Emotion Regulation Tendencies, Achievement Emotions, and Physiological Arousal in a Medical Diagnostic Reasoning Simulation

Jason M. Harley, Ph.D
University of Alberta, Department of Educational Psychology, 5-139, Education North, Edmonton, AB, T6G 2G5, Canada
jharley1@ualberta.ca
Phone: (780) 492-9170
*Corresponding author
ORCID#: 0000-0002-2061-9519

Amanda Jarrell,
McGill University, Department of Educational and Counselling Psychology, 3700 McTavish Street 614, Montréal, QC, CAN, H3A 1Y2
amanda.jarrell@mail.mcgill.ca

Susanne P. Lajoie, Ph.D.
McGill University, Department of Educational and Counselling Psychology, 3700 McTavish Street 614, Montréal, QC, CAN, H3A 1Y2
susanne.lajoie@mcgill.ca

Abstract. Despite the importance of emotion regulation in education there is a paucity of research examining it in authentic educational contexts. Moreover, emotion measurement continues to be dominated by self-report measures. We address these gaps in the literature by measuring emotion regulation and activation in 37 medical students' who were solving medical cases using BioWorld, a computer-based learning environment. Specifically, we examined students' habitual use of emotion regulation strategies as well as electrodermal activation (emotional arousal) from skin conductance level (SCL) or skin conductance response (SCR), as well as appraisals of control and value and self-reported emotional responses during a diagnostic reasoning task in [Blinded]. Our results revealed that medical students reported significantly higher habitual levels of reappraisal than suppression ER strategies. Higher habitual levels of reappraisal significantly and positively predicted learners' self-reported pride. On the other hand, higher habitual levels of suppression significantly and positively predicted learners' self-reported anxiety, shame, and hopelessness. Results also revealed that medical students experienced relatively low SCLs and few SCRs while interacting with BioWorld. Habitual suppression strategies significantly and positively predicted medical students' SCLs, while SCRs significantly and positively predicted their diagnostic efficiency. Findings also revealed a significant, positive predictive relationship between SCL and shame and anxiety and the inverse relationship between SCL and task value. Implications and future directions are discussed.

Key words: emotion; affect; emotion regulation; physiological; arousal; activation; skin conductance

Acknowledgements: This research was supported by funding from the Social Sciences and Humanities Research Council of Canada.

Emotion Regulation Tendencies, Achievement Emotions, and Physiological Arousal in a Medical Diagnostic Reasoning Simulation

Academic achievement emotions are critical because of the impact they have on learners' academic outcomes, including their success and failure (Pekrun, 1992; Pekrun, Lichtenfeld, Marsh, Murayama, & Goetz, 2017; Pekrun & Linnenbrink-Garcia, 2014). Academic achievement emotions influence achievement outcomes by promoting situationally-appropriating information processing and self-regulation strategies (Pekrun, Elliot, & Maier, 2009; Pekrun & Perry, 2014) fostering motivation, and focusing attention and limited cognitive resources on achievement-related activities. Typically, positive activating achievement emotions such as pride, enjoyment, and hope are positively related to achievement and negative deactivating emotions such as hopelessness and boredom are detrimental to achievement (Goetz & Hall, 2013; Pekrun et al., 2017; Pekrun & Linnenbrink-Garcia, 2014). The relationship between achievement and positive deactivating emotions (e.g., relief) as well as negative activating emotions (e.g., anxiety, anger) is, however, more nuanced (Pekrun, 2006; Pekrun, Goetz, Titz, & Perry, 2002). Negative activating emotions are often maladaptive to academic achievement but can help individuals succeed in some cases. Anxiety and shame, for example, can motivate us to invest effort and avoid academic failure (Turner & Schallert, 2001). However, these same emotions can also have negative implications on achievement by consuming cognitive resources needed for the achievement task at hand

(Meinhardt & Pekrun, 2003) and compromising interest and intrinsic motivation (Pekrun & Perry, 2014).

Given the relationship between emotions and academic achievement, it is generally in learners' best interest to control their experience, intensity, and duration of academic achievement emotions, especially negative emotions associated with failure. Emotion regulation (ER), which refers to the attempts people make to influence which emotions they have, when they have them, and how they express and experience them (Gross, 2015), can help learners do this. Students' effective use of (ER) strategies can foster learning by helping them to adapt positively when they experience negative emotions, which can adversely impact performance and cognitive functioning (Chauncey-Strain & D'Mello, 2015; Pekrun, et al., 2011, 2014; Leroy, et al., 2012). Despite the importance of ER in education, there is a paucity of research examining it in educational contexts, particularly in naturalistic settings where participants are engaged in authentic, academically relevant tasks.

There is also a lack of research examining students' ER tendencies where relationships between ER and physiological arousal are examined. The latter being of particular value because physiological arousal provides information about learners' emotions that is challenging for them to mask, unlike behavioral and self-report data (Dan-Glauser & Gross, 2013). Specifically, physiological measures provide information about the *arousal* dimension of emotion that corresponds to the degree of physiological activation in an emotional response. Arousal is one of two central dimensions used to measure and understand emotions (Russell, Weiss, & Mendelsohn, 1989; Pekrun, 2006). The other is valence, which refers to the inherent pleasantness (e.g., enjoyment) or unpleasantness (e.g., anxiety) of an emotion. While physiological measures do not generally assess valence, this dimension can be inferred from either the context of a situation (e.g., arousal being linked to the announcement of a pop quiz vs. receiving an A on a particularly difficult exam) or other emotion expression components such as facial expressions or self-reports of emotion (Harley, Bouchet, Hussain, Azevedo, & Calvo, 2015; Mauss & Robinson, 2009). Studies have also found and replicated findings that physiological data can help differentiate positively and negatively valenced emotions, such as amusement and disgust (Kreibig, Samson, & Gross, 2015). Moreover, physiological measures of emotion have been used to differentiate the effectiveness of ER strategies (Gross, 2002).

The objectives of this study were to contribute to addressing the aforementioned gaps in the literature regarding the relationships between ER and achievement emotions, and the measurement of academic achievement emotions using physiological measures, in particular. This study contributed to addressing these gaps by measuring medical students' habitual (i.e., typical use of) ER strategies as well as their physiological and self-reported emotional responses during a diagnostic reasoning task in BioWorld (Lajoie, 2009), a computer-based learning environment (CBLE) that simulates a medical reasoning situation.

Theoretical Frameworks

Extended emotion process model of ER. Gross's (2015) extended emotion process model of ER was used as the primary theoretical framework for this study. The model posits that emotions are stimuli-directed, goal-related psychological phenomena with experiential, behavioral, and physiological components and that emotions are generated from appraisals of a situation’s goal congruency or misalignment. Gross (2015) conceptualizes ER as a second level valuation system that is used to modify a first emotion-generating system that could otherwise lead to an undesirable negative emotion being elicited. For example, receiving a lower grade than anticipated on a quiz could result in an internal or overt expression of sadness or frustration if the learner does not regulate their emotion. While Gross (2015) describes five different ways that individuals can regulate their emotions, the most widely studied and influential strategies on cognition, learning, and health outcomes are *cognitive reappraisal* and *suppression;* reappraisal being consistently identified as the more effective and adaptive of the two (Chauncey-Strain & D’Mello, 2015; Gross, 2015; Leroy, et al., 2012). Reappraisal is an antecedent-focused strategy and involves altering the way one is thinking about a situation before an emotion is experienced, such as reminding oneself that a bad score on a quiz is an opportunity to identify gaps in understanding, which can lead to a better score on an exam. Suppression is a response-focused strategy and involves deliberately trying to alter the way one is expressing an emotion and typically targets attempts to change one’s behavioral (e.g., avoid frowning or crying) and physiological responses (e.g., taking deep breathes to calm a racing heart).

Physiological measures of emotion have been used to differentiate the effectiveness of ER strategies such as suppression versus reappraisal. For example, a study by Gross (1998a) found that both suppression and reappraisal were effective in reducing emotion-expressive behavior when compared to a control condition, but suppression was associated with increased sympathetic activation whereas reappraisal was not. Another study that measured cardiovascular responses showed that learners’ high in reappraisal had a more adaptive emotional response profile when attempting to down-regulate anger in an experimental study (Mauss, Cook, Cheng, & Gross, 2007). Other research has supported these patterns, identifying reappraisal as a more effective ER strategy than suppression, where the latter impairs memory and increases physiological responding for those attempting to suppress their emotions (for a review see Gross, 2002).

Control-value theory of achievement emotions. A second pertinent and complimentary theoretical framework is Pekrun's control-value theory (CVT) of achievement emotions (Pekrun, 2006, 2011; Pekrun & Perry, 2014). The CVT was selected as an additional theoretical framework to add depth to the framing of the study as well as interpretation of results because of its appropriateness for understanding emotion in academic achievement contexts, as well its description of mechanisms that mediate the generation (i.e., stimulation) of emotions: control and value appraisals. Moreover, the CVT has also been applied to medical education to help guide research and practice in the field (Artino, Holmboe, & Durning, 2012; Artino & Pekrun, 2014; Duffy, Lajoie, Pekrun, & Lachapelle, in press). Research based on the CVT has consistently demonstrated the influence of different appraisal mechanisms on the elicitation of emotions experienced by students in academic achievement situations (e.g., test taking, studying, lectures; for review see Pekrun & Perry, 2014).

Appraisals of control and value are recognized in CVT as the most important appraisals in the direction of an emotion for an activity. Pekrun (2006, 2011; Pekrun & Perry, 2014) defines subjective *control* as one's perceived ability to effectively manage achievement activities and their outcomes, or more broadly, as one's beliefs concerning the causal influence they exert (agency appraisal) over their actions and outcomes (controllability), including the subjective likelihood of obtaining said outcome (probability). Pekrun further defines the term subjective *value* in the context of achievement emotions as the perceived importance of an activity and its outcome(s) to oneself (goal relevance), and more broadly as the perception that an action or outcome is positive or negative in nature (goal congruence—event supports or hinders goal attainment). The CVT also differentiates intrinsic value (e.g., chemistry as important because it is interesting; value stems from activity itself) from the extrinsic value (e.g., chemistry as important because of its instrumental role in admissions for a general bachelor of science program).

Appraisals of control and value can be used by learners in an academic achievement situation to regulate their emotions by changing (reappraising) how they are interpreting a situation with regard to these two dimensions. The CVT describes the relationship between different combinations of control and value (e.g., high control, low value), what the learner is focusing on (object focus: an academic achievement activity, such as writing an exam; or an outcome, such as how they will do on the exam) and different discrete emotions that arise as a result (e.g., boredom or enjoyment). Accordingly, the CVT compliments the extended process model of ER by outlining, for example, which appraisals should be prioritized by reappraisal strategies (control and value) and the direction (e.g., up or down) learners should use to influence their appraisals to regulate their emotion to a more desirable (e.g., positive) emotional state.

Related Work

The majority of published research to-date on ER has examined this phenomenon outside of authentic educational contexts and in highly-controlled experimental settings (Chauncey-Strain & D'Mello, 2015; Leroy, et al., 2012; Gross, 2015; Webb, Miles, & Sheeran, 2012). As such, participants' emotional states are often directed toward tasks that are not necessarily of personal and/or curricular value to them, such as participating in a deliberately tedious memorization task or reading about the U.S. constitution (Chauncey-Strain & D'Mello, 2015; Leroy, et al., 2012). Moreover, much of this research has relied on self-report measures as the only measure of emotions. Findings from research on ER in education and education-related settings is consistent with the results in experimental psychology, where reappraisal strategies are generally more effective and adaptive than suppression strategies (Chauncey-Strain & D'Mello, 2015; Leroy, et al., 2012; Gross, 2015; Jamieson, et al., 2010; Webb, Miles, & Sheeran, 2012).

Other research has sought to externally regulate student emotions by attempting to foster positive emotions (e.g., enjoyment, curiosity) through supportive messages delivered by virtual pedagogical agents (often targeting appraisals). However, these programs of research are too few and too mixed in terms of intervention approaches and results (Arroyo et al., 2013; D'Mello et al., 2010; Robison et al. 2009) to draw substantive conclusions (Harley, Lajoie, Frasson, Hall, 2017). Other research with computer-based learning environments has focused more on the environment (situation); specifically, on creating environments with features and pedagogical interactions that are related to emotions such as enjoyment and curiosity. These two approaches focus on adapting the environment to individual learner characteristics such as personality traits (Harley et al., 2016a) or to general learner preferences, such as the immersion and affective engagement associated with narrative (i.e., story) and other game-like features (Sabourin & Lester, 2014). The aforementioned studies represent the greatest amount of work done to-date: research uncovering design recommendations linked to learners' tendencies to experience postive emotional states. The theoretical frameworks advanced by Pekrun (2006, 2011; Pekrun & Perry, 2014) and Gross (2015) reveal, however, that there are other dimensions of ER besides the environment/situation that warrant empirical examination.

The measurement of emotion is bound to the goal of understanding how emotions can be regulated: one must be able to accurately and reliably detect learners' emotional states as they arise and change in order to select and implement appropriate ER strategies as they are needed. Unfortunately, the majority of educational psychology research continues to rely predominantly on self-report measures of emotion. Aside from the construct-general concerns that self-report measures present, such as social desirability and item fatigue, it is widely agreed that emotions are multicomponential and dynamic, temporally-complex psychological processes (Calvo & D'Mello, 2010; Harley, 2015; Harley et al., 2015; Porayska-Pomsta, Mavrikis, D'Mello, Conati, & Baker, 2013). Most self-report measures only provide a means to measure the experiential component of emotions, and those that include items related to learners' behavior and physiology, such as the achievement emotion questionnaire (AEQ; Pekrun et al., 2002), still cannot directly measure these emotional components and rely on learners' recollections and accurate reporting of such manifestations.

Finally, in expanding the tools of inquiry used to examine emotions, research must also return to investigate and question relationships drawn between psychological traits, processes, and states to determine whether patterns observed with self-report measures hold true for physiological and behavioral manifestations of emotion. If recent studies that examined the degree of coupling between emotional components are any indication (Evers et al., 2014; Harley et al., 2015; Mauss et al., 2005), there is a considerable amount of work to be done to understand if trace measures of emotion, physiological in particular, fit into the patterns of results that research relying on self-report data have woven. A recent meta-analyses of multimodal emotion measurement found substantial differences in the additive predictive value of using different types of emotion measurement methods (D'Mello & Kory, 2015). These findings reinforce the need to expand our understanding of accepted (i.e., assumed) relationships within (e.g., physiological: heart rate and skin conductance) and between emotion expression components (e.g., physiological and self-report; Harley, 2015). Such empirical research is also needed to help develop more granular theories of emotion. Presently, widely used theories of emotion such as the extended process model of emotion regulation (Gross, 2015) and the CVT (Pekrun & Perry, 2014) do not distinguish between experiential and different biological sources or measures of activation (i.e., arousal) nor do they hypothesize on the different relations these assorted expression components may have with learning and individual differences.

Numerous methods exist to measure physiological arousal, such as cortisol sampling from saliva (Jamieson, Mendes, Blackstock, & Schmader, 2010; Spangler, Pekrun, Kramer, & Hofmann, 2002), pupil dilation (Scrimin, Altoè, Moscardino, Pastore, & Mason, 2016), and heart rate (i.e., cardiac vagal tone; Butler, Wilhelm, & Gross, 2006; Dan-Glauser & Gross, 2015; Li et al., 2009). In the current study, we used electrodermal activity because it is one of the most widely-researched physiological channels for measuring emotion (Kapoor, Burleson, & Picard, 2007; Kreibig, Samson, Gross, 2015; Mauss & Robinson, 2009; Picard, Fedor, Ayzenberg, 2016), especially in computer-based learning environments like the present study (Calvo & D'Mello, 2010; Harley, 2015; Woolf, et al., 2009). It is also an appropriate physiological channel for detecting changes in emotion from participants engaging in emotion regulation strategies (Gross, 2002). Within the EDA complex there are two parallel channels: skin conductance level (SCL) and skin conductance response (SRL). SCL is slow changing, smooth, and understood in relation to an individual's physiological baseline (Braithwaite, et al., 2013) whereas SCR is characterized as rapidly changing, meaningful peaks in activation. These aspects of physiological activation are thought to rely on different underlying neurological mechanisms (Dawson et al., 2001; Nagai et al., 2004), which could lead to differences in their patterns and relationships with other psychological processes.

Current Study: Research Questions and Objectives

The current study contributes to addressing the aforementioned gaps in the literature by examining medical students habitual use of ER strategies and physiological arousal while they interacted with a computer-based diagnostic reasoning simulation, BioWorld (Lajoie, 2009). More specifically, relationships between habitual ER strategies, physiological activation, academic achievement (performance on a diagnostic reasoning task), and self-reported emotions and appraisals are examined. The first objective of this study was to examine the predictive relationship between habitual ER strategies, learning and emotions in an authentic and academically relevant task (Jamieson et al., 2010) rather than an artificial, experimental task. This study also extends the investigation of ER in academic and learning contexts to medical students; a demographic for whom the regulation of emotions is a critical part of their future work (Duffy et al., 2014; Lajoie et al., 2015).

The second objective of this study was to meaningfully extend prior research on ER tendencies by measuring emotional activation through the collection and analyses of electrodermal activation (EDA). In doing so this study sought to provide evidence on whether previously identified relationships between habitual ER strategies and emotional activation would carry over to authentic, academic achievement contexts (Gross & Levenson, 1993). In addition, including a physiological measure of emotional activation provided an opportunity to examine an under-examined component of emotion which is challenging for students to mask while also overcoming challenges associated with social desirability and the accuracy of recall (self-report data; Dan-Glauser & Gross, 2013), cultural confounds (Ekman, 1992), and triggering stereotypes through data collection (e.g., activating gender-based math anxiety through self-report; Goetz et al., 2013).

The third objective of this study was to conduct a preliminary, indirect comparative investigation of the relationship between study variables and the two components of the EDA complex: skin conductance levels (SCL; i.e., tonic EDA) versus skin conductance responses (SCR; i.e., phasic EDA); an assumed convergent relationship in theory and research. By measuring and examining the patterns of these two distinct components of the EDA complex we sought to provide preliminary evidence to guide the selection and interpretation of these different measures of physiological activation.

In order to accomplish the above objectives, we used Gross' (2015) extended emotion process model of emotion regulation and Pekrun's (2006, 2011; Pekrun & Perry,

2014) control-value theory of achievement emotions to guide the formation of the following research questions and hypotheses:

Research Question 1: What ER tendencies did medical students report and were they predictors of their (a) performance on, (b) self-reported emotions during a diagnostic reasoning task, and (c) appraisals of task control and value? No prior empirical studies that the authors were aware of had investigated medical students' ER tendencies and appraisals of control and value in a context similar to the current study (a computer-based learning environment). Prior research demonstrates that reappraisal ER strategies are more effective than suppression strategies (Gross, 2015). As such, we hypothesized that medical students, whom have particularly stressful and challenging programs to get into and perform well in (Lajoie et al., 2015) would be more likely to select the more adaptive of the two. Based on studies that have examined learning in a non-medical context, we predicted that habitual use of suppression strategies would relate negatively to higher diagnostic performance scores with BioWorld whereas reappraisal strategies would positively predict higher diagnostic scores (Chauncey-Strain & D'Mello, 2015; Leroy et al., 2012). We predicted the same pattern for positive emotions based on research that has found reappraisal strategies to be more effective than suppression strategies (Gross, 2015). Finally, we expected to find a positive predictive relationship between habitual use of reappraisal strategies and appraisals of control and value because these two appraisal dimensions are the most influential in generating (Pekrun & Perry, 2014), and by extension, regulating academic achievement emotions. In other words: a learner who typically engages in reappraisal to regulate their emotions is likely to reappraise their perceptions of control and value in an academic achievement situation to regulate their achievement emotions. We expected the opposite pattern between habitual use of suppression and control and value appraisals. Specifically, we hypothesized that reappraisal ER tendencies would positively predict control and value appraisals, while suppression ER tendencies would negatively predict control and value appraisals. We further hypothesized that reappraisal ER tendencies would positively predict positive achievement emotions (enjoyment, pride, and hope) and negatively predict negative achievement emotions (anxiety, hopelessness, and shame) while the inverse would be true for suppression ER tendencies.

Research Question 2: Did medical students' respective SCLs or SCRs illustrate the experience of heightened physiological arousal? No prior empirical studies that the authors were aware of have investigated medical students' physiological arousal in a computer-based learning environment. In prior research with computer-based learning environments, levels of physiological arousal are typically low (D'Mello, 2013; Harley, 2015; Harley et al., 2015); however, because in this study medical students engaged in potentially emotionally-laden decision making (Jarrell, Harley, & Lajoie, 2016; Jarrell, Harley, Lajoie, & Naismith, 2017) we expected that participants would experience higher levels of arousal. This RQ was important to address because it provides insight into a specific type of students' affective tendencies while they interact with a special type of learning environment. Understanding learners' over-all physiological arousal tendencies also helps to contextualize the other RQs related to SCL and SCR data.

Research Question 3: Do ER tendencies predict physiological arousal? We predicted that tendencies to use suppression would predict heightened levels of physiological activation (Gross & Levenson, 1993).

Research Question 4: Does physiological arousal predict performance on a diagnostic reasoning task? As physiological activation can indicate both positive activating emotions, such as enjoyment associated with effective performance and learning as well as negative emotions such as anxiety which has a more nuanced relationship with performance (e.g., some anxiety can motivate and focus attention, whereas too much can retract from cognitive resources) a directional hypotheses could not be established for this research question.

Research Question 5: Does physiological arousal predict learners' (a) retrospective self-reported emotions and (b) appraisals of control and value? Physiological activation does not provide any information for the emotional dimension of valence. Therefore, activation could indicate either pleasant activating emotions associated with learning, such as enjoyment, or unpleasant activating emotions such as anxiety and frustration that have a less clear relationship with learning (Pekrun & Perry, 2014). Neither research nor theory was therefore available to develop detailed hypotheses for RQ5 that would effectively differentiate the emotions measured in this study**.** More generally, we hypothesized, based on prior research, that if a predictive relationship emerged between physiological arousal and self-reported emotions, that it would be a positive predictive relationship because all emotions measured are activating emotions (Pekrun, 2006; Pekrun & Perry, 2014). Prior research has indicated that correlations between self-reported emotions and physiological arousal are generally weak to moderate in strength (Evers et al., 2014; Harley et al., 2015; Mauss et al., 2005), therefore we were more confident in the direction of a potential relationship rather than the emergence of the relationship. Typically, emotional intensity increases with task value, so we also hypothesized a predictive relationship between appraisals of value and physiological arousal. The relationship with control is, however, too nuanced in the CVT (Pekrun, 2006; Pekrun & Perry, 2014) to lend itself to a hypothesis for this RQ.

Methods

Participants

Medical students (*N* = 37) from one large North American public university (five were in their first year, 27 were in their second year, and three were in their third and two were in their fourth year) participated in this study[1,2]. None of the participants' diagnostic reasoning scores identified them as outliers. Moreover, when participants were grouped into (1) below year two (*n* = 5), (2) year two (*n* = 27), and (3) above year two (*n* = 5) no significant differences were observed in diagnostic reasoning scores or emotion regulation tendencies. Participants had a mean age of 24.30 (*SD* = 3.50), 57% were female (43% were male), and 41% self-identified as Caucasian (other self-identifying ethnicities included Arab, Asian, Chinese, and others).

Learning Environment

BioWorld (Lajoie, 2009) is a computer-based learning environment (CBLE) designed to help medical students learn how to effectively diagnose a patient through a diagnostic simulation. Each case begins with a patient history, which provides details on

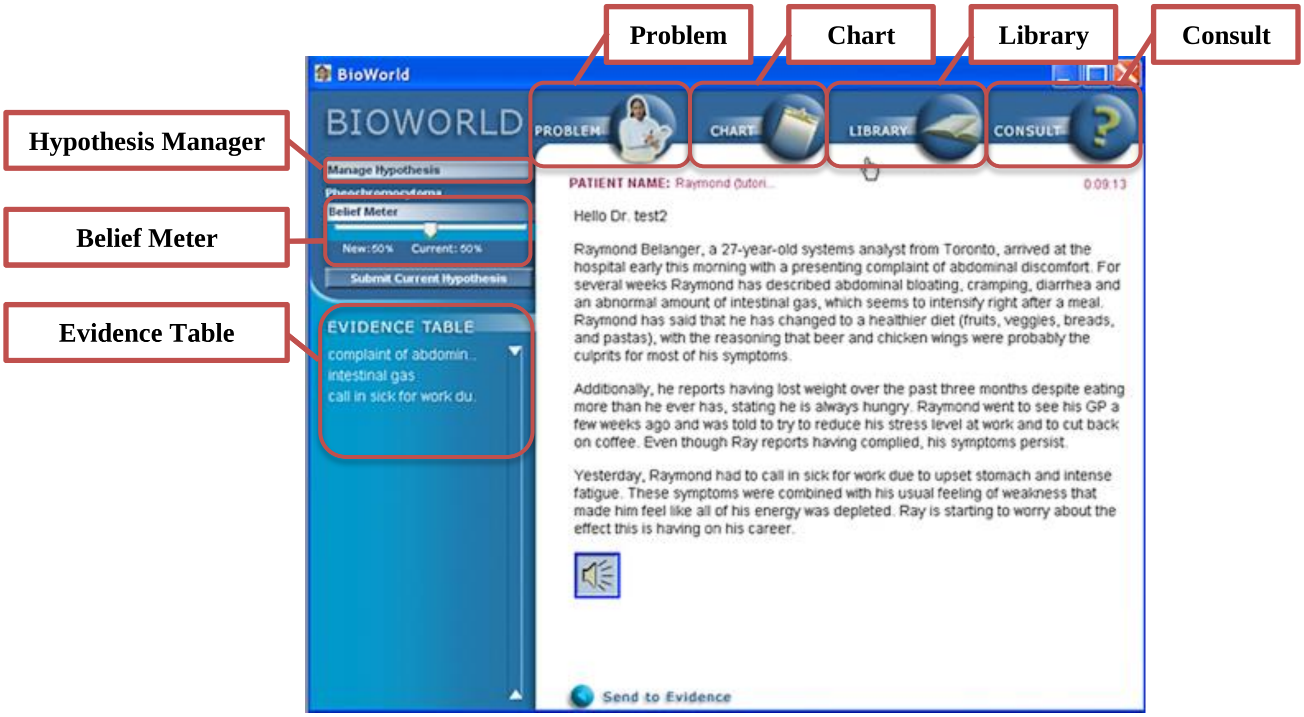
[1] Participants were not asked if they interacted with similar environments to this one, but given the novelty and specificity of BioWorld it is very unlikely.

[2] Students from first year onward in medical school stood to benefit from interacting with BioWorld which provides diagnostic reasoning training, therefore we did not exclude participants based on year of medical studies.

the case including relevant symptoms (see Figure 1). Students propose initial hypotheses (e.g., diagnosis) based on the evidence gathered in the patient history. As participants interact with BioWorld they collect additional information that will allow them to prove or disprove their proposed hypotheses, with the implicit goal of providing a correct diagnosis. During the diagnostic reasoning task different achievement emotions can be evoked. For example, students can obtain further evidence by ordering laboratory tests that immediately confirm or disconfirm a particular hypothesis. In response to the results from these laboratory tests, students can experience different achievement emotions. For instance, a student might become anxious when they receive a non-confirmatory test result (one that fails to provide support for their diagnostic hypothesis). This emotion arises if they appraise the task as being high in value and their appraisals of control drop to become only moderate and involve failure-oriented uncertainty about their ability to solve the case. If the same student focuses on anticipated success rather than failure, however, they may experience hope in the face of uncertainty. A student that comes to believe that they are incapable of solving a case, perhaps after a few non-confirmatory test results, may experience hopelessness. Shame may be experienced if this student attributes their failure to gather support for a hypothesis to themselves. A different student who received a confirmatory test result might feel pride if they value the task and attribute their success in identifying this piece of evidence to themselves.

Students can also obtain more information from searching the digital library. The use of the library can similarly elicit achievement emotions if the information provided by the library tool prompts students to appraise their value and control over the task or their value and expectancy for success and failure. Ordering laboratory tests that confirm or disconfirm hypotheses and searching for information to support and/or form them are just two potential sources of achievement emotions that medical students might experience *during* their interaction with BioWorld. These emotional states were a focus of the current study.

After the final diagnosis is submitted, students receive individualized feedback) on their solution based on an aggregated expert solution. For example, a comparison was provided between their solution with regard to which pieces of evidence they correctly identified or missed. Figure 2 illustrates how BioWorld provided feedback to medical students on their medical diagnosis by listing both the expert's prioritized items of evidence for their medical diagnosis (hypotheses) and the medical students'. Green check marks were used to indicate matching evidence between student and expert lists. Red exclamation marks were used to highlight evidence the expert listed as relevant and the student missed. Finally, evidence a medical student listed as relevant that did not match the experts was highlighted using grey font. Furthermore, students received their own report at the completion of the case that highlighted a written expert case summary as well.



PATIENT NAME: Raymond (tutori... 0:09:13

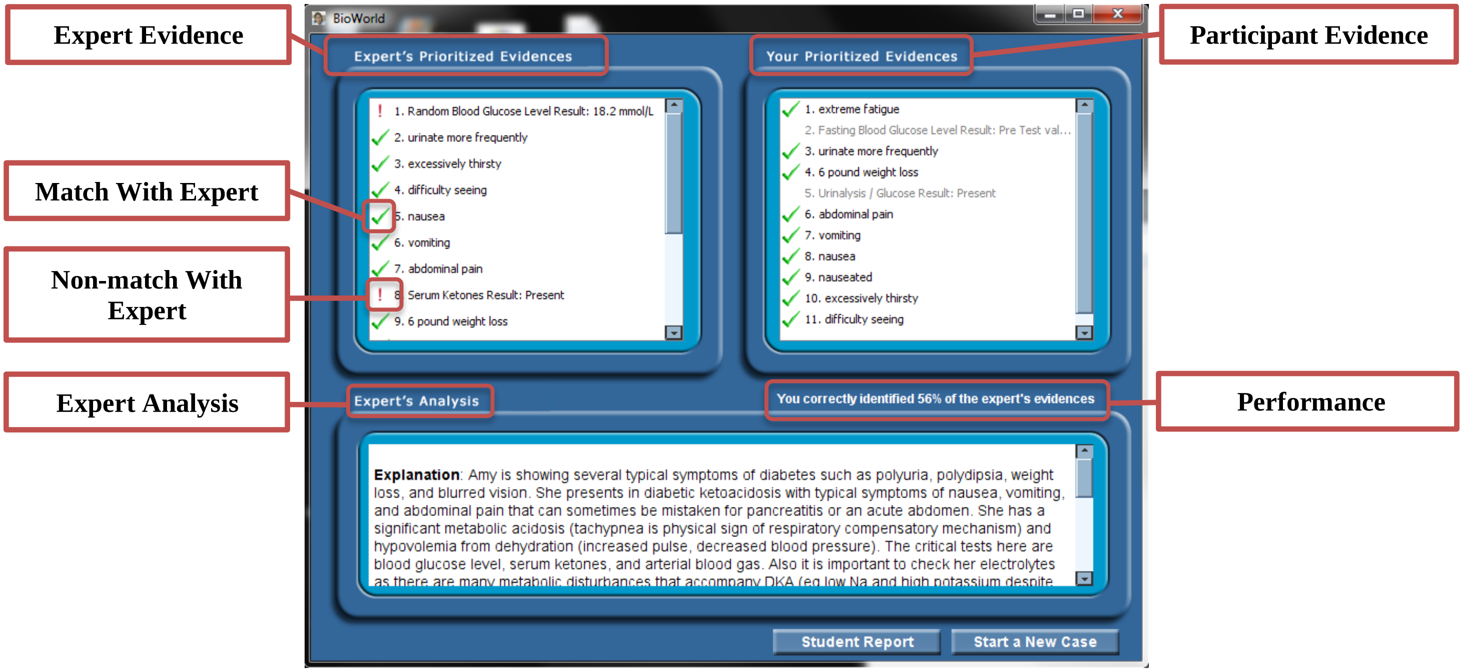
Hello Dr. test2

Raymond Belanger, a 27-year-old systems analyst from Toronto, arrived at the hospital early this morning with a presenting complaint of abdominal discomfort. For several weeks Raymond has described abdominal bloating, cramping, diarrhea and an abnormal amount of intestinal gas, which seems to intensify right after a meal. Raymond has said that he has changed to a healthier diet (fruits, veggies, breads, and pastas), with the reasoning that beer and chicken wings were probably the culprits for most of his symptoms.

Additionally, he reports having lost weight over the past three months despite eating more than he ever has, stating he is always hungry. Raymond went to see his GP a few weeks ago and was told to try to reduce his stress level at work and to cut back on coffee. Even though Ray reports having complied, his symptoms persist.

Yesterday, Raymond had to call in sick for work due to upset stomach and intense fatigue. These symptoms were combined with his usual feeling of weakness that made him feel like all of his energy was depleted. Ray is starting to worry about the effect this is having on his career.

Figure 1. Screenshot of Primary BioWorld Interface: Case Summary Tab.



! 1. Random Blood Glucose Level Result: 18.2 mmol/L
✓ 2. urinate more frequently
✓ 3. excessively thirsty
✓ 4. difficulty seeing
✓ 5. nausea
✓ 6. vomiting
✓ 7. abdominal pain
! 8 Serum Ketones Result: Present
✓ 9. 6 pound weight loss

✓ 1. extreme fatigue
2. Fasting Blood Glucose Level Result: Pre Test val...
✓ 3. urinate more frequently
✓ 4. 6 pound weight loss
5. Urinalysis / Glucose Result: Present
✓ 6. abdominal pain
✓ 7. vomiting
✓ 8. nausea
✓ 9. nauseated
✓ 10. excessively thirsty
✓ 11. difficulty seeing

You correctly identified 56% of the expert's evidences

Explanation: Amy is showing several typical symptoms of diabetes such as polyuria, polydipsia, weight loss, and blurred vision. She presents in diabetic ketoacidosis with typical symptoms of nausea, vomiting, and abdominal pain that can sometimes be mistaken for pancreatitis or an acute abdomen. She has a significant metabolic acidosis (tachypnea is physical sign of respiratory compensatory mechanism) and hypovolemia from dehydration (increased pulse, decreased blood pressure). The critical tests here are blood glucose level, serum ketones, and arterial blood gas. Also it is important to check her electrolytes as there are many metabolic disturbances that accompany DKA (eg low Na and high potassium despite

Figure 2. Screenshot of BioWorld Interface: Expert Feedback Interface

Measures

Electrodermal activation. The skin conductance level (i.e., SCL; tonic) or skin conductance response (i.e., SCR; phasic) aspects of the EDA complex was measured using one of two different devices, Q-sensor 2.0 or Biopac. Q-Sensor 2.0 (Apparatus, 2013) was used to measure learners' SCL and Biopac (Apparatus, 2013) was used to measure learners' SCR. The SCL component of the EDA complex reflects the slower aspect of the EDA signal, whereas the SCR component reflects faster changes. Both components are important dimensions of arousal and are thought to rely on different underlying neurological mechanisms (Dawson et al., 2001; Nagai et al., 2004). Learners did not wear both bracelets simultaneously.

Q-Sensor 2.0 provides eight raw values (microSiemens; uS) every second (sampling frequency). The bracelet was developed by Picard and colleagues who found SCL-EDA to be an effective predictor of affective states in the context of learning (Kapoor et al., 2007). Measurements are understood in relative terms due to individual differences in baseline SCL levels. Arousal is therefore inferred based on a higher or lower level compared to the individuals' average or baseline level.

Biopac (Braithwaite et al., 2013) was used to measure learners' SCR. The accompanying AcqKnowledge software derives the EDA from the bracelet's electrodes and computes SCRs using the guidelines advanced by the Handbook of Psychophysiology (Cacioppo, Rassinary, & Berntson, 2007). One method for assessing the level of arousal is by examining the frequency of significant SCRs. Significance in this context refers to deflections in the signal that exceed a pre-specified threshold. Biopac, like Q-Sensor, is a popular commercial research and development physiological sensor, used in many research studies (Hussain, D'Mello, & Calvo, 2014).

Emotion regulation strategies. The emotion regulation questionnaire (ERQ; Gross & John, 2003) was used to measure medical students' habitual (i.e., typical) use of reappraisal and suppression. Habitual ER strategies represent individual differences in tendencies to use reappraisal and suppression, which have been linked to greater and lesser levels of positive and negative emotions, interpersonal functioning, and wellbeing (Gross & John, 2003). The ERQ is a ten-item questionnaire that uses a seven-point scale (where 1 corresponds to "strongly disagree", 4 corresponds to "neutral", and 7 corresponds to "strongly agree") to measure students' tendencies to engage in reappraisal (6 items) or suppression (4 items) strategies when attempting to regulate positive and negative emotion. Cronbach's Alpha indicated that internal reliability was acceptable for each subscale (reappraisal, α = .85; suppression, α = .84).

Performance on diagnostic reasoning task. Learners' performance on a diagnostic reasoning task was extracted from the BioWorld logfiles using the percent match with the expert solution, which we refer to as diagnostic efficiency (Jarrell, 2016, 2017; Figure 2). The expert solution (provided to the learner at the end of each case) contained a visualization of evidence matches with the expert solution (efficiency score data) and the expert's analysis, which indicated the correct solution (accuracy score, not used; See Fig 2). Red exclamation marks indicate evidence items in the expert's prioritized evidence list that the participant failed to identify; green check marks indicated evidence items correctly identified; and evidence greyed out in the participant's prioritized evidence list indicated evidence items the participant considered relevant but were not pertinent to the final solution.

Academic achievement emotions. The *during* achievement situation subscale of the Academic Achievement Emotions questionnaire (AEQ; Pekrun, et al., 2002) was used to measure the emotions learners' experienced *while* solving the diagnostic reasoning task. The AEQ assesses emotions that commonly occur during achievement settings (Daniels et al., 2008; Pekrun & Perry, 2014; Pekrun et al., 2002, 2011, 2017). The AEQ was an appropriate self-report measure to use because learning with BioWorld is an achievement situation, and one that prior research has shown elicits achievement emotions such as pride and shame (Duffy, in press; Jarell et al., 2016, 2017). The AEQ uses a five-point scale where 1 corresponds to "strongly disagree" and 5 corresponds to "strongly agree." The AEQ was adapted, according to the instructional manual (Pekrun et al., 2002) and previous studies (Jarrell et al., 2017) to measure learners' test-related retrospective state emotions with BioWorld: "For each of the following items, indicate how you felt during the diagnostic reasoning task (solving the case within BioWorld)." An example item measuring hopelessness is "I had given up believing that I could solve the case correctly." The AEQ *during* state test-taking emotion subscale consists of 27 items and measures enjoyment (3 items; α = .55), pride (2 items; α = .74), hope (2 items; α = .54), anxiety (α = .71 after one item dropped; 7 items originally; α = .66), hopelessness (6 items; α = .91), shame (5 items; α = .77), and anger (2 items; α = .08). Cronbach's Alpha indicated that reliability was acceptable, with the exceptions of enjoyment, hope, and anger. Since these subscales were composed of two or three items it was neither advisable (enjoyment) or possible (hope, and anger) to improve the internal reliability through item elimination. Accordingly, enjoyment, hope, and anger variables were disregarded from analyses and pride, anxiety, hopelessness, and shame were retained. Medical students may have experienced pride or shame from attributing their success (pride) or failure (shame) to solve a case to their own effort or ability. On the other hand, they may have felt anxious if they were uncertain whether they would be able to solve a case (or similar case in the future) and were focused on potential failure. Hopelessness may have been aroused in similar situations where potential failure was focused on, but success was deemed to be unlikely rather than just uncertain. In all of these hypothetical situations, medical students would appraise the solving of cases as having value.

While the internal reliability of enjoyment, hope, and anger subscales are lower than those reported in previous studies using the AEQ, it is important to note that this study used only items pertaining to during task emotions. Previous studies collapsed across items that measured the emotion before, during and after the task, resulting in a larger overall number of items per emotion (Daniels et al., 2008; Pekrun et al., 2002, 2011). Less validation work has been done to-date working with subscales, although the AEQ allows for this use and it was theoretically warranted in this study (Pekrun, 2002).

Appraisals of control and task value. Appraisals of task control and value were measured using a single item each to avoid item fatigue: “I felt in control of my performance on the task” and “I valued the task.” Both items used a five-point scale, from 1 “strongly disagree” to 5 “strongly agree”.

Experimental Procedure

After signing the informed consent form learners were asked to complete a pre-session survey on Survey Monkey that included collecting demographic information and the ERQ questionnaire. Next, participants were guided by an experimenter[3] to attach and activate either a Q-Sensor or Biopac bracelet. Next, learners received a tutorial on how to interact with BioWorld in order to solve the diagnostic cases. During the learning session participants solved either two short diagnostic problems or one long problem, depending on the time of their booking (participants were run over the course of about a year *based on their availability*). Eight medical students completed one long problem as part of the piloting process for a newer case that was developed to expand the diagnostic training offered by BioWorld (see Data Analyses section below for analyses examining different subgroups of the study). Time does not permit having medical students complete all three cases in an experimental setting.

[3] Instructors did not have a role in the experiment. Research assistants that were part of the lab that conducted this study managed the experimental protocol.

In either case, the session took 2.5 hours to complete. After solving the last problem, participants were asked to report how they felt by completing a modified version of the AEQ retrospective state emotions test and the single-item self-report measures of control and value appraisals. All self-report data was collectively using an electronic survey: Survey Monkey. The data analyzed in this study was collected as part of a larger, ongoing project. Only the measures relevant to the analyses in this study will be discussed.

Data Analysis

Table 1 summarizes the sample each research question drew upon.

Table 1.

Sample size by research question

Variable	*Total N*	*Groups*		*Research Question Relevance*
EDA		SCR (Biopac bracelet)	*SCL (Q-Sensor)*	*2-4*
	36	*14*	*22*	
ER tendencies	*37*	*15*	*22*	*1, 3*
Self-reported emotions	*21*	*N/A*	*21*	*1, 5*
Diagnostic efficiency Total	*36*	*15*	*21*	*1, 4*
Case 1 and 2[1]		*14*	*13*	
Case 3			*8*	

Notes. [1] = Diagnostic efficiency averaged over case 1 and 2. Analyses used samples from groups, when appropriate, rather than pooling groups in all situations. For example, SRC and SCL were never grouped together in the same SC analyses because they are different measures of skin conductance. Cases were also grouped together when appropriate and possible, but not in all analyses (see BioWorld Cases section starting on page 25). This table therefore represents all the different combinations of samples used in this study and compliments table 3 in providing an overview of the study and study variables.

EDA-SCL. In order to analyze the SCL data from the Q-Sensor 2.0, participant's minimum EDA response value was extracted from their baseline (a period of two-five minutes before the learning session began) and participant's maximum EDA response value was extracted from EDA data collected during the entire session. Each EDA data point logged during the session was then standardized based on each participant's unique minimum and maximum values using the following formula ([Standardized EDA Response = (EDA value – minimum value)/(maximum value – minimum value)] (Dawson et al., 2001). These standardized EDA scores were interpreted based on proximity to zero and one. For example, a standardized EDA value of zero indicates that the EDA score is equal to their minimum EDA value whereas a normalized EDA value of 1 indicates that that the EDA score is equal to their maximum score. Any other value was interpreted relative to its position between 0 and 1 as a standardized proportional scale (i.e. a value close to zero would be interpreted as a relatively low arousal response whereas a value approaching 1 would be interpreted as a relatively high arousal response). The average of these individual standardized scores was then calculated across the session to examine each individual's mean SCL level tendencies across the learning session (i.e. a user-dependent measure of physiological arousal). One participant was dropped from the EDA analyses because of higher baseline than learning session scores, which created an anomalous negative average EDA value (Braithwaite et al., 2013).

EDA-SCR. SCR data from Biopac provided frequency data about the quantity of significant SCRs. In order to calculate a mean value across participants we calculated the frequency of SCRs per minute (across the time learners spent interacting with BioWorld) for each learner (Braithwaite et al., 2013). One participant's Biopac EDA was not recorded properly.

EDA-SCR and SCL. Given that SCL and SCRs represent different and important aspects of the EDA complex, it was not appropriate to amalgamate EDA data from these two different sources. Therefore, to answer the research questions, analyses were run separately for each type of EDA response. Medical students only wore one type of bracelet—depending on the availability of bracelet type— and therefore provided either SCL or SCR data from individual participants. Different bracelets were used because the experimenters did not have enough physiological bracelets for all participants to wear the same model during group data collection (where multiple participants were run at the same time, but on individual computers separated by walls and without collaboration). Comparisons between bracelets was a secondary aim and one that emerged due to the different types of information each bracelet collected: SCL vs SCR. Participants were instructed to wear the physiological bracelet on their non-dominant hand to avoid movement artifacts associated with frequent mouse movement during their interaction with BioWorld. Accordingly, it would not have been appropriate to have participants wear one bracelet on each hand.

Diagnostic efficiency. The number of evidence matches with the expert solution was converted into a percentage score by taking the number of correct evidence items and dividing it by the total number of evidence items in the expert solution. For example, if a participant correctly identified five pieces of evidence but failed to identify three items, then the participant received a percentage of correct matches score of 63% (i.e. 5/8). Converting the number of matches with the expert to a percentage enabled comparisons across cases where the expert solution consisted of varying numbers of evidence items. In order to answer our research questions related to diagnostic efficiency we used an average of medical students' diagnostic efficiency scores across counterbalanced cases when they solved more than one case (i.e., two shorter cases rather than one longer one).

BioWorld Cases. Some of the analyses conducted in this paper pooled medical students who used BioWorld to solve different medical cases. We ran a number of analyses to determine the circumstances that pooling participants would be appropriate.

An independent samples t-test was run examining diagnostic efficiency between all participants who learned about case 1 and 2 versus 3. A significant difference in diagnostic efficiency between students' averaged scores for Case 1 and 2 (M = 52.46; SD = 16.32) and those who learned about Case 3 was observed (M = 11.00; SD = 7.50), $t(16.27) = 10.19$, $p < .05$). Accordingly, those eight participants who interacted with Case 3 only were discarded from the second sub question of RQ1 that examined the relationship between ER strategies and diagnostic efficiency.

We had no reason to believe that medical students ER tendencies, reported prior to their interaction with BioWorld would be affected by their case assignment. None-the-less, we conducted two independent samples t-tests, one for each of the ER tendencies and no significant differences were observed.

An independent t-test analysis was also run to examine whether differences in diagnostic efficiency existed between a subset of participants who wore the SCL bracelet to learn about Case 1 and 2 (n = 13) versus those who learned about Case 3 (n = 8)[4]. The

[4] No comparisons with participants who wore an SCR bracelet were made because these participants were never directly compared (see Table 1; all usable SCR data came from participants who interacted with Case 1 or 2: n = 14)

same pattern as above (with participants who wore both types of bracelets) emerged) A significant difference in diagnostic efficiency between students' averaged scores for Case 1 and 2 (*M* = 51.10; *SD* = 21.56) and those who learned about Case 3 was observed (*M* = 11.00; *SD* = 7.50), *t*(16.13) = 6.13, *p* < .05). Accordingly, we excluded participants who learned about Case 3 from analyses that examined the relationship between diagnostic efficiency and SCL (RQ4).

Next, we ran a series of independent sample t-test analyses to see if any significant differences could be identified from comparing the self-reported emotions and SCLs of participants who learned about Case 1 and 2 (*n* = 12 with full data) versus Case 3 (*n* = 9). No significant differences were observed, therefore we grouped these students together for analyses that examined self-reported emotions (RQ1 and 5) and physiological arousal (RQ5; i.e., analyses where diagnostic accuracy was not assessed).

Data cleaning. Data cleaning was conducted at the variable and group level to limit the undue influence of outlying scores on means by creating and screening distributions using stem and leaf and boxplots generated with SPSS (Tabachnick & Fidell, 2007). Outlying scores were changed to the next most extreme score that was not an outlier (Tabachnick & Fidell, 2007). All variables were also screened for skewness and kurtosis (Tabachnick & Fidell, 2007). None of the variables (including different sample size permutations of them; see Table 3) were skewed or kurtotic (using a threshold of $z = +/-1.96$) with the exceptions of two modest departures in skew ($z = 2.05$ and 2.12) for SCL variables; neither sufficient to warrant variable transformations.

Rational for analyses. In order to answer our research questions, we ran simple linear or multiple regression analyses, depending on the number of predictor variables being examined per question. This article sought to answer five research questions, many of which had sub questions that involved testing separate null hypotheses (H_0). For each of these hypotheses we examined potential relationships between different variables, such as discrete emotions and psychological properties (RQ 1 and 5) or emotions and physiological manifestations (RQ 2-4). Since Type I error is localized for each H_0, family-wise alpha adjustments were neither necessary nor advisable in the context of the current study. Familywise alpha adjustments decrease power and inflates type II error rates (O’Keefe, 2003; Steinfatt, 1979) and therefore, should be conducted only when necessary. While some argue that alpha adjustments for multiple testing is not necessary (O’Keefe, 2003), a more stringent view is that alpha adjustments for multiple testing is not necessary when each statistical test assesses a different null hypothesis (Matsunaga, 2007[5]; Rubin, 2017).

Results

Preliminary Results

We examined the zero-order relationship between self-reported emotion and appraisals (see Table 2). As expected, all negatively-valenced emotions (anxiety, hopelessness, shame) were significantly and positively correlated with one another and negatively correlated with pride. Appraisals of value were significantly negatively correlated with anxiety, $r = -.54$, $p < .05$, and hopelessness, $r = -.55$, $p < .05$. With the exception of the significant negative correlations between value and anxiety and

[5] “Type I error is and should be localized by H_0 because Type I error refers to the error of falsely rejecting a given null hypothesis when it should not be rejected (e.g., Curran-Everett, 2000). In other words, identification of the scope of a given H_0 leads to the proper localization of Type I error, which, in turn, dictates how the respective alpha level should be adjusted.” (Matsunaga, 2007, p. 251).

hopelessness, these results reflect the expected patterns of achievement emotions and appraisals. This provides an additional source of validity for these measures.

Descriptive statistics (see Table 3) revealed that the negative relationship between value and anxiety and value and hopelessness was likely between higher and mid-level appraisals of value rather than low appraisals of task value on account of a high mean *M* = 4.38 (*SD* = 0.67) and a relatively high minimum value of 3.00 (possible minimum value of 1.00 and maximum value of 5.00). CVT (Pekrun, 2006) holds that appraisals of value must be affirmative rather than high for negative emotions such as anxiety to be experienced. If one doesn't care about the outcome, what is there to feel anxious (or hopeless) about? Taken together, the relationships between emotions and between emotions and appraisals are empirically and theoretically aligned.

Table 2.

Zero-order correlations between study variables

	1	2	3	4	5	6	7	8	9
Reappraisal (SCL)	1								
Suppression (SCL)	.23	1							
Pride	.50*	-.07	1						
Anxiety	-.04	.51*	-.41	1					
Hopelessness	-.13	.48*	-.47*	.85**	1				
Shame	-.00	.60**	-.20	.81**	.79**	1			
Control	.17	-.37	.39	-.22	-.30	-.36	1		
Value	-.14	-.51*	.37	-.54*	-.55**	-.36	.46*	1	
Average SCL (C2)[1]	.33	.60**	-.16	.50*	.38	.55*	-.28	-.50*	1

Note. *. Pearson correlation is significant at the 0.05 level (2-tailed); **. Pearson correlation is significant at the 0.01 level (2-tailed). [1] = Average normalized skin conductance level (SCL) during Case 2 (or only case). See Table 3 for more information about variables.

Table 3.

Descriptive statistics for all study variables

Variable[4]	*N*	*Mean*	*SD*	*Observed Min*[4]	*Observed Max*	*Possible Min*	*Possible Max*
Reappraisal	37	5.24	.70	4.33	6.67	1.00	7.00
Suppression	37	3.60	1.29	1.00	6.50	1.00	7.00
Reappraisal (SCL)[1]	22	5.19	0.57	4.33	6.00	1.00	7.00
Suppression (SCL)[1]	22	3.73	1.35	1.00	6.50	1.00	7.00
Reappraisal (SCR)[2]	15	5.14	.97	3.50	7.00	1.00	7.00
Suppression (SCR)[2]	15	3.43	.31	1.75	5.25	1.00	7.00
Reappraisal (NoC3)[3]	28	5.29	.66	4.50	6.67	1.00	7.00
Suppression (NoC3)[3]	28	3.46	1.18	1.75	6.00	1.00	7.00
Diagnostic Accuracy[4]	28	52.70	15.81	21.50	81.50	.00	100.00
Diagnostic Accuracy[4] (SCL)	13	51.08	21.56	15.00	81.00	.00	1.00
DiagnosticAccuracy[5] (SCR)	14	52.36	9.48	34.00	68.50	.00	1.00
Average (SCL)[6]	22	.21	.19	.01	.64	.00	1.00
Average (SCR)[7]	14	.73	.58	.02	1.84	.00	N/A
Average SCL (C2)[8]	22	.21	.21	.01	.60	.00	1.00
Average SCL (NoC3)[9]	13	.10	.08	.01	.20	.00	1.00
Pride[10]	21	2.90	0.96	1.00	4.50	1.00	5.00
Anxiety[10]	21	1.83	0.58	1.17	3.17	1.00	5.00
Hopelessness[10]	21	1.66	0.74	1.00	3.00	1.00	5.00
Shame[10]	21	1.59	0.65	1.00	2.80	1.00	5.00
Control[10]	21	3.61	0.97	2.00	5.00	1.00	5.00
Value[10]	21	4.38	0.67	3.00	5.00	1.00	5.00

Note. Variable values are post-data screening to reflect analyses. [1] = Habitual use of emotion regulation strategies for participants who had their SCL measured and completed the self-reports for Case 2. [2] = Habitual use of emotion regulation strategies for participants who had their SCR measured. [3] = Habitual use of emotion regulation strategies for participants, excluding those who interacted with Case 3. [4] = Includes participants who interacted with case 1 and 2; case 3 excluded to reflect analyses examining diagnostic efficiency. [5] = Diagnostic accuracy for those who had their SCR measured (none interacted with Case 3). [6] = Average normalized skin conductance level (SCL); 22 participants wore the Q-Sensor EDA bracelet and we were therefore able to calculate their SCL. One participants' data was unusable.
[7] = Average frequency per min skin conductance response (SCR); 15 participants wore the Biopac EDA bracelet and we were therefore able to calculate their SCR. [8] = Average normalized skin conductance level during Case 2. [9] = 13 participants wore the Q-Sensor EDA bracelet and interacted with case 1 and 2. [10] = All self-reported emotions and appraisals were directed toward the last (or only) case participants solved.

Research Question 1: What kind of ER strategies did medical students report typically using (i.e., ER tendencies) and were they predictors of their (a)

performance, (b) self-reported emotions during a diagnostic reasoning task, and (c) appraisals of control and task value?

Medical students who participated in this study tended to report moderate habitual levels of reappraisal, *M* = 5.24 (*SD* = 0.70) and lower (intermediate) levels of suppression, *M* = 3.60 (*SD* = 1.29). A paired samples t-test revealed that medical students reported significantly higher habitual levels of reappraisal than suppression ER strategies, *t*(36) = 7.22, *p* < .001, *d* = 1.19 (see Table 3).

In order to identify whether habitual ER strategy use predicted their performance on a diagnostic reasoning task, a multiple linear regression was conducted. No significant predictive relationship of reappraisal or suppression tendencies on diagnostic efficiency was found (See Table 3 for descriptive statistics for: Diagnostic Accuracy, and Reappraisal (NoC3) and Suppression (NoC3); see Table 4 for regression summary and Table 5 for zero-order correlation matrix).

In order to answer the third component of the first research question we ran a series of multiple regression analyses. For each model, reappraisal and suppression were entered as predictive variables and one of the four achievement emotions were entered as the dependent variable (pride, anxiety, hopelessness, shame). The model examining the predictive relationship between ER tendencies and pride was significant (R^2 = .28; *p* = .050[6]), where reappraisal was a significant predictor (*β* = .54, p < .05). The

[6] While *p* = .050 might be considered marginally significant rather than significant, it is squarely on the fence of *p* < .05. Moreover, this result was significant prior to outlier cleaning. Given that no single approach to outlier cleaning can deal perfectly with outliers and their influence on the distribution, the prior significance of this finding to a single outlier being removed, and the difference in reporting versus not reporting being a value of .001, we opted to refer to this variable as significant rather than marginally significant, as we do not typically report marginally significant results.

models examining the relationship ER tendencies anxiety ($R^2 = .28; p < .05$), hopelessness ($R^2 = .29; p < .05$), and shame ($R^2 = .40; p < .01$) were also significant; in all cases, suppression was a significant predictor of negative emotions (see table 4).

In order to answer the fourth component of the first research question, we ran two multiple regression analyses where both ER tendencies were entered as predictive variables and appraisals of value or control were entered as the dependent variable in each model. Neither model was significant. See Table 4 for a summary of multiple regression results for RQ1.

Table 4

Multiple regression results for ERQ on AEQ emotions, appraisals, SCL, and diagnostic efficiency

	Standardized coefficients Emotion regulation tendencies		R^2	Adjusted R^2	*p*
Variable	Reappraisal	Suppression			
Pride[1]	.54*	-.10	.28	.20	.050*
Anxiety[1]	-.17	.55*	.28	.21	.049*
Hopelessness[1]	-.26	.54*	.29	.21	.045*
Shame[1]	-.15	.64**	.38	.31	.013*
Control[1]	.27	-.43	.21	.12	.128
Value[1]	-.02	-.50*	.26	.18	.067
Diagnostic efficiency[2]	.08	-.19	.04	-.04	.60
Average SCL[1]	.20	.52*	.36	.29	.015*
Average SCR[3]	.37	-.41	.30	.17	.15

Note. *. Correlation is significant at the 0.05 level (2-tailed); **. Correlation is significant at the 0.01 level (2-tailed). Variables and sample used to answer Research Questions 1 and 5. [1] = multiple regressions for emotions, appraisals, and Average SCL used Reappraisal (SCL) and Suppression (SCL) as predictor variables. [2] = multiple regression for Diagnostic efficiency used Reappraisal (NoC3) and Suppression (NoC3) as predictor variables. [3] = multiple regression for Average SCR used Reappraisal (SCR) and Suppression (SCR) as predictor variables.

Table 5.

Zero-order correlations between habitual ER strategies and diagnostic efficiency

	1	2	3
Diagnostic efficiency	1		
Reappraisal (No C3)	.05	1	
Suppression (No C3)	-.18	.18	1

Note. *. Pearson correlation is significant at the 0.05 level (2-tailed); **. Pearson correlation is significant at the 0.01 level (2-tailed).

RQ2: Did medical students' respective SCLs or SCRs illustrate the experience of heightened physiological arousal?

Medical students had a low mean frequency of .73 (*SD* = .58) SCRs per minute (Braithwaite et al., 2013; Boucsein, 2012) during their interaction with BioWorld and a low standardized SCL average of .21 (*SD* = 19%). See Table 3.

RQ3: Do ER tendencies predict physiological arousal?

In order to answer our third research question we ran two linear multiple regression analyses where ER tendencies were entered as the predictor variables and physiological arousal (either SCL or SCR) was entered as the dependent variable. The model using habitual ER tendencies to predict SCL was significant (R^2 = .36*; p* < .02), where suppression was a significant predictor (β = .52, p < .05; see Table 4 for regression details and 6 for zero order correlations).

Table 6.

Zero-order correlations between ER tendencies and skin conductance

	1	2	3		1	2	3
1. Average SCL	--			1. Average SCR	--		
2. Reappraisal (SCL)	.32	--		2. Reappraisal (SCR)	.36	--	
3. Suppression (SCL)	.57**	.23	--	***3.*** Suppression (SCR)	-.40	.01	--

Note. *. Pearson correlation is significant at the 0.05 level (2-tailed); **. Pearson correlation is significant at the 0.01 level (2-tailed). SCL correlations differ from Table 2 because Table 6 SCL is computed across cases whereas Table 2 includes SCL for Case 2 only in order to answer RQ 5.

RQ4: Does physiological arousal predict performance on a diagnostic reasoning task?

In order to answer our third research question, we ran two simple linear regression analyses where physiological arousal (either SCL or SCR) was entered as the predictor variable and diagnostic efficiency was entered as the dependent variable. The model using SCR to predict diagnostic efficiency was significant ($R^2 = .33$; $p < .05$; β = .58, $p < .05$; see Table 7 for regression details).

Table 7

Simple linear regression results for skin conductance on diagnostic efficiency

	Standardized coefficients	Model R^2	Model adjusted R^2	*p*
Physiological Arousal	Diagnostic efficiency[1]			
Average SCR	.58*	.33	.28	.03*
Average SCL (No C3)	.09	.01	-.08	.76

Note. *. Significant at the 0.05 level (2-tailed); **. Significant at the 0.01 level (2-tailed). N = 13 for SCL regression because Case 2 participants excluded since diagnostic efficiency is being examined. [1] = Diagnostic (SCL) or (SCR) depending on whether the predictor variable was SCR or SCL.

RQ5: Does physiological arousal predict learners' (a) retrospective self-reported emotions and (b) appraisals of control and value?

In order to investigate whether medical students' physiological activation during their last case (i.e., second or only, depending on case assignment) predicted their self-reported emotions and appraisals of control and value (reported after their last case) we ran a series of simple linear regressions, where SCL[7] was entered as the predictor variable and self-reported achievement emotions and appraisals were entered as the dependent variable. Results revealed that SCL positively predicted anxiety ($R^2 = .25$*;* $p <$ .05; and shame ($R^2 = .30$*;* $p < .05$) and negatively predicted task value ($R^2 = .25$*;* $p < .05$). See Table 8 for regression table and Table 2 for zero-order correlations of variables.

Table 8

Simple linear regression results skin conductance (during second or only case) on achievement emotions and appraisals

	Standardized coefficients	R^2	Adjusted R^2	*p*
Emotion	SCL (C2)			
Pride	-.16	.03	-.03	.48
Anxiety	.50*	.25	.21	.02*
Hopelessness	.38	.14	.10	.09
Shame	.55*	.30	.26	.01*
Control	-.28	.08	.03	.21
Value	-.50*	.25	.21	.02*

Note. *. Significant at the 0.05 level (2-tailed); **. Significant at the 0.01 level (2-tailed).

Discussion

[7] The academic achievement emotion questionnaire data was not collected for participants who wore the Biopac bracelet. See limitations for more details.

The findings from this study indicate that medical students reported significantly higher habitual levels of reappraisal than suppression ER strategies. Higher habitual levels of reappraisal significantly and positively predicted learners' self-reported pride. On the other hand, higher habitual levels of suppression significantly and positively predicted learners' self-reported anxiety, shame, and hopelessness. Results also revealed that medical students experienced relatively low SCLs and few SCRs while interacting with BioWorld. Habitual suppression strategies significantly and positively predicted medical students' SCLs, while SCRs significantly and positively predicted their diagnostic efficiency. Findings also revealed a significant, positive predictive relationship between SCL and shame and anxiety and the inverse relationship between SCL and task value.

Our first main finding (medical students reported significantly higher habitual levels of reappraisal than suppression) aligns with our hypothesis that medical students, whom have particularly stressful and challenging programs to get into and perform well in (Lajoie et al., 2015) would be more likely to select the more adaptive of the two. It is possible (though an inference) that these ER tendencies may have helped the medical students in our sample get into medical school (and will perhaps help get them through, if they continue to favor the use reappraisal to regulate negative emotions).

The significant predictive relationships between ER strategies and discrete self-reported emotions summarized above provided some support for our hypotheses that reappraisal ER tendencies would be positively predictive of positive emotions (pride), while suppression would be positively predictive of negative emotions (anxiety, shame, and hopelessness). These results provide some support for the generalizability of previous research identifying use of, and habitual use of, reappraisal strategies as more effective at regulating emotions than suppression strategies (Gross, 2015) with a sample of medical students interacting with a computer-based learning environment. In other words: habitual reappraisal strategies were associated with positive emotions (pride[8]), whereas habitual suppression strategies were associated with negative emotions (anxiety, shame, and hopelessness). The negative correlation reported in Table 2 (and marginal predictive relationship reported in Table 4) between habitual suppression and appraisal of task value revealed that learners higher in habitual use of suppression were less likely to appraise the task as valuable; perhaps because they didn't attempt to reappraise the task as more valuable. Future research should examine this possibility by measuring context-specific emotion regulation to know if students actually used reappraisal to shape their perceptions of task value.

Our second main finding (RQ2) was that medical students experienced relatively low SCLs and few SCRs per minute (Braithwaite et al., 2013; Boucsein, 2012) revealing low arousal while interacting with BioWorld. While this finding was not aligned with our expectations, it may be explained by the learning session being low-stakes because it was not tied to students' grades in medical school and thus lower in extrinsic and instrumental

[8] According to the CVT, pride can be elicited when students appraise an outcome with positive value (e.g., success) and believes that they are responsible for this outcome (Pekrun, 2006). Pride during a performance task, such as a test, can be attributed to their level of knowledge and performance (Pekrun, 2006; Pekrun, Goetz, Titz, & Perry, 2002). Likewise, during the diagnostic reasoning task, students can feel pride in relation to their knowledge of different diseases, symptoms, and presentations, and their perceived performance as they move through solving the case. For example, a student could feel proud that they were able to correctly select the laboratory test that enabled them to identify an abnormal test result indicative of a particular disease. Given that students can feel pride during diagnostic reasoning, it is possible that individual differences in emotion regulation could be associated with this emotion. Previous research in the context of test-taking and learning have also tested this possibility and found that wishful thinking and self-blame are negatively related to feelings of test pride (Decuir-Gunby, Aultman, & Schutz, 2009) and habitual reappraisal is positively related to feelings of learning-related pride (Burić, Sorić, & Penezić, 2016). Therefore, it is possible and reasonable to suggest that habitual reappraisal could be related to pride in the context of diagnostic reasoning.

value (as opposed to intrinsic value which may have strengthened value appraisals) and overall, lower intensity emotions (Pekrun, 2006; Pekrun & Perry, 2014). Relatedly, while participants reported moderate levels of pride, they reported relatively low levels of all negative emotions. This finding is aligned with other research that has identified low levels of physiological arousal and self-reported emotion in other studies with BioWorld and additional intelligent tutoring systems and computer-based learning environments that do not incorporate game-like features (D'Mello, 2013; Harley, 2015; Harley et al., 2015, 2016a; Jarrell et al, 2016, 2017) in contrast to those that effectively exploit these features (Shute et al., 2016; Sabourin & Lester, 2014). In other words, physiological arousal and emotional intensity is often low in the kind of environment in which this study investigated physiological arousal. While it is generally adaptive to not be confronted with potentially distracting and intense, negative emotions such as anxiety and shame while completing a task, experiencing lower levels of positive emotions can also undermine motivation. Fortunately, the moderate levels of positive emotions and high levels of task value suggest that medical students may have found the task sufficiently pleasant and meaningful, but not so much so as to suggest that they may have been engaging in off-task behavior (e.g., gaming the system). While one would expect high levels of task-value to be associated with higher levels of physiological activation, our single-item measure of task-value was not able to differentiate between intrinsic and extrinsic task value (Pekrun, 2006). It is therefore possible that learners valued the learning task intrinsically because of its relevance to their program of studies and careers as medical doctors, but not extrinsically because they were not graded on their performance. Accordingly, while the high levels of task-value may have been driven by the intrinsic dimension of this appraisal it may not have been enough to evoke high levels of physiological activation.

While the variance in medical students' physiological arousal was low for both SCL and SCR, results suggest that habitual use of suppression strategies may not be adaptive. As expected, suppression strategies were a significant predictor of SCL (though not SCR), which also significantly negatively predicted task value and positively predicted shame and anxiety; a pattern that is generally, negatively associated with learning (Pekrun & Perry, 2014). These findings were consistent with our hypotheses and prior research on the tendency of suppression strategies to lead to decreased positive emotionality and increased physiological activation (Gross 1998a; Gross, 2015). Though we did not find a negative relationship between pride and SCL, pride was negatively correlated with both SCL and negative emotions.

The finding that SCRs—meaningful and occasional, rapid fluctuations in one's physiological activation—was a significant predictor of medical students' diagnostic efficiency indicates that students who experienced higher levels of arousal also performed better on the diagnostic task. These findings are consistent with the Yerkes-Dodson law, where physiological arousal can be beneficial for performance, but only to a certain point, after which arousal can be harmful for performance (Cohen, 2011; Yerks & Dodson, 1908). This critical point varies by type of task and therefore, it is possible that the higher level of arousal attained by students in this study was within a range that was favorable for performance. It is also possible that higher levels of arousal were indicative of cognitive engagement and a lack of boredom, which is an emotional state that is consistently associated with poor achievement (Goetz et al., 2014). The finding that SCL was not associated with learning, on the other hand, as well as the opposing relationships that SCR and SCL had with suppression provides some preliminary evidence that these two components of the EDA complex may have different relationships with individual differences and achievement, despite evidence of them exhibiting parallel global tendencies (see RQ2 results).

Future research should more directly explore SCR and SCL patterns using a single bracelet to measure both components of the EDA complex from the same participants, if such a bracelet becomes available. Future research and analyses should also examine physiological arousal at multiple points in time to advance scientific understanding of the contexts and conditions under which these expression components are and are not likely to align (e.g., intensity, different discrete emotions, latency effects; Harley, 2015; Harley et al., 2015; Mauss et al., 2005). Relatedly, the correlational results between self-reported emotions and SCLs were comparable to prior studies investigating the extent of the coupling between these emotional expression components (Evers et al., 2014; Harley et al., 2015; Mauss et al., 2005).

Another limitation of this study was that the self-report measure of emotion was only completed by participants who wore the Q-Sensor bracelet on account of a study that did not have the Biopac bracelets available and did not have participants fill out the questionnaire due to time constraints (corrected by removing the AEQ and other questionnaires). This prevented us from being able to examine the relationship between SCR and discrete achievement emotions, which we were able to examine for SCL. Moreover, some of the emotion subscales (enjoyment, hope, and anger) had low inter-item reliabilities, but too few items to improve them. Our examination of achievement emotions was therefore limited to four: pride, anxiety, shame, and hopelessness which was a limitation of this study. Fortunately, our preliminary results revealed that the correlational patterns observed between self-reported emotions and self-reported emotions and appraisals conformed with prior research and theory relating to achievement emotions. Another limitation related to the self-report measure used was that the test during-achievement situation subscale does not assess de-activating or low activation/arousal emotions such as boredom or relief. While boredom and relief may seldom occur in typical test-taking situations where grades are on the line, these emotions may be appropriate to assess in contexts such as this study's. Finally, our use of the AEQ to examine students' retrospective achievement emotions did not provide us with an understanding of the object foci (e.g., specific component of the task) that stimulated medical students' emotions. Indeed, asking students "how [they] felt during the diagnostic reasoning task (solving the case within BioWorld)" clarifies the timeframe and achievement situation they are recalling and reporting their emotions during, but not what component(s) of the task generated which emotions. Previous research by Naismith and Lajoie (2018) on medical students' emotions with BioWorld has focused on task-specific elements, including emotions directed toward success or failure feedback from BioWorld. Future research should include questions directed toward a greater variety of specific aspects of complex learning situations, as emotions can differ by object foci as well as over time (Harley, Lajoie, Tressel, & Jarrell, in press; Harley, Bouchet, & Azevedo, 2013; Harley, Poitras, Jarrell, Duffy, & Lajoie, 2016b; Harley et al., 2015, 2016a). In BioWorld more specific emotion questions might measure how the case material (i.e., topic emotions) and aspects of the computer-based learning environment (see Figure 1 and 2) made them feel, for example. Single-item concurrent state emotion questionnaires could also be used to assess how emotions change over time and are elicited in response to specific interactions with BioWorld.

This research was also limited by the in-session as well as general availability of medical students due to their demanding schedule. However, the study does have the strength of measuring learning in the context of an authentic medical task rather than an unrelated laboratory task. This limitation is common in studies that draw their samples from expert and highly-specialized populations, especially when participants are required to spend several hours completing an experimental session with advanced equipment that often fails to collect data properly and leads to data loss for a substantial number of participants (as was the case with this study). Accordingly, although the sample size of this experiment is small by conventional psychological and educational study standards, it is not atypical for this highly-specialized, in-demand population or time-intensive methodology (Duffy et al., 2014; Jarrell et al., 2016, 2017; Lajoie et al., 2015; Taub et al., 2017). The issue of sample size is typically further challenged by data collection failures when multichannel data is collected (Harley, 2015), leading to an increasing number of studies being underpowered by traditional standards (Shute et al., 2015); a limitation that is important to acknowledge in these boundary-advancing studies.

Small sample sizes do limit the power of statistical analyses which can result in failing to reject the null hypothesis when it should have been rejected (i.e., false negatives; Type 2 error). In this study, the smallest number of participants per predictor was 10.5 for multiple regression analyses that examined emotion regulation tendencies (reappraisal and suppression) predictive relationships with self-reported emotion (see part

3 and 4 of research questions 1). The second lowest number of participants per predictor was 14 (for SCR-related analyses; e.g., 2-4) for simple linear regressions (see Table 1). Data screening efforts, such as outlier screening, were conducted to help prevent lone, outlying values from unduly influencing results; small data sets are especially sensitive to outliers. While our analyses lacked statistical power, this does not render the findings invalid. Indeed, the minimum number of subjects per variable in a regression analysis is not absolute. Some studies have cited the number of subjects per variable as small as five (Green, 1991), while more common rules of thumb recommend between ten and twenty (Schmidt, 1971; Harrell, 2015) participants per regression variable (for review see Austin & Steyerberg, 2015). Based on these numbers, our lowest participant per predictor falls within an acceptable range.

In addition to conducting more granular investigations of physiological activation and their relationship with other emotion expression components and other products and processes, another major area of future research is complimenting the collection of students' habitual ER strategies with the ER strategies used during a learning session. At the time data was collected, no widely-validated questionnaire assessing state or retrospective ER strategies was available, making the ERQ the best measure for ER (Gross & John, 2003). Future research should consider designing and administering such surveys about ER to students either during or immediately after they complete their learning session. Alternatively, indicators of ER could be coded from utterances and behaviors, when available (Lajoie et al., 2015).

Future research should also examine physiological arousal in higher-stakes educational contexts such as test-taking or writing term papers where SCRs are likely to be more frequent and SCLs higher and richer in variance. Relatedly, a longer, multi-item assessment of control and value should be applied in order to differentiate appraisal elements, such as intrinsic and extrinsic value, which may provide support for interpretations of appraisals. Another area for future research is the examination of the relationships between a broader array of ER strategies (e.g., attentional deployment; Gross, 2015) and emotions, achievement, and self-regulated learning strategies.

All of the aforementioned lines of research would tremendously benefit from exploring the rich, temporal data that the EDA complex has to offer. Indeed, the fact that physiological data records multiple data points per second provides meaningful opportunities to extend emotion theory, especially in educational research, by observing rapid and dynamic transitions in emotional states (a widely-agreed upon characteristic of emotions). While some studies have examined emotions at multiple time periods with experience sampling approaches (Goetz et al., 2014; Nett, Goetz, & Hall, 2011) or transitions in emotional states with state transition analyses, these studies typically sample learners' emotions every few minutes (at a minimum) through self-report or behavioral observations (Baker, Rodrigo, & Xolocotzin, 2010; Harley et al., 2013; McQuiggan, Robison, Lester, 2010). In doing so, they may fail to capture transitions from one emotion to the next—a phenomena which is more likely to occur at the *second* rather than minute level. While this future direction is promising it is perhaps the most challenging and requires more granular multi-channel analytical procedures to be developed for open-ended educational research studies that differ significantly from most experimental studies and stimulus-response methodologies.

In conclusion, the findings reported in this study make important contributions, particularly, to understanding habitual ER strategies and physiological arousal in computer-based learning environments which can be used to help inform methodologies as well as design recommendations for similar systems. First, this study provides additional evidence that learners tend to experience low levels of physiological arousal while interacting with computer-based learning environments. Our results varied by EDA component, however, regarding whether these low levels of physiological arousal were beneficial for emotional experiences and performance. SCL negatively predicted enjoyment and task-value, and positively predicted anxiety and shame, while SCR was associated with higher diagnostic efficiency scores. Taken together, these preliminary findings highlight the importance of future comparative examinations of SCL and SCR as well as more granular and temporal examinations of this rich data channel that may help reveal crucial contextual information about when the conditions of emotional activation are beneficial for learning and when interventions might be appropriate. The results of this study also contribute to the paucity of research on ER in academic achievement settings and even greater paucity of research examining ER and physiological arousal in academic achievement settings. The findings therefore hold value for theories by supporting propositions that are largely based on self-report data, despite self-reports only constituting one of three major emotion expression components (Gross, 2015; Harley, 2015; Pekrun & Perry, 2014).

This study's findings can also be used to enhanced technology-rich learning environments such as instructional scaffolds or motivating messages. Our findings suggest, in-line with the broader literature, that suppression strategies should not be intelligent tutoring system designers' first choice, for example, in constructing system-delivered prompts. Learner-adaptive features of these environments could be synchronized with unobtrusive physiological bracelets to provide emotional contexts to events to help intelligent systems detect learners' emotions in real time as well as make decisions about when and how to intervene when levels or spikes of arousal that have a high probability of representing negative, activating emotions or deleterious cycles of negative affect are embodied (see D'Mello & Graesser, 2015; Harley et al., 2017, for a review of emotion-supportive features and strategies). Our results reveal that learners may be more or less vulnerable to experiencing certain achievement emotions if their ER tendencies lean toward reappraisal versus suppression strategies. Therefore, it may be appropriate to provide different types and schedules of system-delivered prompts to learners who tend to take different approaches to regulating their emotions. Our findings also provide preliminary evidence that monitoring SCR and SCL levels hold value, but may yield information that is useful to predicting different user outcomes and therefore needs. Thus, it may be appropriate to design intelligent systems to associate different prompts and rules for administering them based on incoming SCL versus SCR data, though more research is required to better understand these EDA components and their relationships to emotions, ER tendencies, and learning, particularly, as no other studies have yet been done.

Finally, this study highlighted an approach for processing and analyzing both SCR and SCL data that is of great interest to the educational psychology community (Harley, 2015; Harley et al. 2017), but not yet well-understood and therefore under-utilized. The authors hope that this study helps guide as well as motivates other researchers to use physiological arousal in their research: theories of emotion have long considered it as one of the prime components of emotion expression (Ekman, 1992; Gross, 1998; Pekrun et al., 2002; Scherer, 1984), it's time we explored this dimension more carefully with respect to its relationship to learning.

References

Arroyo, I., Burleson, W., Tai, M., Muldner, K., & Woolf, B. P. (2013). Gender differences in the use and benefit of advanced learning tech. for mathematics. *Journal of Educational Psychology, 105*, 957-969.

Artino Jr, A. R., Holmboe, E. S., & Durning, S. J. (2012). Can achievement emotions be used to better understand motivation, learning, and performance in medical education?. *Medical teacher*, *34*(3), 240-244.

Artino Jr, A. R., & Pekrun, R. (2014). Using Control-Value Theory to Understand Achievement Emotions in Medical Education. *Academic Medicine*, *89*(12), 1696.

Austin, P. C., & Steyerberg, E. W. (2015). The number of subjects per variable required in linear regression analyses. *Journal of clinical epidemiology*, *68*(6), 627-636.

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: Springer.

Braithwaite, J. J., Watson, D. G., Jones, R., & Rowe, M. A. (2013). Guide for analyzing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. *Psychophysiology*, *49*, 1017-1034.

Boucsein, W. (2012). *Electrodermal activity*. Springer Science & Business Media.

Butler, E. A., Wilhelm, F. H., & Gross, J. J. (2006). Respiratory sinus arrhythmia, emotion, and emotion regulation during social interaction. Psychophysiology, 43(6), 612-622.

Burić, I., Sorić, I., & Penezić, Z. (2016). Emotion regulation in academic domain: Development and validation of the academic emotion regulation questionnaire (AERQ). *Personality and Individual Differences, 96*, 138-147. doi:10.1016/j.paid.2016.02.074

Burleson, W. (2011). Advancing a multimodal real-time affective sensing research platform. In R. A. Calvo, & S. D’Mello (Eds.), *New perspectives on affect and learning technologies* (pp. 97–112). New York: Springer.

Cacioppo, J. T., Tassinary, L. G., & Berntson, G. (Eds.). (2007). *Handbook of Psychophysiology*. Cambridge University Press.

Calvo, R.A., & D’Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing, 1,*18-37.

Chauncey-Strain, A., & D’Mello, S.K. (2015). Affect regulation during learning: The enhancing effect of cognitive reappraisal. *Applied Cognitive Psychology, 29*, 1-19.

Cohen, R. A. (2011). Yerkes–Dodson Law. In J. S. Kreutzer, J. DeLuca, & B. Caplan (Eds.), *Encyclopedia of Clinical Neuropsychology* (pp. 2737-2738). New York, NY: Springer New York.

Curran-Everett, D. (2000). Multiple comparisons: philosophies and illustrations. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, *279*(1), R1-R8.

Dan-Glauser, E. S., & Gross, J. J. (2013). Emotion regulation and emotion coherence: evidence for strategy-specific effects. *Emotion, 13*, 832.

Daniels, L. M., Haynes, T. L., Stupnisky, R. H., Perry, R. P., Newall, N. E., & Pekrun, R. (2008). Individual differences in achievement goals: A longitudinal study of cognitive, emotional, and achievement outcomes. *Contemporary Educational Psychology*, *33*(4), 584-608.

Dawson, M.E., et al (2001) The Electrodermal System. In J. T. Cacioppo, L. G. Tassinary, and G.B. Bernston, (Eds) Handbook of Psychophysiology (2nd Ed), 200–223. Cambridge Press, Cambridge.

Decuir-Gunby, J. T., Aultman, L. P., & Schutz, P. A. (2009). Investigating Transactions Among Motives, Emotional Regulation Related to Testing, and Test Emotions. *The Journal of Experimental Education, 77*(4), 409-438. doi:10.3200/jexe.77.4.409-438

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

D’Mello, S.K., & Graesser, A.C. (2015). Feeling, thinking, and computing with affect-aware learning technologies. In Calvo, R.A., D’Mello, S.K., Gratch, J., Kappas, A. (Eds.) *Handbook of Affective Computing* (pp. 419-434). United Kingdom: Oxford University Press.

D'Mello, S. K., & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)*, *47*(3), 43.

D’Mello, S., Lehman, B., Sullins, J., Daigle, R., Combs, R., Vogt, K., Graesser, A (2010). A time for emoting: When affect-sensitivity is and isn’t effective at promoting deep learning. In V. Aleven, J. Kay & J. Mostow (Eds.), *Lecture Notes in Computer Science: Vol. 6094. Intelligent Tutoring Systems* (pp. 245-254). Berlin, Heidelberg: Springer-Verlag.

Duckworth, A. L., White, R. E., Matteucci, A. J., Shearer, A., & Gross, J. J. (2016). A stitch in time: Strategic self-control in high school and college students. *Journal of educational psychology*, *108*(3), 329.

Duffy, M. C., & Azevedo, R. (2015). Motivation matters: Interactions between achievement goals and agent scaffolding for self-regulated learning within an intelligent tutoring system. *Computers in Human Behavior*, *52*, 338-348.

Duffy, M. C., Lajoie, S. P., Pekrun, R., & Lachapelle, K. (in press). Emotions in medical education: Examining the validity of the Medical Emotion Scale (MES) across authentic medical learning environments. Learning and Instruction. DOI: 10.1016/j.learninstruc.2018.07.001

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, *6*(3), 169-200.

Evers, C., Hopp, H., Gross, J.J., Fischer, A., Manstead, A., & Mauss, I. Emotion response coherence: A dual-process perspective. *Biological Psychology, 98*, 43-49. (2014).

Goetz, T., Frenzel, A. C., Hall, N. C., Nett, U. E., Pekrun, R., & Lipnevich, A. A. (2014). Types of boredom: An experience sampling approach. *Motivation and Emotion*, *38*(3), 401-419.

Goetz, T., & Hall, N. C. (2013). Emotion and achievement in the classroom. In J. Hattie and E. M. Anderman (Eds.), *International guide to student achievement* (pp. 192-195). New York: Routledge.

Green, S. B. (1991). How many subjects does it take to do a regression analysis. *Multivariate Behavioral Research*, *26*(3), 499-510.

Gross, J. J. (1998a). Antecedent-and response-focused emotion regulation: divergent consequences for experience, expression, and physiology. *Journal of personality and social psychology*, *74*(1), 224.

Gross, J. J. (1998b). The emerging field of emotion regulation: An integrative review. *Review of general psychology*, *2*(3), 271.

Gross, J.J. (2002). Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology, 39*, 281-291.

Gross, J. J., & Levenson, R. W. (1993). Emotional suppression: Physiology, self-report, and expressive behavior. *Journal of Personality and Social Psychology, 64,* 970-986.

Gross, J.J., & John, O.P. (2003). Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology, 85,* 348-362.

Gross, J. J. (2015). The extended process model of emotion regulation: Elaborations, applications, and future directions. *Psychological Inquiry*, *26*(1), 130-137.

Gross, J. J. (2013). Conceptualizing emotional labor: An emotion regulation perspective. In A. Grandey., J. Diefendorff., & D. Rupp (Eds.), *Emotional Labor in the 21st Century: Diverse Perspectives on Emotion Regulation at Work* (pp. 288-296). New York, NY: Psychology Press/Routledge.

Harley, J. M. (2015). Measuring emotions: A survey of cutting-edge methodologies used in computer-based learning environment research. In S. Tettegah & M. Gartmeier (Eds.). *Emotions, Technology, Design, and Learning* (pp. 89-114). London, UK: Academic Press, Elsevier.

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 Artificial Intelligence: Vol. 7926. Artificial Intelligence in Education* (pp. 61-70). Berlin, Heidelberg: Springer-Verlag.

Harley, J. M., Bouchet, F., Hussain, S., Azevedo, R., & Calvo, R. (2015). A multi-componential analysis of emotions during complex learning with an intelligent multi-agent system. *Computers in Human Behavior, 48*, 615-625. DOI: 10.1016/j.chb.2015.02.013.

Harley, J. M., Carter, C.K., Papaionnou, N., Bouchet, F., Azevedo, R., Landis, R. L., & Karabachian, L. (2016a). Examining the predictive relationship between personality and emotion traits and students' agent-directed emotions: Towards emotionallyadaptive agent-based learning environments. *User Modeling and User-Adapted Interaction*, *26*, 177-219. DOI: 10.1007/s11257-016-9169-7.

Harley, J.M., Poitras, E. G., Jarrell, A., Duffy, M. C., & Lajoie, S. P. (2016b). Comparing virtual and location-based augmented reality mobile learning: Emotions and learning outcomes. *Educational Technology Research and Development*, *64*(3), 359-388. DOI: 10.1007/s11423-015-9420-7.

Harley, J.M., Lajoie, S. P., Frasson, C., Hall, N.C., & (2017). Developing emotion-aware, advanced learning technologies: A taxonomy of approaches and features. *International Journal of Artificial Intelligence in Education*, *27*(2), 268-297. DOI: 10.1007/s40593-016-0126-8

Harley, J.M., Lajoie, S.P., Tressel, T., & Jarrell, A. (in press). Fostering positive emotions and history learning with location-based augmented reality and tour-guide prompts. *Learning & Instruction*. DOI: 10.1016/j.learninstruc.2018.09.001

Harrell Jr, F. E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

Hussain, S.M., D'Mello, S.K., & Calvo, R.A. (2014). Research and development tools in affective computing. In R.A. Calvo, S.K. D'Mello, J. Gratch, & A. Kappas (Eds.). *The Oxford Handbook of Affective Computing*, (pp. 349-359). Oxford University Press.

Jamieson, J. P., Mendes, W. B., Blackstock, E., & Schmader, T. (2010). Turning the knots in your stomach into bows: Reappraising arousal improves performance on the GRE. *Journal of Experimental Social Psychology*, *46*(1), 208-212.

Jarrell, A., Harley, J.M., & Lajoie, S.P. (2016). The link between achievement emotions, appraisals and task performance: Pedagogical considerations for emotions in CBLEs. *Journal of Computers in Education, 3(3), 289-307*. DOI: *10.1007/s40692-016-0064-3*.

Jarrell, A., Harley, J.M., Lajoie, S.P., & Naismith, L. (2017). Success, failure and emotions: Examining the relationship between performance feedback and emotions in diagnostic reasoning. *Educational Technology Research and Development, 65(5)*, 1263–1284. DOI: 10.1007/s11423-017-9521-6

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

Kreibig, S.D., Samson, A.C., & Gross, J.J. (2015). The psychophysiology of mixed states: Internal and external replicability analysis of a direct replication study. *Psychophysiology, 52*, 873-886.

Lajoie, S. (2009). Developing professional expertise with a cognitive apprenticeship model: Examples from avionics and medicine. In K. A. Ericsson (Ed.), *Development of professional expertise: Toward measurement of expert performance and design of optimal learning environments* (pp. 61–83). Cambridge, UK: Cambridge University Press.

Lajoie, S. P., Lee, L., Poitras, E., Bassiri, M., Kazemitabar, M., Cruz-Panesso, I., ... & Lu, J. (2015). The role of regulation in medical student learning in small groups: Regulating oneself and others' learning and emotions. *Computers in Human Behavior*, *52*, 601-616.

Leroy, V., Gregoire, J., Magen, E., Gross, J.J. & Mikolajczak, M. (2012). Resisting the sirens of temptation while studying: Using reappraisal to increase enthusiasm and performance. *Learning and Individual Differences, 22,* 263-268.

Li, Z., Snieder, H., Su, S., Ding, X., Thayer, J. F., Treiber, F. A., & Wang, X. (2009). A longitudinal study in youth of heart rate variability at rest and in response to stress. *International Journal of Psychophysiology*, *73*(3), 212-217.

Mauss, I. B., Cook, C. L., Cheng, J. Y., & Gross, J. J. (2007). Individual differences in cognitive reappraisal: Experiential and physiological responses to an anger provocation. *International Journal of Psychophysiology*, *66*(2), 116-124.

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*(2), 209-237.

Matsunaga, M. (2007). Familywise error in multiple comparisons: Disentangling a knot through a critique of O'Keefe's arguments against Alpha Adjustment. *Communication Methods and Measures*, *1*(4), 243-265.

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

Meinhardt, J., & Pekrun, R. (2003). Attentional resource allocation to emotional events: An ERP study. *Cognition and Emotion, 17,* 477–500.

Miller, L. H., & Shmavonian, B. M. (1965). Replicability of two GSR indices as a function of stress and cognitive activity. *Journal of personality and social psychology*, *2*(5), 753.

O'Keefe, D. J. (2003). Colloquy: Should familywise alpha be adjusted?. *Human Communication Research*, *29*(3), 431-447.

Nagai, Y., Critchley, H.D., Featherstone, E., Trimble, M.R., & Dolan, R.J., (2004). Activity in ventromedial prefrontal cortex covaries with sympathetic skin conductance level: a physiological account of a ‘‘default mode’’ of brain function, NeuroImage, 22, 243-251.

Naismith, L.M. (2013). Examining motivational and emotional influences on medical students' attention to feedback in a TRE for learning clinical reasoning. McGill University (2013).

Naismith, L. M., & Lajoie, S. P. (2018). Motivation and emotion predict medical students' attention to computer-based feedback. *Advances in Health Sciences Education, 23,* 465-485. DOI: 10.1007/s10459-017-9806-x

Pekrun, R. (1992). The impact of emotions on learning and achievement: Towards a theory of cognitive/motivational mediators. *Applied Psychology, 41,* 359–376.

Pekrun, R. (2006). The control-value theory of achievement emotions. *Educational Psychology Review*, *18*(4), 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., Elliot, A. J., & Maier, M. A. (2009). Achievement goals and achievement emotions: Testing a model of their joint relations with academic performance. *Journal of educational Psychology, 101*(1), 115.

Pekrun, Lichtenfeld, Marsh, Murayama, & Goetz (2017). Achievement emotions and academic performance: A longitudinal model of reciprocal effects. *Child Development.*

Pekrun, R., & Perry, R. 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.

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 Psychology, 36,* 34-48.

Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of quantitative and qualitative research. *Educational Psychologist, 37,* 91-106.

Pekrun, R., Hall, N. C., Goetz, T., & Perry, R. (2014). Boredom and academic achievement: Testing a model of reciprocal causation. *Journal of Educational Psychology, 106,* 696-710.

Pekrun, R. & Linnenbrink-Garcia, L. (2014). *International handbook of emotions in education.* New York, NY: Routledge.

Picard, R. W., Fedor, S., & Ayzenberg, Y. (2016). Multiple arousal theory and daily-life electrodermal activity asymmetry. *Emotion Review, 8,* 62-75.

Porayska-Pomsta, K., Mavrikis, M., D'Mello, S., Conati, C., Baker, R. S. J.d. (2013). Knowledge elicitation methods for affect modeling in education. *International Journal of Artificial Intelligence in Education, 22,* 107-140.

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

Robison, J., McGuiggan, S. W., & Lester, J. (2009). Evaluating the consequences of affective feedback in intelligent tutoring systems. In J. Cohn, A. Nijholt, & M. Pantic (Eds.). *Proceedings of the International Conference on Affective Computing & Intelligent Interaction* (pp. 37-42). Amsterdam, The Netherlands: IEEE Press.

Rubin, M. (2017). Do p values lose their meaning in exploratory analyses? It depends how you define the familywise error rate. *Review of General Psychology*, *21*(3), 269.

Sabourin, J. L., & Lester, J. C. (2014). Affect and engagement in game-based learning environments. *IEEE Transactions on Affective Computing, 5,* 45-55.

Scherer, K.R. (1984). On the nature and function of emotion: A component process approach. In: Scherer, K.R. and Ekman, P. (Eds.), Approaches to emotion, pp. 293-317. Erlbaum, Hillsdale, NJ.

Schmidt, F. L. (1971). The relative efficiency of regression and simple unit predictor weights in applied differential psychology. *Educational and Psychological Measurement, 31*(3), 699-714.

Scrimin, S., Altoè, G., Moscardino, U., Pastore, M., & Mason, L. (2016). Individual differences in emotional reactivity and academic achievement: A psychophysiological study. *Mind, Brain, and Education, 10*(1), 34-46.

Shute, V. J., D'Mello, S., Baker, R., Cho, K., Bosch, N., Ocumpaugh, J., ... & Almeda, V. (2015). Modeling how incoming knowledge, persistence, affective states, and in-game progress influence student learning from an educational game. *Computers & Education, 86*, 224-235

Spangler, G., Pekrun, R., Kramer, K., & Hofmann, H. (2002). Students' emotions, physiological reactions, and coping in academic exams. *Anxiety, Stress & Coping, 15*(4), 413-432.

Steinfatt, T. M. (1979). The alpha percentage and experimentwise error rates in communication research. *Human Communication Research*, *5*(4), 366-374.

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

Taub, M., Mudrick, N. V., Azevedo, R., Millar, G. C., Rowe, J., & Lester, J. (2017). Using multi-channel data with multi-level modeling to assess in-game performance during gameplay with Crystal Island. *Computers in Human Behavior*.

Turner, J.E., & Schallert, D.L. (2001). Expectency-value relationships of shame reactions and shame resilience. *Journal of Educational psychology, 93*, 320-329.

Webb, T. L., Miles, E., & Sheeran, P. (2012). Dealing with feeling: a meta-analysis of the effectiveness of strategies derived from the process model of emotion regulation. *Psychol Bull, 138*(4), 775-808.

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.

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of comparative neurology*, *18*(5), 459-482.