Harley, J. M., Bouchet, F., Hussain, S., Azevedo, R., & Calvo, R. (2014). *A multi-componential analysis of emotions during complex learning with an intelligent multi-agent system*. Paper to be presented at a symposium on Interdisciplinary Approaches for Analyzing Data from Multiple Affective Channels with Computer-Based Learning Environments at the 2014 annual meeting of the American Educational Research Association, Philadelphia, PA.

A Multi-Componential Analysis of Emotions during Complex Learning with an Intelligent Multi-Agent System[1]

Jason M. Harley[1], François Bouchet[2], M. Sazzad Hussain[3],
Roger Azevedo[4], and Rafael A. Calvo[3]
[1]McGill University, Montréal, Québec, Canada
[2]Sorbonne Universités, UPMC Univ Paris 06, Paris, France
[3]The University of Sydney, NSW, Australia
[4]North Carolina State University, Raleigh, North Carolina, USA

Corresponding author: {jason.harley@mail.mcgill.ca}

Abstract. In this paper we discuss the methodology and results of aligning three different emotional measurement methods (automatic facial expression recognition, self-report, electrodermal activation) and their agreement regarding learners' emotions. Data was collected from 67 undergraduate students from a North American university who interacted with MetaTutor, an intelligent, multi-agent, hypermedia environment for learning about the human circulatory system, for a 1 hour learning session (Azevedo et al., 2013, Harley, Bouchet, & Azevedo, 2013). A webcam was used to capture videos of learners' facial expressions, which were analyzed using automatic facial recognition software (FaceReader 5.0). Learners' physiological arousal was measured using Affectiva's Q-Sensor 2.0 electrodermal activation bracelet. Learners self-reported their experience of 19 different emotional states (including basic, learner-centered, and academic achievement emotions) using the Emotion-Value questionnaire (Harley et al., 2013). They did so on five different occasions during the learning session, which were used as markers to align data from FaceReader and Q-Sensor. We found a high agreement between the facial and self-report data (75.6%) when similar emotions were grouped together along theoretical dimensions and definitions (e.g., anger and frustration) (Harley, et al., 2013). However, our new results examining the agreement between the Q-Sensor and these two methods suggests that electrodermal (EDA/physiological) indices of emotions do not have a tightly coupled (Gross, Sheppes, & Urry, 2011) relationship with them. Explanations for this finding are discussed.

Introduction

[1] The research presented in this paper has been supported by a graduate student fellowship from a Joseph-Armand Bombardier Canadian Graduate Scholarship (CGS) from the Social Science and Humanities Research Council (SSHRC) awarded to the first author and funding from the National Science Foundation (IIS 1008282), The Canada Research Chairs program, and the Social Science and Humanities Research Council awarded to the fourth author.

Emotions are a critical component of effective learning and problem solving with computer-based learning environments (CBLEs) (Azevedo & Aleven, 2013; Azevedo & Strain, 2011; D'Mello, 2013; Harley, Bouchet, & Azevedo, 2013; Lester et al., 2013; McQuiggan & Lester, 2009; Pekrun, 2011; Strain, Azevedo, & D'Mello, 2012; Woolf, et al., 2009). Despite its historic neglect, there has been a surge in interdisciplinary research, which has led to a plethora of new tools and technologies to measure emotions (Calvo & D'Mello, 2011, 2012). This surge in research and technology has, however, led to a variety of emerging conceptual, theoretical, methodological and measurement issues that need to be resolved before educational prescriptions can reliably and validly be used to improve learners' emotions (e.g., adaptive emotions). Adaptive emotions facilitate students' learning and include both positive emotions, such as engagement and curiosity, as well as neutral states in which students can still concentrate on learning (Harley & Azevedo, under review; Harley et al., 2013; Pekrun, 2011). In contrast, negative emotions have typically been found to impair attentional and motivational processes (e.g., boredom, frustration, high anxiety; Harley & Azevedo, in press; Harley et al., 2013; Pekrun, 2011). One key area in the development of educational prescriptions that would target learners' adaptive emotions is the use of multiple data channels to measure their emotions during interactions with CBLEs (see Azevedo et al., 2013).

Using multiple channels (e.g., facial expressions, self-report measures, physiological signals) to analyze learners' emotional states is well aligned with theories that define emotions as multi-componential (behavioral, physiological, experiential / feeling) appraisal-driven responses to objectives which have valence (positive/negative) and arousal (high/low) dimensions (Gross, 2010, 2013; Pekrun, 2006, 2011). Multimodal approaches also afford researchers the opportunity to circumvent the constraints of individual channels (e.g., Hawthorne effect; physiological channels cannot be socially masked) and therefore achieve greater construct validity and reliability.

Recently, an increasing number of CBLEs have incorporated multiple emotional measurement channels (e.g., physiological sensors, facial expression coding) in order to detect, measure, and adapt to learners' emotional states (Baker et al., 2012; D'Mello & Graesser, 2013; McQuiggan & Lester, 2009). However, many challenges exist which make the development and use of research platforms that include multiple emotional channels a formidable challenge, including: (1) differences in the sampling rate of emotional data (e.g., frame rate for automatic facial recognition vs. pre-determined time intervals for self-report measures); (2) variation in the detail and kind of emotional information that different channels can record (e.g., one dimension [arousal] for EDA bracelets vs. discrete emotional states from facial expressions); and (3) disagreement amongst theories regarding how tightly or loosely coupled emotional responses should be, when data comes from different psychological components (behavioral, physiological, experiential / feeling; Gross et al., 2011).

The purpose of this paper is to address some of these challenges. The first and second issues can be summarized by the following research question: How can we use emotion measurement methods, which have different characteristics, in combination? This question is answered though a detailed description of the methodological approaches used in this study to extract, treat, and align data from in-session self-reports, automatic facial expression detection, and electro dermal activation (EDA) data. A second question captures the third issue: Do our results, which compare the agreement between channels, support a tight or loose coupling of psychological components? In other words, do different channels identify the same emotion (e.g., anger) or provide complementary emotional information at a given point in time (e.g., high arousal)? This question is addressed through a theoretical and contextually situated discussion of the study's results.

Methods and Data Sources

This section has been structured to provide details about the participants of this study as well as the learning environment, MetaTutor (Azevedo et al., 2013) (and apparatus) before describing the different methods used to measure learners' emotions. The Experimental Procedure describes the context and process through which the data were collected. How the data were extracted and aligned for the purposes of comparison is described in the Data Analysis section.

Participants

Sixty-seven (*N*= 67) undergraduate students from a large, public university in North America participated in this study. Participants (82.8% female, 72.4% Caucasian) were randomly assigned to either of the two conditions tested.

MetaTutor and Apparatus

MetaTutor (Azevedo et al., 2010, 2011, 2013; Azevedo, Behnagh, Duffy, Harley, & Trevors, 2012) is a multi-agent Intelligent Tutoring System (ITS) and hypermedia learning environment which consists of 38 pages of text and static diagrams organized by a table of contents displayed in the left pane of the environment. The version of MetaTutor used in this experiment is comprised of material on the human circulatory system, which it is designed to teach participants about during their interactions with four embedded, pedagogical agents (PAs). The four PAs' instructional scaffolding varied depending on the experimental condition learners were assigned to (aside from PA scaffolding, the conditions were identical). In the prompt and feedback condition (PF) condition, learners were prompted by the PAs to use specific self-regulatory processes (e.g., to metacognitively monitor their emerging understanding of the topic or deploy a specific cognitive learning strategy such as re-reading or coordinating informational sources), and were given feedback about their use of those processes. In the control (C) condition, participants did not receive prompts or feedback and could only perform these self-regulatory processes on their own initiative.

A Logitech Orbit AF webcam was used to record the participants' faces during their interaction with MetaTutor. In accordance with FaceReader's guidelines, the camera was mounted above the monitor of the computer participants were using, in order to capture their faces, but not obstruct the screen. Videos were recorded as WMV files with a resolution of 1600x1200, and 12.1 frames per second on average.

Measures and Materials

Q-Sensor 2.0. Q-sensor (Affectiva, 2013) was used to measure learners' electrodermal activation (EDA). EDA refers to electrical changes at the surface of the skin caused by sympathetic activity which alters sweating. EDA is commonly used to measure physiological arousal. One method of measuring EDA is to measure the variations of electrical conductance of the skin (expressed in micro Siemens (μS))**.** The Q-Sensor accomplishes this by passing a small amount of current between two electrodes placed on the skin. Measurements are understood in relative terms because individuals' EDA baseline varies. Arousal is therefore inferred based on a higher or lower level than an individuals' average or baseline resting level. Higher levels may be induced by excitatory stimuli, for example, a bad score on a quiz could provoke anxiety. Conversely, an interesting piece of information may engage the learner, having the same effect, but with an adaptive emotional outcome (e.g., curiosity) rather than a negative one. Lower levels of arousal suggest that the learner may be relaxed or bored, perhaps from reading a page of content that the learner isn't interested in or doesn't find particularly challenging.

Participants were asked to put the Q-Sensor bracelet on before beginning their learning session with MetaTutor and before other recording devices were set up (i.e., webcams positioned). This combined with the videos introducing the learning environment typically afforded 10-15 min. of baseline data collection before participants began interacting with MetaTutor. Q-Sensor 2.0 provides eight values every second. The Q-Sensor was developed by Picard and colleagues who have examined EDA in the context of learning and intelligent tutoring systems (ITSs) and found it to be an effective predictor of affective states (Kapoor, Burleson, & Picard, 2007; Woolf et al., 2009).

FaceReader 5.0. FaceReader (VicarVision, 2013) analyzes participants' facial expressions and provides a classification of their emotional states. It uses an Active Appearance Model which models participants' facial expressions, and an artificial neural network with seven discrete outputs (corresponding to Ekman and Friesen's six basic emotions, in addition to neutral; Ekman & Friesen, 1992), that classifies participants' constellations of facial expressions. FaceReader has been validated through comparison with human coders (Terzis, Moridis, & Economides, 2010).

FaceReader provides a score between 0 and 1, for each frame of each participant's video for each of Ekman's six basic emotions, in addition to neutral. FaceReader also provides information about the dominant emotional state (computed with a proprietary algorithm using the scores of the seven emotional states in the previous frames) and timestamp information regarding the on and offset of the hierarchical rankings of these states.

Emotions-Value Questionnaire (EV). During the learning session, participants were asked on five occasions (see section 2.4) by a PA to complete the EV questionnaire, for which each participant responded to 20 items: 19 items on emotions and 1 item on task value which was not considered in this analysis. These items were on a 5-point Likert scale ranging from “Strongly Disagree” to “Strongly Agree.” One example item is: “right now I feel bored.” The 19 emotions that are measured using the EV represent an exhaustive list of discrete basic and learner-centered emotions that appear in the research and theories of a variety of emotion researchers (e.g., D’Mello, Lehman, Person, 2010; Pekrun, Goetz, Frenzel-Anne, Petra, & Perry, 2011). Definitions, based on these researchers’ work and operationalizations of these emotions, were used to create a digital definition handout that was provided in a side panel to participants every time they filled out an electronic version of the EV embedded in MetaTutor. The instructions and wording of the questions were based on a subscale of Pekrun and colleagues’ academic emotions questionnaire (AEQ; Pekrun, Goetz, Titz, Perry, 2002) which assesses participants’ concurrent, ‘right now’ state-emotions as opposed to emotions generated from prospective or retrospective focal points. The majority of the 19 emotions can be conceptualized into different quadrants along the axis of valence (positive/negative) and activation (activating/deactivating) (Pekrun, et al., 2002, 2011).

Experimental Procedure

During Day One of the experiment, which took approximately 30 minutes, participants read and signed the informed consent form, took a pretest on the human circulatory system, completed a demographics questionnaire, and several self-report measures (e.g., AEQ trait emotions; Pekrun et al., 2002) on a computer with their face being video recorded. For Day Two, we collected video, audio, and physiological data on each participant while they used MetaTutor for about 90 min to learn about the human circulatory system. At the beginning of the learning session participants set up two sub goals for learning about the human circulatory system and proceeded to interact with MetaTutor and its learning content for one hour; half-way through, they were asked to complete the concurrent state AEQ and then invited to take a five-minute break. At the end of their learning session, learners filled out the post-test measure and a series of self-report measures, including the retrospective state AEQ. Days One and Two occurred at least one hour apart from each other and no more than four days apart. The first time participants filled out the EV was at the beginning of the learning session after they had successfully set two sub goals. The following occasions occurred regularly every 14 minutes during the one hour learning session, with the fifth EV being administered just before learners took the post-test. Participants had as much time as necessary to fill out the EV on each occasion.

Data Analysis

This section describes the steps that were taken in order to treat and extract data from the individual channels (EV, FaceReader, Q-Sensor). The processes we used to align them in order to calculate their agreement rates are also described.

Treating and extracting data from individual channels.

EV. Several scores on different emotions on the EV measure were identified as univariate outliers with standardized scores exceeding z = +/- 3.29 and were therefore replaced with the next most outlying values for each variable (Tabachnick & Fidell, 2007). Several variables were identified as being skewed with values exceeding z = +/- 3.20. Only emotion variables that were skewed across all five EVs were transformed, including fear, shame, hopelessness, disgust, sadness, and eureka. Square root, logarithmic, and inverse transformations were performed, but did not normalize the distributions for all variables (only hopelessness and eureka). Two to three of the five EV variables for anger, contempt, surprise, and confusion were skewed, but were not transformed in order to maintain consistency across the measures of each emotion.

FaceReader 5.0. Data was exported from the FaceReader program to CSV files. FaceReader data was collected for analysis ten seconds prior to the administration of the EV measures. Videos recorded during the two sessions of the experiment (with an average length of 40 and 100 minutes respectively) were imported and used to calibrate FaceReader with General or Asian face models. Videos of the second session (when the learning occurred) were then analyzed with the "smoothen classification" parameter enabled.

Q-Sensor. Similar to the FaceReader data, EDA data was exported from the Q-Sensor 2.0 into CSV files and was collected ten seconds prior to the administration of the EV measures. The average microSiemens (μS) value was considered during these five periods of 10 secs. The features extracted (using the 10 second window) in these models included the EDA means and ranges of individual participants. Features were normalized on a 1-10 scale based on a user-dependent model that took participants' baseline values into consideration. The Augsburg Biosignal Toolbox (AubT)[2] in Matlab was used for extracting the features.

Aligning individual channels.

Aligning FaceReader and EV Data. We aligned FaceReader's dominant state with the EV by extracting log information corresponding to the 10 seconds of video footage of participants' right before they were asked to fill in each of the EVs. This period of time was selected because it was short enough to capture the emotion participants were experiencing at the moment, which change rapidly. It was also long enough to provide additional data that would prevent "noise", such as a participant blinking or rubbing their face, from eliminating the data point.

We selected the primary dominant state defined as the state reported as dominant during the majority of the 10 seconds. In 80.7% of the cases, no other unique emotion was dominant for more than 3s, which makes it unnecessary to consider the possibility of a secondary co-occurring emotion (Harley, Bouchet, & Azevedo, 2012). Moreover, in 92.9% of the remaining situations, neutral was either the primary or secondary dominant emotion.

67 participants were analyzed, but nine of them were excluded from our sample because their dominant state in the 10s for at least three of the five EVs were identified as "Unknown" by FaceReader (this situation generally occurs when the participant's face is not sufficiently oriented towards the webcam, e.g. when they look down to type on the keyboard).

[2] http://www.informatik.uni-augsburg.de/lehrstuehle/hcm/projects/tools/aubt/

In order to evaluate the agreement between the self-reported emotions in the 5 EVs and the dominant emotion identified by FaceReader during the 10s before, we started by defining a mapping between the 13 non-basic emotions from the EV onto the 6 basic emotions in addition to neutral that are used by FaceReader to classify participants' emotions. Using work by Pekrun and colleagues (2002, 2011) on the AEQ, (1) all positively valenced activating emotions (enjoyment, hope, pride, curiosity and eureka) were associated with happy; among the negatively-valenced activating emotions, (2) frustration was grouped with anger, (3) anxiety with fear and (4) contempt with disgust, and (5) all negatively-valenced deactivating emotions (hopelessness and boredom) were associated with sadness, while the (6 and 7) non-valenced emotions (neutral and surprise) were kept as two distinct categories. Two additional emotions (confusion and shame) used in the EV could not be associated to any basic emotions and were therefore discarded for this analysis.

Given these seven groups of emotions, we defined that there was an agreement between FaceReader's dominant emotion and the EV if and only if one of the emotions associated to FaceReader's dominant emotion was rated with a score of 3 or more (out of 5) in the EV (e.g., if the dominant emotion according to FaceReader is anger, either anger or frustration need to have a score of 3 or more in the EV). The 20 (out of 290) occurrences of "Unknown" were excluded from this analysis.

Aligning FaceReader and Q-Sensor Data. In order to compare the EDA and FaceReader data, Q-sensor data was dichotomized into high and low using the standardized 10-point scale. Values of five and lower were classified as low levels of arousal, while values six and above were classified as high arousal. The seven emotions

FaceReader detects were each labeled as high or low arousal states. Neutral and sadness were classified as low-arousal states, while happiness, anger, surprise, disgust, and fear were classified as high-arousal states based on operationalizations of these and other emotions by D'Mello and colleagues (2010) and Pekrun (2011). Agreement was calculated by identifying how often the emotional states FaceReader classified fit the expected high or low levels of arousal.

EV and Q-Sensor. Similar to our alignment of the EV with FaceReader, we defined an EV emotion as present if it was given a value of three or more (out of five) by learners. Boredom, hopelessness, sadness, and neutral were classified as low arousal emotions. Shame, Surprise, Confusion, and Eureka were not examined. All other emotions were classified as high arousal. As learners sometimes reported more than a single emotion as present (i.e. with a score superior or equal to three), we calculated the agreement between each individual emotion and the Q-Sensor arousal value for that EV. For instance, if a learner reports Neutral with a 5 and Happy with a 3 in the EV while the Q-Sensor measures a low-arousal value, it will count as an agreement on Neutral and a disagreement on Happy. The overall agreement is then calculated based on the weighted mean of each of the 15 emotions considered.

Results

FaceReader and EV

Using this approach we have found a high agreement between the facial and self-report data (75.6%) when similar emotions were grouped together along theoretical dimensions and definitions (e.g., anger and frustration).

FaceReader and Q-Sensor

We found an agreement rate of 60.1% (κ = 0.07) between the Q-Sensor and FaceReader.

EV and Q-Sensor

We found an agreement of 41.9% (κ = .003) between Q-Sensor and the self-report measure of emotions. The highest agreement between the Q-sensor discrete emotions was between learners' self-reported experience of boredom and low arousal (67.5%) and neutral and low arousal (69.59%).

Conclusion and Discussion

This paper has addressed two research questions. The first, *how can we use emotion measurement methods, which have different characteristics, in combination?* was answered though a detailed description of the methodological approaches used in this study to extract, treat, and align data from three different methods of measuring emotions. Our results reveal that the answer to our second research question, *do our results, which compare the agreement between channels, support a tight or loose coupling of psychological components?* varies depending on which channels are being compared.

The high level of agreement between the EV and FaceReader provides evidence that facial expressions and learners' experience of emotions are tightly coupled (possess common emotional characteristics; Gross et al., 2011). In other words, if someone feels and expresses that they are happy, they will probably also have a matching facial expression (e.g., smile). This finding is in line with theories of emotion that hold that the different channels through which emotions are expressed will have coordinated responses (Ekman, 1992, Pekrun, 2011). For example, Pekrun describes a student's anxiety before an exam as comprising of "nervous, uneasy feelings (affective); worries about failing the exam (cognitive); increased heart rate or sweating (physiological); impulses to escape the situation (motivation); and an anxious facial expression (expressive)" (Pekrun, 2011, pp. 24). While theories of emotion vary in the number of discrete components that emotions are expressed through, this quote illustrates the idea that they are expected to be congruent.

Congruency between channels is not, however, supported by our results which examined the agreement between the Q-Sensor and these two methods. Rather, they suggest that the physiological component (i.e., EDA data) of emotions do not have a tightly coupled relationship with facial expressions and self-reported emotions; at least in the context of MetaTutor.

There are several potential explanations for this finding. First, it is possible that theoretically driven expectations that data from three different channels would be tightly coupled are not always appropriate. Instead, a tight coupling between all three channels may not necessarily exist, as other theorists posit (Barret, Mesquita, Ochsner, & Gross, 2007). Alternatively, how closely related emotional responses are from different channels may be a question of context. In a laboratory setting, for example, the levels of arousal detected by the EDA device may not possess enough variance to reliably differentiate between emotional states. An examination of both the self-report data and the facial expression data reveal that learners' experienced moderate to low levels of most emotions and a strong tendency toward a neutral emotional state. Since arousal levels are relative, the higher range of arousal experienced by students may not have been as high as it may be in other experimental contexts, such as playing a video game or viewing emotion eliciting photos. As such, skin conductance would not be as sensitive to changes in emotional states as the other channels.

Other contexts may elicit higher levels of arousal because of the cognitive appraisals that students make while interacting with them. Pekrun (2006, 2011) has identified two types of appraisals that exert a strong influence on the academic achievement emotions learners will experience. Learners' appraisals of subjective control include one's perception of the causal influence they exert over their actions and outcomes. Appraisals of value include the merit of an activity and its outcome(s), or more broadly, the perception that an action or outcome is positive or negative in nature. A recent selective review by Harley and Azevedo (in press) identified a tendency for learners' to experience greater proportions of positive emotions (e.g., engagement, curiosity) when interacting with computer-based learning environments that possessed game-like elements, afford students choice, and are based on content that is related to their studies. It also indicated that students tend to experience relatively few instances of the types of negative emotions that would be characterized as high arousal (e.g., anger, anxiety) while interacting with CBLEs. Therefore, CBLEs, such as MetaTutor, may represent a more challenging educational context in which to collect meaningful information from EDA data than other higher-stakes ones (e.g., studying for a unit related to the students academics, medical students practicing making diagnosis).

Another possibility for the lack of agreement between the EDA data and the other two channels relates to the methodology of this study. While guided by research on emotions in psychology, educational, and affective computing, many of the decisions regarding data analyses were made independent of analytic precedents (which have not been published) and therefore require further study and potential calibration. For example, it could be revealing to examine a more sophisticated categorization of the EDA data (beyond a dichotomization) in order to attempt to capture intermediate levels of arousal that may better represent emotions of different arousal levels. For example, anger and curiosity are both labeled as high arousal emotions, but differences between their typical arousal levels may exist and, if so, could help improve agreement between channels. The same situation applies to emotions labeled as low in arousal, such as neutral and boredom.

Although using a more sophisticated categorization of the EDA data this was not possible with this data set, future analyses with additional participants in the newer version of MetaTutor may yield a higher absolute arousal range and provide more variance. Additionally the application of more sophisticated machine learning techniques may yield more detailed parameters to categorize EDA levels. In the new version of MetaTutor, participants are asked to make a forced-choice self-report of their emotional state, which will make future alignment easier.

In conclusion, our paper provides a methodological description of how we have measured and aligned emotion data using three different methods. The high agreement rate we found between automatic facial recognition and self-report methods bolsters the validity of our emotion assessments with these two channels and provides a strong foundation to make valid and reliable diagnostic examinations of learners' emotions at discrete points during learning with MetaTutor. The agreement between these channels and the EDA device suggest that future research should be conducted, in particular, in environments expected to elicit higher arousal levels from students (e.g., serious game environments). Conceptually and theoretically, our results provide evidence that the experiential and behavioral components of emotions are tightly coupled. Educationally, improved measurement methods of emotions will lead to better informed interventions that can be designed to support and sustain adaptive emotional states during learning with CBLEs.

References

Affectiva (2013). Q-Sensor (2.0) [physiological measurement hardware]. Waltham, MA: Affectiva.

Azevedo, R., & Aleven, V. (2013). Metacognition and learning technologies: An overview of the current interdisciplinary research. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 1–16). Amsterdam, the Netherlands: Springer.

Azevedo, R., & Chauncey Strain, A. D. (2011). Integrating cognitive, metacognitive, and affective regulatory processes with MetaTutor. In R. A. Calvo & S. K. D'Mello (Eds.), *New perspectives on affect and learning technologies* (pp. 141–*1*54). Amsterdam, the Netherlands: Springer.

Azevedo, R., Johnson, A., Chauncey, A., & Burkett, C. (2010). Self-regulated learning with MetaTutor: Advancing the science of learning with MetaCognitive tools. In M. Khine & I. Saleh (Eds.), *New science of learning: Cognition, computers, and collaboration in education* (pp. 225–247). Amsterdam, the Netherlands: Springer.

Azevedo, R., Johnson, A., Chauncey, A., & Graesser, A. (2011). Use of hypermedia to convey and assess self-regulated learning. In B. Zimmerman & D. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 102–121). New York, NY: Routledge.

Azevedo, R., Harley, J., Trevors, G., Feyzi-Behnagh, R., Duffy, M., Bouchet, F., & Landis, R.S. (2013). Using trace data to examine the complex roles of cognitive, metacognitive, and emotional self-regulatory processes during learning with multi-agent systems. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 427-449). Amsterdam, The Netherlands: Springer-Verlag.

Azevedo, R., Behnagh, R., Duffy, M., Harley, J., & Trevors, G. (2012). Metacognition and self-regulated learning with advanced learning technologies. In D. Jonassen & S. Land (Eds.), *Theoretical foundations of learning environments* (2nd ed.) (pp. 171-197). Mahwah, NJ: Erlbaum.

Baker, R. S. D., Gowda, S. M., Wixon, M., Kalka, J., Wagner, A. Z., Salvi, A.,…Rossi, L. (2012). Towards sensor-free affect detection in cognitive tutor algebra. In K. Yacef, O. Zaiane, H. Hershkovitz, & J. Stamper (Eds.), *Proceedings of the 5th International Conference on Educational Data Mining,* Crete, Greece*,* 126-133.

Barrett, L. F., Mesquita, B., Ochsner, K. N., & Gross, J. J. (2007). The experience of emotion. *Annual Review of Psychology*, *58*, 373–403.

Calvo R. A., & D'Mello, S. (Eds.) (2011). *New Perspectives on Affect and Learning Technologies.* New York: Springer.

Calvo, R. A., & D'Mello, S. (2012). Frontiers of affect-aware learning technologies. *IEEE Intelligent Systems, 27*, 86-89.

Chauncey Strain, A., Azevedo, R., & D'Mello, S. (2013). Using a false biofeedback methodology to explore relationships among learners' affect, metacognition, and performance. *Contemporary Educational Psychology, 38*, 22-39.

D'Mello, S.K., Lehman, B., Person, N. (2010). Monitoring affective states during effortful problem solving activities. *International Journal of Artificial Intelligence in Education 20,* 361–389.

D'Mello, S. K., & Graesser, A. C. (2013). AutoTutor and affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems, 2*, 23-39.

D'Mello, S. K. (2013). A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. *Journal of Educational Psychology, 105(4), 1082-1099.*

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion, 6*, 169.

Gross, J. J. (2010). The future's so bright, I gotta wear shades. *Emotion Review, 2*, 212-216.

Gross, J. J. (2013). Emotion regulation: taking stock and moving forward. *Emotion,13*(3), 359–365.

Gross, J. J., Sheppes, G., & Urry, H. L. (2011). Emotion generation and emotion: A distinction we should make (carefully). *Cognition & Emotion, 25*, 765-781.

Harley, J. M., & Azevedo, R. (under review). Understanding students' emotions during interactions with advanced agent-based learning environments: A selective review. Paper submitted for presentation at the 2014 *Intelligent Tutoring Systems Conference*, Honolulu, Hawaii.

Harley, J., Bouchet, F., & Azevedo, R. (2012). Measuring learners' co-occurring emotional responses during their interaction with a pedagogical agent in MetaTutor. In S. A. Cerri, W. J. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Lecture*

Notes in Computer Science: Vol: 7315. Intelligent Tutoring Systems (pp. 40-45). Berlin, Heidelberg: Springer-Verlag.

Harley, J. M., Bouchet, F., & Azevedo, R. (2013). Aligning and comparing data on learners' emotions experienced with MetaTutor. In C. H. Lane, K. Yacef, J. Mostow, P. Pavik (Eds.), *Lecture Notes in Computer Science: 7926. Artificial Intelligence in Education* (pp. 61-70). Berlin, Heidelberg: Springer-Verlag.

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

Lester, J. C., Mott, B. W., Robison, J., Rowe, J., & Shores, L. (2013). Supporting self-regulated science learning in narrative-centered learning environments. In R. Azevedo & V. Aleven (Eds.), *International handbook on metacognition and learning technologies* (pp. 471–483). Amsterdam, the Netherlands: Springer.

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review, 18,* 315-341.

Pekrun, R. (2011). Emotions as drivers of learning and cognitive development. In R. A. Calvo & S. D'Mello (Eds.), *New Perspectives on Affect and Learning Technologies* (pp. 23-39). New York: Springer.

Pekrun, R., Goetz, T., Frenzel-Anne, C., Petra, B., Perry, R.P (2011). Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary Educational Psychologist. 36*, 34–48 (2011)

Pekrun, R., Goetz, T., Titz, W., Perry, R. (2002). Academic achievement emotions in students' self-regulated learning and achievement: a program of quantitative and qualitative research. Educational Psychologist 37, 91–206.

McQuiggan, S. W., & Lester, J. C. (2009). Modeling affect expression and recognition in an interactive learning environment. *International Journal of Learning Technology 4*, 216-233.

Tabachnick, B. G., & Fidell, L. S. Using Multivariate Statistics (5th ed.). Pearson Education/ Allyn and Bacon, Boston, MA (2007)

Terzis, V., Moridis, C. N., & Economides, A. A. (2010). Measuring instant emotions during a self-assessment test: The use of FaceReader. In A.J. Spink, F. Grieco, O. E. Krips, L. W. S. Loijens, L.P.J.J. Noldus, & P. H. Zimmerman (Eds.). *Proceedings of Measuring Behavior 2010* (pp. 192-195). Eindoven, The Netherlands: ACM.

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

VicarVision (2013). FaceReader (5.0) [Computer software]. Wageningen, The Netherlands: Noldus Information Technology.