Published in final edited form as:
Educ Psychol. 2017 ; 52(2): 104–123. doi:10.1080/00461520.2017.1281747.

Advanced, Analytic, Automated (AAA) Measurement of Engagement During Learning

Sidney D'Mello[*],
Departments of Psychology and Computer Science and Engineering, University of Notre Dame

Ed Dieterle, and
Summit Consulting, LLC, Washington, DC

Angela Duckworth
Department of Psychology, University of Pennsylvania

Abstract

It is generally acknowledged that engagement plays a critical role in learning. Unfortunately, the study of engagement has been stymied by a lack of valid and efficient measures. We introduce the advanced, analytic, and automated (AAA) approach to measure engagement at fine-grained temporal resolutions. The AAA measurement approach is grounded in embodied theories of cognition and affect, which advocate a close coupling between thought and action. It uses machine-learned computational models to automatically infer mental states associated with engagement (e.g., interest, flow) from machine-readable behavioral and physiological signals (e.g., facial expressions, eye tracking, click-stream data) and from aspects of the environmental context. We present15 case studies that illustrate the potential of the AAA approach for measuring engagement in digital learning environments. We discuss strengths and weaknesses of the AAA approach, concluding that it has significant promise to catalyze engagement research.

Keywords

engagement; measurement; machine learning; digital learning

In the popular 1999 Hollywood film *The Matrix,* the character Trinity learns to fly a helicopter in a matter of seconds by downloading the training program directly into her brain. Another character, Neo, learns Kung-Fu in much the same way. If only learning could be this efficient and effortless. Alas, most meaningful learning takes considerable time and effort (but see Shibata, Watanabe, Sasaki, and Kawato (2011) who appear to have made initial progress towards Matrix-style learning). It also requires sustained engagement, a point widely recognized by researchers, practitioners, and policy-makers (Loveless, 2015; PISA, 2012). Researchers have also made significant advances in conceptualizing *student engagement* or *academic engagement* as a complex multi-componential, multitemporal construct involving a diverse range of phenomena, such as momentary affective states of

[*]Correspondence should be addressed to Sidney D'Mello, Departments of Psychology and Computer Science and Engineering, University of Notre Dame, 118 Haggar Hall, Notre Dame, IN 46556. sdmello@nd.edu.

interest and enjoyment to long-term dispositions about school (Christenson, Reschly, & Wylie, 2012; Linnenbrink-Garcia & Pekrun, 2011; Sinatra, Heddy, & Lombardi, 2015). Unfortunately, methodological advances have lagged theoretical developments (Azevedo, 2015; Sinatra et al., 2015). Traditional measures of engagement include self-report questionnaires, experience-sampling methods, online observations, video coding, teacher ratings, and discourse analysis (Fredricks & McColskey, 2012; Henrie, Halverson, & Graham, 2015). Methodological advances have so far been limited to iterative refinement of traditional measures or combining methods (Greene, 2015). In our view, radical improvements require a qualitatively different measurement approach.

The digital revolution has fundamentally transformed how students engage in learning. In parallel, a new and exciting digital measurement approach is emerging as a viable complement to traditional measures. This approach uses *advanced* computational techniques for the *analytic* measurement of fine-grained components of engagement in a fully *automated* fashion. This advanced, analytic, and automated (AAA) measurement approach is theoretically-grounded in the embodied affective and cognitive sciences, while its methodological footing stems from the fields of digital signal processing and machine learning. The AAA approach espouses measures that are fine-grained and contextually coupled with unfolding learning events, so these measures can answer questions about *why* a learner is engaged, *what* an engaged interaction looks like, and *how* engagement changes over time. This information, in turn, can be used to develop interventions that dynamically respond to periods of waning engagement, thereby facilitating change in tandem with measurement.

We believe that AAA-based measures fill a critical gap in educational measurement. Contributors to a recent special issue of *Educational Psychologist* on “The Challenges of Defining and Measuring Student Engagement in Science” highlighted the need for new and innovative measures of engagement, especially micro-level measures to complement existing macro-level measures. For example, in their introductory article, the guest editors Sinatra et al. (2015) noted, “Also absent [from the special issue] are studies using more micro-level analyses of engagement such as eye tracking, physiology measures, and even brain imaging work” (p. 15). Such measures have been in development for over a decade in specialized research areas (e.g., affective computing and augmented cognition) that might be unfamiliar to most educational psychologists. However, an interdisciplinary approach is precisely what is needed to catalyze innovation in measurement of a complex construct like engagement. This point is aptly made by Azevedo (2015) in his commentary on the special issue and his perspective on the future of the field: “It is important to explicitly highlight that the first path [to develop an overarching and unifying theoretical framework to account for the majority of critical elements of the construct] is challenging and that many researchers may not be willing to pursue it for a variety of reasons (…). Such a challenge will require *interdisciplinary* research efforts currently witnessed in several fields” (p. 88). We respond to this call to action by providing an accessible introduction, selective review, and analysis of the AAA measurement approach that has emerged at the intersection of the psychological and computing sciences.

What is Engagement?

A scientific definition of engagement remains elusive. Reschly and Christenson (2012) note that the term *engagement* has been used to describe diverse behaviors, thoughts, perceptions, feelings, and attitudes, and at the same time, diverse terms have been used by different authors to refer to similar constructs. Theorists generally agree that engagement is a multidimensional construct, although the number and nature of the dimensions are unclear. Fredricks, Blumenfeld, and Paris (2004) proposed three components of engagement. *Emotional engagement* encompasses feelings and attitudes about the learning task or learning context, such as feelings of interest towards a particular subject, teacher (Renninger & Bachrach, 2015), or general satisfaction about school. *Behavioral engagement* broadly refers to learners’ participation in learning, including effort, persistence, and concentration. C*ognitive engagement* pertains to learners’ investment in the learning task, such as how they allocate effort toward learning, and their understanding and mastery of the material.

Reeve and Tseng (2011) recently suggested a fourth dimension: *agentic* engagement, characterized by learners proactively contributing to the learning process. Alternatively, Pekrun and Linnenbrink-Garcia (2012) proposed a five component model that includes cognitive (e.g., attention and memory processes), motivational (e.g., intrinsic and extrinsic motivation), behavioral (e.g., effort and persistence), social-behavioral (e.g., participating with peers), and cognitive-behavioral (e.g., strategy use and self-regulation) aspects of engagement.

We can trace the diverse components of engagement to different theoretical traditions. Theories of motivation, including self-determination theory (Deci & Ryan, 1985; Ryan & Deci, 2000), expectancy-value theory (Eccles & Wigfield, 2002), and self-efficacy theory (Bandura, 1986, 1997; Schunk & Pajares, 2005), focus on precursors of engagement, such as self-efficacy, interest in and value of a learning activity, autonomy, and the alignment between skill and challenge. Cognitive theories focus instead on the extent to which the learning activity engages the cognitive system (Eastwood, Frischen, Fenske, & Smilek, 2012). For example, the Interactive-Constructive-Active-Passive (ICAP) framework proposes four levels of cognitive engagement based on the level of interactivity afforded by the learning activity (Chi & Wylie, 2014). The levels, in decreasing order of expected engagement and learning, are Interactive (e.g., reciprocal teaching), Constructive (e.g., self-explanation), Active (e.g., verbatim note taking), and Passive (e.g., viewing a lecture). Author (year) extend ICAP to ICAP-A (attention) by suggesting that attentional control follows a similar pattern in that learners will maximally attend to interactive tasks and minimally to passive tasks (i.e., I > C > A > P). Finally, affective theories, including the control-value theory of academic emotions (Pekrun & Linnenbrink-Garcia, 2012), the assimilation-accommodation framework (Fiedler & Beier, 2014), and discrepancy-interruption and goal appraisal theories (Author, year; Mandler, 1990; Stein & Levine, 1991) emphasize the role of physiological arousal and cognitive appraisal in triggering emotions during learning and on the influence of affect on cognition and instrumental action.

Thus, engagement has emerged as a broad and complex construct pertaining to diverse aspects of the educational experience (e.g., showing up, completing homework, feelings of belongingness, graduating) and across multiple time scales (e.g., momentary affective episodes, stable dispositions such as general disengagement with school, and life-altering outcomes like dropping out of school). As Eccles and Wang (2012) note, these broad all-encompassing definitions make the construct more accessible for policy-makers and the educated lay person, but less useful for scientific research where precise definitions are of greater value, especially when it comes to elucidating cause and effect relationships. Thus, measuring "general" engagement might be as theoretically diffuse as measuring "cognition" or "emotion." It may be more fruitful to study specific aspects of this complex construct with an eye for broader assimilation across measures.

In this vein, Sinatra et al. (2015) conceptualize engagement along a continuum, anchored by person-oriented perspectives at one extreme, context-oriented at the other, and person-in-context in between. Person-oriented perspectives focus on the cognitive, affective, and motivational *states* of the student at the moment of learning and are best captured with fine-grained physiological and behavioral measures (e.g., electrodermal activity, facial expressions, actions). The context-oriented perspective emphasizes the environmental context as the analytic unit. Here, the focus is on macro-level structures like teachers, classrooms, schools, and the community, rather than the individual student. Finally, the intermediate-grain size, person-in-context perspective conceptualizes engagement at the level of the interaction between student and context (e.g., how students interact with each other or with technology).

We adopt a multi-componential perspective. For this we operationalize engagement in terms of affective states, cognitive states, and behaviors that arise from interactions with the learning environment. We conceptualize *engagement* as a goal-directed state of active and focused involvement in a learning activity. It is temporally constrained in that we are concerned with the state (not trait) of engagement across micro-level time scales ranging from seconds to minutes. Thus, our operationalization of engagement, and the AAA measurement approach derived from it, aligns with the person-oriented level of analysis of Sinatra et al. (2015). We should clarify that the term person-oriented does not imply that engagement is stable over time; rather, it refers to a micro-level analysis centered on the thoughts, feelings, and behaviors that emerge from a person's interaction with his or her environment. It is also distinct from a person-in-context level of analysis because the focus is on the *person* rather than his or her *interaction* with the environment.

Contemporary Engagement Measures

The most widely used measures of engagement are self-report questionnaires; see Fredricks and McColskey (2012); Greene (2015); Henrie et al. (2015) for reviews. Although relatively inexpensive, easy to administer, and generally reliable, questionnaires have well-known limitations (Author, year; Krosnick, 1999). For instance, when endorsing items, respondents must compare the target (e.g., a teacher rating a student, a student rating himself or herself) to some implicit standard, and standards may vary from respondent to respondent. To one student, "I am a hard worker" may be exemplified by doing five hours of homework each day; for others, the same statement may be exemplified by simply showing up for class. For both informant-report and self-report questionnaires, biases that arise from heterogeneous frames of reference reduce validity (Heine, Lehman, Peng, & Greenholtz, 2002). For self-report questionnaires, social desirability bias is another important limitation (Krosnick, 1999), both when respondents aim to appear admirable to others and also when they inflate responses to preserve their own self-esteem. Likewise, memory recall limitations and acquiescence bias can influence self-reports, and halo effects can influence informant-reports (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003).

Several non-questionnaire engagement measures have also been developed. Examples include experience-sampling methods (ESM) (Csikszentmihalyi & Larson, 1987), day reconstruction (Kahneman, Krueger, Schkade, Schwarz, & Stone, 2004), and interviews (Turner & Meyer, 2000). However, because they still rely on self- and informer-reports, they are subject to similar biases as questionnaires.

Observational methods are an attractive alternative to self- and informer-reports because they are arguably more objective (Nystrand & Gamoran, 1991; Pianta, Hamre, & Allen, 2012; Renninger & Bachrach, 2015; Ryu & Lombardi, 2015; Volpe, DiPerna, Hintze, & Shapiro, 2005). Unfortunately, these methods entail considerable human effort, which might not be a major limitation for small scale studies, but poses a significant challenge for repeated long-term measurement at scale. Further, observations cannot be conducted in some learning contexts, such as students’ homes.

Researchers have attempted to circumvent some of the limitations of observational methods by combining automated data collection with semi-automated or manual data coding. For example, the Electronically Activated Recorder (EAR) is a device that randomly samples audio clips in naturalistic environments (Mehl, Pennebaker, Crow, Dabbs, & Price, 2001). Data collection with the EAR is efficient and cost-effective; however, the data still need to be transcribed and coded by humans, which increases cost and reduces scalability. Similarly, engagement can be coded from videos by researchers (Author, year) or even teachers (Author, year), but video coding is a labor- and time-intensive effort.

Finally, engagement can be adduced from academic and behavior records, such as homework completion, absences, achievement test scores, and teacher ratings of classroom conduct (Lehr, Sinclair, & Christenson, 2004; Skinner & Belmont, 1993), but these measures are limited in what they can reveal about engagement at the micro-analytic level espoused here.

The Advanced, Analytic, Automated (AAA) Measurement Approach

An AAA-based measure provides continual assessments of person-oriented components of engagement at a fine-grained temporal resolution, all with no human involvement. These measures have several advantages over counterparts. They are uniquely suited to track person-oriented components of engagement since they operate at fine-grained time scales ranging from seconds to a few minutes. They are more objective because computers provide the measurements, thereby partially obviating reference, social desirability, acquiescence, and other biases associated with self- and observer-reports. AAA-based measures are also unaffected by momentary lapses in attention or by fatigue, as can occur with humans. They vastly reduce time and effort, which is a limitation of ESM, day-reconstruction, video coding, and observations.

In this paper we introduce the theoretical and methodological foundation of the AAA approach, highlight exemplary AAA-based measures, and analyze the approach and measures derived from it. To keep the scope manageable, we emphasize measures that are nonintrusive, cost-effective, and are usable in the near-term. These include analyzing machine-readable aspects of a learning session, such as log files recorded during interactions with digital learning environments, facial features, eye gaze, and physiology. Several of these signals have a long history in the psychological sciences, including the measurement of cognitive engagement (Miller, 2015). However, they have mainly been used as passive data sources that humans analyze offline. The AAA approach stands apart because it combines machine-sensing and machine-analysis to provide measurement that is real-time and fully-automated.

Theoretical and Methodological Foundations

We ground the AAA measurement approach in the aforementioned person-oriented operationalization of engagement as the momentary affective and cognitive states that arise throughout the learning process. Embodied theories of cognition and affect posit that these mental states manifest in the body in multiple ways because cognition and affect are in the service of action and bodies are the agents of action (Barsalou, 2008; deVega, Glenberg, & Graesser, 2008; Ekman, 1992; Niedenthal, 2007; Russell, Bachorowski, & Fernandez-Dols, 2003). For example, there is increased activation in the sympathetic nervous system during fight or flight responses (Larsen, Berntson, Poehlmann, Ito, & Cacioppo, 2008). Similarly, there are well-known relationships between facial expressions and affective states (Ekman, 1984; Keltner & Ekman, 2000; Matsumoto, Keltner, Shiota, O'Sullivan, & Frank, 2008), for example the furrowed brow during experiences of confusion (Author, year; Darwin, 1872). Researchers have also identified bodily/physiological correlates of cognitive states like attention and cognitive load. Eye movements are an invaluable tool to investigate visual attention due to the so called *eye-mind* link (Deubel & Schneider, 1996; Hoffman & Subramaniam, 1995; Rayner, 1998), while electroencephalography (EEG) can index mental workload via a *brain-mind* link (Berka et al., 2007). The mind-body link suggests that observable bodily responses can be used to *infer* unobservable mental states, which is at the heart of the AAA measurement approach.

Here's the basic assumption: cognitive and affective states reflecting different components of engagement are associated with responses at multiple levels (neurobiological, physiological, bodily expressions, overt actions, metacognitive, and subjective), which in turn influence the states themselves in a form of circular causality (Lewis, 2005). Some of these responses are implicit (e.g., neurobiological, some physiological changes) in that they occur outside of conscious awareness, while others are more explicit (e.g., metacognitive reflections, subjective feelings). The states are modulated by individual differences as well as contextual, social, and cultural influences (Elfenbein & Ambady, 2002; Kappas, 2013; Mesquita & Boiger, 2014).

Some of these responses are detectable by machine sensors and human observers, but others are only accessible to the self. In particular, external observers only have access to visible behaviors (e.g., facial expressions, gestures, actions), information on the environmental context, and physiological changes (e.g., respiration rate), and must rely more heavily on inference to decode a person’s mental state (Mehu & Scherer, 2012). In contrast, the self has privileged access to subjective feelings, memories, meta-cognitive reflections, and some physiological changes, but not to other responses (e.g., involuntary expressions and neurobiological changes).. Machine sensors can measure neurobiological, bodily, physiological, and action-oriented responses beyond what is available to humans (e.g., thermal cameras, electroencephalogram), but they cannot infer the mental state from the measurements nor can they interpret contextual cues on par with humans. Thus, the core problem faced by machines is to *infer* the latent mental states associated with engagement (e.g., concentration, interest) from machine-readable signals and from aspects of the environmental context.

AAA measurement begins when sensors record low-level *signals*. Signals are then processed to obtain high-level abstractions, called *features*. For instance, a video is the signal recorded from a web-cam (the sensor). Sample features, computed by applying computer vision techniques to video, include activations of specific facial muscles (also called action units, such as inner brow raise or lip pucker (Ekman & Friesen, 1978)), facial textures, and head position and orientation (Pantic & Patras, 2006; Valstar, Mehu, Jiang, Pantic, & Scherer, 2012). Similarly, digital signal processing techniques in the speech domain (Eyben, Wöllmer, & Schuller, 2010) are used to extract paralinguistic (also called acoustic-prosodic) features such as pitch and amplitude from an audio signal recorded with microphones (the sensor). Researchers can also use this paradigm to analyze spoken content; in this case they’ll leverage automatic speech recognition and natural language understanding techniques. In general, signal processing methods (denoising, filtering, smoothing, feature extraction, etc.) are required to compute features from the raw signals (see Author, year; St. John, Kobus, Morrison, and Schmorrow (2004) for details on these methods).

The next step in the AAA measurement approach entails inferring mental states from the corresponding features. This is done with machine learning, which prescribes methods to *learn a program* (or computational model) from *data* (Domingos, 2012). Machine learning has many subfields, of which supervised learning is most widely used in the AAA approach. Supervised learning (see Figure 1) requires *training data,* consisting of features (extracted from signals recorded by sensors as noted above) along with temporally synchronized *annotations* of mental states (e.g., from self-reports or observer judgments), collected at multiple points in a learning session. In a *training* phase, supervised learning methods automatically *model* (learn) the relationship between the features and human annotations to yield a computational model. The degree of overlap between the model-generated and human-provided annotations is assessed in a *validation* phase. The model can then take sensor data collected at some future time and/or from a new set of students and automatically generate estimates of mental states *without* needing human annotations.

The computational model can take on many forms depending on the supervised learning method. Selecting a computational model is a design decision with multiple tradeoffs -

separability of feature space (i.e., data), transparency of internal representations, accuracy, generalizability, computational efficiency, robustness to noisy data, and others not discussed here. Hence, in contemplating the question of how to select an *appropriate computational model?* it is prudent to first ask, appropriate for *what purpose*?

One important factor involves the linear separability of the data; i.e., whether the different classes (e.g., bored vs. curious vs. confused) as represented in feature space can be discriminated by linear functions, such as lines for two-dimensional data or hyperplanes for higher-dimensional data. Linear models are attractive in their simplicity, but are ineffective when the data is non-linearly separable, which is usually the case. These situations require more sophisticated models; for example, support vector machines transform a non-linearly separable feature space into a linearly separable space by projecting it into higher dimensions (Cortes & Vapnik, 1995).

The added sophistication does incur a price, especially for some of the more complex models which have internal representations that are not inspectable. This so called "black box" problem is a frequent critique of machine learning. Although the concerns are valid for some models (e.g., neural networks), other models are much more transparent - for instance, models that operate by rule induction (e.g., If blink rate is high and heart rate is low then Boredom = high), organize rules into decision trees, or compute conditional probabilities of mental states given features (e.g., Probability [Boredom | {Blink rate = high and Heart rate = low]).

In most cases, it is sufficient to select models with inspectable representations and with sufficient performance. However, priorities might shift when models are intended for real-time measurement, such as to trigger technological interventions aimed at re-engaging a bored learner (Author, year). Here, computational efficiency (both in terms of clock time and computational resources) and robustness (in the face of noisy or missing data) might take precedence over transparency and performance.

This leads to another issue: how to quantify performance. Given that the goal is to use the model to provide *accurate* estimates of engagement on *unseen* data, two key performance metrics are accuracy and generalizability. Accuracy (similar to convergent validity) is measured as the alignment between automated estimates and an external standard, typically self- or observer-annotations. The alignment can be quantified by a number of standard metrics (e.g., recognition rate, kappa, correlation). 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).

Generalizability is concerned with the robustness of the model when applied to data beyond what was used to train the model. It is usually established by dividing the data into two sets (A and B), training the model on one set (A or B), and testing it on its complement (B or A). Cross-validation is a widely used variant of this procedure in which each set serves as training and testing sets across multiple *folds*. For example, in 3-fold cross-validation, the data is divided into three sets A, B, and C, and folds are created as follows: Fold 1: train A and B, test C; Fold 2: train A and C, test B; Fold 3: train B and C, test A. This method ensures that every data point is tested at least once.

The level of generalizability achieved depends on the data and how the folds are constructed. *Instance-level validation* ensures that individual cases are either in the training or testing set, but data points (albeit different ones) from the same person can be in both sets. The resulting model risks over-fitting to individual characteristics and may not generalize to new people. In contrast, *person- or student-level validation* ensures that data from the same person are either in the training or testing set but never both. This provides more confidence that the model will generalize to new people with *similar* characteristics. In *population-level validation,* the data are split on some population characteristic (e.g., gender) and tested on its complement (e.g., train males and test females, and vice versa).

Case Studies

We now discuss representative case studies featuring the AAA approach to measure person-oriented components of engagement during learning with technology. We have selected 15 studies to emphasize key dimensions of the measurement approach, including sensor-free vs. sensor-based measurement, annotations by the self vs. external observers, unimodal vs. multimodal sensing, lab vs. classroom research, learning activities with varying levels of interactivity, and different validation methods. We prioritized studies that can be considered as pioneering in the field, such as the first study showcasing multimodal engagement measurement in real-world classrooms (Arroyo et al., 2009), the first study emphasizing generalizability beyond the individual (Ocumpaugh, Baker, Gowda, Heffernan, & Heffernan, 2014), or the first person-independent automated measure of mind wandering (Author, year). We acknowledge that our choice of case studies is both subjective and incomplete, but our goal is to provide an overview of a promising new approach rather than review a well-established paradigm. We hope that the studies covered here will pique interest and inspire further inquiry into AAA-based measures.

Table 1 provides an overview of the studies. Despite the considerable variability, each study followed the basic approach discussed above and summarized in Figure 1. Step 1 consists of recording signals (video, physiology, log files, etc.) as students complete a learning activity within a particular learning context (Step 1a) followed by computing features from the raw signals (Step 1b). In Step 2, annotations of mental states reflecting various components of engagement are obtained, from the students themselves, from external observers, or via some other method (see Author (year) for a review of methods to annotate mental states in learning contexts). In Step 3, supervised learning methods computationally model the relationship between the features and temporally synchronized annotations. In Step 4, the resulting model produces computer-generated engagement estimates that are compared to human-provided annotations for validation.

In the interest of brevity, we discuss eight case studies below and present the remaining seven in Supplementary Material A. We organize the case studies by sensors used. *Sensor-free* measures analyze digital traces recorded in log-files while sensor-based measures use physical sensors. We further categorize the sensor-based measures as *sensor-light* if they use sensors that are readily available in contemporary digital devices (e.g., webcams, microphones) or *sensor-heavy* if they require nonstandard sensors like eye trackers, pressure pads, and physiological sensors (see Figure 2).

Accuracy metrics varied considerably across studies. Several studies reported recognition rate (RR), which is the proportion of cases where computer estimates match the human-provided annotations and is highly flawed when there is class skew (i.e., uneven division among categories being discriminated) (Hayes & Krippendorff, 2007). Some studies that made binary discriminations (e.g., mind wandering or not) reported the area under the receiver operating characteristic curve as the accuracy metric (Hanley & McNeil, 1982) (AU ROC or AUC or A-prime, or 2AFC [two-alternative forced choice]). The ROC curve is obtained by plotting true-positives vs. false-positives at various decision thresholds ranging from 100% true positives to 100% false positives (Hanley & McNeil, 1982). The AUC ranges from 0.5 (chance model) to 1.0 (perfect discrimination). Studies that attempted categorical discriminations (e.g., focused, anxious, neutral) reported recognition rate and/or Cohen’s kappa, a chance-corrected metric ranging from 0 (chance) to 1 (perfect accuracy) (Cohen, 1960). To ameliorate these differences, we provide the *percent improvement in accuracy above chance* as a rough metric that permits comparison and aggregation across studies (see footnote in Table 1).

Sensor-free Measures

A person interacting with a digital learning environment leaves a rich digital trace stored in log-files. The logs reflect student actions in response to system queries, system feedback, content covered, student preferences, and so on. Sensor-free measures analyze these log-files without utilizing sensors beyond standard input devices (e.g., keyboard, computer mouse, trackpad).

Interaction patterns and contextual cues from AutoTutor—Author (year) developed one of the first sensor-free measures of affective engagement. They collected training data in a lab study where 28 students completed a 32-minute tutoring session on computer literacy with a natural language intelligent tutoring system (ITS) called AutoTutor (Graesser, Chipman, Haynes, & Olney, 2005). AutoTutor mimics human tutors by posing challenging questions, using hints and prompts to elicit student responses, providing feedback and elaborations on their responses, and summarizing answers. Students type their responses in conversational English. AutoTutor uses natural language processing techniques to analyze the response and adapts the tutorial dialog based on its assessment of student progress.

The researchers used an offline video-coding procedure to identify momentary episodes of boredom, flow, confusion, frustration, delight, and neutral (no affect). Soon after the AutoTutor session, students viewed videos of their faces and computer screens recorded during the session and self-reported their affective states at 20-second intervals indexed into the videos. At a later time, an untrained peer and two trained human judges assessed student affect at the same points in the video. Agreement between the different judges varied widely

(Author, year), so a majority vote was taken to reflect the affective state at each judgment point.

AutoTutor maintains a log file that captures the student’s response, assessments of the response, the feedback provided, the tutor’s next move, reaction and response times, and so on. The researchers computed 17 features from the log files after temporally aligning the logs with the affect judgments. They built supervised learning models to discriminate each affective state from neutral (no affect). The models were validated using 10-fold instance-level cross-validation and achieved a mean recognition rate of 0.71, equivalent to a 41% improvement over chance.

This early study demonstrated the potential for measuring affective engagement from log-file data. However, the use of instance-level cross-validation and the lab-based data collection protocol reduced generalizability of the measure.

Interaction patterns from ASSISTments—ASSISTments is an ITS for middle- and high-school mathematics used by approximately 50,000 students in the Northeast U.S. (Razzaq et al., 2005). Pardos, Baker, San Pedro, and Gowda (2013) developed an AAA-based measure of cognitive and affective engagement for ASSISTments. They collected training data from 229 students who used ASSISTments in their school computer lab as part of their mathematics classes. Researchers made online observations (annotations) of students’ boredom, frustration, engaged concentration, and confusion using the Baker-Rodrigo Observation Method Protocol (BROMP) (Ocumpaugh, Baker, & Rodrigo, 2012), recently renamed the Baker-Rodrigo Observation *Monitoring* Protocol (Ocumpaugh, Baker, & Rodrigo, 2015). The observations were based on explicit actions towards the interface, interactions with peers and teachers, body movements, gestures, and facial expressions. Observers had to achieve a minimum kappa of 0.6 with a BROMP expert prior to making the observations.

The researchers focused on discriminating each affective state from the others (e.g., boredom vs. engaged concentration, confusion, and frustration) using features distilled from ASSISTments log files (e.g., performance on problems, hints, timing information). They achieved a 30% above-chance accuracy after averaging across the four affective states (mean of 0.68 measured with A-prime metric) using 5-fold student-level cross-validation.

Student-level cross-validation ensures generalizability to new students with *similar* demographics as those in the training set. Ocumpaugh et al. (2014) studied the measure’s ability to generalize to new students with *different* demographics by retraining the models on a more diverse data set encompassing urban, suburban, and rural students. The models yielded mean above-chance improvements of 18%, 16%, and 6% when tested on independent samples of urban, suburban, and rural students, respectively. Thus, the model appeared to generalize to urban and suburban, but not to rural students.

The team also provided some evidence of the models’ predictive validity. Pardos et al. (2013) showed that model-generated engagement estimates, obtained from log files of a *different* set of 1,393 students collected during the 2004–2006 school years (several years *before* the models were even developed), correlated with standardized achievement test scores. Subsequently, San Pedro, Baker, Bowers, and Heffernan (2013) showed that model-based estimates of confusion and boredom obtained from the log files of 3,707 students who used ASSISTments in the 2004 to 2009 years predicted college enrollment (as recorded in the National Student Clearinghouse (NSC, 2016)) several years later.

The ASSISTments studies are significant because they provide evidence that their AAA-based measure has a degree of population generalizability and predictive validity. They also demonstrate that it is feasible to retrospectively measure affective and cognitive components of engagement from log files collected years before the measures even existed.

Interaction patterns from Inq-ITS—Gobert, Baker, and Wixon (2015) developed an AAA-based measure for Inq-ITS, a computer-based learning environment to help students develop scientific inquiry skills. In Inq-ITS, students generate hypotheses of scientific phenomena, collect data via simulated experiments embedded in micro-worlds, and evaluate their hypotheses in light of collected data. Students have considerable agency in how they interact with Inq-ITS, which sometimes leads to unproductive behaviors that the researchers term *disengaged from the task goal* (DTG). They define DTG as "engaging with the task, but in a fashion unrelated to the learning task's design goals or incentive structure" (p. 48). For example, running an unusually large number of simulations with identical parameters would be considered DTG as each simulation would produce an identical result.

Researchers collected training data from 144 middle-school U.S. students who used Inq-ITS as part of their science classes. Two humans coded DTG from human-readable excerpts (called clips) of Inq-ITS log files which were presented in sequence to preserve contextual information. The coders achieved a kappa of .66 (deemed "acceptable agreement" by the authors) in classifying each clip as being an instance of DTG or not.

Features, such as the total number of actions, time between actions, duration of the longest pause, and number of simulations run, were computed from each clip. Using 6-fold student-level cross validation, the researchers obtained a 41% improvement above chance accuracy (A-prime [similar to AUC as noted above] of .81) in discriminating DTG from non-DTG clips. This study is significant for its nuanced conceptualization of disengagement as DTG rather than being merely distracted or off-task.

Sensor-based Measures

The measures below used one or more physical sensors in isolation or in tandem with interaction logs to measure specific components of engagement.

Facial features during cognitive skills training—Whitehill, Serpell, Lin, Foster, and Movellan (2014) automatically measured behavioral engagement from videos. They collected training data from 34 undergraduate students who interacted with an in-house cognitive skills training software system on an Apple iPad. The researchers recorded video with a commercial webcam aimed directly at students' faces and used computer vision techniques to detect facial expressions (Littlewort et al., 2011) and facial textures. Trained coders annotated the videos for behavioral engagement using the following ordinal scale: (1)

not engaged at all (2) nominally engaged (3) engaged in tasks and (4) very engaged. Machine-learned models that discriminated each engagement level from the others (e.g., level 4 vs. levels 1, 2, and 3) using the facial features yielded an average 31% above-chance accuracy (2AFC [two-alternative forced-choice] of .73) using 4-fold student-level cross-validation.

The model’s estimate of behavioral engagement (level 4 vs. levels 1, 2, and 3) correlated (r = 0.27) with performance gains assessed before and after training, providing some evidence for its predictive validity. Further, the researchers tested the model’s population generalizability by retraining on the 26 African American students who were recruited from a historically Black College/University and testing on the eight Asian-American or Caucasian-American students. This yielded an average 23% above-chance improvement, which implies some, but not severe, degradation compared to training and testing on the combined sample (above-chance improvement of 31%).

Body movements and posture while playing Fripples Place—Mota and Picard (2003) developed a measure of interest (an affective component of engagement) while ten 8–11-year-old children interacted with a challenging constraint satisfaction game called *Fripples Place*. They tracked body movements with the Tekscan™ Body Pressure Measurement System (BPMS), which consists of a thin-film pressure pad (or mat) that can be mounted on a variety of surfaces. Videos of the children’s faces and computer screens were also recorded. Three teachers annotated the videos for high, medium, or low interest, taking a break, bored, and “other,” and achieved an average kappa of .79. Due to insufficient bored and other annotations and difficulties in discriminating medium interest from high or low interest, the researchers focused on discriminating among high interest, low interest, and taking a break.

They first used a neural network to classify each pressure map into one of nine static postures (e.g., leaning back, sitting upright), achieving an accuracy of 87.6%. Next, Hidden Markov Models discriminated among the interest levels from 3-second sequences of static postures. Validating the model on two held-out students (student-level validation) yielded a recognition rate of 0.77 (above-chance improvement of 61%). Although replication with a larger sample is warranted, the finding that interest can be detected from posture alone is significant.

Eye gaze and contextual cues during computerized reading—Mind wandering (or zoning out) occurs when attention drifts away from the learning task to task-unrelated thoughts (Smallwood, McSpadden, & Schooler, 2008). It is a critical indicator of cognitive engagement, but poses a unique challenge for the AAA approach as it is a deeply internal state that might not be perceivable by observers.

Author (year) addressed this challenge by using eye gaze to measure mind wandering during computerized reading. They collected data in a lab study where 178 undergraduate students from two U.S. universities read four instructional texts on scientific research methods. Participants’ eye gaze was recorded with Tobii T60 and TX 300 eye trackers (one at each university). Participants self-reported mind wandering by responding “yes” or “no” to an auditory probe (i.e., a beep) triggered on pseudo-random pages (screen of text), 4 to 12 seconds from the time the page appeared.

The researchers extracted two sets of gaze features across windows of variable length (3 seconds, 5 seconds, etc.) immediately preceding each thought probe. Global features were independent of the words read and focused instead on general eye gaze patterns, such as the number of fixations, fixation durations, variability in fixation durations, and saccade lengths. Local features were sensitive to the words read, such as the length of fixations of certain types (e.g., first fixation on a word vs. re-fixating on a word previously read). The gaze features were complemented by contextual features that encoded general information about the reading context, such as the current page being read, reading time, and text difficulty.

Supervised learning models discriminated the presence or absence of mind wandering from the gaze and context features. When validated by randomly sampling students into training and testing sets (student-level validation), the models yielded a mean recognition rate of 0.70 (25% improvement over chance). Interestingly, the model-based mind wandering rates more strongly correlated with learning (*Spearman’s rho* = −.33) and transfer (*rho* = −.21) than self-reported mind wandering (*rhos* of −.07 and −.12, respectively) after controlling for prior knowledge.

Author (year) used a variant of the model to trigger real-time interventions based on predicted mind wandering. When evaluated on a new sample of 104 participants, model-predicted mind wandering rates negatively correlated with performance on comprehension questions interspersed during reading (*r* = −.30) as well as on a subsequent posttest (*r* = −.32). These results suggest the possibility of objectively measuring a highly internal component of attentional engagement in real-time and in a manner that generalizes to new students.

Interaction features, facial features, skin conductance, mouse pressure, and body movement during learning with Wayang Outpost—Arroyo et al. (2009) developed the first sensor-based engagement measure for use in computer-enabled classrooms. They collected interaction features from log files, facial features from video, body movements from pressure pads, skin conductance from a wrist sensor, and pressure exerted on a pressure-sensitive mouse. These data streams were collected from 38 high-school and 29 female undergraduate students (potential elementary school teachers) who used the Wayang Outpost geometry ITS as part of regular mathematics instruction over the course of 4–5 days (Arroyo, Beal, Murray, Walles, & Woolf, 2004). Students were prompted to self-report levels of interest, confidence, excitement, and frustration on 1 to 5 point scales at 5-minute intervals and after completing each problem.

The researchers extracted a number of features from their sensor suite. Examples include facial expressions and head movements, postures such as leaning forward and movement variability, amount of physiological arousal, and pressure exerted on the mouse sensor. The sensor-based features were supplemented with interaction features, such as number of hints viewed and time spent interacting with the tutor. Multiple linear regression models were used to predict self-reported levels of each state from different combinations of features. The best results (average R^2 of 0.47) were obtained by combining the facial and interaction features.

This result should be interpreted with some caution. Models were trained on very small samples (between 20 to 36 cases) because valid data were only obtained for about 50% of the students due to complications with the use of sensors in classrooms. Further, generalizability is unclear as the models were not validated with a separate testing set. Despite these caveats, this study is pioneering because it incorporated sensors in an authentic classroom environment.

Facial features, body moments, and interaction patterns while playing Physics Playground—Author (year) also developed an AAA-based measure for use in computer-enabled classrooms. They collected training data from 137 8th and 9th grade U.S. students during interactions with a physics educational game called Physics Playground (Shute, Ventura, & Kim, 2013). Students played the game in two 55-minute sessions across two days. Trained observers performed live annotations of boredom, engaged concentration, confusion, frustration, and delight using the BROMP field observation protocol as in the ASSISTments study discussed above (Pardos et al., 2013). The researchers also recorded videos of students' faces and upper bodies, which were synchronized with the affect annotations.

The videos were processed using FACET, a computer-vision program (Emotient, 2014) that estimates the likelihood of 19 facial action units along with head pose and position. Body movement was also estimated from the videos using motion filtering algorithms (Author, year). Supervised learning models were trained to discriminate each affective state from the others (e.g., boredom vs. confusion, frustration, engaged concentration, and delight), and were validated by randomly sampling students into training and testing sets (student-level validation). The models yielded an average accuracy of 0.69 (measured with the AUC metric), which reflects an approximate 37% improvement above chance. Follow-up validation analyses confirmed that the models generalized across multiple days (i.e., training on subset of students from day 1 testing on different students from day 2 and vice versa), class periods (i.e., training on a random five of the seven class periods, testing on the remaining two, and repeating across multiple iterations), genders (i.e., training on males, testing on females and vice versa), and perceived ethnicity (i.e., coded by humans as no demographics were available).

Measures derived from video-based facial feature tracking are limited in that they can only be used when the face can be automatically detected in the video. This was about 65% of the time in the current study due to occlusions, gestures that mask the face, poor lighting, and other complicating factors. To address this, Author, (year) developed a second AAA measure that utilized contextual and interaction features stored in log files, such as the difficulty of the current game level, students' actions, the feedback received, and response times. Then, logistic regression models were trained to adjudicate between the estimates of the face- and interaction-models, essentially weighting their relative influence on the final outcome. The multimodal models were almost as accurate as the video-based models but could be used 98% of the time (compared to 65% for face-only models). This is notable given the noisy nature of the classroom environment with students incessantly fidgeting, talking with one another, and occasionally using their cellphones despite it being against classroom policy.

General Discussion

Advancing the scientific study of engagement requires advancing measurement. We propose an advanced, analytic, automated (AAA) approach that capitalizes on the proliferation of digital learning environments. In Supplementary Materials B, we compare the AAA approach with self-report questionnaires and online observations as these are two of the most popular traditional measurement approaches. In what follows, we analyze the AAA approach beginning with how engagement has been defined.

How have the AAA-based measures defined engagement?

The lack of consensus on how to define engagement within educational psychology (Reschly & Christenson, 2012) is mirrored in the AAA-based measures. The case studies reviewed operationalized engagement in a myriad of ways, including engaged concentration, engagement/flow, zone-outs, mind wandering, interest, boredom, focus, and disengagement from task goals. Zone-outs and mind wandering align with the cognitive component of engagement, positive affect and interest with the affective component, on-task behaviors with the behavioral component, while engagement/flow and engaged concentration are more holistic amalgamations. Several studies also considered affective states that are associated with cognitive engagement (e.g., confusion and frustration) or lack thereof (e.g., disengagement from task goals). While the approach has been to focus on individual components, future work should consider modeling multiple components (e.g., interest, mind wandering, on-task behaviors) in order to investigate interdependencies and to more accurately capture an inherently multicomponential construct.

Relatedly, while the AAA-based measures tended to focus on the presence (1) or absence (0) of the mental *states* at each time point, their fine-grained temporal resolution affords a *process-level* account of engagement, disengagement, and re-engagement. For example, binary interest (1) or disinterest (0) estimates obtained every 10 seconds could be aggregated across longer time periods (i.e., every two minutes) to construct time series that reflect moments where interest was first captured, periods of maintained interest, when interest appears to diminish, when it is re-captured, and so on (Hidi & Renninger, 2006; Renninger & Bachrach, 2015). When multiple components are modeled, time series analyses can illustrate how individual components interact over time. For example, cross-correlational analyses of interest and mind wandering over time can provide an array of insights: lag-lead relationships among the two, how these relationships unfold over time, and how they influence behaviors and other mental states. Thus, when coupled with multi-componential measurement, the fine-grained temporal resolution of AAA-based measures holds considerable promise to model engagement as a dynamic process with components that interact over time.

How accurate are the AAA-based measures?

Convergent validity is assessed by computing agreement between computer-generated and human-provided estimates of engagement components. The measures we reviewed yielded an average 39% improvement in agreement over chance (95% confidence interval of 31% to 47%). To put this score in context, recall that the goal is to *infer* covert mental states from overt bodily signals and aspects of the environmental context. The inference process is rife with ambiguity at multiple levels. First, there are weak relationships between mental states and bodily signals; for example, reviews and meta-analyses on correlations between facial expressions and self-reported emotions have yielded small to medium sized effects for spontaneous expressions (Camras & Shutter, 2010; Ruch, 1995; Russell et al., 2003). Second, other than the rare case of extreme or prototypical emotions, there are weak relationships among the various bodily signals, perhaps because different situations demand different bodily responses (Barrett, 2006; Larsen et al., 2008; Russell, 2003). Third, the machine learning approach relies on supervision in the form of annotations provided by humans, which introduces additional errors that permeate the learned models and computation of accuracy metrics.

On account of these factors, there is unlikely to be a magical device that can “read-out” covert mental states from behavior and physiology. Contrary to popular belief, humans are not all that accurate. In fact, the average 39% above-chance improvement obtained by the AAA-based measures is identical to the average 39% above-chance inter-rater agreement reported in an analysis of 14 studies involving fine-grained coding of affective (and some cognitive) states by external observers (Author, year). Further, agreement between external observers and self-reports is very low (Author, year) and frame-of-reference training (Bernardin & Buckley, 1981) increases agreement between observers but not between the self and an observer (Author, year).

To summarize, measuring mental states associated with engagement involves inference, and inference involving complex psychological constructs involves a degree of error. This is irrespective of whether the inference is performed by a human or a computer, and in some cases the computer even outperforms humans (Bartlett, Littlewort, Frank, & Lee, 2014). Researchers increasingly appreciate that the underlying cognitive-affective-bodily system is not a rigid, deterministic machine that unfailingly produces the same outputs for a given set of inputs; rather, the system is loosely coupled and dynamic, and it rapidly self-organizes to external influences (Author, year; Camras & Witherington, 2005; Coan, 2010; Lewis, 2005). If the system itself produces “beyond-chance probabilistic” output (Roseman, 2011 p., 440), then so will attempts to computationally model it.

What are some key limitations of AAA-based measures?

The measures have a few key limitations. First, the supervised learning method used to develop the measures requires mental state annotations from learners themselves or external observers. The two sources (self vs. observers) have access to different types of information and are influenced by different biases (see Introduction), resulting in low agreement (Afzal & Robinson, 2011; Author, year). This in turn leads to less reliable measures as well as reduced confidence in their validation. The best approach would be to incorporate annotations from multiple sources to reduce error, but most studies rely on either the self or observers–seldom both.

Second, validity of AAA-based measures needs to be established more precisely. Researchers have focused on convergent validity at the expense of other forms of validity. Discriminant validity of AAA-based measures is not well studied, nor is predictive and external validity (generalizability). In only three of the case studies (Author, year; Ocumpaugh et al., 2014; Whitehill et al., 2014) did researchers show that the measures predict meaningful outcomes, such as learning gains and college enrollment. Similarly, although 11 of the studies addressed generalizability to new students, only three (Author, year; Pardos et al., 2013; Whitehill et al., 2014) considered additional forms of generalizability, such as generalizability across time and across student demographics.

Third, because the underlying computational techniques are compromised by noisy data, robustness is another aspect that needs more attention. For example, changes in background illumination complicate face detection, seemingly benign behaviors like face-occluding gestures or chewing gum pose a challenge for facial feature tracking, glasses or contact lenses reduce precision of some eye-trackers, and speech and background noise hinder automatic speech recognition. Robustness to noise has received insufficient attention because missing or noisy data are usually discarded prior to the modeling process (e.g., Author (year)). A different approach is needed for measures intended for real-time use in ecological settings, where noisy data is more the norm than the exception (Arroyo et al., 2009; Author, year). Robustness to noisy data needs to be a fundamental design constraint rather than an afterthought, and in some cases systematically modeling rather than discarding noisy data can improve measurement accuracy (Author, year).

Fourth, there are privacy concerns for measures that use biometric signals (e.g., an image of a face, an audio sample). Some sensors can also inadvertently record compromising background information, as in the recent “WebcamGate” scandal (Martin, 2010). An effective strategy to protect privacy is to only retain non-identifiable features from the signals while discarding the signals themselves. There are also ethical concerns with respect to how the measures are used. We strongly recommend *against* their use for teacher or student evaluation since the measures are imperfect and engagement is influenced by factors out of the control of teachers and students. At this time, the measures are best suited for basic and applied research on engagement. They can also be used to improve learning technologies, either by passively measuring periods of sustained disengagement for retrospective review and refinement (e.g., Miller, Petsche, Baker, Labrum, & Wagner, 2014) or by dynamically re-engaging disengaged students (Author, year).

Finally, whereas self-report questionnaires can be used at scale assuming computer-based administration and scoring, the scalability of the AAA-based measures varies as a function of equipment cost, the sensors used, availability of computer/internet access for students, privacy/ethical considerations, and technical expertise needed to deploy the measures with sufficient fidelity. Sensor-based measures are currently limited to small-scale IRB-approved research studies conducted by interdisciplinary teams of psychologists, computer scientists, and education researchers (e.g., Author, year; Wang et al., 2014). We anticipate that these measures will be more scalable as computing becomes increasingly ubiquitous, wearable, and cost-effective, safeguards to protect privacy are established, and interdisciplinary collaborations between education and computer science researchers increase. In the near term, sensor-free measures can be (and have been) applied at scale when digital technologies are integrated in the curriculum. For example, Baker and Ocumpaugh (2014) report that the engagement measure developed for ASSISTments (Pardos et al., 2013) has been applied to 231,543 hours of tutoring data from 54,401 students. They estimate a measuring cost of $0.28 per student/hour after factoring the costs required to develop the measure.

What are some future developments in AAA-based measurement?

We see many opportunities moving forward. For one, the rapid advent of cost-effective consumer-off-the-shelf (COTS) devices (e.g., Fitbit, Apple Watch, EyeTribe) that afford wearable and wireless sensing, coupled with pervasive computing via smartphones and tablets, has ushered forth exciting new opportunities. For example, Hutt et al. (in review) used multiple COTS eye tracker to collect eye gaze data as entire classes (14 to 30 at a time) of high-school students interacted with an intelligent tutoring system in their regular classroom. To further ameliorate sensing costs, researchers have been replacing sensors with scalable proxies (aka soft sensors or virtual sensors). Webcams are an ideal "proxy" sensor due to their widespread availability in contemporary computing devices. They have been used to estimate bodily movement Author (year), eye gaze (Sewell & Komogortsev, 2010), and peripheral physiological signals like heart rate, respiratory rate, cardiac R-R intervals, and blood oxygen saturation (Poh, McDuff, & Picard, 2011; Scully et al., 2012).

We also see considerable potential in multilevel measurement. Aside from a few exceptions (Arroyo et al., 2009; Author, year; Kapoor & Picard, 2005), extant studies have either focused on low-level bodily/physiological signals *or* higher-level interaction/contextual cues. Physiological/bodily responses reflect rapid internal changes, but the responses are nonspecific (e.g., an increase in autonomic arousal can signal a range of affective states) or are subject to social masking (Ekman, 2002) and cannot be taken at face value. Interaction/contextual approaches can help resolve ambiguity in these responses, though they might not be sufficiently diagnostic in and of themselves. Thus, there is much to be gained by combining top-down predictive models derived from interactional/contextual features with bottom-up diagnostic models based on physiological/bodily signals (Conati, 2002).

Multimodal measurement is another promising area. Although such measures are not guaranteed to improve measurement accuracy (Author, year), they have other advantages. From a practical perspective, a secondary modality can compensate when the primary is unavailable (e.g., the speech signal is unavailable when the person is silent) or unusable (e.g., the face is occluded). From a theoretical perspective, the assumption that engagement is a multi-componential construct necessitates multimodal measurement as different modalities optimally index specific components. In particular, eye gaze and central physiology are best suited for cognitive engagement (Berka et al., 2007; Deubel & Schneider, 1996; Hoffman & Subramaniam, 1995; Marshall, 2005; Rayner, 1998), facial features and peripheral physiology for affective engagement (Ekman, 1984; Keltner & Ekman, 2000; Larsen et al., 2008; Matsumoto et al., 2008), and interaction features for behavioral engagement (Baker & Ocumpaugh, 2015; Baker & Rossi, 2013; Bulger, Mayer, Almeroth, & Blau, 2008; Gobert et al., 2015). Multimodal measures 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 mental states that manifest across different timescales (Author, year).

Finally, successful re-engagement strategies require identifying causes of disengagement, which is a complicated prospect. For instance, boredom can stem from multiple factors, including understimulation, a sense that effort is coerced, underchallenge, excess challenge, lack of interest, lack of value, and dislike of teacher (Daschmann, Goetz, & Stupnisky, 2011; Pekrun, Goetz, Daniels, Stupnisky, & Perry, 2010). Re-engaging a bored learner requires going beyond merely measuring boredom: it requires assessing what led to each boredom episode. Thus, the ability to model the antecedents of (dis)engagement is an important future challenge.

Concluding Remarks

Unlike the fictitious characters Trinity and Neo from the 1999 Hollywood film *The Matrix*, who learn by directly downloading knowledge/skills directly into their brains, learning still requires engagement. A student who is engaged is primed to learn; a student who is disengaged is not. The last decade has produced a flurry of theoretical and empirical research aimed at defining engagement, identifying its causes and effects, and devising interventions to re-engage disengaged learners. The digital revolution can catalyze similar advances on the measurement front. We are very optimistic about the potential for AAA-based measures to complement traditional measures, especially for micro-level measurement of person-oriented components of engagement. The AAA approach goes beyond measurement alone, because its fine-grained, contextually-coupled, person-oriented foci afford dynamic interventions to re-engage disengaged learners. Developing effective interventions that promote engagement so that learning is enjoyable, efficient, and effective is the next frontier.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

Afzal, S., Robinson, P. Natural affect data: Collection and annotation. In: Calvo, R., D'Mello, S., editors. New Perspectives on Affect and Learning Technologies. New York, NY: Springer; 2011. p. 44-70.

Arroyo, I., Beal, CR., Murray, T., Walles, R., Woolf, BP. Web-based intelligent multimedia tutoring for high stakes achievement tests. In: Lester, JC.Vicari, RM., Paraguaçu, F., editors. Proceedings of the 7th International Conference on Intelligent Tutoring Systems (ITS 2004); Berlin: Springer; 2004. p. 468-477.

Arroyo, I., Woolf, B., Cooper, D., Burleson, W., Muldner, K., Christopherson, R. Emotion sensors go to school. In: Dimitrova, V.Mizoguchi, R.Du Boulay, B., Graesser, A., editors. Proceedings of the 14th International Conference on Artificial Intelligence In Education; Amsterdam: IOS Press; 2009. p. 17-24.

Azevedo R. Defining and measuring engagement and learning in science: Conceptual, theoretical, methodological, and analytical issues. Educational Psychologist. 2015; 50(1):84–94.

Baker, R., Ocumpaugh, J. Interaction-based affect detection in educational software. In: Calvo, R.D'Mello, S.Gratch, J., Kappas, A., editors. The Oxford Handbook of Affective Computing. New York: Oxford University Press; 2015. p. 233-245.

Baker, R., Rossi, LM. Assessing the disengaged behaviors of learners. In: Sottilare, R.Graesser, A.Hu, X., Holden, H., editors. Design Recommendations for Intelligent Tutoring Systems. Orlan do, FL: Army Research Laboratory; 2013. p. 153

Baker, RS., Ocumpaugh, J. Cost-effective, actionable engagement detection at scale. Proceedings of the 7th International Conference on Educational Data Mining; 2014. p. 345-346.

Baker, RSJ., Kalka, J., Aleven, V., Rossi, L., Gowda, SM., Wagner, AZ., … Ocumpaugh, J. Towards sensor-free affect detection in cognitive tutor algebra. In: Yacef, K.Zaïane, O.Hershkovitz, H.Yudelson, M., Stamper, J., editors. Proceedings of the 5th International Conference on Educational Data Mining; International Educational Data Mining Society; 2012. p. 126-133.

Bandura, A. Social foundations of thought and action: A social cognitive theory. Englewood Cliffs, NJ: Prentice-Hall, Inc; 1986.

Bandura, A. Self-efficacy: The exercise of control. New York: Freeman; 1997.

Barrett L. Are emotions natural kinds? Perspectives on Psychological Science. 2006; 1:28–58. [PubMed: 26151184]

Barsalou LW. Grounded cognition. Annual Review of Psychology. 2008; 59(1):617–645.

Bartlett MS, Littlewort GC, Frank MG, Lee K. Automatic decoding of facial movements reveals deceptive pain expressions. Current Biology. 2014; 24(7):738–743. [PubMed: 24656830]

Berka C, Levendowski DJ, Lumicao MN, Yau A, Davis G, Zivkovic VT, … Craven PL. EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. Aviation, space, and environmental medicine. 2007; 78(Supplement 1):B231–B244.

Bernardin HJ, Buckley MR. Strategies in rater training. Academy of Management Review. 1981; 6(2): 205–212.

Bulger ME, Mayer RE, Almeroth KC, Blau SD. Measuring learner engagement in computer-equipped college classrooms. Journal of Educational Multimedia and Hypermedia. 2008; 17(2):129–143.

Camras L, Shutter J. Emotional facial expressions in infancy. Emotion Review. 2010; 2(2):120–129.

Camras LA, Witherington DC. Dynamical systems approaches to emotional development. Developmental Review. 2005; 25(3–4):328–350. DOI: 10.1016/j.dr.2005.10.002

Chi M, Wylie R. The ICAP framework: Linking cognitive engagement to active learning outcomes. Educational Psychologist. 2014; 49(4):219–243.

Christenson, SL., Reschly, AL., Wylie, C. Handbook of research on student engagement. Springer; 2012.

Coan JA. Emergent ghosts of the emotion machine. Emotion Review. 2010; 2(3):274–285.

Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960; 20(1):37–46.

Conati C. Probabilistic assessment of user's emotions in educational games. Applied Artificial Intelligence. 2002; 16(7–8):555–575. DOI: 10.1080/08839510290030390

Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995; 20(3):273–297.

Csikszentmihalyi M, Larson R. Validity and reliability of the experience-sampling method. The Journal of nervous and mental disease. 1987; 175(9):526–536. [PubMed: 3655778]

Darwin, C. The expression of the emotions in man and animals. London: John Murray; 1872.

Daschmann EC, Goetz T, Stupnisky RH. Testing the predictors of boredom at school: Development and validation of the precursors to boredom scales. British Journal of Educational Psychology. 2011; 81(3):421–440. [PubMed: 21770913]

Deci, EL., Ryan, RM. Self-Determination. Wiley Online Library; 1985.

Deubel H, Schneider WX. Saccade target selection and object recognition: Evidence for a common attentional mechanism. Vision research. 1996; 36(12):1827–1837. [PubMed: 8759451]

deVega, M.Glenberg, A., Graesser, A., editors. Symbols, embodiment, and meaning. Oxford: Oxford University Press; 2008.

Domingos P. A few useful things to know about machine learning. Communications of the ACM. 2012; 55(10):78–87.

Drummond, J., Litman, D. In the zone: Towards Detecting student zoning out using supervised machine learning. In: Aleven, V.Kay, J., Mostow, J., editors. Intelligent Tutoring Systems. Vol. 6095. Berlin / Heidelberg: Springer-Verlag; 2010. p. 306-308.

Eastwood JD, Frischen A, Fenske MJ, Smilek D. The unengaged mind: Defining boredom in terms of attention. Perspectives on Psychological Science. 2012; 7(5):482–495. [PubMed: 26168505]

Eccles, J., Wang, M-T. Part I commentary: So what is student engagement anyway?. In: Christenson, S.Reschly, A., Wylie, C., editors. Handbook of research on student engagement. New York: Springer; 2012. p. 133-145.

Eccles JS, Wigfield A. Motivational beliefs, values, and goals. Annual Review of Psychology. 2002; 53(1):109–132.

Ekman, P. Expression and the nature of emotion. In: Scherer, K., Ekman, P., editors. Approaches to emotion. Hillsdale, NJ: Erlbaum; 1984. p. 319-344.

Ekman P. An argument for basic emotions. Cognition & Emotion. 1992; 6(3–4):169–200.

Ekman, P. Darwin, deception, and facial expression. Paper presented at the Conference on Emotions Inside Out, 130 Years after Darwins the Expression of the Emotions in Man and Animals; New York, NY. 2002 Nov 16–17.

Ekman, P., Friesen, W. The Facial Action Coding System: A technique for the measurement of facial movement. Palo Alto, CA: Consulting Psychologists Press; 1978.

Elfenbein H, Ambady N. On the universality and cultural specificity of emotion recognition: A meta-analysis. Psychological Bulletin. 2002; 128(2):203–235. DOI: 10.1037//0033-2909.128.2.203 [PubMed: 11931516]

Emotient. FACET-Facial Expression Recognition Software. 2014.

Eyben, F., Wöllmer, M., Schuller, B. Opensmile: the munich versatile and fast open-source audio feature extractor. Proceedings of the international conference on Multimedia; New York, NY: ACM; 2010. p. 1459-1462.

Fiedler, K., Beier, S. Affect and cognitive processes in educational contexts. In: Pekrun, R., Linnenbrink-Garcia, L., editors. International handbook of emotions in education. New York, NY: Routledge; 2014. p. 36-56.

Fredricks JA, Blumenfeld PC, Paris AH. School engagement: Potential of the concept, state of the evidence. Review of Educational Research. 2004; 74(1):59–109.

Fredricks, JA., McColskey, W. The measurement of student engagement: A comparative analysis of various methods and student self-report instruments. In: Christenson, S.Reschly, A., Wylie, C., editors. Handbook of research on student engagement. New York: Springer; 2012. p. 763-782.

Gobert JD, Baker RS, Wixon MB. Operationalizing and detecting disengagement within online science microworlds. Educational Psychologist. 2015; 50(1):43–57.

Graesser A, Chipman P, Haynes B, Olney A. AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. IEEE Transactions on Education. 2005; 48(4):612–618. DOI: 10.1109/TE.2005.856149

Greene BA. Measuring cognitive engagement with self-report scales: Reflections from over 20 years of research. Educational Psychologist. 2015; 50(1):1–17.

Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; 143(1):29–36. [PubMed: 7063747]

Hayes AF, Krippendorff K. Answering the call for a standard reliability measure for coding data. Communication methods and measures. 2007; 1(1):77–89.

Heine SJ, Lehman DR, Peng K, Greenholtz J. What's wrong with cross-cultural comparisons of subjective Likert scales?: The reference-group effect. Journal of Personality and Social Psychology. 2002; 82(6):903–918. [PubMed: 12051579]

Henrie CR, Halverson LR, Graham CR. Measuring student engagement in technology-mediated learning: A review. Computers & Education. 2015; 90:36–53.

Hidi S, Renninger KA. The four-phase model of interest development. Educational Psychologist. 2006; 41(2):111–127.

Hoffman JE, Subramaniam B. The role of visual attention in saccadic eye movements. Attention, Perception, & Psychophysics. 1995; 57(6):787–795. [PubMed: 7651803]

Kahneman D, Krueger AB, Schkade DA, Schwarz N, Stone AA. A survey method for characterizing daily life experience: The day reconstruction method. Science. 2004; 306(5702):1776–1780. [PubMed: 15576620]

Kapoor, A., Picard, R. Multimodal affect recognition in learning environments. Proceedings of the 13th annual ACM international conference on Multimedia; New York: ACM; 2005. p. 677-682.

Kappas A. Social regulation of emotion: messy layers. Frontiers in Psychology. 2013; 4(51)doi: 10.3389/fpsyg.2013.00051

Keltner, D., Ekman, P. Facial expression of emotion. In: Lewis, R., Haviland-Jones, JM., editors. Handbook of emotions. Vol. 2. New York: Guilford; 2000. p. 236-264.

Krosnick JA. Survey research. Annual Review of Psychology. 1999; 50(1):537–567.

Larsen, J., Berntson, G., Poehlmann, K., Ito, T., Cacioppo, J. The psychophysiology of emotion. In: Lewis, M.Haviland-Jones, J., Barrett, L., editors. Handbook of emotions. 3. New York, NY: Guilford; 2008. p. 180-195.

Lehr CA, Sinclair MF, Christenson SL. 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. 2004; 9(3):279–301.

Lewis MD. Bridging emotion theory and neurobiology through dynamic systems modeling. Behavioral and Brain Sciences. 2005; 28(2):169–245. [PubMed: 16201458]

Linnenbrink-Garcia L, Pekrun R. Students' emotions and academic engagement: Introduction to the special issue. Contemporary Educational Psychology. 2011; 36(1):1–3.

Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., Bartlett, M. The computer expression recognition toolbox (CERT). Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition; Washington, DC: IEEE; 2011. p. 298-305.

Loveless, T. How well are American students learning?. Washington DC: Brookings Institution; 2015.

Mandler, G. Interruption (discrepancy) theory: Review and extensions. In: Fisher, S., Cooper, CL., editors. On the Move: The Psychology of Change and Transition. Chichester: Wiley; 1990. p. 13-32.

Marshall, SP. Assessing cognitive engagement and cognitive state from eye metrics. In: Schmorrow, DD., editor. Foundations of Augmented Cognition. Mahwah, NJ: Lawrence Erlbaum Associates; 2005. p. 312-320.

Martin, JP. Lower Merion district's laptop saga ends with $610,000 settlement. The Philadelphia Inquirer. 2010. Retrieved from http://articles.philly.com/2010-10-12/news/24981536_1_laptop-students-district-several-million-dollars

Matsumoto, D., Keltner, D., Shiota, MN., O'Sullivan, M., Frank, M. Facial expressions of emotion. In: Lewis, M.Haviland-Jones, J., Barrett, L., editors. Handbook of emotions. Vol. 3. New York, NY: Guilford Press; 2008. p. 211-234.

Mehl MR, Pennebaker JW, Crow DM, Dabbs J, Price JH. The Electronically Activated Recorder (EAR): A device for sampling naturalistic daily activities and conversations. Behavior Research Methods, Instruments, & Computers. 2001; 33(4):517–523.

Mehu M, Scherer K. A psycho-ethological approach to social signal processing. Cognitive Processing. 2012; 13(2):397–414. [PubMed: 22328016]

Mesquita B, Boiger M. Emotions in context: A sociodynamic model of emotions. Emotion Review. 2014; 6(4):298–302.

Miller BW. Using reading times and eye-movements to measure cognitive engagement. Educational Psychologist. 2015; 50(1):31–42.

Miller, WL., Petsche, K., Baker, RS., Labrum, MJ., Wagner, AZ. Boredom Across Activities, and Across the Year, within Reasoning Mind. Paper presented at the Workshop on Data Mining for Educational Assessment and Feedback (ASSESS 2014); New York, NY. 2014.

Mota, S., Picard, R. Automated posture analysis for detecting learner's interest level. Paper presented at the Computer Vision and Pattern Recognition Workshop; 2003; 2003.

Niedenthal PM. Embodying emotion. Science. 2007; 316(5827):1002–1005. DOI: 10.1126/science.1136930 [PubMed: 17510358]

NSC. National Student Clearinghouse. 2016. from http://www.studentclearinghouse.org/

Nystrand M, Gamoran A. Instructional discourse, student engagement, and literature achievement. Research in the Teaching of English. 1991; 25(3):261–290.

Ocumpaugh J, Baker R, Gowda S, Heffernan N, Heffernan C. Population validity for educational data mining: A case study in affect detection. British Journal of Educational Psychology. 2014; 45(3):487–501.

Ocumpaugh, J., Baker, RS., Rodrigo, MMT. Baker-Rodrigo Observation Method Protocol (BROMP) 1.0. Training Manual version 1.0. New York, NY: 2012.

Ocumpaugh, J., Baker, RS., Rodrigo, MMT. 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; 2015.

Pantic M, Patras I. Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. IEEE Transactions on Systems, Man, and Cybernetics, Part B. 2006; 36(2):433–449. DOI: 10.1109/tsmcb.2005.859075

Pardos, Z., Baker, RSJd, San Pedro, MOCZ., Gowda, SM. Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. In: Suthers, D.Verbert, K.Duval, E., Ochoa, X., editors. Proceedings of the 3rd International Conference on Learning Analytics and Knowledge; New York, NY: ACM; 2013. p. 117-124.

Pekrun R, Goetz T, Daniels L, Stupnisky RH, Perry R. Boredom in achievement settings: Exploring control–value antecedents and performance outcomes of a neglected emotion. Journal of Educational Psychology. 2010; 102(3):531–549. DOI: 10.1037/a0019243

Pekrun, R., Linnenbrink-Garcia, L. Academic emotions and student engagement. In: Christenson, S.Reschly, A., Wylie, C., editors. Handbook of research on student engagement. New York: Springer; 2012. p. 259-282.

Pianta, RC., Hamre, BK., Allen, JP. Teacher-student relationships and engagement: Conceptualizing, measuring, and improving the capacity of classroom interactions. In: Christenson, S.Reschly, A., Wylie, C., editors. Handbook of research on student engagement. New York, NY: Springer; 2012. p. 365-386.

PISA. Ready to Learn: Students' Engagement, Drive and Self-Beliefs. 2012; III

Podsakoff PM, MacKenzie SB, Lee JY, Podsakoff NP. Common method biases in behavioral research: A critical review of the literature and recommended remedies. Journal of Applied Psychology. 2003; 88(5):879–903. [PubMed: 14516251]

Poh MZ, McDuff DJ, Picard RW. Advancements in noncontact, multiparameter physiological measurements using a webcam. IEEE Transactions on Biomedical Engineering. 2011; 58(1):7–11. [PubMed: 20952328]

Rayner K. Eye movements in reading and information processing: 20 years of research. Psychological Bulletin. 1998; 124(3):372–422. [PubMed: 9849112]

Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, NT., Koedinger, KR., Junker, B., … Choksey, S. The Assistment project: Blending assessment and assisting. In: Loi, C., McCalla, G., editors. Proceedings of the 12th International Conference on Artificial Intelligence In Education; Amsterdam: IOS Press; 2005. p. 555-562.

Reeve J, Tseng CM. Agency as a fourth aspect of students' engagement during learning activities. Contemporary Educational Psychology. 2011; 36(4):257–267.

Renninger KA, Bachrach JE. Studying triggers for interest and engagement using observational methods. Educational Psychologist. 2015; 50(1):58–69.

Reschly, A., Christenson, S. Jingle, jangle, and conceptual haziness: Evolution and future directions of the engagement construct. In: Christenson, S.Reschly, A., Wylie, C., editors. Handbook of research on student engagement. Berlin: Springer; 2012. p. 3-19.

Roseman IJ. Emotional behaviors, emotivational goals, emotion strategies: Multiple levels of organization integrate variable and consistent responses. Emotion Review. 2011; 3(4):434–443.

Ruch W. Will the real relationship between facial expression and affective experience please stand up: The case of exhilaration. Cognition & Emotion. 1995; 9(1):33–58.

Russell J. Core affect and the psychological construction of emotion. Psychological Review. 2003; 110:145–172. [PubMed: 12529060]

Russell JA, Bachorowski JA, Fernandez-Dols JM. Facial and vocal expressions of emotion. Annual Review of Psychology. 2003; 54:329–349.

Ryan R, Deci E. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist. 2000; 55(1):68–78. [PubMed: 11392867]

Ryu S, Lombardi D. Coding classroom interactions for collective and individual engagement. Educational Psychologist. 2015; 50(1):70–83.

Sabourin, J., Mott, B., Lester, J. Modeling learner affect with theoretically grounded dynamic bayesian networks. In: D'Mello, S.Graesser, A.Schuller, B., Martin, J., editors. Proceedings of the Fourth International Conference on Affective Computing and Intelligent Interaction; Berlin Heidelberg: Springer-Verlag; 2011. p. 286-295.

San Pedro, M., Baker, RS., Bowers, AJ., Heffernan, NT. Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. In: D'Mello, S.Calvo, R., Olney, A., editors. Proceedings of the 6th International Conference on Educational Data Mining (EDM 2013); International Educational Data Mining Society; 2013. p. 177-184.

Schunk, DH., Pajares, F. Competence perceptions and academic functioning. In: Elliot, AJ., Dweck, CS., editors. Handbook of competence and motivation. New York: Guilford; 2005. p. 141-163.

Scully CG, Lee J, Meyer J, Gorbach AM, Granquist-Fraser D, Mendelson Y, Chon KH. Physiological parameter monitoring from optical recordings with a mobile phone. IEEE Transactions on Biomedical Engineering. 2012; 59(2):303–306. [PubMed: 21803676]

Sewell, W., Komogortsev, O. Real-time eye gaze tracking with an unmodified commodity webcam employing a neural network. Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems; New York: ACM; 2010. p. 3739-3744.

Shibata K, Watanabe T, Sasaki Y, Kawato M. Perceptual learning incepted by decoded fMRI neurofeedback without stimulus presentation. Science. 2011; 334(6061):1413–1415. [PubMed: 22158821]

Shute VJ, Ventura M, Kim YJ. Assessment and learning of qualitative physics in Newton's playground. The Journal of Educational Research. 2013; 106(6):423–430.

Sinatra GM, Heddy BC, Lombardi D. The challenges of defining and measuring student engagement in science. Educational Psychologist. 2015; 50(1):1–13.

Skinner EA, Belmont MJ. Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. Journal of Educational Psychology. 1993; 85(4):571.

Smallwood J, McSpadden M, Schooler JW. When attention matters: The curious incident of the wandering mind. Memory & Cognition. 2008; 36(6):1144–1150. DOI: 10.3758/mc.36.6.1144 [PubMed: 18927032]

St John M, Kobus DA, Morrison JG, Schmorrow D. Overview of the DARPA augmented cognition technical integration experiment. International Journal of Human-Computer Interaction. 2004; 17(2):131–149.

Stein, N., Levine, L. Making sense out of emotion. In: Kessen, W.Ortony, A., Kraik, F., editors. Memories, thoughts, and emotions: Essays in honor of George Mandler. Hillsdale, NJ: Erlbaum; 1991. p. 295-322.

Turner JC, Meyer DK. Studying and understanding the instructional contexts of classrooms: Using our past to forge our future. Educational Psychologist. 2000; 35(2):69–85.

Valstar M, Mehu M, Jiang B, Pantic M, Scherer K. Meta-analysis of the first facial expression recognition challenge. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics. 2012; 42(4):966–979.

Volpe RJ, DiPerna JC, Hintze JM, Shapiro ES. Observing students in classroom settings: A review of seven coding schemes. School Psychology Review. 2005; 34(4):454.

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., … Campbell, AT. Student Life: assessing mental health, academic performance and behavioral trends of college students using smartphones. In: Kientz, J.Scott, J., Song, J., editors. Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (ACM UbiComp'14); New York, NY: ACM; 2014. p. 3-14.

Whitehill J, Serpell Z, Lin YC, Foster A, Movellan J. The faces of engagement: Automatic recognition of student engagement from facial expressions. IEEE Transactions on Affective Computing. 2014; 5(1):86–98.

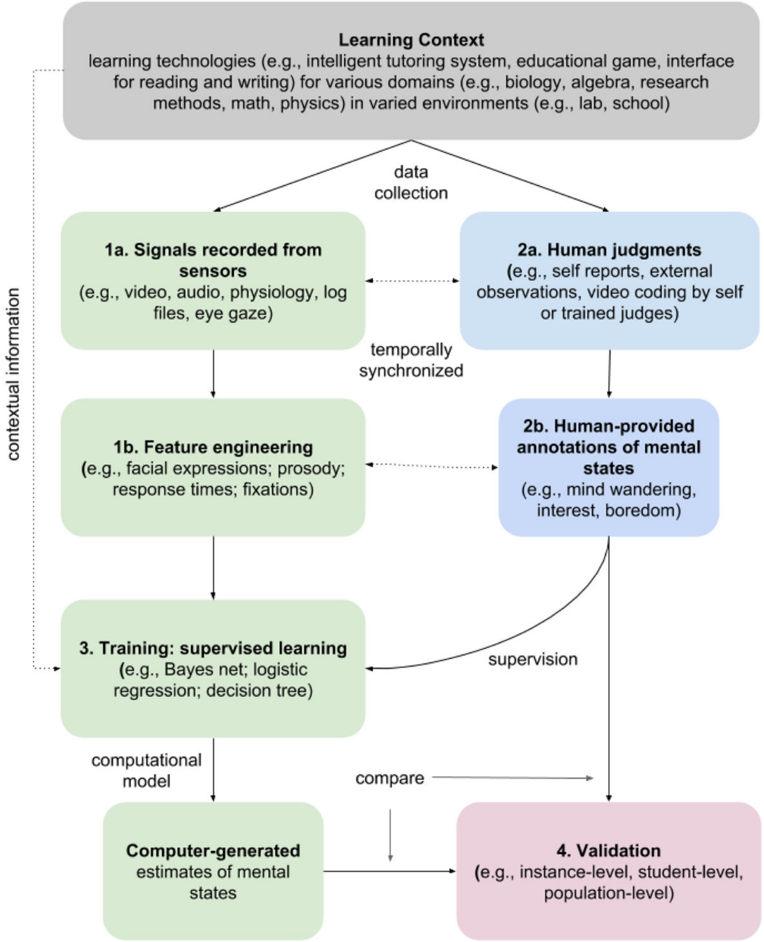
Learning Context
learning technologies (e.g., intelligent tutoring system, educational game, interface for reading and writing) for various domains (e.g., biology, algebra, research methods, math, physics) in varied environments (e.g., lab, school)
data collection
1a. Signals recorded from sensors
(e.g., video, audio, physiology, log files, eye gaze)
2a. Human judgments
(e.g., self reports, external observations, video coding by self or trained judges)
temporally synchronized
contextual information
1b. Feature engineering
(e.g., facial expressions; prosody; response times; fixations)
2b. Human-provided annotations of mental states
(e.g., mind wandering, interest, boredom)
3. Training: supervised learning
(e.g., Bayes net; logistic regression; decision tree)
supervision
computational model
compare
Computer-generated estimates of mental states
4. Validation
(e.g., instance-level, student-level, population-level)

Figure 1.

Major steps involved in building an automated engagement measure

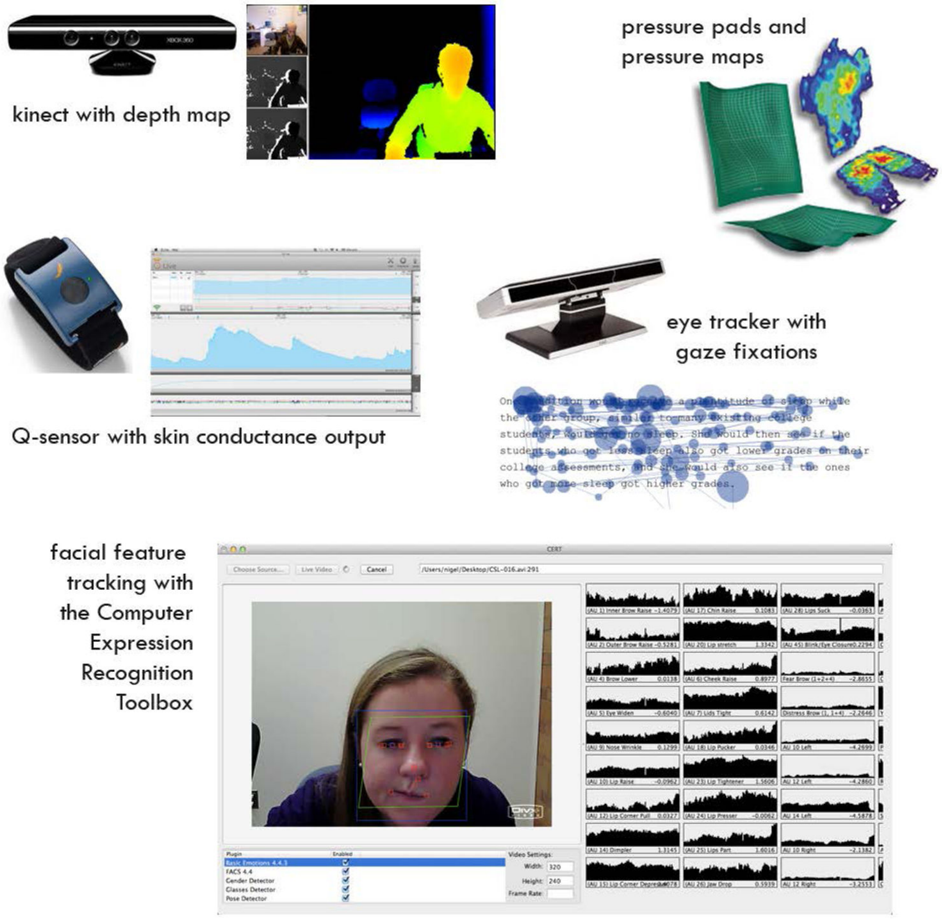
kinect with depth map
pressure pads and pressure maps
Q-sensor with skin conductance output
eye tracker with gaze fixations
One condition would require a multitude of sleep while the other group, similar to many existing college students, would get no sleep. She would then see if the students who got less sleep also got lower grades on their college assessments, and she would also see if the ones who got more sleep got higher grades.
facial feature tracking with the Computer Expression Recognition Toolbox
CERT
Choose Source...
Live Video
Cancel
/Users/nigel/Desktop/CSL-016.avi:291
Plugin
Enabled
Basic Emotions 4.4.3
FACS 4.4
Gender Detector
Glasses Detector
Pose Detector
Video Settings:
Width: 320
Height: 240
Frame Rate:

Figure 2.

Sensors and signals (output)

Table 1

Overview of case studies

Study	Learning Context	Component of engagement	Annotation Method	Sensor	Signal	Features	Supervised Learning Method	Generalizes to New Students	Accuracy (Metric)	Improvement over chance
Discussed in main text										
Author (year)	Computer literacy from AutoTutor	Boredom, flow, confusion, frustration	Offline video coding by self, peers, trained judges	None	Log files	Discourse features and interaction patterns	Varied	No	0.71 (RR)	42% [a]
Pardos et al. (2013)	Math with ASSISTments	Boredom, frustration, confusion, engaged concentration	Online observations by researchers	None	Log files	Interaction patterns	Varied	Yes	0.68 (A′)	30% [b]
Gobert et al. (2015)	Science microworlds with Inq-ITS	Disengaged from task goal	Offline coding of logs	None	Log files	Interaction patterns	PART	Yes	0.81 (A′)	41% [b]
Whitehill et al. (2014)	Cognitive skills training on iPad	Behavioral engagement (4 levels)	Video coding by researchers	Webcam	Video	Facial expressions	Support vector machine	Yes	0.73 (2AFC)	31% [b]
Mota and Picard (2003)	Constraint satisfaction game	Interest (3-levels)	Video coding by teachers	Pressure-sensitive pads	Pressure maps	Body movements/ posture	Hidden Markov Models	Yes	0.77 (RR)	61% [c]
Author (year)	Research methods from text	Probe-caught mind wandering (yes or no)	Online self-reports	Eye tracker	Eye gaze & log files	Eye movements, contextual cues	Bayesian	Yes	0.70 (RR)	25% [b]
Arroyo et al. (2009)	Math with Wayang Outpost	Interest, confidence, excitement, frustration (1–5 scale)	Online self-reports	Webcam, physiological sensor, pressure-mouse, pressure-pads	Log files, video, pressure maps, time series	Interaction features, facial expressions, skin conductance, pressure exerted on mouse, body movements/ posture	Linear Regression	No	0.47 (R^2)	47% [d]
Bosch et al. (2016)	Newtonian Physics with Physics Playground	Boredom, engaged concentration, confusion, frustration, delight	Online observations by researchers	Webcam	Video	Facial expressions and body movements	Varied	Yes	0.69 (AUC)	37% [e]
Discussed in supplementary review (Online Supplement A)										
Baker et al. (2012)	Algebra with a Cognitive Tutor	Boredom, engaged concentration, confusion, frustration	Online observations by researchers	None	Log files	Interaction patterns	Varied	Yes	0.85 (A′)	30% [b]
Author (year)	Writing proficiency with computer interface	Engagement/flow, boredom	Offline self-reports	None	Log files	Keystrokes, individual attributes, task appraisals	REP Tree	Yes	0.87 (RR)	37% [b]
Sabourin, Mott, and Lester (2011)	Microbiology with Crystal Island	Positive affect (curious, focused, excited) vs. other (anxious, bored, confused, frustrated)	Online self-reports	None	Log files	Interaction patterns and individual attributes	Dynamic Bayesian Network	Yes	0.73 (RR)	43% [c]
Drummond and Litman (2010)	Biology from text	Zone outs (high vs. low)	Online self-reports	Microphone	Audio	Acoustic-prosodic feaures	J48 Decision Trees	No	0.64 (RR)	22% [f]
Author (year)	Research methods from text	Probe-caught mind wandering (yes or no)	Online self-reports	Wearable physiological sensors	Time series	Skin conductance and skin temperature (&context)	Filtered Classifier; LAD Tree	Yes	0.58 (RR)	19% [b]
Kapoor and Picard (2005)	Constraint satisfaction game	Interest (3-levels)	Offline video coding by teachers	Infrared camera, pressure-pads	Log files, video, pressure maps	Interaction context, facial expressions, body movements/posture	Mixture of Gausian Processes	No	0.87 (RR)	73% [f]
Author (year)	Creative writing	Behavioral engagement (2-levels)	Online self-reports, offline video-coding by self	Depth-camera	Depth maps, color video	Facial expressions, facial textures, heart rate	Updatable Naive Bayes	Yes	0.75 (AUC)	49% [e]

Note. RR = recognition rate; 2AFC = 2-alternative forced choice. AUC = area under the receiver operating characteristic (ROC) curve. A′ = A-prime. R^2 = coefficient of determination.

[a,b,c,d,e,f] denote the method used to obtain percent improvement above chance: [a] estimated as proportion improvement in RR over base rate; [b] Cohen’s kappa as reported; [c] Cohen’s kappa as estimated from reported classification tables; [d] R^2 as reported in paper (assumes that a chance model will yield an R^2 of 0); [e] estimated as improvement of achieved AUC over chance AUC of 0.5; [f] estimated as proportion improvement in RR over majority baseline (classifying all instances as the majority class).