Computers in Human Behavior 48 (2015) 615–625

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

A multi-componential analysis of emotions during complex learning with an intelligent multi-agent system ☆

Jason M. Harley [a,b,*], François Bouchet [c,d], M. Sazzad Hussain [e], Roger Azevedo [f], Rafael Calvo [e]

[a] Université de Montréal, Department of Computer Science and Operations Research, 2920 Chemin de la Tour, Pavillon André-Aisenstadt 2194, Montréal, QC H3C 3J7, Canada
[b] McGill University, Department of Educational and Counselling Psychology, 3700 McTavish Street, 614, Montréal, QC H3A 1Y2, Canada
[c] Sorbonne Universités, UPMC Univ. Paris 06, UMR 7606, LIP6, F-75005 Paris, France
[d] CNRS, UMR 7606, LIP6, F-75005 Paris, France
[e] The University of Sydney, School of Electrical and Information Engineering, Sydney, NSW 2006, Australia
[f] North Carolina State University, Department of Psychology, 2310 Stinson Drive, Poe Hall 640, Raleigh, NC 2765-7650, USA

ARTICLE INFO

Article history:
Available online 4 March 2015

Keywords:
Emotions
Affect
Computer-based learning environments
Intelligent tutoring systems (ITS)

ABSTRACT

This paper presents the evaluation of the synchronization of three emotional measurement methods (automatic facial expression recognition, self-report, electrodermal activity) and their agreement regarding learners' emotions. Data were collected from 67 undergraduates enrolled at a North American University whom learned about a complex science topic while interacting with MetaTutor, a multi-agent computerized learning environment. Videos of learners' facial expressions captured with a webcam were analyzed using automatic facial recognition software (FaceReader 5.0). Learners' physiological arousal was recorded using Affectiva's Q-Sensor 2.0 electrodermal activity measurement bracelet. Learners' self-reported their experience of 19 different emotional states on five different occasions during the learning session, which were used as markers to synchronize data from FaceReader and Q-Sensor. We found a high agreement between the facial and self-report data (75.6%), but low levels of agreement between them and the Q-Sensor data, suggesting that a tightly coupled relationship does not always exist between emotional response components.

1. Introduction

Emotions are a critical component of effective learning and problem solving, especially when it comes to interacting with computer-based learning environments (CBLEs; multi-agent systems, intelligent tutoring systems, serious games; Azevedo & Aleven, 2013; Baker et al., 2012; Calvo & D'Mello, 2011; D'Mello and Graesser, 2012; Graesser, D'Mello, & Strain, 2014; Grafsgaard, Wiggins, Boyer, Wiebe, & Lester, 2014; Harley, Bouchet, & Azevedo, 2013; Pekrun, 2011; Sabourin & Lester, 2014). In recent years there has been a surge in interdisciplinary research leading to a plethora of new approaches (including tools/devices and analytical techniques) to measure emotions (e.g., physiological sensors, automatic facial expression analysis software, concurrent state self-report measures; Alzoubi, Hussain, D'Mello, & Calvo, 2011; Baker et al., 2012; Calvo & D'Mello, 2011; D'Mello and Graesser, 2012; Grafsgaard et al., 2014; Harley et al., 2013). The variety of tools and analytical techniques available to researchers enables studies to examine emotions from different modalities (e.g., physiological signals, audio, and video). *Multimodal* approaches (using more than one modality to measure emotions) are aligned with theories that define emotions as multi-componential; in other words, that emotions are expressed and experienced in different ways (e.g., an open mouth, elevated heart rate, *feeling* surprised; Gross, 2010, 2013; Pekrun, 2006, 2011). Multimodal approaches also afford researchers the opportunity to circumvent the constraints of individual channels; particularly those associated with self-report data (e.g., Hawthorne effect; physiological channels cannot be socially masked), and therefore achieve greater construct validity and reliability (Harley, in press; Pantic & Rothkrantz, 2003; Utthara, Suranjana, Sukhendu, & Pinaki, 2010).

☆ *Note.* An earlier version of the synchronization approach used and the agreement rate reported in this manuscript for FaceReader and the EV self-report measure was published in: 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: Vol. 7926. Artificial Intelligence in Education* (pp. 61–70). Berlin, Heidelberg: Springer-Verlag. This manuscript extends our work synchronizing different methods to a physiological measurement device, provides more detailed results for agreement rates, and elaborates upon our discussion of them.

* Corresponding author at: Université de Montréal, Department of Computer Science and Operations Research, 2920 Chemin de la Tour, Pavillon André-Aisenstadt 2194, Montréal, QC H3C 3J7, Canada. Tel.: +1 (514) 561 3724.

E-mail addresses: jason.harley@umontreal.ca (J.M. Harley), francois.bouchet@lip6.fr (F. Bouchet), sazzad.hussain@sydney.edu.au (M.S. Hussain), razeved@ncsu.edu (R. Azevedo), rafael.calvo@sydney.edu.au (R. Calvo).

http://dx.doi.org/10.1016/j.chb.2015.02.013
0747-5632/

The use of multiple methods to measure emotions in the context of student-CBLE interactions has, however, led to several emerging conceptual, theoretical, methodological, and measurement issues that need to be resolved before empirically driven prescriptions pertaining to learners' emotions can reliably and validly be made (Harley, in press). Challenges include: (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 methods provide (e.g., dimensional [activation and valence information] for bracelets measuring electrodermal activity (EDA) vs. discrete emotional states from facial expressions); (3) disagreement among theories regarding whether data from different emotional responses should implicate the same emotional state (e.g., if a participant is biting his lip and reports that he is experiencing anxiety should there also be a spike in his physiological arousal data?; Gross, Sheppes, & Urry, 2011); and, (4) day variations in physiological measures due to factors such as environmental changes and sensor placements. The purpose of this paper is to address challenges one through three in the context of research with CBLEs using emotion data from learners' interactions with MetaTutor, a multi-agent-adaptive hypermedia learning environment (Azevedo, Behnagh, Duffy, Harley, & Trevors, 2012; Azevedo et al., 2013; Taub, Azevedo, Bouchet, & Khosravifar, 2014; Trevors, Duffy, & Azevedo, 2014; see Section 2.2).

1.1. Theoretical framework

We view emotions as goal-related and appraisal-driven multicomponential psychological processes that mediate effective learning (Gross, 2010, 2013; Pekrun, 2011). In line with other widely accepted qualities of emotions, we assert that discrete emotions can be categorized by the broad dimensions of arousal (i.e., activation) and valence (Pekrun, 2011; Russell, Weiss, & Mendelsohn, 1989). Valence refers to the intrinsic pleasantness (e.g., enjoyment) or unpleasantness of an emotional state (anger), while arousal refers to the physiologically activating (i.e., arousing; anxiety) or de-activating nature of an emotion (e.g., boredom).

We use Pekrun's (2006; Pekrun & Perry, 2014) control-value theory of achievement emotions, which highlights the role of learners' appraisals of value and subjective control in eliciting emotions that are related to and influential regarding the success of students' academic achievement activities such as, taking tests, studying, and attending class. Pekrun (2006, 2011) distinguishes these two types of appraisals as follows: learners' *appraisals of subjective control* include one's perception of the causal influence they exert over their actions and outcomes. In contrast, *appraisals of value* concern 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. For example, it is expected that students who make appraisals of both positive value and sufficient (e.g., high) control will have the most positive emotions while engaging in an academic activity (e.g., enjoyment). On the other hand, students who make appraisals of negative value and high control will experience negative emotional states such as anger. Learners who make appraisals of positive or negative value and low control will experience negative emotional states such as frustration. Students who don't appraise an academic situation as possessing any value are likely to feel bored irrespective of their appraisals of control (Pekrun, 2006; Pekrun & Perry, 2014).

Another factor that informs Pekrun's theory is the object focus, in other words, where a learner's attention is being focused regarding an academic situation that will take place (prospective), has already taken place (retrospective), or is presently taking place (concurrent or activity). The object focus has implications for the appraisals a student will make and the emotions they will subsequently experience. Object foci delineate whether an outcome is being reflected upon or whether an action (that may lead to an outcome) is the focal point. In this study we measured activity emotions: emotions that students report feeling while interacting with a CBLE. We do, however, draw on other emotional states (beyond those Pekrun lists as academic achievement activity emotions) because of the relevance of examining emotions that pertain to appraisals other than achievement standards, including epistemic emotions that relate to the cognitive and learning components of an academic task (e.g., information processing) and include curiosity and confusion (Pekrun, 2011). Examining a more comprehensive set of emotional states also allowed us to compare our findings (1) between modalities (which measure different types of emotions and emotional characteristics such as arousal) and (2) with the results of other researchers whom have identified a large number of emotional states in interactions with computer-based learning environments (Harley & Azevedo, 2014).

Although theories of emotion have different labels and numbers of emotional components, there is indication of agreement in behavioral (e.g., facial expressions), experiential (e.g., how an emotion makes one feel), and physiological (e.g., electrodermal activation) expressions of emotional states (Gross, 2010; Pekrun, 2006). Accordingly, there is growing recognition among researchers of emotions that there is a need to move beyond experiential, self-report measures to inform our theoretical and empirical understanding of emotions in the context of learning (Calvo & D'Mello, 2011; Graesser, D'Mello, & Strain, 2014; Harley, in press; Pekrun, 2006; Pekrun & Linnenbrink-Garcia, 2014). One of the caveats in this area of research is disagreement between theories of emotion regarding whether different components of emotions should provide similar or dissimilar emotional information; in other words, whether different components of emotions should provide a coherent (i.e., coordinated) response (Ekman, 1992; Pekrun, 2011). Pekrun's description of a student's anxiety before an exam illustrates a coherent response among emotional expression components comprised 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, p. 24). Other researchers, on the other hand, argue that a tight coupling (i.e., high level of coherence) between all components may not necessarily exist (D'Mello, Dale, & Graesser, 2012; Gross et al., 2011). This empirical study therefore contributes to the body of research being conducted using multiple modalities to examine emotions with educational, technology-rich environments as well as more experimental contexts. Both are briefly reviewed below.

1.2. Brief review of multi-modal emotion research

A number of empirical studies have used multiple modalities to examine different emotional components during learners' interactions with computer-based learning environments (AlZoubi, D'Mello, & Calvo, 2012; Arroyo et al., 2009; D'Mello, Dale, & Graesser, 2012; D'Mello and Graesser, 2010; Grafsgaard et al., 2014; Kapoor & Picard, 2005). Most, however, have focused on using them to predict learners' emotions compared to a single, separate modality that is used as the grounded truth measure (i.e., standard). Grounded truth measures are typically self-report measures or classifications of facial expressions from video data that are compared with other modalities post-experiment. These studies have focused on optimizing the cumulative accuracy ratings of multimodal measurement approaches and weeding out extraneous individual methods that bring little or no additive gain to the combined ratings. Taken together, results from these studies and a meta-analysis conducted by D'Mello and Kory (2012; that included studies from the

affective computing literature that did not use computer-based learning environments) suggest that: facial expressions may be the best single method for accurately identifying emotional states; using additional methods to accurately classify an emotional state typically results in only modest additive gains to accuracy ratings; and that measures of posture are likely the weakest method for accurately detecting emotional states.

The main limitation of the approaches deployed in the affective computing field is that they do not reveal the extent to which multiple methods agree or not on the emotional state in question at a particular moment in time. Instead, all comparisons are made between the grounded truth (e.g., training data) vs. all other methods either individually or collectively through the use of machine learning classification approaches (e.g., Bayesian models; Grafsgaard et al., 2014). In other words: in a multimodal study where self-report measures are used as the grounded truth measure for comparisons, the agreement rate between posture and electrodermal activation or between facial expressions and heart rate is not typically presented. A related limitation is that the agreement rates between individual modalities for different emotions is not known, although multimodal studies reveal different predictive accuracy ratings for different emotions (e.g., frustration vs. engagement; Grafsgaard et al., 2014). These limitations hinder insights being drawn with regard to which modalities are most complimentary for detecting certain emotions.

A separate body of literature in experimental psychology contains more research studies that address these shortcomings by focusing on correlations between different modalities for measuring emotions (for a review see Mauss & Robinson, 2009). In their review of empirical research that examined coherence between modalities Mauss and Robinson (2009) found that even those studies with the most sound psychometric properties (valid and reliable measures and within-subject designs) typically found only modest correlations between measures, the stronger between self-report and facial expressions (e.g., Mauss, Levenson, McCarter, Wilhelm, & Gross, 2005). Mauss and Robinson (2009) concluded their review noting that current research on coordinated emotional responses produces weaker correlations between components because different modalities are sensitive to different dimensions (e.g., facial EMG to valence; EDA to arousal) that are not strongly related to one another, and that psychological mechanisms, such as emotion regulation tendencies, can mediate or moderate (i.e., influence) the expression of emotions across components. They also point out the importance of context, which is a limitation in many of the studies in both literatures where emotions are often experimentally elicited (e.g., Mauss et al., 2005; by film) or the data used to train the emotion classifiers stems from databases of posed (e.g., actors imitating emotional states behaviorally; D'Mello and Kory, 2012) rather than naturally occurring emotions.

Both cases raise questions about the generalizability of results to more authentic contexts where emotions occur naturally and may, therefore, be subtler in their expression. For example, the range of intensity of amusement or sadness elicited from a film selected in order to produce these emotional responses may be different (e.g., broader or higher), even after accounting for individual differences, than students' experience of sadness or amusement in an educational context that is commonly not linked to a course grade. Relatedly, many multimodal studies measure a limited number of academic achievement and epistemic emotional states that commonly exclude those that are important for educational outcomes and observed to be among the most commonly occurring (e.g., boredom, confusion; Harley & Azevedo, 2014; Pekrun, 2011). This is particularly pertinent for interpreting the state of the art because studies have found that response coordination is higher for certain emotions, such as amusement, than others (e.g., sadness; Mauss et al., 2005).

1.3. Current study: overview and research questions

The current study presents a novel methodological approach for synchronizing emotional data from modalities that correspond to three different emotion expression components and making comparisons regarding the agreement between each (experiential component, self-report measure; behavioral component, automatic facial expressions analyses; physiological component, electrodermal activity). Unlike previous research that has synchronized multiple emotion measurement modalities from different components, a large number of emotional states are examined that are relevant to the learning context in which the data is collected. Moreover, the emotion data that is collected is naturally occurring, in that emotions arise from interactions with a computer-based learning environment, MetaTutor, designed to teach learners about the human circulatory system and to effectively use self-regulated learning skills (see Section 2.2). Another crucial difference between this study and previous research is that a theoretically based framework is used to synchronize emotion data from different modalities that measure different sets of related emotional states (e.g., fear and anxiety) and emotional information (arousal). As such, one of the primary contributions of this paper is the detailed description of the methodological approaches developed and used to extract, treat, and synchronize data from in-session self-reports, automatic facial expression detection, and electrodermal activity.

The second objective of this study was to use our novel methodology for synchronizing emotion data to determine if different modalities identify the same emotions (e.g., anger) and/or provide complementary emotional information at a given point in time (e.g., high arousal). In line with Pekrun's (2006, 2011) and other's theoretical assertions as well as the experimental and methodological contexts in which several previous shortcomings (e.g., experimental stimulated or posed emotions, lack of examination of agreement between methods) are overcome (Gross, 2010, 2013; Mauss & Robinson, 2009), we expected to see convergence across emotion measurement methods in terms of the emotional states identified at a particular point in time. More specifically, because of the methodology we employed in this study (e.g., measurement of naturalistic rather than stimulated emotions and inclusion of a broader set of emotional states) we expected to see greater convergence (in line with theoretical assumptions) than that previously found in empirical literature.

In addition to our superordinate hypothesis that we would identify higher agreement levels between emotional components we also made three further hypotheses about the congruency for specific emotions between components. (1) First, we expected to find higher agreement rates, especially with regard to electrodermal activation, between emotional states that had higher activation (i.e., arousal) levels, such as happiness and anger, because of the greater potential variance in these emotions' data (e.g., it being easier to detect emotions with stronger than weaker physiological signals). (2) We also expected stronger congruency between positive emotional states because of the tendency for participants to control their expression of negative emotions (Mauss et al., 2005). (3) Finally, we hypothesized that non-basic emotions (e.g., boredom, curiosity) could be less coordinated than basic emotions because they are evolutionarily less associated with fight or flight tendencies (Mauss et al., 2005).

2. Methods

This section provides details about the participants of this study (see Section 2.1) as well as the learning environment, MetaTutor (see Section 2.2). It then describes the apparatuses and modalities used for measuring emotions (see Section 2.3). The experimental

procedure (see Section 2.4) describes how data was collected. The novel approach we used to synchronize data between modalities is described in the data analysis section (see Section 2.5).

2.1. Participants

Sixty-seven undergraduate students (82.1% female) whom were enrolled at a large, public North American University participated in this study. The sample had a mean age of 21.00 (SD = 1.90) and mean GPA of 3.14 (SD = 0.69). The majority of participants were Caucasian (74.60%) and in their senior (40.30%) year of university. Participants represented a variety of disciplines, including math or engineering (10.4%), social sciences (21.00%), sciences (32.80%), business (9.00%), and arts (7.50%). About 54% of the participants had prior experience with biology-related topics from university courses and/or work experience. Learners' mean pretest score was 78% (SD = 0.15), which was comparable across experimental groups (Prompt and Feedback, $M = 0.75$%; SD = 0.03; Control, $M = 0.82$, SD = 0.03) participants were randomly assigned to for their interaction with MetaTutor (see Section 2.2 for details). In order to participate in the study participants had to be eighteen years of age or older and full time undergraduate students. Due to facial recognition software requirements, participants also had to be able to tie long hair and bangs back, not wear any type of head covering that could obstruct or cast a shadow over their face, and have normal vision (e.g., not wear glasses).

2.2. MetaTutor

MetaTutor is a multi-agent, intelligent tutoring hypermedia system that consists of 38 pages of text and static diagrams that students can navigate through using a table of contents (Azevedo et al., 2012, 2013; Taub et al., 2014; Trevors et al., 2014). MetaTutor is both a learning and research tool that teaches students about the human circulatory system and provides training on how to use self-regulated learning strategies. Self-regulation involves learners actively and efficiently managing one's own learning of a topic (e.g., body systems) through setting subgoals, using learning strategies (e.g., taking notes, re-reading), and monitoring and regulating aspects of their cognition, behavior, emotions, and motivation in order to achieve their learning objective(s) (Azevedo et al., 2013; Pintrich, 2000; Winne & Hadwin, 2013; Zimmerman & Schunk, 2011).

Instructional scaffolding was provided by four pedagogical agents (PAs; 3D virtual characters) and varied depending on the experimental condition to which learners were assigned (aside from PAs' 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), and were given feedback about their use of those processes. In the control (C) condition, participants did not receive prompts or feedback from the PAs and could only perform these self-regulatory processes on their own initiative. The PAs include Gavin the Guide, Pam the Planner, Mary the Monitor, Sam the Strategizer. Gavin provides guidance for participants in the learning environment and administers pretest and posttest knowledge assessments and self-report measures. Gavin's interactions with learners did not vary between conditions. Pam prompts and scaffolds planning processes such as encouraging students to activate (i.e., recall) relevant prior knowledge about the topic and to set two subgoals at the beginning of their learning session which help them approach the learning task and achieve their overall learning goal: to learn all they can about the human circulatory system. In addition to prompting students to activate their prior knowledge in the PF condition, Pam provides learners with feedback regarding the appropriateness of their proposed subgoals and offers them the opportunity to try again (when proposed subgoal are inappropriate). This additional affordance involves them more in the goal setting process than when they are immediately offered a more suitable alternative (in the C condition). Mary prompts and supports participants in their monitoring processes (e.g., judgment of learning) and Sam prompts participants to engage in learning strategies and ensures their use (e.g., note-taking, summarizing). In the C condition Mary and Sam only respond to learners' self-initiated monitoring and self-regulated learning strategies in an instructional manner (e.g., acknowledging completion of a summery), when appropriate, rather than providing feedback on the quality of their responses or recommending that they engage in them.

MetaTutor's main interface (see Fig. 1) consists of a timer that indicates how much time remains in the learning session, and an SRL palette where participants can initiate interactions with one of the four PAs depending on the action chosen. An integrated notepad is embedded into MetaTutor and available for participants to take notes and access them at any time during the learning session, except during the posttest. Participants' subgoals are displayed during the learning session directly below their overall learning goal within progress bars that are automatically filled as learners navigate through those pages that are relevant to the currently active subgoal. One of the four PAs is always visible in the upper right-hand corner of the learning environment and audibly communicates with the learner through the use of a text-to-speech engine (interactions are also available in text format in a dialog box that participants can choose to open, and which also allows them to re-read previous interactions with the PAs). Self-report questionnaires are administered using a Google Docs form embedded in the MetaTutor learning environment.

2.3. Apparatuses and measures

2.3.1. Q-Sensor 2.0

Q-Sensor 2.0 [Apparatus (2013)] was used to measure learners' electrodermal activity (EDA), which is a signal commonly used to measure physiological arousal. Q-Sensor 2.0 provides eight values every second and was developed by Picard and colleagues who found EDA to be an effective predictor of affective states in the context of learning and intelligent tutoring systems (Kapoor, Burleson, & Picard, 2007; Woolf et al., 2009). EDA refers to electrical changes at the surface of the skin that are caused by sympathetic activity and alter sweating. One method of measuring EDA is to measure the variations of electrical conductance of the skin (expressed in micro Siemens or 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 due to individual differences in baseline EDA levels. Arousal is therefore inferred based on a higher or lower level compared to the individuals' average or baseline 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 outcome. Lower levels of arousal suggest that the learner may be relaxed or bored, perhaps from reading a page of content of little interest or not challenging enough to them.

2.3.2. Emotions-value questionnaire (EV)

During the learning session, participants were asked on five occasions (see Section 2.4) by MetaTutor to complete the EV questionnaire in which each participant responded to 20 items: 19 items on emotions (cf. Table 4 for the list) and 1 item on task value that was not considered in this analysis. These items were rated on a 5-point Likert scale ranging from 1 = "Strongly Disagree" to

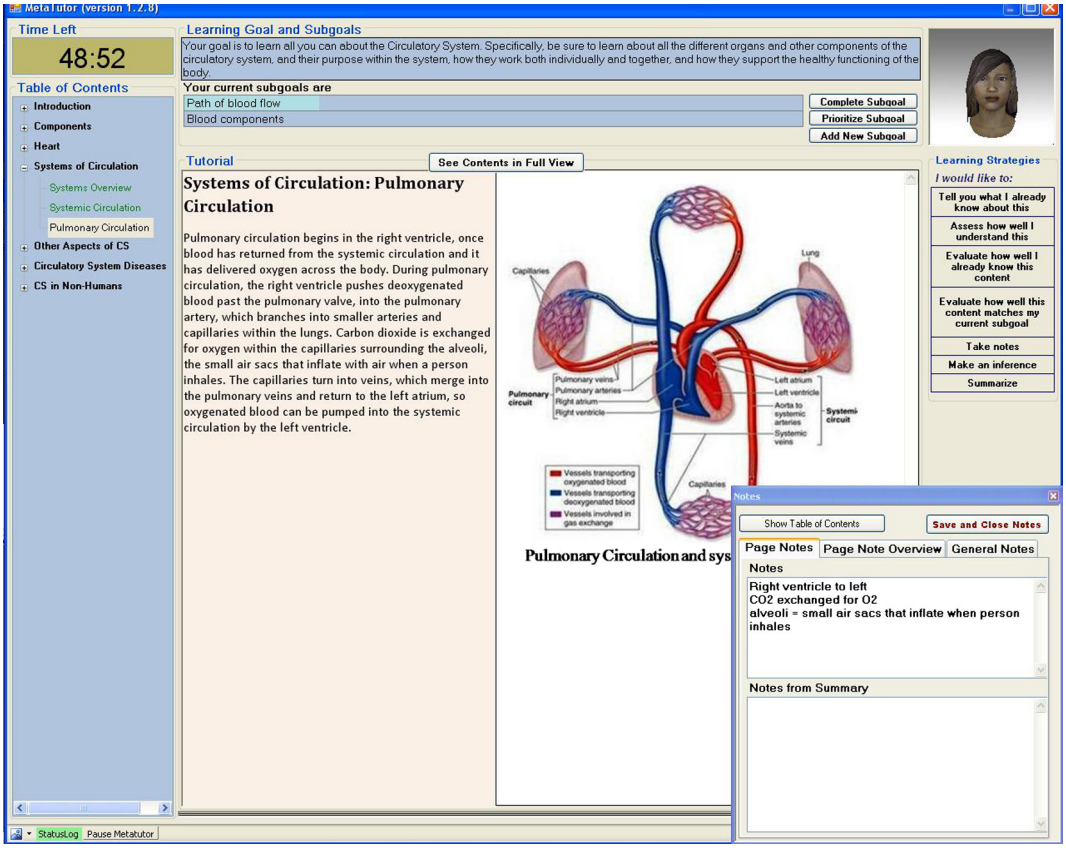

Fig. 1. Screenshot of MetaTutor interface.

Table 1
Emotions sorted by definitional proximity, arousal, and valence.

		Valence		Other
		Positive	Negative	
Arousal	High	[1]**Happy** (enjoyment, hope, pride, curiosity and eureka)	[2]**Anger** (frustration), [3]**Fear** (anxiety), [4]**Disgust*** Contempt*	[6]**Neutral*** [7]**Surprise***
	Low		[5]**Sadness** (hopelessness, boredom*)	

Note. Bold emotions are basic emotions + neutral (FaceReader). Emotions with an asterisk are discussed in this note. Numbers correspond to emotions and emotion groups analyzed in comparisons between the EV questionnaire and FaceReader in section 3.1. Neutral* is non-valenced and in between high arousal (i.e., activating) and low arousal (i.e., de-activating), though more toward low arousal. Surprise* is non-valenced and activating (high arousal). Boredom* was subject to discussion and therefore we tried both to associate it with sadness and to completely exclude it (cf. Table 2). Contempt* was also subject to discussion because of its nature as a social emotion (Pekrun, 2011) and the inclusion of a separate artificial neural network to differentiate it from disgust* in FaceReader 6.0. Given, however, the low incidence of disgust, (cf. Table 2) we did not have the opportunity to contrast different approaches like we did for sadness and boredom and therefore decided to exclude it (contempt) from comparisons with disgust (from FaceReader data). Shame and confusion were not included in the above table because these states could not be properly associated with high or low arousal levels.

5 = "Strongly Agree". One example item is: "Right now I feel bored". The instructions and wording of the questions were based on a subscale of the academic emotions questionnaire developed by Pekrun and colleagues (AEQ; Pekrun, Goetz, Titz, & Perry, 2002) that assesses participants' concurrent, state-emotions as opposed to emotions reported on prospective or retrospective measures. The 19 emotions that are measured using the EV questionnaire represent a comprehensive list of discrete basic and learner-centered emotions that appear in the research and theories from several emotion researchers (e.g., D'Mello et al., 2010; Pekrun, Goetz, Frenzel-Anne, Petra, & Perry, 2011). The majority of the 19 emotions can be conceptualized according to the four quadrants defined according to the axes of valence (positive/negative) and arousal (high/low or activating/deactivating; Pekrun et al., 2002., 2011; Russell et al., 1989; see Table 1).

Definitions based on researchers' work and operationalizations of these emotions (D'Mello et al., 2010; Pekrun et al., 2011) were used to create a digital definition handout that was provided in a side panel to participants every time they completed an electronic version of the EV embedded in MetaTutor. For example, the definition for happiness was "satisfaction with performance, general feeling of pleasure". An example was also provided for each definition (e.g., "I'm really happy with how I did on that quiz! That's great that we'll be learning about frogs!"). Definitions were pilot tested with lab and non-lab members and students from various cultural backgrounds in focus groups to ensure that they accurately reflected and differentiated each emotional state.

2.3.3. FaceReader 5.0

FaceReader (5.0) analyzes participants' facial expressions and provides a classification of their emotional states. It uses an Active Appearance Model that models participants' facial expressions, and an artificial neural network with seven discrete outputs that

classifies participants' constellations of facial expressions corresponding to Ekman and Friesen's six basic emotions (happiness, anger, fear, sadness, disgust, surprise), in addition to neutral (Ekman, 1992). FaceReader has been validated through comparison with human coders and used in a number of empirical psychological studies (Chentsova-Dutton & Tsai, 2010; Harley et al., 2013; 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 onset and offset of the hierarchical (i.e., goodness of fit) rankings of these states.

2.3.4. Logitech Orbit AF webcam

A webcam was used to record the participants' faces during their interaction with MetaTutor. In accordance with FaceReader 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 1600 × 1200 at an average rate of 12.1 frames per second.

2.4. Experimental procedure

During the first session, participants were provided 30 min to read and sign the informed consent form and subsequently complete a pretest on the human circulatory system, a demographics questionnaire, and self-report measures (e.g., AEQ trait emotions) on a computer with their face being video recorded. In the second session, participants set up two subgoals for learning about the human circulatory system and proceeded to interact with MetaTutor, and spent approximately 90 min learning about the human circulatory system. During the session video, screen capture, audio, eye-tracking, and physiological data were collected for each participant while they used MetaTutor. Halfway through the session, participants were invited to take an optional 5-min break. At the end of their learning session, learners completed the post-test measure, and additional self-report measures (e.g., AEQ retrospective emotions).

Both sessions took place at least 1 hour apart from each other and no more than 4 days apart. The first time participants filled out the EV was at the beginning of the learning session after they had successfully set two subgoals. The following occasions occurred regularly every 14 minutes during the 1-hour learning session, with the fifth EV questionnaire being administered just before the post-test. Participants were permitted as much time as necessary to complete the EV on each occasion. Prior to the start of the MetaTutor learning session, the webcam was positioned and commenced recording, and participants were asked to put the Q-Sensor bracelet on. This typically provided 10–15 min of baseline data. Participants were compensated up to $50 for completing the study.

2.5. Data coding and analyses

2.5.1. Treating and extracting data from individual channels

This section describes the steps taken in order to treat and extract data from the individual channels (EV, FaceReader, Q-Sensor).

2.5.1.1. EV questionnaire data. Several scores on different emotions assessed in the EV questionnaire were identified as univariate outliers with standardized scores exceeding $z = \pm 3.29$ and were therefore replaced with the next most outlying values for each variable (Tabachnick & Fidell, 2007).

2.5.1.2. FaceReader 5.0. Data was exported from the FaceReader program to CSV files. FaceReader data considered corresponded to the analysis of the 10 seconds prior to the administration of each EV measures (50 seconds overall for each participant). 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 depending on participants' self-declared ethnicity. Videos of the second session (when the learning occurred) were then analyzed with the "smoothen classification" parameter enabled. Sixty-seven participants were analyzed, but nine of them were excluded from our sample because their dominant state in the 10 seconds prior to their completion of the EV questionnaire was identified as "Unknown" by FaceReader for at least three of the five EV questionnaires. This situation generally occurs when the participant's face is not sufficiently oriented toward the webcam (e.g. when they look down to type on the keyboard).

2.5.1.3. Q-Sensor. Similar to the FaceReader data, EDA data was exported from the Q-Sensor 2.0 into CSV files and the segments considered correspond to the 10 seconds prior to the administration of the EV measures. The average microSiemens (μS) value was considered during these five periods of 10 s. The features extracted (using the 10 seconds 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 Bio signal Toolbox (AubT) in Matlab was used for extracting the features (Wagner, Kim, & André, 2005).

2.5.2. Synchronizing individual channels

The processes we used to synchronize results from the different methods in order to calculate their agreement rates are described below.

2.5.2.1. Synchronizing FaceReader and EV data. We synchronized the dominant emotional state as identified by FaceReader with the EV questionnaire data by extracting log information corresponding to the 10 seconds of video footage of participants immediately before they completed each of the EV questionnaires. This period of time was selected because it was short enough to capture the rapidly changing emotions participants were experiencing at the moment. 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 then selected the primary dominant state that was defined as the state reported as dominant during the majority of the 10-s segment. In 80.70% of the cases, no other unique emotion was dominant for more than 3 s, making 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.

To evaluate the agreement between the self-reported emotions in the 5 EV questionnaires and the dominant emotion identified by FaceReader during the 10 s prior, we defined the correspondence between the 13 non-basic emotions assessed in the EV questionnaire with the 6 basic emotions in addition to neutral used by FaceReader to classify participants' emotions. Using work from Pekrun et al. (2002, 2011) the correspondence was determined as follows (see Table 1): (1) All positively valenced high arousal (i.e., activating) emotions (enjoyment, hope, pride, curiosity and eureka) were associated with happiness; and among the negatively valenced high arousal emotions, (2) frustration was grouped with

anger, (3) and anxiety with fear. (4) Disgust was not associated with any other emotions. (5) All negatively valenced, low arousal (i.e., deactivating emotions; hopelessness and boredom) were associated with sadness, while the non-valenced emotions ([6]neutral and [7]surprise) were kept as two distinct categories because of differences in arousal. Three additional emotions (contempt, confusion, and shame) evaluated in the EV questionnaire could not be associated with any basic emotions and were therefore discarded for this analysis.

Using these seven groups of emotions, we defined that there was an agreement between FaceReader's dominant emotion and the EV questionnaire data if and only if one of the emotions associated with 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 needed to have a score of 3 or more in the EV). The 20 (out of 290) occurrences of "Unknown" were excluded from this analysis.

2.5.2.2. Synchronizing FaceReader and Q-Sensor data. In order to compare the EDA and FaceReader data, Q-Sensor data was dichotomized into high and low levels 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 detected by FaceReader 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 et al. (2010) and Pekrun (2011). Agreement was calculated by identifying how often the emotional states classified by FaceReader fit the expected high or low levels of arousal.

2.5.2.3. Synchronizing EV and Q-Sensor. Similar to our synchronizing of the EV data with FaceReader, we defined an EV-assessed emotion as present if it was rated three or higher (out of five) by learners. Boredom, hopelessness, sadness, and neutral were classified as low arousal emotions. Shame and Confusion were not examined because they could not be properly associated with high or low arousal levels. All other emotions were classified as high arousal (see Table 1). As learners occasionally reported more than a single emotion (i.e., with a score higher than 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 a neutral level of five and a happy level of three on the EV questionnaire while the Q-Sensor measures a low-arousal value, it counted as an agreement on Neutral and a disagreement on Happy. The overall agreement was then calculated based on the weighted mean of each of the 17 emotions considered.

3. Results

3.1. FaceReader and EV agreement

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). We could not calculate a kappa score for the agreement between FaceReader and the EV because of differences between the scales that the two measures used (e.g., seven options vs. 19). Table 2 provides the agreement rates between measures for each of the emotions and administrations of the EV questionnaire, which (excluding disgust and fear) ranged from

Table 2
Agreement between FaceReader and EV by emotion.

EV admin.	Happy		Anger		Sadness[1]		Sadness[2]		Fear		Surprise		Disgust		Neutral	
Method	EV	FR	EV	FR	EV	FR	EV	FR	EV	FR	EV	FR	EV	FR	EV	FR
1	5	5	0	2	2	2	1	2	0	0	0	2	0	0	41	45
2	8	11	0	0	4	4	1	4	0	0	1	7	0	0	22	30
3	11	12	0	2	4	7	1	7	0	0	0	1	0	0	25	31
4	15	17	0	2	2	3	0	3	0	0	0	1	0	1	29	33
5	8	11	2	3	2	3	1	3	0	0	0	3	0	0	23	32
Total	47	56	2	9	14	19	5	19	0	0	1	14	0	1	140	171
Agreement (%)	84.00		22.22		73.68		21.05		–		7.14		0		81.87	

Note. The numbers below represent the number of times each of the seven emotional states were (1) reported by learners with a score of three or higher on the EV questionnaire and/or (2) classified as the dominant emotional state by FaceReader (FR). Sadness[1] includes self-reported boredom in agreement rate calculation (with sadness and hopelessness) whereas sadness[2] excludes it (only including self-reported sadness and hopelessness).

Table 3
Agreement between FaceReader and EV by discrete emotion.

FaceReader								
Happy			Anger			Sad		
EV	%	Sum	EV	%	Sum	EV	%	Sum
Happy	.18	32	Anger	.33	1	Sad	.06	1
Enjoyment	.18	33	Frustration	.67	2	Boredom	.72	13
Hope	.21	38				Hopeless	.2	4
Pride	.17	31						
Eureka	.06	11						
Curiosity	.19	34						
Total	1.00	179		1.00	3		1.00	18

Note. The numbers below the Sum column correspond to the number of times an emotion classified by FaceReader matches an emotion from the self-report EV measure of the same category. The % column represents the proportion of the common category states (cf. Table 1) that a discrete emotion is represented by. For example, enjoyment represented 33 (18%) of 179 states that FaceReader classified as happy. Fear was excluded because of a lack of classification (i.e., presence; see Table 2) as well as surprise, neutral, and disgust because they did not have multiple corresponding discrete emotional states. The following example illustrates the relationship between Tables 2 and 3: Table 2 illustrates that FaceReader classified participants' facial expressions as angry on nine occasions overall, and that participants self-reported feeling angry or frustrated on two occasions (fifth administration of the EV). Table 3 reveals that for one of these two occasions a participant reported feeling both angry and frustrated, whereas in the other, a participant only reported feeling frustrated.

about 7% (surprise) to 84% (happiness). Table 3 illustrates the agreement between the basic emotions measured with FaceReader and the discrete emotions measured with the EV.

Table 4
Agreement between FaceReader/EV and the Q-Sensor by arousal level.

		Q-Sensor	
		Low	High
FaceReader	Low	111	47
	High	43	25
	Total	154	72
	Agreement (%)	0.60	
	K	0.07	
EV	Low	246	119
	High	768	378
	Total	1014	497
	Agreement (%)	0.41	
	K	0.00	

Note. The numbers below the Low and High columns represent the number of times a state was classified as low or high by (1) Q-Sensor and (2) categorized as low or high using the emotional state reported by FaceReader (neutral and sad were categorized as low level of arousal states; happiness, anger, fear, surprise, and disgust were categorized as high arousal states). For the 19 emotional states reported by learners with a score of three or higher on the EV questionnaire: boredom, hopelessness, sadness, and neutral were classified as low arousal emotions; shame, and confusion were not examined because these states could not be properly associated with high or low arousal levels; and all other emotions were classified as high arousal.

Table 5
Agreement between FaceReader/EV and the Q-Sensor by emotion.

Channel	Emotion	Q-Sensor		Sum	Agreement (%)
		Low	High		
FaceReader	Happy	33	16	49	32.65
	Sadness	9	5	14	64.29
	Anger	3	2	5	40
	Fear	0	0	0	–
	Surprise	6	7	13	53.5
	Disgust	1	0	1	0
	Neutral	102	42	144	70.83
	Overall	*N/A*	*N/A*	*226*	*60.18*
EV	Happy	105	48	153	31.37
	Enjoyment	92	46	138	33.33
	Hope	120	47	167	28.14
	Pride	91	43	134	32.09
	Anger	33	22	55	40
	Frustration	55	36	91	39.56
	Anxiety	56	31	87	35.63
	Fear	6	4	10	40
	Shame	14	16	30	–
	Hopelessness	25	14	39	64.1
	Boredom	79	38	117	67.52
	Surprise	30	17	47	36.17
	Contempt	43	20	63	31.75
	Disgust	8	5	13	38.46
	Confusion	30	19	49	–
	Curiosity	102	50	152	32.89
	Sadness	7	8	15	46.67
	Eureka	27	9	36	33.33
	Neutral	135	59	194	69.59
	Overall	*N/A*	*N/A*	*1590*	*41.30*

Note. The numbers below the Low and High columns represent the number of times a state was classified as low or high by Q-Sensor for each the seven emotional states classified by FaceReader (neutral and sad were categorized as low level of arousal states; happiness, anger, fear, surprise, and disgust were categorized as high arousal states). For the 19 emotional states reported by learners with a score of three or higher on the EV questionnaire: boredom, hopelessness, sadness, and neutral were classified as low arousal emotions; shame and confusion were not examined because these states could not be properly associated with high or low arousal levels; and all other emotions were classified as high arousal.

3.2. FaceReader and Q-Sensor agreement

We found an agreement rate of 60.1% ($\kappa = 0.07$) between the Q-Sensor and FaceReader when comparing arousal levels (see upper part of Table 4). The upper part of Table 5 provides the agreement rates between measures for individual emotional states reported with FaceReader and arousal levels from Q-Sensor, which (excluding fear and disgust) ranged from about 33% (happy) to 71% (neutral).

3.3. EV and Q-Sensor agreement

We found an agreement of 41.3% ($\kappa = .00$) between Q-Sensor and the self-report measure of emotions (see lower part of Table 4). The highest agreement between the Q-sensor and the EV questionnaire for discrete emotions was between learners' self-reported experience of boredom and low arousal (67.5%) and neutral and low arousal (69.6%), while hope and high arousal (28.1%) had the lowest (see lower part of Table 5). We also examined the relationship between emotional intensity and EDA arousal level in Table 6. Overall, however, the relationship between arousal remains weak even when we consider the level of self-reported emotional presence or intensity. The percentage of agreement is better when considering emotions reported as clearly indicative of a learner's state (e.g., "strongly agree that they feel X"; 5 on a Likert scale). This phenomenon is confirmed by the kappa (not reported in Table 4) but which is as follows for the values from 3 to 5 (we did not consider the emotion to be present if it was not rated as 3 or higher): $\kappa = 0.01$, $\kappa = -0.02$, and $\kappa = 0.05$. The agreement is

Table 6
Agreement between EV intensity and the Q-Sensor by emotion.

Emotion\Q-Sensor	EV self-report level								
	3			4			5		
	L	H	Ag. (%)	L	H	Ag. (%)	L	H	Ag. (%)
Happy	68	37	35.2	31	9	22.5	6	2	25.0
Enjoyment	56	25	30.9	28	16	36.4	8	5	28.5
Hope	77	30	28.0	34	13	27.7	9	4	20.8
Pride	66	35	34.7	22	7	24.1	3	1	25
Anger	24	17	41.5	8	5	38.5	1	0	0
Frustration	17	11	39.3	30	18	37.5	8	7	46.7
Anxiety	28	17	37.8	23	12	34.3	5	2	28.6
Fear	6	4	40	0	0	N/A	0	0	N/A
Shame	11	10	–	3	6	–	0	0	–
Hopelessness	18	6	75.0	3	6	33.3	4	2	66.7
Boredom	28	15	65.1	34	16	68	17	7	70.8
Surprise	21	11	34.0	9	5	36.0	0	1	100.
Contempt	28	7	20	10	12	54.6	5	1	16.7
Disgust	8	5	38.5	0	0	N/A	0	0	N/A
Confusion	24	12	–	6	4	–	0	3	–
Curiosity	41	18	30.5	49	25	33.8	12	7	36.8
Sadness	7	7	50	0	1	0	0	0	N/A
Eureka	13	4	23.5	14	5	26.3	0	0	–
Neutral	63	25	71.6	32	17	65.3	40	17	70.2
Overall	–	–	40.0	–	–	39.7	–	–	52.3

Note. The numbers below the Low and High columns represent the number of times a state was classified as low or high by Q-Sensor when a learner reported experiencing the emotion on the left hand column with a score from 1 to 5. The agreement rate for Likert scale values 3–5 for each emotion corresponds to whether the EDA level matched the expected EV value. Likert scale values 1–2 were not reported because we did not consider the emotion to be present if it was not rated as 3 or higher. Of the 19 emotional states reported by learners with a score of three or higher on the EV questionnaire boredom, hopelessness, sadness, and neutral were classified as low arousal emotions; shame and confusion were not examined because these states could not be properly associated with high or low arousal levels; all other emotions were classified as high arousal. For example, when learners reported that they strongly agreed (Likert value of 5) with the statement that they were feeling bored their EDA levels were classified as low (as would be expected) in over 70% of the cases.

further improved when we consider the EDA max level instead of the mean level (used in the analyses), with the following kappa values: $\kappa = 0.01$, $\kappa = 0.01$, and $\kappa = 0.14$. Nonetheless, these are low values and therefore considering "stronger" emotional states does not (unfortunately) circumvent the low agreement rate.

4. Discussion

This paper addressed two research objectives. The first, through a detailed description of the multimodal emotion measurement approach used to extract, treat, and synchronize data from three different modalities. This approach to measuring multimodal data provides a means of overcoming challenges related to (1) differences in the sampling rate of emotional data and (2) variation in the detail and kind of emotional information that different modalities provide. Our second objective was to determine whether different modalities identified the same emotions (e.g., anger) and/or provided complementary emotional information at a given point in time (e.g., high arousal). Results revealed that the agreement varied depending on which modalities (i.e., emotion expression components) were being compared. These results are valuable because they provide empirical evidence to help inform theories of emotion regarding whether coherence exists among different emotional expression components, and do so with a large set of naturally occurring emotions rather than a small corpus of experimentally elicited emotions.

The high level of agreement between the EV data 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 our superordinate hypothesis, prior research (examining agreement between facial expressions and self-reported emotions), and theories of emotion that hold that the different channels through which emotions are expressed will have coherent responses (Ekman, 1992; Mauss et al., 2005; Pekrun, 2011).

This paper also found preliminary evidence that self-reported experiences of boredom may match traditional facial expressions associated with sadness. While these emotional states are distinct at the discrete emotion level and research has been conducted examining facial expressions for each, the authors were unable to find published empirical research examining both in a single study (e.g., D'Mello and Graesser, 2010; Mauss et al., 2005; McDaniel et al., 2007). As such, prior research has not examined whether there is an overlap in participant's behavioral expression of these two states which our results (cf. Table 2) suggest may exist. Specifically, Table 2 shows that the agreement rate between the EV and FaceReader drops by a large proportion when boredom is excluded from negative valence, low arousal agreement comparisons with sadness (from 74% to 21%). This finding, though exploratory, and in need of replication, is in line with research that has been conducted with posture (also a behavioral component, like facial expressions) that found similarities between expressions of sadness and boredom (Wallbott, 1998).

Our hypothesis that coherence between channels would be found between the Q-Sensor and the EV and FaceReader was not, however, supported by our results. Rather, these results suggest that the physiological component of emotions (i.e., EDA data) does 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 postulate that data from three different expression components should be tightly coupled are not always appropriate. Instead, a tight coupling between all three may not necessarily exist, as other researchers posit (D'Mello, Dale, & Graesser, 2012; Gross et al., 2011).

Alternatively, how closely related emotional responses are from different emotional expression components 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 (see Table 2). Since arousal levels are relative, the higher range of arousal experienced by students may not have been as high as in other experimental contexts, such as playing a serious game or watching a film clip intended to elicit amusement or sadness (Harley & Azevedo, 2014; Mauss et al., 2005). As such, electrodermal activation would not be as sensitive to changes in emotional states as the other modalities. This may help explain some of the lack of agreement between higher arousal emotions, such as anger, surprise, and disgust where they are classified, but often physiologically experienced at lower levels. Other contexts may elicit higher levels of arousal because of the cognitive appraisals that students make while interacting with them. A recent selective review by Harley and Azevedo (2014) 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, and in line with Pekrun's control-value theory of achievement emotions (2006) afforded students choice, and were based on content related to their studies. This review also indicated that students tended to experience relatively few instances of negative emotions characterized as high in 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 as compared to other higher-stakes, academically meaningful situations (e.g., studying for a unit related to the students' academics; medical students practicing making diagnosis) and/or those environments that deploy engaging elements (e.g., narrative, gamification) where a greater proportion of positive, activating emotions that may be easier to observe with EDA data are more likely to be observed.

Another possibility for the lack of agreement between the EDA data and the self-report and facial recognition data 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 independently of analytic precedents (that 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 EDA data (beyond a dichotomization) 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 could help improve agreement between channels if assessed. The same situation applies to emotions labeled as low in arousal, such as neutral and boredom.

Although more sophisticated categorizations of the EDA data were examined (e.g., trichotomization), they failed to yield stronger agreement rates, likely due to the data's variance. Future analyses using additional physiological and behavioral modalities (e.g., heart rate and posture) could potentially either strengthen this study's findings related to coherence among emotional expression components and/or reveal nuances between their agreement/lack of for different emotional states. In other words, we might not observe (for example) the same coherence patterns between posture and (1) self-report and (2) EDA data as we did with facial

expression data. These results would have important implications for theories of emotion and further refine the methodological techniques advanced in this study. Future analyses could also be conducted in more high-stakes or engaging environments as mentioned above to overcome the contextual limitations of this study with regard to EDA levels and potential appraisals of task value.

In response to our discrete emotion-specific hypothesis, we found mixed results where higher intensity emotions tended to have lower agreement rates when comparisons between EDA and other methods were made, and that congruency was highest between negative and non-valenced emotions when comparisons were made between both self-report and facial expressions, and physiological data (e.g., Neutral, Sadness, Boredom, Hopelessness). These findings also provided evidence counter to our hypothesis about stronger coordination between basic when compared to non-basic emotions. While these findings are interesting, it is more likely that the basic and typically higher intensity emotions of anger, anxiety (and others) had low levels of agreement between methods because they weren't experienced at sufficiently high levels to elicit expression across components. In other words, even when learners' rated these states as a 3 or higher on the Likert scale, their absolute experience of this emotion may not have been as strong as in another academic setting (e.g., studying for an exam the next day).

Congruency between positive emotional states was supported as hypothesized between self-report and facial expression data, however, which makes sense given that these two expression components are controllable while physiological ones are less so. Relatedly, most of the different discrete emotional states that were related to basic emotions were not relevant to our third discrete emotion-specific hypothesis, but were instead related to Happiness and Sadness. The distribution of the agreement rates for these emotional states suggest, as mentioned above for Sadness and Boredom, that it can be valuable to expand the set of emotions one examines in association with facial expressions because of potential overlap. From an evolutionary perspective this makes sense as it is less important to understand whether a colleague is happy vs. curious or bored vs. sad than it does to distinguish anger from anxiety. In an educational setting, however, these distinctions can prove valuable for determining the most appropriate instructional intervention.

In conclusion, this study provides a methodological description of how to measure and synchronize emotion data obtained from learners interacting with MetaTutor using three different modalities. The high agreement between automatic facial recognition and self-report methods bolsters the validity of our emotion assessments with respect to these two modalities and provides a strong foundation for incorporating these measures as valid and reliable diagnostic indicators of learners' emotions at discrete points while learning. The agreement between these modalities and the EDA data suggests that future research should be conducted, specifically in environments expected to elicit higher arousal levels from students (e.g., serious game environments). Future research should also explore the use of multi-level modeling to examine agreement between modalities for individuals over time, which would advance the analytics used in affective computing beyond the use of affective-state transitions and within subject analysis of variance for emotions (Baker, Rodrigo, & Xolocotzin, 2007; D'Mello and Graesser, 2012; Harley et al., 2013; McQuiggan, Robison, & Lester, 2010).

Conceptually and theoretically, these results provide evidence that the behavioral and experiential components of emotions are tightly coupled. Educationally, improved measurement methods of emotions should lead to better informed interventions that can be designed to support and sustain adaptive emotional states during learning with computer-based learning environments.

Acknowledgements

The research presented in this paper has been supported by a doctoral and postdoctoral fellowship from the Fonds Québécois de recherche – Société et culture (FQRSC) and a Joseph-Armand Bombardier Canada Graduate Scholarship for Doctoral research from the Social Sciences and Humanities Research Council (SSHRC) of Canada awarded to the first author. This research has also been supported by funding awarded to the fourth author from the National Science Foundation (IIS 1008282), the Social Sciences and Humanities Research Council of Canada, and the Canada Research Chairs program. The authors would like to thank Lauren Agnew, Kelsey Anderson, Valérie Bélanger-Cantara, Reza Feyzi-Behnagh, Sophie Griscom, Nicholas Mudrick, Nicole Pacampara, Alejandra Segura, Victoria Stead, Gregory Trevors, Grace Wang, and Wook Yang for assisting in running participants.

References

Alzoubi, O., Hussain, M. S., D'Mello, S., & Calvo, R. A. (2011). Affective modeling from multichannel physiology: Analysis of day differences. In S. D'Mello, A. Graesser, B. Schuller, J.-C. Martin (Eds.), *Lecture notes in computer science: Vol: 6974. Affective computing and intelligent interaction* (pp. 4–13). Berlin, Heidelberg: Springer-Verlag.

Arroyo, I., Woolf, B., Cooper, D., Burleson, W., Muldner, K., & Christopherson, R. (2009). Emotion sensors go to school. In V. Dimitrova, R. Mizoguchi, B. du Boulay, & A. Graesser (Eds.), *Proceedings of the international conference on artificial intelligence in education* (pp. 17–24). Amsterdam, Netherlands: IOS Press.

Azevedo, R., & Aleven, V. (Eds.). (2013). *International handbook of metacognition and learning technologies*. Amsterdam, The Netherlands: Springer.

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.

Azevedo, R., Harley, J., Trevors, G., Feyzi-Behnagh, R., Duffy, M., Bouchet, F., et al. (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.

Baker, R. S. D., Gowda, S. M., Wixon, M., Kalka, J., Wagner, A. Z., Salvi, A., et al. (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* (pp. 126–133). Crete, Greece: International Educational Data Mining Society.

Baker, R., Rodrigo, M., & Xolocotzin, U. (2007). The dynamics of affective transitions in simulation problem-solving environments. In A. R. Paiva, R. Prada, & R. Picard (Eds.). *Affective Computing and Intelligent Interaction* (Vol. 4738, pp. 666–677). Berlin, Heidelberg: Springer.

Calvo, R. A., & D'Mello, S. (Eds.). (2011). *New perspectives on affect and learning technologies*. New York: Springer.

Chentsova-Dutton, Y. E., & Tsai, J. L. (2010). Self-focused attention and emotional reactivity: The role of culture. *Journal of Personality and Social Psychology, 98*, 507–519.

D'Mello, S. K., Dale, R., & Graesser, A. (2012). Disequilibrium in the mind, disharmony in the body. *Cognition and Emotion, 22*, 362–374.

D'Mello, S. K., & Graesser, A. (2010). Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-Adapted Interaction, 20*(2), 147–187.

D'Mello, S. K., & Graesser, A. C. (2012). Dynamics of affective states during complex learning. *Learning and Instruction, 22*, 145–157.

D'Mello, S. K., & Kory, J. (2012). Consistent but modest: A meta-Analysis on unimodal and multimodal affect detection accuracies from 30 studies. In G. Castellano & K. Masse (Eds.), *Proceedings of the 14th ACM international conference on multimodal interaction* (pp. 31–38). New York: ACM.

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.

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

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

Graesser, A. C., D'Mello, S. K., & Strain, A. (2014). Emotions in advanced learning technologies. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *Handbook of emotions and education* (pp. 473–493). New York, NY: Taylor & Francis.

Grafsgaard, J. F., Wiggins, J. B., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2014). Predicting learning and affect from multimodal data streams in task-oriented

tutorial dialogue. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th international conference on educational data mining* (pp. 122–129). London, England: International Educational Data Mining Society.

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*, 359–365.

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

Harley, J. M. (in press). Measuring emotions: A survey of cutting-edge methodologies used in advanced agent-based learning environment research. In S. Tettegah & M. Gartmeier (Eds.). *Emotions, technology and learning: Communication of feeling for, with, and through digital media (Vol.1: Emotions, Technology, Learning and Design)*. London, UK: Elsevier.

Harley, J. M., & Azevedo, R. (2014). Toward a feature-driven understanding of students' emotions during interactions with agent-based learning environments: A selective review. *International Journal of Games and Computer Mediated Simulation, 6*(3), 17–34.

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, & Picard, R. W. (2005). Multimodal affect recognition in learning environments. In *Proceedings of the 13th annual ACM international conference on multimedia* (pp. 677–682). New York, NY: ACM.

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

Mauss, I. B., Levenson, R. W., McCarter, L., Wilhelm, F. H., & Gross, J. J. (2005). The tie that binds? Coherence among emotion experience, behavior, and physiology. *Emotion, 5*(2), 175–190.

Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition and Emotion, 23*, 209–237.

McDaniel, B. T., D'Mello, S. K., King, B. G., Chipman, P., Tapp, K., & Graesser, A. C. (2007). Facial features for affective state detection in learning environments. In D. McNamara & G. Trafton (Eds.), *Proceedings of the 29th annual cognitive science society* (pp. 467–472). New York, NY: Erlbaum.

McQuiggan, S. W., Robison, J. L., & Lester, J. C. (2010). Affective transitions in narrative-centered learning environments. *Journal of Educational Technology & Society, 13*(1), 40–53.

AlZoubi, O., D'Mello, S. K., & Calvo, R. A. (2012). Detecting naturalistic expressions of nonbasic affect using physiological signals. *IEEE Transactions on Affective Computing, 3*, 298–310.

Pantic, M., & Rothkrantz, L. J. M. (2003). Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE, 91*(9), 1370–1390.

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., & Linnenbrink-Garcia, L. (Eds.). (2014) *Handbook of emotions in education*. New York, NY: Routledge.

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.

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.

Pekrun, R., & Perry, P. P. (2014). Control-value theory of achievement emotions. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *International handbook of emotions in education* (pp. 120–141). New York: Routledge.

Pintrich, P. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 451–502). San Diego, CA: Academic Press.

Q-Sensor 2.0 [Apparatus and software]. (2013). Waltham, MA: Affectiva.

Russell, J. A., Weiss, A., & Mendelsohn, G. A. (1989). Affect grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology, 57*(3), 493–502.

Sabourin, J., & Lester, J. (2014). Affect and engagement in game-based learning environments. *IEEE Transactions on Affective Computing, 5*(1), 45–56.

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

Taub, M., Azevedo, R., Bouchet, F., & Khosravifar, B. (2014). Can the use of cognitive and metacognitive self-regulated learning strategies be predicted by learners' levels of prior knowledge in hypermedia-learning environments? *Computers in Human Behavior, 39*, 356–367.

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* (pp. 192–195). Eindoven, The Netherlands: ACM.

Trevors, G., Duffy, M., & Azevedo, R. (2014). Note-taking within MetaTutor: Interactions between an intelligent tutoring system and prior knowledge on note-taking and learning. *Educational Technology Research and Development*, 1–22.

Utthara, M., Suranjana, S., Sukhendu, D., & Pinaki, C. (2010). A survey of decision fusion and feature fusion strategies for pattern classification. *IETE Technical Review, 27*(4), 293–307.

Wagner, J., Kim, J., & André, E. (2005, July). From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. In *IEEE international conference on multimedia and expo* (pp. 940–943). Amsterdam, The Netherlands: IEEE Press.

Wallbott, H. G. (1998). Bodily expression of emotion. *European Journal of Social Psychology, 28*(6), 879–896.

Winne, P. H., & Hadwin, A. F. (2013). NStudy: Tracing and supporting self-regulated learning in the internet. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 293–308). Amsterdam, The Netherlands: Springer-Verlag.

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

Zimmerman, B., & Schunk, D. (2011). *Handbook of self-regulation of learning and performance*. New York: Routledge.