Supplementary Materials

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

Sidney D’Mello, Ed Dieterle, and Angela Duckworth

Supplement A. Additional Case Studies on the Advanced Analytic Automated (AAA) Measurement Approach

We briefly review seven case studies to complement the eight reviewed in the main text.

Interaction patterns with Cognitive Tutor. Baker et al. (2012) developed an AAA-based measure of engagement for Cognitive Tutor Algebra, an intelligent tutoring system (ITS) used by tens of thousands of students in the U.S. as part of a blended math curriculum (Koedinger & Corbett, 2006). In Cognitive Tutor, students solve problems that are automatically selected to match their skill based on a model of student knowledge (Corbett & Anderson, 1995). Based on student responses, the ITS provides learning supports in the form of feedback, hints, and explanations. The researchers obtained training data from 89 high school students who used Cognitive Tutor Algebra 1 in a school computer lab as part of their regular mathematics instruction. Using the Baker–Rodrigo Observation Method Protocol (BROMP), trained observers provided live annotations of engaged concentration, boredom, confusion, and frustration (Ocumpaugh, Baker, & Rodrigo, 2012) (also see ASSISTments study in main text). The annotations were temporally aligned with interaction features extracted from Cognitive Tutor’s log files (e.g., student actions, tutor feedback, use of help resources, and response times). Supervised classification methods were used to discriminate each state from all the others (e.g., boredom vs. confusion, frustration, engaged concentration). The models, validated with 6-fold student-level cross-validation, yielded a mean accuracy of 0.85 (measured with A-prime - similar
to area under the receiver operating characteristic curve [AUC]; see main text) – or a 30% improvement over chance. The results suggest the possibility of automatically measuring engagement of the tens of thousands of students who use Cognitive Tutor daily.

Keystrokes, task appraisals, and individual attributes during essay writing. Bixler and D'Mello (2013) developed an AAA-based measure of engagement during essay writing. They conducted a lab study where 44 undergraduate students wrote three 10-minute essays on personal emotional experiences, socially charged issues (e.g., death penalty), and academic topics (e.g., the use of class discussion). The writing software recorded each keystroke as well as videos of students’ faces and computer screens. Immediately after the writing session, students annotated their own videos for boredom, engagement, and several other affective and cognitive states (see below) at approximately 15-second intervals using a video-based retrospective affect judgment procedure (similar to the AutoTutor study as discussed in main text).

The researchers focused on automatically discriminating engagement from boredom as the other states were infrequent. They considered a combination of three feature sets: (1) individual differences in scholastic aptitude, writing apprehension, and exposure to print, (2) task appraisals (e.g., perceived interest, level of comfort with essay topic, prospective effort) immediately before writing each essay, and (3) keystroke features, such as verbosity (amount of text generated), latency between keystrokes, and length of pauses. Supervised classification models were trained to differentiate engagement from boredom and achieved a recognition rate of 0.87 (37% improvement over chance) with student-level validation (random sampling of students into training and testing sets). The results illustrate the utility of multilevel analyses consisting of individual differences that vary by student, task appraisals that vary by essay topic, and keystroke feature that vary across 15-second annotation intervals.
**Interaction patterns and individual attributes with CRYSTAL ISLAND.** CRYSTAL ISLAND is a narrative-centered educational game for microbiology that is aligned with the 8th grade curriculum in North Carolina (Spires, Rowe, Mott, & Lester, 2011). In CRYSTAL ISLAND, the learner assumes the role of a protagonist named Alex, who discovers that members of a research team are falling ill and is charged with identifying the source of the infectious disease. Alex visits various areas of the island (dining hall, lab, inferminary, dorm), interviews other islanders, and manipulates objects in the virtual world (e.g., performing lab tests to diagnose the disease).

Sabourin, Mott, and Lester (2011) built an AAA-based measure of affective engagement in CRYSTAL ISLAND. They collected training data from 260 middle-school students who interacted with CRYSTAL ISLAND in a 55-minute in-class session. Students self-reported their affective states at multiple points during gameplay by selecting one out of seven states: anxious, bored, confused, curious, excited, focused, and frustrated. They also completed self-report questionnaires of personality, goal orientation, and emotion regulation tendencies. The researchers extracted interaction features from CRYSTAL ISLAND’s log files (e.g., note taking in the software, number of tests run to diagnose potential diseases) and aligned them with the trait-based measures. They used Dynamic Bayesian networks to discriminate positive affective states (curious, excited, and focused) from the other states and validated the models using 10-fold student-level validation. They achieved a recognition rate of 0.73 (43% improvement over chance), thereby demonstrating the feasibility of automatically measuring affective engagement during interactions with an educational game in a computer-enabled classroom.

**Speech during learning biology from text.** Drummond and Litman (2010) adopted a cognitive perspective on engagement by developing a measure of *zone-outs* during reading. They collected data from 37 undergraduate students who read short biology passages aloud and then
paraphrased or self-explained each passage. Students reported zone-outs on a 1 (all the time) to 7 (not at all) scale in response to the following question that appeared at set intervals during reading: “I found myself zoning out and thinking about other things when reading this text.” The responses were dichotomized as high (1-3) and low (5-7) zone-outs; responses of 4 were discarded. The researchers extracted paralinguistic features (e.g., percent of silence, minimum pitch, and minimum energy - loudness) from speech segments corresponding to each zone-out response. Supervised learning methods achieved a 0.64 recognition rate (22% improvement over chance) in discriminating between high vs. low zone-outs. Despite the use of instance-level validation, which does not ensure generalizability to new students, the study is significant as it was the first to automatically detect zone-outs.

**Peripheral physiology and context cues during computerized reading.** Blanchard, Bixler, Joyce, and D'Mello (2014) developed a physiological-based measure of mind wandering during computerized reading. They collected training data from a subset of 70 students from the Bixler and D'Mello (2016) study (see main text) who reported instances of mind wandering in response to thought probes while reading instructional texts. The researchers recorded skin conductance (SC) and skin temperature (ST) with an Affectiva Q wireless physiological sensor affixed to the student’s non-dominant wrist. They computed a number of features (e.g., mean activation levels, changes in activation across time, peaks, and autocorrelations) from the SC and ST signals in 3 to 30 second windows preceding each mind wandering report. Similar to Bixler and D'Mello (2016), they considered contextual features (e.g., total reading time, text difficulty) in addition to the physiological features. Supervised learning models with student-level validation (random sampling of students into training and testing sets), achieved a mind wandering recognition rate of 0.58 (19% above-chance improvement). These results suggest the
potential for combining the physiological-based measure with other measures (e.g., eye gaze) for more accurate mind wandering detection.

**Facial expressions, body movements, and interaction context while playing Fripples Place.** Kapoor and Picard (2005) developed an AAA-based measure of children’s interest while solving constraint satisfaction problems with a game called *Fripples Place*. The researchers used the same data as the Mota and Picard (2003) posture-based interest measurement study (see main text). However, they considered a much broader set of features, including facial features (e.g., likelihoods of nods, head shakes, eye blinks, fidgets, and smiles), aspects of the interaction context (difficulty level and state of the game), and posture features (current posture and level of activity). The researchers used a mixture-of-Gaussian Processes approach to discriminate high interest from “uninterest” (low interest and taking a break). They achieved a recognition rate of 0.87 (above-chance improvement of 71%) on the combined feature set, which exceeded recognition rates from models that individually considered the face (0.67 upper face, 0.53 lower face), body (0.82), and interaction context (0.57). Although using instance-level validation limits claims of generalizability to new students, the study is significant since it was one of the earliest attempts at multimodal measurement during learning.

**Heart rate, facial expressions, and facial textures during creative writing.** Monkaresi, Bosch, Calvo, and D'Mello (in press) used standard 2-dimensional color and 3-dimensional depth videos from the Microsoft Kinect Sensor to develop an AAA-based measure of behavioral engagement. The researchers collected data in a lab study where students wrote a draft of an essay for 30-minutes, received feedback on their essay (about 10-minutes), and revised their initial draft for another 20-minutes. Students were probed to report whether or not they were engaged every 2-minutes during writing/revision (concurrent annotations). After the
writing/revision session, a researcher segmented the recorded videos based on visible changes, such as facial expressions, posture, etc. Students returned a week later to rate their levels of engagement in these video segments (retrospective annotations).

The researchers tracked six animation units and the position and movement of the head from the depth videos. The color-videos were used to extract facial textures around the left-eye, right-eye, and mouth. Heart rate was estimated using video-based photoplethysmography (Monkaresi, Calvo, & Yan, 2014; Poh, McDuff, & Picard, 2010). An Updatable Naïve Bayes classifier that combined animation units and facial textures achieved recognition rates of 0.76 and 0.73 (average of 49% greater than chance) for detecting engagement based on concurrent and retrospective judgments, respectively. The validation method consisted of sampling students into training and testing sets (student-level validation). These results reflect the state of the art in student-independent video-based engagement measurement.

Supplement B. Comparison of Advanced Analytic Automated (AAA) and Traditional Measurement Approaches

We analyze the AAA-based measures based on affordances, applicability, and scalability and compare them to self-report questionnaires and live observations by human judges – two widely used traditional measurement approaches. We focus on the sensor level (sensor-free, sensor-light, and sensor-heavy) rather than on individual studies, since our goal is to highlight trends among families of AAA-based measures (see Table below).

Affordances. Affordances broadly refer to what a measure supports or permits. We consider four affordances here: whether conscious and/or unconscious components of engagement can be measured, the temporal granularity of measurement, whether dynamic
interventions are possible, and whether after-action reviews are supported. With respect to the first affordance, sensor-based measures can tap both conscious (e.g., verbal responses) and unconscious (e.g., neurophysiological) components of engagement, whereas sensor-free measures are more suited for conscious components (e.g., actions reflected in interaction logs). In contrast, self-reports are restricted to information that can be consciously accessed. Observers can monitor some unconscious components, such as facial expressions, but are less apt at detecting lower level physiological responses (e.g., heart rate).

The fine temporal granularity (milliseconds) is a distinguishing feature of AAA-based measures, which contrasts with the coarse-grained (tens of secs) self-reports that can be disruptive and reactive if administered too frequently. Observers can provide intermediate-grained measurement (seconds), but only if students are individually monitored.

AAA-based measures are automatic and operate in real time, so they can be used to dynamically trigger interventions aimed at re-engaging disengaged students (D'Mello, Blanchard, Baker, Ocumpaugh, & Brawner, 2014). This is not feasible with self-reports and online observations as these measures require human input.

AAA measures also afford retrospective reviews because they can provide continual estimates of engagement. For instance, if several students appear bored during part of a lesson (as detected by an AAA-based measure), the corresponding instructional activities could be redesigned (Hawkins, Heffernan, & Baker, 2013). Self-reports and observers only provide measurements at discrete intervals, so they only partially support retrospective reviews.

**Applicability.** Applicability refers to the contexts in which a measure can be used. The age of the students is one key factor. Sensor-free measures have no age constraints as long as students can use a digital device. Likewise, assuming devices are available, sensor-light and
sensor-heavy measures are applicable across age groups. Self-reports are not appropriate for very young children, while online observations are feasible for any age group.

Applicability also depends on the learning context. Self-reports and online observers are applicable in both digital and non-digital learning contexts (e.g., in-class lecture, face-to-face tutoring), which is a key advantage. Sensor-free AAA-based measures require a digital trace and are not applicable in non-digital contexts. In contrast, sensor-light measures are applicable in multiple digital and non-digital contexts; for example, a camera can track engagement during a lecture (Raca, Kidzinski, & Dillenbourg, 2015) or a microphone can capture speech during class discussions (D'Mello et al., 2015). Some sensor-heavy measures, such as wearable physiological devices, can also be used in non-digital contexts (Hernandez, Morris, & Picard, 2011), whereas others (e.g., pressure-sensitive mouse) are less applicable.

AAA-based measures are trained on data collected during interactions with a specific learning technology. Can they be applied to different technologies? This is unlikely for sensor-free measures because their features are derived from technology-specific log-files. Sensor-light and sensor-free measures are more applicable across technologies because they track more generalizable behavioral/physiological relationships (e.g., the relationship between a lowered brow and confusion (McDaniel et al., 2007) should be more or less consistent across technologies). Self-report questionnaires and human observations are decoupled from any specific technology and should generalize widely.

There is also the issue of applicability to diverse environments beyond computer-enabled classrooms and research labs, such as students’ homes, museums, libraries, or even the subway. Self-report questionnaires can be used in a variety of environments, as can sensor-free and sensor-light measures, provided there is access to computing devices in these settings. Some
sensor-heavy measures can be used in diverse environments (e.g., wireless physiological-tracking bracelets (Picard, Fedor, & Ayzenberg, 2015)), but others (e.g., EEG) are still confined to more restrictive contexts, though this is rapidly changing with the prominence of more portable sensing (Mills et al., in press). Observation-based measures cannot be used in private spaces like students homes.

**Scalability.** Scalability varies by human labor costs, measured in person-hours, and equipment costs, measured in currency. The human labor required for sensor-free and sensor-light measures is low and comparable to computer-administered self-report questionnaires. Some human labor is needed for sensor-heavy measures (e.g., during setup), but is negligible compared to human observer labor costs.

Equipment costs for sensor-free and sensor-light measures are moderate in that they require computing devices (sensor-free) along with webcams and microphones (sensor-light), whereas sensor-heavy measures accrue significant equipment costs. Self-reports and online observations do not require any equipment costs for paper and pencil administration.

Due to minimal human and equipment costs, self-report questionnaires are scalable to large numbers of students if administered electronically. Similarly, the scalability of sensor-free and sensor-light measures varies based on computer and internet access. Scalability of sensor-heavy measures is limited by equipment costs, while human labor reduces scalability of online observations.

Another dimension to consider is temporal scalability or repeated measurement across extended time frames (e.g., semester or school year). The sensor-light and sensor-free measures can continually sense in the background without being reactive and without accruing additional human or equipment costs. Thus, their temporal scalability depends on access to computing
devices and the internet. In contrast, sensor-heavy measures require maintenance of the sensing equipment, so they are less scalable in this regard. The temporal scalability of self-report questionnaires is moderate – although repeated assessments incur little human/labor costs, they might be fatiguing and/or reactive. Because there is an incremental cost for each additional session observed, the temporal scalability of online observations is also moderate.
### Overview of comparisons of measures of engagement

<table>
<thead>
<tr>
<th>Dimension</th>
<th>AAA Approaches</th>
<th>Traditional Approaches</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sensor</td>
<td>Sensor</td>
</tr>
<tr>
<td></td>
<td>Free</td>
<td>Light</td>
</tr>
</tbody>
</table>

#### Affordances

- **Conscious and/or unconscious components**
  - **Mostly conscious**
  - **Both**
  - **Both**
  - **Only conscious**
  - **Mostly conscious**

- **Temporal granularity of measurement**
  - **Fine**
  - **Fine**
  - **Very fine**
  - **Coarse**
  - **Intermediate**

- **Dynamic intervention supported**
  - **Yes**
  - **Yes**
  - **Yes**
  - **No**
  - **No**

- **Retrospective review supported**
  - **Yes**
  - **Yes**
  - **Yes**
  - **Partially**
  - **Partially**

#### Applicability

- **Age restrictions**
  - **None**
  - **None**
  - **None**
  - **Not for very young children**
  - **None**

- **Digital and/or non-digital learning contexts**
  - **Digital**
  - **Both**
  - **Mostly digital**
  - **Both**
  - **Both**

- **Generalizes to learning technologies beyond training data**
  - **No**
  - **Yes**
  - **Yes**
  - **Yes**
  - **Yes**

- **Use in diverse environments (e.g., students home, subway)**
  - **Varies\(^1\)**
  - **Varies\(^1\)**
  - **Somewhat limited**
  - **Yes**
  - **Severely limited**

#### Scalability

- **Human labor costs**
  - **Low**
  - **Low**
  - **Medium**
  - **Low**
  - **High**

- **Equipment costs**
  - **Moderate**
  - **Moderate**
  - **High**
  - **None**
  - **None**

- **Use with large numbers of students**
  - **Varies\(^1\)**
  - **Varies\(^1\)**
  - **No**
  - **Varies\(^2\)**
  - **No**

- **Use over long timeframes**
  - **Varies\(^1\)**
  - **Varies\(^1\)**
  - **Limited\(^1\)**
  - **Moderate**
  - **Moderate**

---

Note. \(^1\)Depends on reliable computer and internet access. \(^2\)Depends on whether the self-reports are administered via computers or if paper and pencil format is used.
References


Hawkins, W., Heffernan, N., & Baker, R. S. (2013). Which is more responsible for boredom in intelligent tutoring systems: students (trait) or problems (state)? In A. Nijholt, S. D'Mello & M. Pantic (Eds.), *Proceedings of the 5th International Conference on Affective Computing and Intelligent Interaction (ACII 2013)* (pp. 618-623). Washington, DC: IEEE.


