engagement: the inputs and the outputs: Conference overview," in *Proceedings of the 2013 Inputs-Outputs Conference: An Interdisciplinary Conference on Engagement in HCI and Performance*, 2013, pp. 1–4.

[54] S. Christenson, A. L. Reschly, C. Wylie *et al.*, *Handbook of research on student engagement*. Springer, 2012, vol. 840.

[55] J. Eccles and M.-T. Wang, "Part I commentary: So what is student engagement anyway?" in *Handbook of Research on Student Engagement*. Springer, 2012, pp. 133–145.

[56] E. L. Deci and R. M. Ryan, *Intrinsic Motivation and Self-determination in Human Behavior*. Springer Science & Business Media, 2013.

[57] R. M. Ryan and E. L. Deci, "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being." *American Psychologist*, vol. 55, no. 1, p. 68, 2000.

[58] A. Bandura, *Social Foundations of Thought and Action: A Social Cognitive Theory*. Cambridge University Press, 1986.

[59] A. Bandura, W. H. Freeman, and R. Lightsey, "Self-efficacy: The exercise of control," *Journal of Cognitive Psychotherapy*, pp. 158–166, 1997.

[60] D. H. Schunk, F. Pajares *et al.*, "Competence perceptions and academic functioning," *Handbook of Competence and Motivation*, vol. 85, p. 104, 2005.

[61] Z. Wang, L. Chen, and T. Anderson, "A framework for interaction and cognitive engagement in connectivist learning contexts," *International Review of Research in Open and Distributed Learning*, vol. 15, no. 2, pp. 121–141, 2014.

[62] A. Barlow, S. Brown, B. Lutz, N. Pitterson, N. Hunsu, and O. Adesope, "Development of the student course cognitive engagement instrument (SCCEI) for college engineering courses," *International Journal of STEM Education*, vol. 7, no. 1, pp. 1–20, 2020.

[63] A. M. Olney, E. F. Risko, S. K. D'Mello, and A. C. Graesser, "Attention in educational contexts: The role of the learning task in guiding attention." *Grantee Submission*, 2015.

[64] K. Fiedler and S. Beier, "Affect and cognitive processes in educational contexts," in *International handbook of Emotions in Education*. Routledge, 2014, pp. 46–65.

[65] M. Csikszentmihalyi, *Flow: The Psychology of Optimal Experience*. HarperCollins, 2009.

[66] Y. Douglas and A. Hargadon, "The pleasure principle: Immersion, engagement, flow," in *Proceedings of the Eleventh ACM Conference on Hypertext and Hypermedia*, 2000, pp. 153–160.

[67] L. Gros, N. Debue, J. Lete, and C. Van De Leemput, "Video game addiction and emotional states: Possible confusion between pleasure and happiness?" *Frontiers in Psychology*, vol. 10, p. 2894, 2020.

[68] S. K. D'Mello, B. Lehman, R. Pekrun, and A. C. Graesser, "Confusion can be beneficial for learning," *Learning and Instruction*, vol. 29, pp. 153–170, 2014.

[69] J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris, "School engagement: Potential of the concept, state of the evidence," *Review of educational research*, vol. 74, no. 1, pp. 59–109, 2004.

[70] K. A. Renninger and J. E. Bachrach, "Studying triggers for interest and engagement using observational methods," *Educational Psychologist*, vol. 50, no. 1, pp. 58–69, 2015.

[71] J. P. Gee, "What video games have to teach us about learning and literacy," *Computers in Entertainment (CIE)*, vol. 1, no. 1, pp. 20–20, 2003.

[72] G. M. Sinatra, B. C. Heddy, and D. Lombardi, "The challenges of defining and measuring student engagement in science," pp. 1–13, 2015.

[73] N. Hunkins, S. Kelly, and S. K. D'Mello, ""Beautiful work, you're rock stars!": Teacher analytics to uncover discourse that supports or undermines student motivation, identity, and belonging in classrooms," in *LAK22: 12th International Learning Analytics and Knowledge Conference*, 2022, pp. 230–238.

[74] R. S. Baker, S. K. D'Mello, M. M. T. Rodrigo, and A. C. Graesser, "Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments," *International Journal of Human-Computer Studies*, vol. 68, no. 4, pp. 223–241, 2010.

[75] R. Pekrun, T. Goetz, L. M. Daniels, R. H. Stupnisky, and R. P. Perry, "Boredom in achievement settings: Exploring control–value antecedents and performance outcomes of a neglected emotion." *Journal of educational psychology*, vol. 102, no. 3, p. 531, 2010.

[76] A. S. Wasson, "Susceptibility to boredom and deviant behavior at school," *Psychological Reports*, vol. 48, no. 3, pp. 901–902, 1981.

[77] V. Tze, L. M. Daniels, and R. M. Klassen, "Evaluating the relationship between boredom and academic outcomes: A meta-analysis," *Educational Psychology Review*, vol. 28, no. 1, pp. 119–144, 2016.

[78] M. Theobald, "Self-regulated learning training programs enhance university students' academic performance, self-regulated learning strategies, and motivation: A meta-analysis," *Contemporary Educational Psychology*, vol. 66, p. 101976, Jul. 2021.

[79] K. Manikandan and A. Neethu, "Student engagement in relation to academic stress and self-efficacy," *Guru Journal of Behavioral and Social Sciences*, vol. 6, no. 1, pp. 775–784, 2018.

[80] M. J. Van Ryzin and C. J. Roseth, "The cascading effects of reducing student stress: cooperative learning as a means to reduce emotional problems and promote academic engagement," *The Journal of Early Adolescence*, vol. 41, no. 5, pp. 700–724, 2021.

[81] S. B. Whiting, S. V. Wass, S. Green, and M. S. Thomas, "Stress and learning in pupils: Neuroscience evidence and its relevance for teachers," *Mind, Brain, and Education*, vol. 15, no. 2, pp. 177–188, 2021.

[82] A. Grodner and N. G. Rupp, "The role of homework

in student learning outcomes: Evidence from a field experiment," *The Journal of Economic Education*, vol. 44, no. 2, pp. 93–109, 2013.

[83] G. A. Tetteh, "Effects of classroom attendance and learning strategies on the learning outcome," *Journal of International Education in Business*, vol. 11, no. 2, pp. 195–219, 2018.

[84] M. Hu and H. Li, "Student engagement in online learning: A review," in *2017 International Symposium on Educational Technology (ISET)*, Jun. 2017, pp. 39–43.

[85] Q. Li and R. Baker, "The different relationships between engagement and outcomes across participant subgroups in Massive Open Online Courses," *Computers & Education*, vol. 127, pp. 41–65, Dec. 2018.

[86] J. R. Buelow, T. A. Barry, and L. E. Rich, "Supporting learning engagement with online students," *Online Learning*, vol. 22, no. 4, Jan. 2019.

[87] M. B. Butler and R. J. Zerr, "The use of online homework systems to enhance out-of-class student engagement," *International Journal for Technology in Mathematics Education*, vol. 12, no. 2, pp. 51–58, 2005.

[88] N. Suárez, B. Regueiro, I. Estévez, M. del Mar Ferradás, M. A. Guisande, and S. Rodríguez, "Individual precursors of student homework behavioral engagement: The role of intrinsic motivation, perceived homework utility and homework attitude," *Frontiers in Psychology*, vol. 10, 2019.

[89] D. Yang, T. Sinha, D. Adamson, and C. P. Rosé, "Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses," in *Proceedings of the 2013 NIPS Data-Driven Education Workshop*, vol. 11, 2013, p. 14.

[90] C. C. Gray and D. Perkins, "Utilizing early engagement and machine learning to predict student outcomes," *Computers & Education*, vol. 131, pp. 22–32, Apr. 2019.

[91] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagementfrom facial expressions," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 86–98, 2014.

[92] B. A. Greene, "Measuring cognitive engagement with self-report scales: Reflections from over 20 years of research," *Educational Psychologist*, vol. 50, no. 1, pp. 14–30, 2015.

[93] J. D. Wammes, B. C. W. Ralph, C. Mills, N. Bosch, T. L. Duncan, and D. Smilek, "Disengagement during lectures: Media multitasking and mind wandering in university classrooms," *Computers & Education*, vol. 132, pp. 76–89, 2019.

[94] P. Pham and J. Wang, "AttentiveLearner: Improving mobile MOOC learning via implicit heart rate tracking," in *Proceedings of the 17th International Conference on Artificial Intelligence in Education*. Cham, CH: Springer, Jun. 2015, pp. 367–376.

[95] J. A. Fredricks, M. Filsecker, and M. A. Lawson, "Student engagement, context, and adjustment: Addressing definitional, measurement, and methodological issues," *Learning and Instruction*, vol. 43, pp. 1–4, Jun. 2016.

[96] T. Fütterer, K. Scheiter, X. Cheng, and K. Stürmer, "Quality beats frequency? Investigating students' effort in learning when introducing technology in classrooms," *Contemporary Educational Psychology*, vol. 69, p. 102042, Apr. 2022.

[97] I. H. Robertson and R. G. O'Connell, "Vigilant attention," in *Attention and Time*, A. C. Nobre, K. Nobre, and J. T. Coull, Eds. New York, NY: Oxford University Press, 2010, pp. 79–88.

[98] S. Mann and A. Robinson, "Boredom in the lecture theatre: An investigation into the contributors, moderators and outcomes of boredom amongst university students," *British Educational Research Journal*, vol. 35, no. 2, pp. 243–258, 2009.

[99] B. C. Patrick, E. A. Skinner, and J. P. Connell, "What motivates children's behavior and emotion? Joint effects of perceived control and autonomy in the academic domain." *Journal of Personality and social Psychology*, vol. 65, no. 4, p. 781, 1993.

[100] A.-J. Griffiths, E. Lilles, M. J. Furlong, and J. Sidhwa, "The relations of adolescent student engagement with troubling and high-risk behaviors," in *Handbook of research on student engagement*. Springer, 2012, pp. 563–584.

[101] L. M. Daniels, R. H. Stupnisky, R. Pekrun, T. L. Haynes, R. P. Perry, and N. E. Newall, "A longitudinal analysis of achievement goals: From affective antecedents to emotional effects and achievement outcomes." *Journal of Educational Psychology*, vol. 101, no. 4, p. 948, 2009.

[102] J. D. Finn and K. E. Voelkl, "School characteristics related to student engagement," *The Journal of Negro Education*, vol. 62, no. 3, pp. 249–268, 1993.

[103] J. A. Fredricks and W. McColskey, "The measurement of student engagement: A comparative analysis of various methods and student self-report instruments," in *Handbook of Research on Student Engagement*. Springer, 2012, pp. 763–782.

[104] C. R. Henrie, L. R. Halverson, and C. R. Graham, "Measuring student engagement in technology-mediated learning: A review," *Computers & Education*, vol. 90, pp. 36–53, 2015.

[105] S. R. Hart, K. Stewart, and S. R. Jimerson, "The student engagement in schools questionnaire (SESQ) and the teacher engagement report form-new (TERF-N): Examining the preliminary evidence," *Contemporary School Psychology: Formerly "The California School Psychologist"*, vol. 15, no. 1, pp. 67–79, 2011.

[106] D. Kahneman, A. B. Krueger, D. A. Schkade, N. Schwarz, and A. A. Stone, "A survey method for characterizing daily life experience: The day reconstruction method," *Science*, vol. 306, no. 5702, pp. 1776–1780, 2004.

[107] J. C. Turner and D. K. Meyer, "Studying and understanding the instructional contexts of classrooms: Using our past to forge our future," *Educational Psychologist*,

vol. 35, no. 2, pp. 69–85, 2000.

[108] M. Csikszentmihalyi and R. Larson, "Validity and reliability of the experience-sampling method," in *Flow and the Foundations of Positive Psychology*. Springer, 2014, pp. 35–54.

[109] K. Mundnich, B. M. Booth, M. l'Hommedieu, T. Feng, B. Girault, J. L'hommedieu, M. Wildman, S. Skaaden, A. Nadarajan, J. L. Villatte *et al.*, "Tiles-2018, a longitudinal physiologic and behavioral data set of hospital workers," *Scientific Data*, vol. 7, no. 1, pp. 1–26, 2020.

[110] S. M. Mattingly, J. M. Gregg, P. Audia, A. E. Bayraktaroglu, A. T. Campbell, N. V. Chawla, V. Das Swain, M. De Choudhury, S. K. D'Mello, A. K. Dey *et al.*, "The Tesserae project: Large-scale, longitudinal, in situ, multimodal sensing of information workers," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–8.

[111] S. Hutt, J. F. Grafsgaard, and S. K. D'Mello, "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*, 2019, pp. 1–14.

[112] N. Bosch and S. K. D'Mello, "Automatic detection of mind wandering from video in the lab and in the classroom," *IEEE Transactions on Affective Computing*, vol. 12, no. 4, pp. 974–988, 2019.

[113] K. E. Arnold and M. D. Pistilli, "Course signals at purdue: Using learning analytics to increase student success," in *Proceedings of the 2nd international conference on learning analytics and knowledge*, 2012, pp. 267–270.

[114] C. A. Lehr, M. F. Sinclair, and S. L. Christenson, "Addressing student engagement and truancy prevention during the elementary school years: A replication study of the check & connect model," *Journal of Education for Students Placed at Risk*, vol. 9, no. 3, pp. 279–301, 2004.

[115] J. D. Gobert, R. S. Baker, and M. B. Wixon, "Operationalizing and detecting disengagement within online science microworlds," *Educational Psychologist*, vol. 50, no. 1, pp. 43–57, 2015.

[116] J. Ocumpaugh, "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*, vol. 60, 2015.

[117] M. G. Meany-Daboul, E. M. Roscoe, J. C. Bourret, and W. H. Ahearn, "A comparison of momentary time sampling and partial-interval recording for evaluating functional relations," *Journal of Applied Behavior Analysis*, vol. 40, no. 3, pp. 501–514, 2007.

[118] J. Bidwell and H. Fuchs, "Classroom analytics: Measuring student engagement with automated gaze tracking," *Behavior Research Methods*, vol. 49, no. 113, 2011.

[119] A. Dhall, A. Kaur, R. Goecke, and T. Gedeon, "Emotiw 2018: Audio-video, student engagement and group-level affect prediction," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 653–656.

[120] M. Dewan, M. Murshed, and F. Lin, "Engagement detection in online learning: A review," *Smart Learning Environments*, vol. 6, no. 1, pp. 1–20, 2019.

[121] K. Porayska-Pomsta, M. Mavrikis, S. K. D'Mello, C. Conati, and R. S. Baker, "Knowledge elicitation methods for affect modelling in education," *International Journal of Artificial Intelligence in Education*, vol. 22, no. 3, pp. 107–140, 2013.

[122] S. K. D'Mello, S. D. Craig, A. Witherspoon, B. McDaniel, and A. C. Graesser, "Automatic detection of learner's affect from conversational cues," *User Modeling and User-Adapted Interaction*, vol. 18, no. 1, pp. 45–80, 2008.

[123] J. L. Sabourin and J. C. Lester, "Affect and engagement in game-based learning environments," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 45–56, 2013.

[124] N. Siddiqui, K. Miah, and A. Ahmad, "Peer to peer synchronous interaction and student engagement: A perspective of postgraduate management students in a developing country," *American Journal of Educational Research*, vol. 7, no. 7, pp. 491–498, 2019.

[125] K. J. Hsu, K. N. Babeva, M. C. Feng, J. F. Hummer, and G. C. Davison, "Experimentally induced distraction impacts cognitive but not emotional processes in think-aloud cognitive assessment," *Frontiers in Psychology*, vol. 5, p. 474, 2014.

[126] S. Hutt and S. K. D'Mello, "Evaluating calibration-free webcam-based eye tracking for gaze-based user modeling," in *Proceedings of the 2022 International Conference on Multimodal Interaction*, 2022, pp. 224–235.

[127] R. S. Baker and J. Ocumpaugh, "Interaction-based affect detection in educational software," *The Oxford Handbook of Affective Computing*, pp. 233–245, 2015.

[128] M. Iman, K. Rasheed, and H. R. Arabnia, "A review of deep transfer learning and recent advancements," *arXiv preprint arXiv:2201.09679*, 2022.

[129] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *International Journal of Computer Vision*, vol. 127, no. 6, pp. 907–929, 2019.

[130] T. Feng, B. M. Booth, B. Baldwin-Rodríguez, F. Osorno, and S. Narayanan, "A multimodal analysis of physical activity, sleep, and work shift in nurses with wearable sensor data," *Scientific Reports*, vol. 11, no. 1, pp. 1–12, 2021.

[131] S. K. D'Mello, E. Dieterle, and A. Duckworth, "Advanced, analytic, automated (AAA) measurement of engagement during learning," *Educational psychologist*, vol. 52, no. 2, pp. 104–123, 2017.

[132] A. M. Aung, A. Ramakrishnan, and J. R. Whitehill, "Who are they looking at? Automatic eye gaze following for classroom observation video analysis," in *Proceedings of the 11th International Conference on*

Educational Data Mining. International Educational Data Mining Society, 2018, pp. 252–258.

[133] R. Klein and T. Celik, "The wits intelligent teaching system: Detecting student engagement during lectures using convolutional neural networks," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 2856–2860.

[134] M. Raca, L. Kidzinski, and P. Dillenbourg, "Translating head motion into attention-towards processing of student's body-language," in *Proceedings of the 8th International Conference on Educational Data Mining*, no. CONF, 2015.

[135] American Educational Research Association, American Psychological Association, Joint Committee on Standards for Educational and Psychological Testing (US), National Council on Measurement in Education *et al.*, *Standards for Educational and Psychological Testing*. American Educational Research Association, 2014.

[136] J. A. Krosnick, "Survey research," *Annual Review of Psychology*, vol. 50, no. 1, pp. 537–567, 1999.

[137] P. M. Podsakoff, S. B. MacKenzie, J.-Y. Lee, and N. P. Podsakoff, "Common method biases in behavioral research: a critical review of the literature and recommended remedies." *Journal of Applied Psychology*, vol. 88, no. 5, pp. 879–903, 2003.

[138] S. J. Heine, D. R. Lehman, K. Peng, and J. Greenholtz, "What's wrong with cross-cultural comparisons of subjective likert scales?: The reference-group effect." *Journal of Personality and Social Psychology*, vol. 82, no. 6, pp. 903–918, 2002.

[139] A. A. Gorin and A. A. Stone, "Recall biases and cognitive errors in retrospective self-reports: A call for momentary assessments," *Handbook of Health Psychology*, vol. 23, pp. 405–413, 2001.

[140] M. Nystrand and A. Gamoran, "Instructional discourse, student engagement, and literature achievement," *Research in the Teaching of English*, pp. 261–290, 1991.

[141] R. C. Pianta, B. K. Hamre, and J. P. Allen, "Teacher-student relationships and engagement: Conceptualizing, measuring, and improving the capacity of classroom interactions," in *Handbook of Research on Student Engagement*. Springer, 2012, pp. 365–386.

[142] S. Ryu and D. Lombardi, "Coding classroom interactions for collective and individual engagement," *Educational Psychologist*, vol. 50, no. 1, pp. 70–83, 2015.

[143] R. J. Volpe, J. C. DiPerna, J. M. Hintze, and E. S. Shapiro, "Observing students in classroom settings: A review of seven coding schemes," *School Psychology Review*, vol. 34, no. 4, pp. 454–474, 2005.

[144] S. P. Vecera and M. Behrmann, "Attention and unit formation: A biased competition account of object-based attention," in *Advances in Psychology*. Elsevier, 2001, vol. 130, pp. 145–180.

[145] K. Mahtani, E. A. Spencer, J. Brassey, and C. Heneghan, "Catalogue of bias: observer bias," *BMJ Evidence-based Medicine*, vol. 23, no. 1, pp. 23–24, 2018.

[146] P. Chipman, S. K. D'Mello, B. Gholson, A. Graesser, B. McDaniel, and A. Witherspoon, "Detection of emotions during learning with autotutor," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 28, no. 28, 2006.

[147] S. K. D'Mello, R. Taylor, K. Davidson, and A. Graesser, "Self versus teacher judgments of learner emotions during a tutoring session with autotutor," in *International Conference on Intelligent Tutoring Systems*. Springer, 2008, pp. 9–18.

[148] A. L. Duckworth and D. S. Yeager, "Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes," *Educational Researcher*, vol. 44, no. 4, pp. 237–251, 2015.

[149] S. Afzal and P. Robinson, "Natural affect data: Collection and annotation," in *New Perspectives on Affect and Learning Technologies*. Springer, 2011, pp. 55–70.

[150] S. K. D'Mello, "A selective meta-analysis on the relative incidence of discrete affective states during learning with technology." *Journal of Educational Psychology*, vol. 105, no. 4, p. 1082, 2013.

[151] A. C. Graesser, Z. Cai, M. M. Louwerse, and F. Daniel, "Question understanding aid (QUAID): a web facility that tests question comprehensibility," *Public Opinion Quarterly*, vol. 70, no. 1, pp. 3–22, 2006.

[152] S. K. D'Mello, "On the influence of an iterative affect annotation approach on inter-observer and self-observer reliability," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 136–149, 2015.

[153] M. R. Mehl, J. W. Pennebaker, D. M. Crow, J. Dabbs, and J. H. Price, "The electronically activated recorder (EAR): A device for sampling naturalistic daily activities and conversations," *Behavior Research Methods, Instruments, & Computers*, vol. 33, no. 4, pp. 517–523, 2001.

[154] B. Lehman, M. Matthews, S. K. D'Mello, and N. Person, "What are you feeling? investigating student affective states during expert human tutoring sessions," in *International Conference on Intelligent Tutoring Systems*. Springer, 2008, pp. 50–59.

[155] R. Bixler and S. K. D'Mello, "Automatic gaze-based user-independent detection of mind wandering during computerized reading," *User Modeling and User-Adapted Interaction*, vol. 26, no. 1, pp. 33–68, 2016.

[156] J. Ocumpaugh, R. S. Baker, S. Gowda, N. Heffernan, and C. Heffernan, "Population validity for educational data mining models: A case study in affect detection," *British Journal of Educational Technology*, vol. 45, no. 3, pp. 487–501, 2014.

[157] C. Stone, A. Quirk, M. Gardener, S. Hutt, A. L. Duckworth, and S. K. D'Mello, "Language as thought: Using natural language processing to model noncognitive traits that predict college success," in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 2019, pp. 320–329.

[158] N. Bosch, S. K. D'Mello, R. S. Baker, J. Ocumpaugh, V. Shute, M. Ventura, L. Wang, and W. Zhao, "Detecting student emotions in computer-enabled classrooms." in *IJCAI*, 2016, pp. 4125–4129.

[159] Z. A. Pardos, R. S. Baker, M. O. San Pedro, S. M.

Gowda, and S. M. Gowda, "Affective states and state tests: Investigating how affect and engagement during the school year predict end-of-year learning outcomes." *Journal of Learning Analytics*, vol. 1, no. 1, pp. 107–128, 2014.

[160] M. E. Dale, A. J. Godley, S. A. Capello, P. J. Donnelly, S. K. D'Mello, and S. P. Kelly, "Toward the automated analysis of teacher talk in secondary ela classrooms," *Teaching and Teacher Education*, vol. 110, p. 103584, 2022.

[161] P. H. Le-Khac, G. Healy, and A. F. Smeaton, "Contrastive representation learning: A framework and review," *IEEE Access*, vol. 8, pp. 193 907–193 934, 2020.

[162] J. Wang, X. Xie, J. Shi, W. He, Q. Chen, L. Chen, W. Gu, and T. Zhou, "Denoising autoencoder, a deep learning algorithm, aids the identification of a novel molecular signature of lung adenocarcinoma," *Genomics, Proteomics & Bioinformatics*, vol. 18, no. 4, pp. 468–480, 2020.

[163] M. Faber, R. Bixler, and S. K. D'Mello, "An automated behavioral measure of mind wandering during computerized reading," *Behavior Research Methods*, vol. 50, no. 1, pp. 134–150, 2018.

[164] B. M. Booth, K. Mundnich, T. Feng, A. Nadarajan, T. H. Falk, J. L. Villatte, E. Ferrara, and S. Narayanan, "Multimodal human and environmental sensing for longitudinal behavioral studies in naturalistic settings: Framework for sensor selection, deployment, and management," *Journal of Medical Internet Research*, vol. 21, no. 8, p. e12832, 2019.

[165] B. M. Booth, H. Vrzakova, S. M. Mattingly, G. J. Martinez, L. Faust, and S. K. D'Mello, "Toward robust stress prediction in the age of wearables: Modeling perceived stress in a longitudinal study with information workers," *IEEE Transactions on Affective Computing*, 2022.

[166] S. K. DMello, "Getting really wild: Challenges and opportunities of real-world multimodal affect detection," *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, pp. 1–1, 2021.

[167] N. Bosch, "Multimodal affect detection in the wild: Accuracy, availability, and generalizability," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 2015, pp. 645–649.

[168] E. Jensen, S. Hutt, and S. K. D'Mello, "Generalizability of sensor-free affect detection models in a longitudinal dataset of tens of thousands of students." *International Educational Data Mining Society*, 2019.

[169] N. Bosch, S. K. D'Mello, R. Baker, J. Ocumpaugh, and V. Shute, "Temporal generalizability of face-based affect detection in noisy classroom environments," in *International Conference on Artificial Intelligence in Education*. Springer, 2015, pp. 44–53.

[170] G. Vilagut, "Test-retest reliability," *Encyclopedia of quality of life and well-being research*, pp. 6622–6625, 2014.

[171] N. Bosch, H. Chen, S. D'Mello, R. Baker, and V. Shute, "Accuracy vs. availability heuristic in multimodal affect detection in the wild," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 2015, pp. 267–274.

[172] J. P. Martin, "Lower merion district's laptop saga ends with $610,000 settlement," Oct 2010. [Online]. Available: https://web.archive.org/web/20101016003945/http://www.philly.com/philly/news/20101012_Lower_Merion_district_s_laptop_saga_ends_with__610_000_settlement.html

[173] T. Feng, A. Nadarajan, C. Vaz, B. Booth, and S. Narayanan, "Tiles audio recorder: An unobtrusive wearable solution to track audio activity," in *Proceedings of the 4th ACM Workshop on Wearable Systems and Applications*, 2018, pp. 33–38.

[174] L. Tay, S. E. Woo, L. Hickman, B. M. Booth, and S. D'Mello, "A conceptual framework for investigating and mitigating machine-learning measurement bias (MLMB) in psychological assessment," *Advances in Methods and Practices in Psychological Science*, vol. 5, no. 1, p. 25152459211061337, 2022.

[175] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on Fairness, Accountability and Transparency*. PMLR, 2018, pp. 77–91.

[176] J. E. Flege, "Second language speech learning: Theory, findings, and problems," *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, vol. 92, pp. 233–277, 1995.

[177] J. Anderson-Hsieh and H. Venkatagiri, "Syllable duration and pausing in the speech of Chinese ESL speakers," *TESOL quarterly*, vol. 28, no. 4, pp. 807–812, 1994.

[178] O. Kang, "Relative salience of suprasegmental features on judgments of l2 comprehensibility and accentedness," *System*, vol. 38, no. 2, pp. 301–315, 2010.

[179] G. J. DuPaul, T. D. Pinho, B. L. Pollack, M. J. Gormley, and S. D. Laracy, "First-year college students with ADHD and/or LD: Differences in engagement, positive core self-evaluation, school preparation, and college expectations," *Journal of Learning Disabilities*, vol. 50, no. 3, pp. 238–251, 2017.

[180] R. E. V. Junod, G. J. DuPaul, A. K. Jitendra, R. J. Volpe, and K. S. Cleary, "Classroom observations of students with and without adhd: Differences across types of engagement," *Journal of School Psychology*, vol. 44, no. 2, pp. 87–104, 2006.

[181] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[182] J. R. Anderson, "Acquisition of cognitive skill." *Psychological Review*, vol. 89, no. 4, pp. 369–406, 1982.

[183] J. S. Brown and K. VanLehn, "Repair theory: A generative theory of bugs in procedural skills," *Cognitive Science*, vol. 4, no. 4, pp. 379–426, 1980.

[184] D. Sleeman and J. Brown, *Intelligent Tutoring Systems*, ser. Computers and people series. Academic Press,

1982.

[185] A. Pardino, I. Gleyzer, I. Javed, J. Reid-Hector, A. Heuer *et al.*, "The best pedagogical practices in graduate online learning: A systematic review," *Creative Education*, vol. 9, no. 07, p. 1123, 2018.

[186] L. Ceelen, A. Khaled, L. Nieuwenhuis, and E. de Bruijn, "Pedagogic practices in the context of students' workplace learning: a literature review," *Journal of Vocational Education & Training*, pp. 1–33, 2021.

[187] K. A. Ericsson, R. T. Krampe, and C. Tesch-Römer, "The role of deliberate practice in the acquisition of expert performance." *Psychological Review*, vol. 100, no. 3, p. 363, 1993.

[188] A. L. Duckworth, J. L. Taxer, L. Eskreis-Winkler, B. M. Galla, and J. J. Gross, "Self-control and academic achievement," *Annual Review of Psychology*, vol. 70, no. 1, pp. 373–399, 2019.

[189] S. K. D'Mello, "What do we think about when we learn?" in *Deep Comprehension: Multi-Disciplinary Approaches to Understanding, Enhancing, and Measuring Comprehension*, D. L. K. Millis, J. Magliano and K. Wiemer, Eds. Routledge, 2018, pp. 52–67.

[190] A. Y. Wong, S. L. Smith, C. A. McGrath, L. E. Flynn, and C. Mills, "Task-unrelated thought during educational activities: A meta-analysis of its occurrence and relationship with learning," *Contemporary Educational Psychology*, vol. 71, p. 102098, 2022.

[191] J. G. Randall, F. L. Oswald, and M. E. Beier, "Mind-wandering, cognition, and performance: A theory-driven meta-analysis of attention regulation." *Psychological Bulletin*, vol. 140, no. 6, pp. 1411–1431, 2014.

[192] E. F. Risko, D. Buchanan, S. Medimorec, and A. Kingstone, "Everyday attention: Mind wandering and computer use during lectures," *Computers & Education*, vol. 68, pp. 275–283, 2013.

[193] D. Charsky, "From edutainment to serious games: A change in the use of game characteristics," *Games and Culture*, vol. 5, no. 2, pp. 177–198, 2010.

[194] S. Papert, "Does easy do it? Children, games, and learning," *Game Developer*, vol. 5, no. 6, p. 88, 1998.

[195] W. Ma, O. O. Adesope, J. C. Nesbit, and Q. Liu, "Intelligent tutoring systems and learning outcomes: A meta-analysis." *Journal of Educational Psychology*, vol. 106, no. 4, p. 901, 2014.

[196] S. Steenbergen-Hu and H. Cooper, "A meta-analysis of the effectiveness of intelligent tutoring systems on k–12 students' mathematical learning." *Journal of Educational Psychology*, vol. 105, no. 4, pp. 970–987, 2013.

[197] ——, "A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning." *Journal of educational psychology*, vol. 106, no. 2, pp. 331–347, 2014.

[198] S. Craig, A. Graesser, J. Sullins, and B. Gholson, "Affect and learning: an exploratory look into the role of affect in learning with autotutor," *Journal of Educational Media*, vol. 29, no. 3, pp. 241–250, 2004.

[199] W. J. Hawkins, N. T. Heffernan, and R. Baker, "Which is more responsible for boredom in intelligent tutoring systems: Students (trait) or problems (state)?" *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 618–623, 2013.

[200] T. Del Solato and B. Du Boulay, "Implementation of motivational tactics in tutoring systems," *Journal of Artificial Intelligence in Education*, vol. 6, pp. 337–378, 1995.

[201] J. A. Schmidt, J. M. Rosenberg, and P. N. Beymer, "A person-in-context approach to student engagement in science: Examining learning activities and choice," *Journal of Research in Science Teaching*, vol. 55, no. 1, pp. 19–43, 2018.

[202] C. Bovill, C. J. Bulley, and K. Morss, "Engaging and empowering first-year students through curriculum design: perspectives from the literature," *Teaching in Higher Education*, vol. 16, no. 2, pp. 197–209, 2011.

[203] C. Walkington and M. L. Bernacki, "Personalizing algebra to students' individual interests in an intelligent tutoring system: Moderators of impact," *International Journal of Artificial Intelligence in Education*, vol. 29, no. 1, pp. 58–88, 2019.

[204] D. Gibson, N. Ostashewski, K. Flintoff, S. Grant, and E. Knight, "Digital badges in education," *Education and Information Technologies*, vol. 20, no. 2, pp. 403–410, 2015.

[205] K. M. Kapp, *The gamification of learning and instruction: Game-based methods and strategies for training and education*. John Wiley & Sons, 2012.

[206] S. Ahmad, U. R. Hashim *et al.*, "The effectiveness of gamification technique for higher education students engagement in polytechnic Muadzam Shah Pahang, Malaysia," *International Journal of Educational Technology in Higher Education*, vol. 15, no. 1, pp. 1–16, 2018.

[207] A. Bruckman, "Can educational be fun," in *Game Developers Conference*, vol. 99, 1999, pp. 75–79.

[208] J. L. Plass, S. Heidig, E. O. Hayward, B. D. Homer, and E. Um, "Emotional design in multimedia learning: Effects of shape and color on affect and learning," *Learning and Instruction*, vol. 29, pp. 128–140, 2014.

[209] E. Um, J. L. Plass, E. O. Hayward, B. D. Homer *et al.*, "Emotional design in multimedia learning." *Journal of Educational Psychology*, vol. 104, no. 2, pp. 485–498, 2012.

[210] S. D'Mello and A. Graesser, "Inducing and tracking confusion and cognitive disequilibrium with breakdown scenarios," *Acta Psychologica*, vol. 151, pp. 106–116, 2014.

[211] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information." *Psychological Review*, vol. 63, no. 2, p. 81, 1956.

[212] M. Fidan, "The effects of microlearning-supported flipped classroom on pre-service teachers' learning performance, motivation and engagement," *Education and Information Technologies*, pp. 1–28, 2023.

[213] T. Gao and L. Kuang, "Cognitive loading and knowledge hiding in art design education: Cognitive engage-

ment as mediator and supervisor support as moderator," *Frontiers in Psychology*, vol. 13, 2022.

[214] M. Ninaus and S. Nebel, "A systematic literature review of analytics for adaptivity within educational video games," in *Frontiers in Education*, vol. 5. Frontiers Media SA, 2021, p. 611072.

[215] R. Kadel, K. Paudel, and M. P. Gurung, "A review on educational games design, development and effectiveness measurement," in *2019 IEEE International Conference on Engineering, Technology and Education (TALE)*. IEEE, 2019, pp. 1–7.

[216] D. Vlachopoulos and A. Makri, "The effect of games and simulations on higher education: a systematic literature review," *International Journal of Educational Technology in Higher Education*, vol. 14, no. 1, pp. 1–33, 2017.

[217] M. J. Dondlinger, "Educational video game design: A review of the literature," *Journal of Applied Educational Technology*, vol. 4, no. 1, pp. 21–31, 2007.

[218] M. Faber and S. K. D'Mello, "How the stimulus influences mind wandering in semantically rich task contexts," *Cognitive Research: Principles and Implications*, vol. 3, no. 1, 2018.

[219] S. D'Mello and A. Graesser, "Dynamics of affective states during complex learning," *Learning and Instruction*, vol. 22, no. 2, pp. 145–157, 2012.

[220] J. A. DeFalco, J. P. Rowe, L. Paquette, V. Georgoulas-Sherry, K. Brawner, B. W. Mott, R. S. Baker, and J. C. Lester, "Detecting and addressing frustration in a serious game for military training," *International Journal of Artificial Intelligence in Education*, vol. 28, no. 2, pp. 152–193, 2018.

[221] K. Forbes-Riley and D. Litman, "Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor," *Speech Communication*, vol. 53, no. 9-10, pp. 1115–1136, 2011.

[222] T. W. Acee, H. Kim, H. J. Kim, J. I. Kim, H. N. R. Chu, M. Kim, Y. J. Cho, and F. W. Wicker, "Academic boredom in under-and over-challenging situations," *Contemporary Educational Psychology*, vol. 35, no. 1, pp. 17–27, 2010.

[223] E. Okur, N. Alyuz, S. Aslan, U. Genc, C. Tanriover, and A. Arslan Esme, "Behavioral engagement detection of students in the wild," in *Artificial Intelligence in Education*, E. André, R. Baker, X. Hu, M. M. T. Rodrigo, and B. du Boulay, Eds. Cham, CH: Springer International Publishing, 2017, pp. 250–261.

[224] N. Alyuz, E. Okur, E. Oktay, U. Genc, S. Aslan, S. E. Mete, B. Arnrich, and A. A. Esme, "Semi-supervised model personalization for improved detection of learner's emotional engagement," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ser. ICMI '16. New York, NY, USA: Association for Computing Machinery, Oct. 2016, pp. 100–107.

[225] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE transactions on biomedical engineering*, vol. 58, no. 1, pp. 7–11, 2010.

[226] C. G. Scully, J. Lee, J. Meyer, A. M. Gorbach, D. Granquist-Fraser, Y. Mendelson, and K. H. Chon, "Physiological parameter monitoring from optical recordings with a mobile phone," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 2, pp. 303–306, 2011.

[227] W. Sewell and O. Komogortsev, "Real-time eye gaze tracking with an unmodified commodity webcam employing a neural network," in *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, 2010, pp. 3739–3744.

[228] C. Berka, D. J. Levendowski, M. N. Lumicao, A. Yau, G. Davis, V. T. Zivkovic, R. E. Olmstead, P. D. Tremoulet, and P. L. Craven, "EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks," *Aviation, space, and environmental medicine*, vol. 78, no. 5, pp. B231–B244, 2007.

[229] H. Deubel and W. X. Schneider, "Saccade target selection and object recognition: Evidence for a common attentional mechanism," *Vision Research*, vol. 36, no. 12, pp. 1827–1837, 1996.

[230] J. E. Hoffman and B. Subramaniam, "The role of visual attention in saccadic eye movements," *Perception & Psychophysics*, vol. 57, no. 6, pp. 787–795, 1995.

[231] S. P. Marshall, "Assessing cognitive engagement and cognitive state from eye metrics," *Foundations of augmented cognition*, vol. 11, p. 312À320, 2005.

[232] K. Rayner, "Eye movements in reading and information processing: 20 years of research." *Psychological Bulletin*, vol. 124, no. 3, p. 372, 1998.

[233] P. Ekman, "Expression and the nature of emotion," *Approaches to emotion*, vol. 3, no. 19, p. 344, 1984.

[234] D. Keltner and P. Ekman, "Emotion: An overview." 2000.

[235] J. T. Larsen, G. G. Berntson, K. M. Poehlmann, T. A. Ito, and J. T. Cacioppo, "The psychophysiology of emotion," 2008.

[236] D. Matsumoto, D. Keltner, M. N. Shiota, M. O'Sullivan, and M. Frank, "Facial expressions of emotion." 2008.

[237] S. Kai, L. Paquette, R. S. Baker, N. Bosch, S. D'Mello, J. Ocumpaugh, V. Shute, and M. Ventura, "A comparison of video-based and interaction-based affect detectors in physics playground." in *Proceedings of the 8th International Conference on Educational Data Mining*. International Educational Data Mining Society, 2015, pp. 77–84.

[238] R. S. Baker and L. M. Rossi, "Assessing the disengaged behaviors of learners," *Design Recommendations for Intelligent Tutoring Systems*, vol. 1, p. 153, 2013.

[239] M. E. Bulger, R. E. Mayer, K. C. Almeroth, and S. D. Blau, "Measuring learner engagement in computer-equipped college classrooms," *Journal of Educational Multimedia and Hypermedia*, vol. 17, no. 2, pp. 129–143, 2008.

[240] S. D'Mello and A. Graesser, "The half-life of cognitive-affective states during complex learning," *Cognition & Emotion*, vol. 25, no. 7, pp. 1299–1308, 2011.

[241] E. C. Daschmann, T. Goetz, and R. H. Stupnisky, "Testing the predictors of boredom at school: Development and validation of the precursors to boredom scales," *British Journal of Educational Psychology*, vol. 81, no. 3, pp. 421–440, 2011.

[242] T. Kucher, "Principles and best practices of designing digital game-based learning environments," *International Journal of Technology in Education and Science (IJTES)*, vol. 5, no. 2, pp. 213–223, 2021.

[243] V. Shute, S. Rahimi, and G. Smith, "Game-based learning analytics in Physics Playground," in *Data analytics approaches in educational games and gamification systems*. Springer, 2019, pp. 69–93.

[244] S. Rahimi, V. J. Shute, C. Fulwider, K. Bainbridge, R. Kuba, X. Yang, G. Smith, R. S. Baker, and S. K. D'Mello, "Timing of learning supports in educational games can impact students' outcomes," *Computers & Education*, vol. 190, p. 104600, 2022.

Brandon M. Booth is a Research Scientist in the Institute of Cognitive Science at the University of Colorado Boulder and will be an Assistant Professor of Computer Science at the University of Memphis starting in the fall of 2023. He received his PhD in Computer Science from the University of Southern California. His research focuses on human-centered information processing using interdisciplinary approaches for robust computational modeling of human perception and behaviors in contexts with a societal need. His work spans several application domains including health and wellbeing, education, and AI ethics. He has a diverse industry background researching and developing video games, serious games, robots, computer vision and human-computer interactions systems.

Nigel Bosch is an Assistant Professor in the School of Information Sciences and the Department of Educational Psychology at the University of Illinois at Urbana-Champaign. Previously, he received his PhD in Computer Science from the University of Notre Dame, and was a postdoctoral researcher for two years at the National Center for Supercomputing Applications. His research interests include machine learning methods for personalization and intervention, and as a means to study human behavior. His machine learning work spans several application domains, including education, human–computer interaction, and addiction science.

Sidney K. D'Mello (PhD in Computer Science) is a Professor in the Institute of Cognitive Science and Department of Computer Science at the University of Colorado Boulder. He is interested in the dynamic interplay between cognition and emotion while individuals and groups engage in complex real-world activities. He applies insights gleaned from this basic research program to develop intelligent technologies that help people achieve to their fullest potential by coordinating what they think and feel with what they know and do. D'Mello has co-edited seven books and has published more than 300 journal papers, book chapters, and conference proceedings. His research has received 17 awards at international conferences and has been funded by numerous grants. D'Mello serves(d) as Associate Editor and on the Editorial Boards of 11 journals. He leads the NSF National Institute for Student-Agent Teaming (2020-2025), which aims to develop AI technologies to facilitate rich socio-collaborative learning experiences for all students.

APPENDIX

TABLE A1
SELECTIVE SURVEY OF PAST SIX YEARS OF ENGAGEMENT RESEARCH FOR LEARNING

Paper	Engagement continuum	Engagement components	Engagement construct	Learning context	Data modalities	Measurement approach	Modeling approach	Enhancement approach
(Ninaus et al., 2019)	person/person-context	Affective	PANAS, flow short scale (FKS), affect	Individual, digital activity, lab	Face	Face, surveys, Microsoft emotion API (face)	Stats	None
(Li et al., 2021)	person	Cognitive		Individual, digital activity, lab	Face	Survey	SVM	None
(Chang et al., 2018)	person	Behavioral	Visible engagement (Whitehill)	Individual, remote learning, remote	Face	Annotation	NN, AdaBoost	None
(Yun et al., 2020)	person	Behavioral	Visible engagement (Whitehill)	Individual, digital activity, in situ controlled	Face	Annotation	NN	None
(Sümer et al., 2021)	person-context	Behavioral	Visible engagement	Cohort, classroom activities, in situ	Face	Annotation	NN	None
(Fujii et al., 2018)	context/person-context	Behavioral	Visible engagement, posture synchrony	Cohort, classroom activities, in situ	Face	Expert rules, annotation	NN	Real-time monitor for teachers
(Thomas & Jayagopi, 2017)	person/person-context	Behavioral	Visible engagement (Whitehill)	Cohort, video, lab	Face, gaze (via head pose)	Annotation	SVM, log. reg.	None
(Kaur et al., 2018)	person	Behavioral	Visible engagement (Whitehill)	Individual, remote learning, remote	Face	Annotation	NN, RF, SVR	None
(Savchenko et al., 2022)	person-context/person	Affective, Behavioral	AffectNet labels	Individual, remote learning, remote	Face	Annotation	NN, RF, SVR, ridge reg.	None
(Psaltis et al., 2018)	person-context	Affective, Behavioral, Cognitive	Flow, success in game, visible engagement	Cohort, digital activity, in situ controlled	Face, whole body	Surveys, acted emotions	NN	None
(Zaletelj & Košir, 2017)	person	Behavioral	Visible engagement (gaze, writing)	Cohort, classroom activities, in situ	Face, whole body	Annotation	KNN, trees	None
(Grawemeyer et al., 2017)	person	Affective	Flow, affect	Cohort, classroom activities, in situ	Audio, logs	Annotation	Bayes net	Supportive feedback
(Ahuja et al., 2019)	person-context/person	Affective, Behavioral	Smiling, posture, visible activity	Cohort, classroom activities, in situ	Audio, face, whole body	Annotation	NN, SVM	None
(Park et al., 2019)	person-context	Affective, Behavioral	Activity in task (Q's answered), affect	Individual, intelligent tutoring systems, in situ	Audio, whole body	Expert rules	RL	Curriculum sequencing
(Aslan et al., 2019)	person-context/context	Affective, Behavioral	On/off-task, affect	Cohort, classroom activities, in situ	Face, logs	Annotation	RF	Teacher dashboard
(Lu et al., 2017)	person-context	Behavioral	Activity in task	Individual, remote learning, remote	Logs	Expert rules	Stats	Instructor email

Paper	Engagement continuum	Engagement dimensions	Engagement construct	Learning context	Data modalities	Measurement approach	Modeling approach	Enhancement approach
(Seo et al., 2021)	context	Behavioral	SRL activities	Individual, video, online/in situ	Logs	Survey	Stats	None
(Crues et al., 2018)	person-context/context	Behavioral	Persistence in task	Individual, remote learning, remote	Logs	Expert rules	Stats	None
(Soffer & Cohen, 2019)	person-context/person	Behavioral	Activity in task	Individual, remote learning, remote	Logs	Expert rules	Stats	None
(Li et al., 2020)	person	Cognitive	Information processing behaviors	Individual, digital activity, lab	Logs	Clustering	Stats	None
(Rodriguez et al., 2021)	context/person-context	Behavioral	Activity in task, time management	Individual, remote learning, remote	Logs	Clustering	Stats	None
(Mills et al., 2021)	person-context/person	Cognitive	Mind wandering	Individual, digital activity, lab	Gaze	Self-report	SVM, naïve Bayes, RF	Self-explanation, rereading
(Chauhan et al., 2020)	person/person-context	Behavioral	Gaze alignment w/saliency map	Individual, remote tutoring, lab	Gaze	Annotation	NN	None
(Chorianopoulou et al., 2017)	context/person-context	Affective, Behavioral	Gaze at interaction partner, interact	Individual, structured interactions, home	Gaze (from video), audio	Annotation	SVM	None
(Hutt et al., 2021)	person-context	Cognitive	Mind wandering	Individual/cohort, intelligent tutoring system, lab/in situ	Gaze	Self-report	Bayes net	Reiteration, questions, personalization
(Hutt et al., 2019)	person-context	Cognitive	Mind wandering	Individual and cohort, intelligent tutoring system, lab	Gaze	Self-report	Bayes net	None
(Apicella et al., 2022)	person	Affective, Cognitive	Cognitive effort, cheerful/sad	Individual, digital activity, lab	EEG	Expert rules, music-induced	SVM, KNN, NN	None
(Khedher et al., 2019)	person/person-context	Cognitive	Time in EEG frequency bands	Individual, digital activity, lab	EEG	Expert rules	Stats	None
(Eldenfria & Al-Samarraie, 2019)	person	Cognitive	Time in EEG frequency bands	Individual, digital activity, lab	EEG	Expert rules	Stats	Change presentation of materials
(Gao et al., 2020)	context/person-context	Affective, Behavioral, Cognitive	Attention, enjoyment, self-reflection	Cohort, classroom activities, in situ	EDA, PPG, accelerometer, temperature, environment (CO2, humidity, ambient sound)	Self-report	LightGBM	None
(Monkaresi et al., 2017)	person	Affective, Behavioral, Cognitive	Broadly "were you engaged"	Individual, digital activity, lab	Face, heart rate	Self-report	Naïve Bayes, other classics	None
(Yue et al., 2019)	person	Affective, Behavioral, Cognitive	Task performance, attention	Individual, remote learning, remote	Face, gaze, logs	Self-report, expert rules	NN	None

The engagement continuum entries denote a location along the continuum, where for instance "context/person-context" corresponds to studies focused mostly on context-oriented engagement but also considered person-in-context to some extent. These entries were determined by the authors based on subjective assessment of each paper's focus on highly person-oriented research questions, in which the purpose was to learn about signals of individuals' engagement, to highly context-oriented questions, in which the purpose was to learn something about the engaging properties of the context itself. Most research projects fall between the extremes of the continuum, where they take contextual factors into account for predicting individual engagement or consider a mix of both contextual and person-oriented research questions. EEG = electroencephalography, EDA = electrodermal activity, CO2 = carbon dioxide levels.