Human Expert Labeling Process (HELP)

Towards a Reliable Higher-Order User State Labeling Process and Tool to Assess Student Engagement

Sinem Aslan, Sinem Emine Mete, Eda Okur, Ece Oktay, Nese Alyuz, Utku Ergin Genc, David Stanhill, Asli Arslan Esme

In a series of longitudinal research studies, researchers at Intel Corporation in Turkey have been working towards an adaptive learning system automatically detecting student engagement as a higher-order user state in real-time. The labeled data necessary for supervised learning can be obtained through labeling conducted by human experts. Using multiple labelers to label collected data and obtaining agreement among different labelers on the same samples of data, it is critical to train all to use the engagement model accurately. Addressing these challenges, the researchers developed a rigorous human expert labeling process (HELP) specific to the educational context, with multi-faceted labels and multiple expert labelers. HELP has three distinct stages: (1) *Pre-Labeling*, including planning, labeler recruitment, training, and evaluation steps; (2) *Labeling,* involving actual labeling conducted by multiple labelers, and related steps for formative assessment of their performance; and (3) *Post-Labeling*, generating final labels and agreement measures through processing multiple decisions. In this article, the researchers outline proposed methods in HELP and describe the developed labeling tool.

Sinem Aslan (Ph.D.) is a Research Scientist and the Principal Investigator of the Adaptive Learning Project at Intel Corporation. As a researcher, she is interested in understanding how advanced technologies can be leveraged for enabling personalized experiences (e-mail: sinem.aslan@intel.com). **Sinem Emine Mete** is an Educational Researcher at Intel Corporation. Her research interests include interactive multimedia applications in education and personalized learning. **Eda Okur** is a Research Scientist at Intel Corporation. Her research interests include artificial intelligence, machine learning, and natural language processing. **Ece Oktay** is a Software Development Engineer at Intel Corporation. Her research interests include software application development, embedded software development, and system design. **Nese Alyuz** (Ph.D.) is a Research Scientist at Intel Corporation. Her research interests include computer vision, image processing, artificial intelligence, pattern recognition, and machine learning; specifically for human-computer interaction applications. **Utku Ergin Genc** is a Research Scientist at Intel Corporation. He is interested in autonomous machines, data analytics, sensors, and embedded systems. **David Stanhill** (D.Sc.) is a Research Scientist at the Perceptual Computing Group at Intel Corporation. His research interests include computer vision and machine learning, especially topics related to human-machine interface. **Asli Arslan Esme** is a Research Director at Intel Corporation. Her research interests include cognitive computing, neuromorphic computing, emotional intelligence, the Internet of Things (IoT), and wearables.

Introduction

In a series of longitudinal research studies, researchers at Intel Corporation in Turkey have been working towards an adaptive learning system incorporating machine learning and perceptual computing to detect student engagement as a higher-order user state in real-time (Aslan *et al.*, 2014). To classify the engagement level of a student, supervised machine learning is preferred, where a large amount of data is needed to be labeled for the training engagement model. For traditional object classification problems, the labeling process involves coding of scenes indicating presence/absence of a specific object/activity. Such problems are rather objective and can be handled by any trained non-expert. In such contexts, guidelines to train labelers can be very obvious: If you see a car in the picture, label the scene as a picture with a car.

As opposed to such problems, labeling student states in a classroom environment (i.e., in the wild) requires a lot of cognitive processing, as there are many variables for labelers to consider before selecting a certain label. For example, a student playing with her hair can signify either a state of boredom or frustration, and this gesture itself cannot be used to decide on a label. On top of this context-complexity, this research focuses on multi-dimensional student states—higher-order user states. For student engagement, the researchers incorporate two dimensions of labeling: behavioral and emotional (i.e., affective). With behavioral labeling, they aim to capture how much a student is into a learning task, whereas with emotional labeling, they try to understand the student's emotional experience during the learning task. A composite of behavioral and emotional labels would give a holistic picture of the student's engagement state. By nature, such states are ambiguous, and labeling those states is relatively subjective. Thereby, to obtain ground-truth labels as accurately as possible, rigorous labeling methods should be implemented.

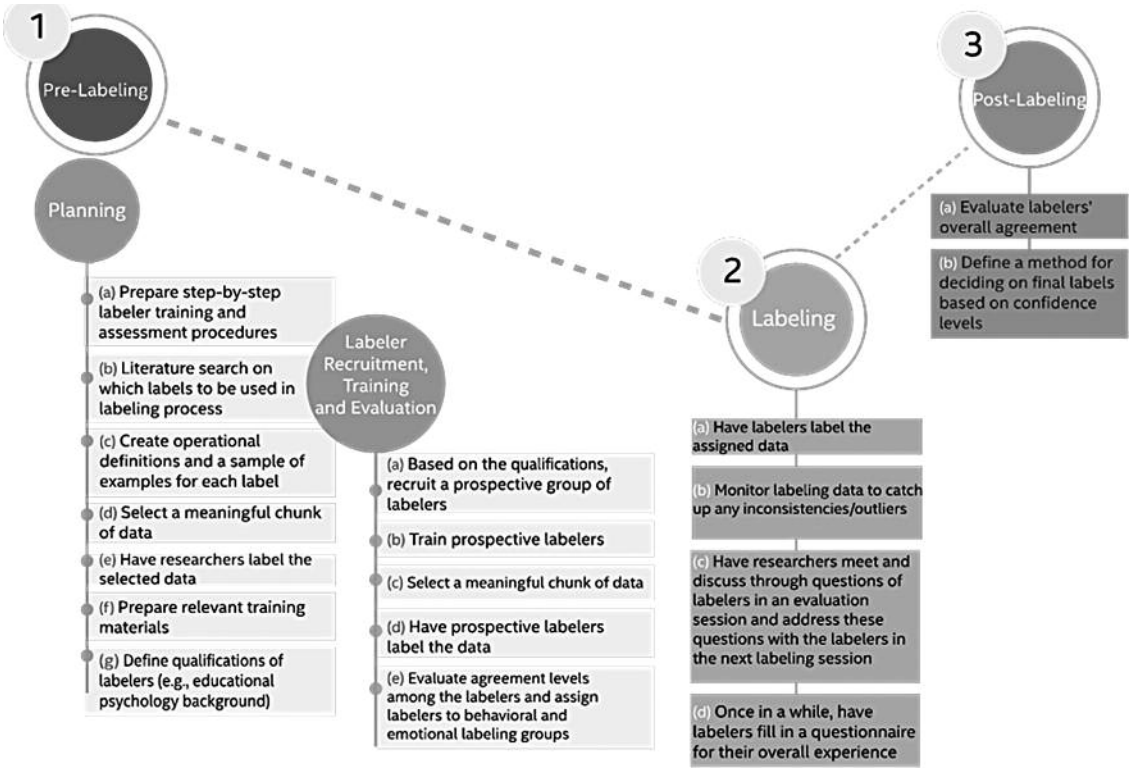

Figure 1. Overview of HELP.

In this line of research, the majority of studies to date have implemented a labeling process requiring one labeler only—some conducting *post facto labeling*—labeling after data collection (Salmeron-Majadas *et al.*, 2015; Saneiro, Santos, Salmeron-Majadas, & Boticario, 2014), with some others doing *in vivo labeling*—labeling during data collection. A well-known example for the latter is BROMP (Bosch *et al.*, 2015; Ocumpaugh, 2015; Ocumpaugh, Baker, & Rodrigo, 2012). BROMP is a highly preferred labeling protocol used to record observations of student behaviors and/or emotions in authentic field settings. In BROMP, students are observed in their classrooms one at a time through a round-robin technique by an observer (see the study by Bosch *et al.* (2015) to review BROMP in action). BROMP has advantages in terms of labeling process, such as time (i.e., completing data labeling at the end of a lesson) and resource (i.e., decreasing labeling cost—any instance of a student is labeled by only one observer). However, there are some challenges unaddressed:

1. *Limited chance for revisions:* As BROMP requires an observer to complete real-time labeling within a time frame of a class, there is minimal chance for the observer to make any changes backward.
2. *Real-time, complex decision making:* For a holistic judgment of a student's state, BROMP suggests monitoring facial expressions, speech, body posture, gestures, and the student's interactions with a computer or other device. However, from a cognitive-processing perspective, it seems impossible to perform all of these in real-time—without having any option to stop the time and think about a final label to assign.
3. *Fragmented labeling experience:* BROMP requires a labeler to conduct observations using a round-robin technique—spending a short amount of time with one student and moving to the next. This results in a fragmented labeling experience instead of observing one student throughout the whole time.
4. *Limited labels for training a model:* The round-robin technique in BROMP results in loss of data and thereby labels. There is a high risk of disregarding important state changes. This signifies a need for continuous labeling.
5. *Observer effect:* Although having an observer label student data within a classroom can be advantageous in terms of labelers' having a more authentic labeling experience, there is an inevitable observer effect associated with it (Abikoff, Gittelman-Klein, & Klein, 1977).

To address the above challenges, the researchers at Intel have developed a set of guidelines enabling a reliable higher-order user state labeling by human experts:

Human Expert Labeling Process—HELP. Throughout a series of longitudinal, three-year research studies with three pilots conducted with students in authentic classrooms, this process has been consolidated and refined (Aslan et al., 2014). The researchers have been conducting various experiments for improving the labeling methods and tools. This article outlines the details of HELP together with a description of the labeling tool as well as the preliminary experiments and results.

The HELP Process

HELP has three distinct stages (see ***Figure 1***): (1) *Pre-Labeling*—planning, labeler recruitment, training, and evaluation steps; (2) *Labeling*—actual labeling conducted by multiple labelers along with steps for formative assessment of their performance; and (3) *Post-Labeling*—generating final labels and agreement measures through processing multiple labelers' decisions.

Stage 1: Pre-Labeling

Planning

a. At this very first stage, it was necessary to design how to train prospective labelers, what sort of materials and tools to use during training, and criteria to select the best subset of labelers to be recruited. Labeler training has two main targets: (1) train labelers on labeling process, and disqualify those with low agreement levels; and (2) fine-tune the labeling process according to feedback from labelers to maximize the agreement level.

b. Researchers reviewed related literature and leveraged their own educational expertise to decide on appropriate labels for labeling of behavioral and emotional states. Based on their in-depth research, the researchers decided to use the labels given in the next step (c), taking the circumplex model of affect (Russell, 1980) as a reference.

c. The definitions and examples of each selected label were created based on the literature (Bosch *et al.*, 2015; D'Mello, 2013; Kapoor, Burleson, & Picard, 2007; Woolf *et al.*, 2009), discussions with teachers, and students' observable behaviors.

Behavioral labels:

- ***On-Task:*** If the student is active in the learning task (e.g., s/he is watching relevant instructional videos/ solving questions, etc.).
- ***Off-Task:*** If the student is not active in the learning task.
- ***Can't Decide:*** If the labeler cannot decide on the behavioral state.
- ***Not Available (N/A):*** If the data cannot be labeled (e.g., while the student is preparing to leave the class at the end of the session).

Emotional labels:

- ***Satisfied:*** If the student is not having any emotional problems during the learning task. This can include all positive states of the student, from being neutral to being excited during the learning task.
- ***Bored:*** If the student is feeling bored during the learning task.
- ***Confused:*** If the student is getting confused during the learning task—in some cases, this state might include some other negative states, such as frustration.
- ***Can't Decide:*** If the labeler cannot decide on the emotional state.
- ***Not Available (N/A):*** If the data cannot be labeled (e.g., while the student is preparing to leave the class at the end of the session).

d. After determining the labels and their operational definitions, a sample chunk of data was selected to be labeled in the labeling practice by researchers. While selecting these data, certain considerations were made:

- i. For behavioral labeling practice, 2 x ~6 min. segments were chosen, as representatives of the two major labels (i.e., On-Task, Off-Task), from both assessment and instructional activities.
- ii. For emotional labeling practice, 3 x ~10 min. segments were chosen, as representatives of the three major emotional labels (i.e., Satisfied, Bored, Confused), selecting both from assessment and instruction.

e. A selected chunk of data was labeled by the researchers to check the validity of the examples and the definitions of the given labels prior to sharing them with prospective labelers.

f. Next, training handouts including the definitions and examples of each label for both behavioral and emotional labeling were prepared. Additionally, a labeling tool user manual was prepared to guide labelers during the training session (see the labeling tool section).

g. Lastly, the qualifications of the labelers in terms of their professional background were identified. As the researchers incorporated higher-order user states in their research, they used educational psychologists for making valid labeling of students' complex states.

Labeler Recruitment, Training, and Evaluation

a. A group of prospective labelers with educational psychology background was invited to take place in the training session.

b. Researchers described the project and labeling job requirements, and explained the definitions of each label by giving and demonstrating some examples from the project pilots. Then, the labeling tool and its func-

Figure 2. Labeling tool visualized for the behavioral engagement.

tionalities were explained to the prospective labelers. The detailed procedure followed at this stage, for both behavioral and emotional labeling, is described below. Note that these five steps were conducted individually for each prospective labeler, and each candidate did labeling for both behavioral and emotional states:

1. Describe setup and context of the collected data.
2. Give (printed) definitions of labels to prospective labelers, and explain these states within the context of 1:1 learning.
3. Demonstrate how to use the labeling tool.
4. Using the tool, present examples for different students' states. Demonstrate what clues were used (e.g., face, head motion, posture).
5. Discuss the examples and make sure all candidates are on the same page at a high level for definitions. Note that labeler training requires a more descriptive approach rather than prescriptive, as higher-order user states are ambiguous in nature. The researchers empower expert labelers to make decisions using their expertise in educational psychology.

c. In the practice part of the training session, a specific chunk of data was labeled by individual prospective labelers.

d. The practice part was divided into two rounds. First, prospective labelers labeled the selected chunk of data for behavioral labeling, and then they continued to label the selected data for emotional labeling. After each round, a group discussion was created to get feedback from labelers and provide them guidance to achieve a mutual agreement about labeling specific cases.

e. At the end of this training, researchers analyzed prospective labelers' data and their labels using a rigorous evaluation procedure. Based on these results, labelers were recruited and assigned to behavioral and emotional labeling tasks separately. At the end, the researchers assigned three labelers for behavioral labeling and five labelers for emotional labeling. Assigning labelers with top agreement levels to emotional labeling is suggested due to relative complexity.

Stage 2: Labeling

a. Labelers started labeling the scheduled data on a daily basis. For this case, the researchers had them work 2.5 days a week. Whole-week labeling is not legitimate, as it is a cognitively-tiring process.

b. Throughout the labeling process, labelers were monitored regularly to catch any outliers. Towards this end, the researchers created a module within the tool to visually monitor labels across different labelers for the same data. The researchers used this module on a weekly basis to check any discrepancies, together with the overall inter-rater agreement measures calculated over the available data using the Krippendorff's Alpha (Krippendorff, 1995). Based on the findings, the researchers gave formative feedback to labelers.

c. To monitor and keep track of labelers' questions during the process, the researchers used a Q&A document to input such questions along with how the researchers addressed them. On a weekly basis, the

researchers shared the updated document with all labelers. This way, they could see some of the specific questions coming from different labelers and how the researchers addressed them.

d. At the end of the labeling process, a questionnaire was delivered to understand labelers' experiences (e.g., strengths and weaknesses of the labeling process, labeling tool, etc.). The researchers used this feedback to improve the next labeling cycle.

Stage 3: Post-Labeling

a. The inter-rater agreement measures both for behavioral and emotional labelers were calculated using the Krippendorff's Alpha (Krippendorff, 1995). In case of any significant outlier, the corresponding labelers were discarded from the subsequent process of deciding on the final labels.

b. After filtering out any significantly inconsistent labelers' data, it was necessary to define a method to decide on the final labels. In general, when there is a majority among the labelers, the mostly voted label is assigned as the final label. As for all of the labels, the instances with majority of 'Can't Decide' are labeled accordingly. However, the instances of class 'Can't Decide' can easily be extended: If there is a strong disagreement, these instances are labeled as 'Can't Decide.'

The Labeling Tool

In the data collection sessions, students used a content platform which enabled them to conduct instructional activities (e.g., watching instructional videos, reading instructional articles) and doing exercises (i.e., assessment activities) on a laptop computer. During these activities, the video of the students with a 3D camera (i.e., Intel® RealSense™ Camera F200), desktop videos, and context and performance logs from the content platform were recorded.

Visualization of the tool enables displaying information from different modalities, such as the data collected by the camera and data collected from the content platform's event logs. The labeling tool enables an external observer to label the data by assigning predefined labels to the labeler-defined session segments (see ***Figure 2***).

The collected data are displayed in the form of two video streams: (1) RGB videos and (2) desktop videos of individual students. The tool incorporates playback controllers to facilitate data visualization and labeling. Moreover, different contextual data segments are displayed along the timeline with different colors: blue for instruction and grey for assessment segments. Contextual data such as student ID, session number, question number in exercise segment, or attempt number to solve questions are visualized in the text fields together with date and time of the data collection. The ability to jump to the next/previous video/article or exercise segment is enabled. To improve labeling experience and increase accuracy of labeling, audio data recorded during the data collection sessions are also integrated into the labeling tool. The tool requires labelers to assign labels to the segments they define based on the state changes. The assigned labels are then visualized along the timeline.

Preliminary Experiments and Results

The researchers carried out a number of preliminary experiments towards consolidating and refining procedures of HELP. The experiments aimed to answer three major research questions:

1. Is appearance modality (i.e., video-recording of students) in itself enough for labelers to accurately label students' emotional states, or is it necessary to include contextual modality (i.e., desktop videos of students)?
2. For emotional states, is it practical to define a separate state for each quadrant of the circumplex model (Kapoor, Burleson, & Picard, 2007) (i.e., the arousal-valence graph)?
3. Using HELP, can an acceptable agreement be achieved among labelers?

Note that the list of experiments is not limited to these. The research is still on-going with various experiments to improve the proposed process. In the following sections, each experiment will address each research question noted above.

Experiment 1: Combination of Modalities

For this experiment, the researchers used six hours of student data (recorded while students were reading articles and solving exercises on a digital platform) and asked three labelers to label the data based on emotional states using two different labeling tools. In the first tool, the labelers labeled the data where the tool included both the students' recorded videos and their desktop videos. A week after that, the labelers labeled the data using a different version of the tool, where they could only see the recorded videos of the students, without their desktop videos.

In this experiment, the researchers aimed to observe the effects of using contextual information (i.e., students' desktop videos) during the labeling process. The researchers incorporated a qualitative method to understand such effects, as they were mostly interested in how labelers made sense of information in the tool. Towards this end, the researchers applied "think-aloud-protocol" (Groves *et al.*, 2009) during the labeling. They recorded the labelers' speech and videos while using the two different labeling applications (i.e., with and without desktop videos). The researchers conducted thematic analysis on these data (Boyatzis, 1998).

Table 1. Theoretical foundations for engagement labeling (adapted from Woolf *et al.*, 2009).

Behavioral State	Emotional State	Desirability Value	Engagement State
On-Task	Highly Motivated/ Excited	Highly Desirable	Engaged
On-Task	Calm/ Satisfied	Highly Desirable	Engaged
On-Task	Confused/ Frustrated	Maybe Desirable	Engaged
On-Task	Bored	Not Desirable	Not Engaged
Off-Task	Highly Motivated/ Excited	Not Desirable	Not Engaged
Off-Task	Calm/ Satisfied	Not Desirable	Not Engaged
Off-Task	Confused/ Frustrated	Not Desirable	Not Engaged
Off-Task	Bored	Not Desirable	Not Engaged

The results show that when there is no desktop videos, the labelers mostly had difficulty to decide between *'Satisfied'* and *'Bored'* states. Additionally, the labelers indicated that they benefited from contextual information during the labeling process. The labelers stated having difficulty when labeling without seeing the desktop videos of the students. This implies that providing students' desktop videos as context information is important for labeling.

Experiment 2: Selection of Learning-Related Emotional States

The researchers based the foundations of engagement modeling on the work by Woolf *et al.* (2009), where the researchers define engagement as a combination of behavioral and emotional states based on their desirability value (see ***Table 1***). For emotional states, previously the researchers assigned one state for each quadrant of the circumplex model (Russell, 1980), having four states in total: 'Excited,' 'Calm,' 'Bored,' and 'Confused.' However, during the initial labeling trials, the researchers had post-interviews with the labelers where the results revealed that the distinction between positive and negative arousal for positive valence states was not clear (e.g., 'Excited' vs. 'Calm'). In addition to this feedback, as illustrated in ***Table 1***, the two positive valence quadrants can be treated in the same way, considering either desirability levels (as in Woolf *et al.*, 2009) or engagement states. Based on these findings, the researchers merged the positive valence states ('Excited,' 'Calm') into a single one: 'Satisfied.'

To reinforce this decision, the researchers performed an experiment, where 10 hours of student data* (recorded while the students were watching math videos and solving related exercise questions on a content platform) were labeled by three labelers with two different label sets: One with six states where each quadrant is represented separately ('Excited,' 'Calm,' 'Bored,' 'Confused,' 'Can't Decide,' 'N/A'); and the other with five states, where positive valence states are merged ('Satisfied,' 'Bored,' 'Confused,' 'Can't Decide,' 'N/A'). When the subset of data was labeled with a single 'Satisfied' state, the inter-rater agreement level computed using the Krippendorff's Alpha was approximately doubled. Although an increase in agreement is expected by switching from six to five states, the substantial improvement indicates that the better labeling was achieved by having one positive valence state.

Experiment 3: Inter-Rater Agreement Measures

On a dataset of approximately 30 hours collected from 12 students* in four one-hour sessions in an authentic classroom (where students were watching math videos and solving related exercise questions on a content platform), the researchers utilized HELP to obtain both behavioral and emotional engagement states of the students. The agreement among the labelers is expected to highlight the subjective nature of the task, both for the behavioral and the emotional labeling. Therefore, for both types of labeling, the researchers computed the overall agreement: For behavioral, the inter-rater agreement of three labelers over four states; and for emotional, the inter-rater agreement of the five labelers over five states were calculated. The researchers experimented with four different measures, namely Krippendorff's Alpha, Fleiss' Kappa, Cohen's Kappa, and Scott's Pi (Gwet, 2014), to investigate whether the choice of metric will affect the results. Agreement measures are summarized in ***Table 2***. As similar results were achieved by different metrics, the researchers utilized Krippendorff's Alpha in HELP (as given in Post-Labeling (b)) due to its applicability to

* Four 9th grade students: Three males and one female.

* Twelve 9th grade students: Nine males and three females.

Table 2. Inter-rater agreement measures for engagement labels.

Engagement	State Count	Labeler Count	Krippendorff's Alpha	Fleiss' Kappa	Cohen's Kappa	Scott's Pi
Behavioral	4	3	0.814	.835	.824	.824
Emotional	5	5	0.542	.559	.544	.545

multiple labelers. The agreement measures reported in ***Table 2*** show that, especially for emotional labeling, low-to-moderate agreement is achieved. This indicates that emotional labeling is a subjective task as expected, and it is necessary to have multiple number of labelers for each instance and to apply majority voting to obtain final decisions.

Conclusion

For an improved performance in the supervised strategies utilized, the researchers developed a rigorous labeling process specific to educational context, with multi-faceted labels and multiple expert labelers. In this article, the researchers outlined the details of this process, along with a labeling tool developed as a part of their longitudinal research. As the research is on-going, from a design-based research perspective, the researchers will continue refining this process and the labeling tool towards a more reliable higher-order user state labeling by human experts. □

References

Abikoff, H., Gittelman-Klein, R., & Klein, D. F. (1977). Validation of a classroom observation code for hyperactive children. *Journal of Consulting and Clinical Psychology, 45*(5), p. 772.

Aslan, S., Cataltepe, Z., Diner, I., Dundar, O., Esme, A. A., Ferens, R., ... & Yener, M. (2014). Learner engagement measurement and classification in 1:1 learning. In The 13th International Conference on Machine Learning and Applications (ICMLA) (pp. 545–552). New York: IEEE.

Bosch, N., D'Mello, S., Baker, R., Ocumpaugh, J., Shute, V., Ventura, M., ... & Zhao, W. (2015). Automatic detection of learning-centered affective states in the wild. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (pp. 379–388). New York: ACM.

Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. Thousand Oaks, CA: Sage.

D'Mello, S. (2013). A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. *Journal of Educational Psychology, 105*(4), p. 1082.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology (2nd ed.)*. Hoboken, NJ: John Wiley.

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters (4th edition)*. Gaithersburg, MD: Advanced Analytics, LLC.

Kapoor, A., Burleson, W., & Picard, R. W. (2007). Automatic prediction of frustration. *International Journal of Human-Computer Studies, 65*(8), 724–736.

Krippendorff, K. (1995). On the reliability of unitizing continuous data. *Sociological Methodology,* 47–76.

Ocumpaugh, J. (2015). *Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual*. Technical Report. New York: Teachers College, Columbia University; Manila, Philippines: Ateneo Laboratory for the Learning Sciences.

Ocumpaugh, J., Baker, D., & Rodrigo, M. A. (2012). *Baker-Rodrigo Observation Method Protocol (BROMP) 1.0. Training Manual*. New York: Teachers College, Columbia University.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161–1178.

Salmeron-Majadas, S., Arevalillo-Herráez, M., Santos, O. C., Saneiro, M., Cabestrero, R., Quirós, P., ... & Boticario, J. G. (2015). Filtering of spontaneous and low intensity emotions in educational contexts. In *International Conference on Artificial Intelligence in Education* (pp. 429–438). New York: Springer.

Saneiro, M., Santos, O. C., Salmeron-Majadas, S., & Boticario, J. G. (2014). Towards emotion detection in educational scenarios from facial expressions and body movements through multimodal approaches. *The Scientific World Journal; doi:10.1155/2014/484873* .

Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., & Picard, R. (2009). Affect-aware tutors: Recognising and responding to student affect. *International Journal of Learning Technology, 4*(3), 129–164.