The Impact of Contextualized Emotions on Self-Regulated Learning and Scientific Reasoning during Learning with a Game-Based Learning Environment

Michelle Taub[1*]*;* **Robert Sawyer**[2], **James Lester**[2], **Roger Azevedo**[1]

[1]*University of Central Florida, Department of Learning Sciences and Educational Research, 12494 University Blvd., Orlando, FL 32816*

[2]*North Carolina State University, Department of Computer Science, 890 Oval Drive, Raleigh, NC 27606*

{michelle.taub, roger.azevedo}@ucf.edu

{rssawyer, lester}@ncsu.edu

Abstract. The goal of this study was to examine college students' (n = 61) contextualized emotions during in-game actions while playing CRYSTAL ISLAND, a game-based learning environment where students are tasked with solving the mystery of what illness impacted all island inhabitants. We examined emotions during in-game actions: during book reading, after scanning food items for the transmission source, and after submitting a final diagnosis. We dichotomized each activity's feedback into a positive or negative outcome: a relevant or irrelevant book for solving the mystery, testing food items that generate a positive or negative result, or submitting a correct or incorrect final diagnosis. Results revealed that expressing joy while reading a relevant book and expressing confusion after a positive scan significantly positively predicted overall game score, which we used as a proxy for problem-solving performance. Implications include understanding different levels of emotions students express during learning with all advanced learning technologies.

Keywords. Context; Facial expressions of emotions; Game-based learning; Scientific reasoning; Self-regulated learning

INTRODUCTION

Game-based learning environments (GBLEs) have been shown to foster effective learning of complex topics (Mayer, 2014, Plass, Homer, & Kinzer, 2015). Many different types of games have been developed with the aim of fostering different learning processes. For example, games that foster self-regulated learning (SRL) teach students to monitor and control their cognitive, affective, metacognitive, and motivational processes during learning (Taub, Mudrick, Bradbury, & Azevedo, in press). Because students often have a difficult time self-regulating their learning as they learn challenging material, game-based learning provides these students the opportunity to learn and practice their self-regulatory skills while learning with environments that aim to sustain high levels of motivation and engagement (Mayer, 2014).

GBLE designs can incorporate components that foster affective, behavioral, cognitive, and social/cultural engagement (Plass et al., 2015) to ensure these games are not only effective for learning but also maintain high levels of positive affect during learning (Mayer, 2014). Plass et al. (2015) provide a framework for game design where affective (emotions, attitudes), motivational (self-efficacy, interest),

cognitive (scaffolding, feedback), and social/cultural (social context, social agency) foundations inform decisions for particular design elements (narrative, assessment), which should then influence learners' affective, behavioral, cognitive, and social/cultural engagement. Theories of game-based learning, therefore, focus on the structure of GBLEs, where research can inform the design of game elements.

When examining learners' emotions during gameplay, we need to examine their contextual nature, as research examining emotions during learning with GBLEs focus on the general impact of emotions. Studies investigate the role of specific emotions spanning entire learning sessions. However, emotions are likely to fluctuate during learning based on appraisals related to specific learning processes. These processes include metacognitive monitoring, cognitive strategy use, knowledge acquisition, scientific reasoning, hypothesis generation, and examining evidence. As such, this study examines the specific contextual nature of emotions during complex learning and scientific reasoning with a GBLE.

CRYSTAL ISLAND: a GBLE that fosters SRL and scientific reasoning during learning

CRYSTAL ISLAND is a narrative-centered GBLE that centers on a tropical island where a mysterious illness has developed and impacted the island inhabitants. Participants play the role of an agent who has been brought to the island to solve the mystery of what illness has spread throughout the camp (Rowe, Shores, Mott, & Lester, 2011). To solve the mystery, students have to gather clues by navigating to the five different locations (see Figure 1, left) and talking to non-player characters and engaging in in-game activities. Students could read books (see Figure 1, top right) about different illnesses and respond to factual questions about the content. They could also pick up different food items dispersed in the camp and use the scanner (see Figure 1, middle right) to test them for different pathogenic substances (viruses, bacteria, carcinogens, or mutagens). Students could monitor their steps by making entries into their diagnosis worksheet (see Figure 1, bottom right) where they could note their test results, mark down patient symptoms, deduce the likelihood of different, and form a final diagnosis, where they have to indicate the illness type, transmission source, and a treatment plan. Once students have submitted a correct final diagnosis to the camp nurse, they will have solved the mystery and completed the game.

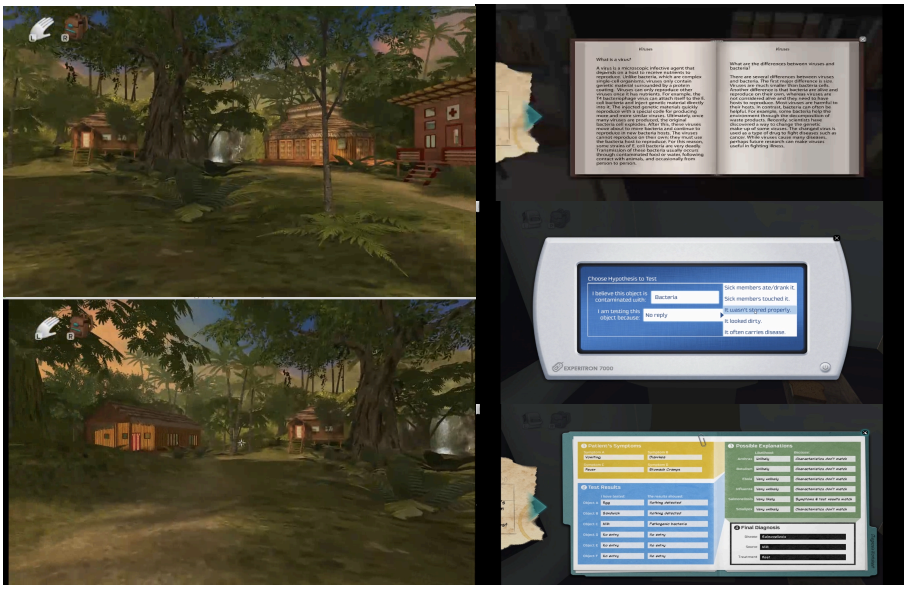

Fig. 1. Images of CRYSTAL ISLAND'S camp (left) and book (top right), scanner (middle right), diagnosis worksheet (bottom right).

This game fosters self-regulated learning because students are required to control which activities to engage in and monitor their gameplay by keeping track of the evidence they've gathered. This involves students reading books that are relevant to solving the mystery (e,g., read books with content about illnesses

that match the symptoms reported by the sick patients, scan and test food items that sick patients reported eating). Studies have investigated how students engage in SRL during learning with CRYSTAL ISLAND (Sabourin et al., 2013; Taub et al., 2017, 2018), and have demonstrated how middle school students make reflective statements demonstrating goal setting and monitoring, and categorizing them into low, medium, and high SRL groups based on these statements (Sabourin et al., 2013). Other studies have investigated the use of SRL processes by demonstrating eye-tracking behaviors indicative of monitoring text relevancy when reading books and completing the associated assessment (Taub et al., 2017), and how monitoring scanning behaviors results in more efficient gameplay (Taub et al., 2018). Thus, these studies demonstrate how these in-game actions are impacted by students' use of SRL processes. Additionally, the game fosters scientific reasoning because students are required to form hypotheses about potential illnesses based on reading the books and talking to the sick patients, and then scanning the food item and testing it for what they hypothesized might be the pathogenic substance spreading the illness. Studies have demonstrated this behavior by investigating sequences of scanning behaviors, and how students who are more efficient at solving the mystery demonstrated strategic hypothesis testing (Taub & Azevedo, 2018; Taub et al., 2018). Another study demonstrated the relationship between information gathering, hypothesis generation, and problem-solving behaviors (Sabourin et al., 2012). Thus, the goal of the game is not only for students to learn about microbiology, but also to teach them how to engage in effective self-regulatory and scientific reasoning skills, which has been empirically demonstrated by prior research.

In this study, we investigated students' learner-centered emotions (confusion, frustration, joy) during learning with CRYSTAL ISLAND. We wanted to determine which emotions students experienced during learning and gameplay differed by context. We were interested in contextualizing emotions surrounding in-game actions that can have positive or negative feedback to determine differences in emotions. These actions include during reading a book that is relevant or irrelevant to solving the mystery, after testing food items that generate a positive or negative result, or after submitting a final diagnosis that is correct or incorrect. After examining these differences, we investigated whether these contextual emotions were predictive of participants' overall game score, which we used a proxy for problem-solving performance.

We generated the following research questions: (RQ1) Which emotions do students experience differently based on the context (i.e., positive vs. negative valences) of in-game action outcomes? (RQ 2): How are differences in contextual emotion outcomes (i.e., the difference between positive vs. negative outcomes) related to overall game score? (RQ 3): What features of contextualized emotions and in-game behaviors are included in a more predictive model of overall game score?

Theoretical frameworks

Our work is grounded in theories of self-regulated learning, scientific reasoning, and affect, as theories of game-based learning focus on game design, where the theoretical component related to learning will depend on each particular game and which processes it fosters (Plass et al., 2015). As such, for our theoretical frameworks, we relied on theories that can explain the processes fostered during learning about microbiology in the GBLE, CRYSTAL ISLAND. These processes include self-regulated learning, emotions, and scientific reasoning.

The information processing theory of self-regulated learning (Winne, 2018; Winne & Hadwin, 1998) states that there are four phases of self-regulated learning (understanding the task, setting goals and making plans to achieve them, engaging in learning strategies, and making adaptations), where students engage in monitoring and control processes to complete a task. Although there are distinct phases, they are not independent of each other. For example, if a student determines testing every food item is not an effective strategy, they can make an adaptation to this plan and decide to only test food items reported being eaten by the sick patients. The model posits that during these phases there are different mechanisms for engaging in cognitive processes (COPES: conditions, operations, products, evaluations, and standards), and there are different types of cognitive processes (SMART: searching, monitoring, assembling, rehearsing, and translating) a student can engage in during self-regulated learning. This model also views self-regulated learning as events that unfold during learning, which is appropriate for this study because we aimed to

investigate the changing nature of emotions during different self-regulatory activities while solving a mystery in a game-based learning environment.

There are different types of self-regulatory processes, which can be cognitive, affective, metacognitive, or motivational in nature (Azevedo et al., 2018, 2019). Although the information processing theory does view SRL as different events, the theory does not, however, account for the affective nature of self-regulatory processes. As such, we use the model of affective dynamics (D'Mello & Graesser, 2012) because this model focuses exclusively on learner-centered emotions, which are emotions typically expressed in learning contexts (D'Mello, 2013). According to this model, confusion results from an impasse encountered during learning with an intelligent tutoring system. This confusion can be resolved by engaging in problem-solving strategies, which will restore the cognitive disequilibrium that stemmed from the confusion. However, if the confusion is not resolved, this can lead to a state of frustration, which can then lead to boredom (i.e., complete disengagement). Although this model does focus on affect dynamics, we note that we use this model because it focuses on specific learner-centered emotions, and do not focus on affect transitions for this study. In addition, this model focuses on four emotional states: engagement, confusion, frustration, and boredom/disengagement. For our study, we were unable to detect engagement and boredom (due to restrictions with our facial detection software), and so we used joy as our proxy measure of a positive emotional state (instead of engagement) with confusion and frustration as the other emotions (excluding boredom).

The model of scientific discovery as dual search (Klahr & Dunbar, 1988) is a model of scientific reasoning that can be applied to any task involving scientific reasoning (i.e., not contextually specific) where the task involves forming hypotheses and collecting data. The fundamental assumption is that scientific reasoning involves search in both the hypothesis space and the experimental space. The hypothesis space involves engaging in discovery, leading to hypothesis generation. The experimental space involves conducting experiments to test the hypotheses, where results can then be used to formula more hypotheses. The model proposes a series of three main components that take place during scientific reasoning. First, students engage in search where the goal is to gather enough information to from a hypothesis. Next, students experiment, where the goal is to test their hypothesis and collect evidence they will use to accept or reject their hypothesis. Last, students make the decision to accept or reject the hypothesis where they make an evaluation that has them reach a decision. Within these three components are multiple sub-components demonstrating there are multiple ways to engage in these processes (see Klahr & Dunbar, 1988) to engage in successful scientific reasoning.

All three of these constructs have been tested and empirically demonstrated during learning with CRYSTAL ISLAND, emphasizing the importance of studying these behaviors, and specifically how SRL and scientific reasoning differ based on different emotional outcomes. It is also important to note that we are not directly testing these theories, rather we used them to select the in-game actions we assessed in this study (book reading, scanning food items, making a final diagnosis) and to help interpret our findings (see discussion section).

Literature review

Over the past decade many researchers have conducted meta-analyses to determine when games are the most effective for learning. Two recent meta-analyses conducted by Mayer (2014) and Clark, Tanner-Smith, and Killingsworth (2016) have found that in general, games can lead to increased learning outcomes, however this depends on certain factors, such as what the comparison is, age, and game type. For example, Clark et al. found that games are more effective than not playing games (g = .33), but they did not compare games to other advanced learning technologies (e.g., intelligent tutoring systems). Mayer found games to be more effective than traditional classroom instruction, however this was only for science and second language learning. Clark et al. included participants aged 6 to 26 in their findings, however Mayer included age as a factor and determined that games were the most beneficial for college students (d = .74) compared to middle school (d = .58) and elementary school (d = .34) students. The effectiveness of games also depends on the type of game, where Mayer found adventure games to be more effective (d = .72) than

simulation games (d = .62) or puzzle games (d = .45). Clark et al. found that the enhanced version of the game (g = .34) was better than the standard version, which was even more apparent when the enhanced version included scaffolding (g = .41). Thus, it is apparent that games are beneficial for learning, but more development is needed to ensure that they are effective for more populations and for all subjects.

Given one of the overarching goals of game-based learning is to foster high levels of positive emotions (Plass et al., 2015), studies are investigating the relationship between emotions and other affective processes (e.g., engagement) during game-based learning with different types of games, different populations, and different detectors of emotions. Sabourin & Lester (2014) assessed affect among middle school students while playing CRYSTAL ISLAND (see above) using self-report measures. They asked students to report on their affect every 7 minutes, where they selected how they were currently feeling from a list of 7 affective states. Results found that students reported feeling focused and curious the most, followed by being confused, frustrated, and excited, then bored, and lastly anxious. In addition, results found that confusion and boredom were negatively correlated with learning gain (Sabourin & Lester, 2014).

Another technique to examine affect is the BROMP observation method (Baker-Rodrigo-Ocumpaugh Monitoring Protocol; Ocumpaugh, Baker, & Rodrigo, 2015), which includes making field observations of students' affect. Once the observations are complete, a follow-up method is feeding that data to develop automated models for affect detection. Many studies have used these methods to examine affect during learning with games. In one study conducted by Andres and colleagues (2014), they examined how high school students played Newton's Playground (now called Physics Playground), where they solved Physics problems by drawing simple machines. Students earned a gold or silver badge after solving a problem, where a gold badge meant solving the problem under a predefined set of steps and a silver badge meant solving the problem over the preset number of steps. Results revealed there was a significant negative correlation between confusion and earning a gold badge, but a significant positive correlation between confusion and earning a silver badge, and between confusion and engaging in stacking, a behavior indicative of gaming the system (Andres et al., 2014). In another study conducted by Andres and colleagues (2015), they investigated confusion and boredom during learning with Physics Playground, where they examined middle school students' sequences of behaviors. Their results revealed confusion was positively correlated with actions indicative of engaging in inefficient behaviors (i.e., earning a silver badge), and boredom was positively correlated with action sequences where students did not solve the problem (Andres et al., 2015). Thus, in both studies, confusion is associated with less efficient gameplay behaviors.

In a study conducted by Ocumpaugh and colleagues (2017), they used BROMP to investigate cadets' affect while playing vMedic, a game for combat training. They investigated affect transitions, and found 4 significant transitions during gameplay. These included concentrated engagement to confusion and confusion to concentrated engagement, concentrated engagement to boredom, and boredom to confusion (Ocumpaugh et al., 2017). These studies demonstrate how we can examine affective processes using different methodologies, using different GBLEs, and across different populations. All of these studies demonstrate the prevalence of confusion during gameplay, and how it can have a negative impact on learning.

Although these studies have found evidence for the impact of emotions during game-based learning, there are some methodological constraints that we feel could be addressed using multichannel data. For example, the BROMP method is a validated approach, however it requires vigorous human coding and extensive training, as opposed to using facial detection software that analyzes videos of facial expressions of emotions automatically. Affect labels from the BROMP method require observation by a human coder, meaning these labels are instantaneous measures of student affect at intervals the human coder is observing, rather than continuous measures of student affect throughout gameplay. This does not allow for fine-grained analysis of the context of emotions that a continuous measure provides. Specifically, the emotions investigated in previous works were not contextualized to specific in-game actions, and so we do not know if the emotions detected occurred within a specific aspect of gameplay. This prevents useful insights regarding the emotional response and regulation of students surrounding in-game actions and feedback. Researchers have emphasized the importance of designing learning environments that provide feedback based on student affect (Poryaska-Pomsta et al., 2013) and studies have begun investigating contextualized

emotions during learning with advanced learning technologies such as open-ended learning environments and intelligent tutoring systems (e.g., Munshi et al, 2018; Taub et al., in press/online first 2019), and we aim to expand this research to GBLEs. Thus, the goal of this study was to use facial detection software to examine automatically detected emotions contextualized within in-game activities in CRYSTAL ISLAND.

Current study

In this study, we define context in terms of different in-game outcomes (i.e., a relevant vs. irrelevant book for solving the mystery, a positive vs. negative scan outcome, a correct vs. incorrect final diagnosis). Thus, context is defined in terms of different event outcomes. Alternatively, different contexts can be seen as different instances of the same type of event (i.e., reading multiple relevant books, scanning multiple food items). We therefore did not include the temporal component of context in this study, as our approach for this analysis was to first determine if we could, in fact, detect differences in emotions based on different action outcomes. As we were able to do so, it is important for future work to examine the temporality of context in terms of SRL, emotions, and scientific reasoning. As this line of research is fairly new, we felt our first approach should globally examine whether we could first detect this or not.

In addition, as previously mentioned, we used the model of affective dynamics to select which emotions to examine for our study; however since the model and previous research does not specify the conditions under which confusion can be the most beneficial (e.g., as seen in D'Mello et al., 2014), we used a macro-level approach to form our hypotheses, where positively valenced emotions have a beneficial effect on learning, and negatively valenced emotions have a more detrimental effect on learning (Pekrun et al., 2017), with the assumption that these emotions cannot be resolved.

Based on our macro-level approach, we hypothesized: (H1) Higher levels of joy and lower levels of confusion and frustration for positive, compared to negative in-game activity outcomes. (H2) Students who experience more joy after positive outcomes will be positively correlated with overall game score and students who experience more confusion and frustration after negative outcomes will be negatively correlated with overall game score (H3): The more predictive model of overall game score will include joy after positive outcomes as a positive predictor, confusion and frustration after negative outcomes as negative predictors, and efficient gameplay behaviors as a positive predictor of overall game score.

These analyses are key for investigating the complex nature of SRL because it includes assessing students' emotions stemming from their use of cognitive and metacognitive learning processes. Including game score in addition to emotions highlights the multicomponential nature of SRL and emotions such that all these different factors play a role in impacting learning and performance. For example, research has shown that the more predictive model of in-game assessment performance included both eye-tracking and log-file variables (Taub et al., 2017). The regression model for this study that included data from different channels (e.g., trace data and facial expressions) also highlight the importance of including multimodal multichannel data to understand how an individual engages in learning processes.

Although we are not directly testing our theoretical frameworks, these three research questions and hypotheses shed insight into how students' emotions are affected by their use of self-regulatory and scientific reasoning processes. First, these in-game actions foster the use of self-regulation and scientific reasoning because reading books involves monitoring the information they read and integrating it with other sources of information, such as mapping patient symptoms to the ones read in the books. Matching symptoms can also be used to form hypotheses of the illness type, which is part of the hypothesis space (i.e., hypothesis generation). Scanning food items is part of the experimental space, which involves testing hypotheses. This also requires students to monitor the results they are obtaining. Submitting a final diagnosis requires students to first coordinate all the information they have gathered to form a hypothesis of the final solution, and then testing that hypothesis by making a submission. Adding the valence (i.e., positive vs. negative outcome) to these actions allows us and students to determine if they were effective at using these processes. For example, if students express higher levels of confusion after a negative scan result, this can indicate they did not form a correct hypothesis and need to gather more information to form a new one.

In addition, such results can help inform the design of intelligent educational systems, such as games, that cater to individual learning needs. For example, if we know that higher levels of confusion is a result of forming inaccurate hypotheses, the system can scaffold the student to ensure the hypothesis is correct prior to scanning the food item or submitting the final diagnosis. Furthermore, as the current game does not allow students to explicitly state their hypotheses, perhaps designing a tool that allows them to do so can be advantageous. This will also allow them to monitor their progress in forming and testing hypotheses in the game during learning.

METHODS

Participants and materials

Participants included 61 undergraduate students (68.9% female) from a large North American university. Their ages ranged from 18 to 26 (M_{age} = 20.0, SD = 1.5). They were compensated $10/hour.

Materials used included a pre- and post-test, which consisted of 21-item multiple-choice tests on microbiology that include 12 factual and 9 procedural questions. We also administered self-report questionnaires prior to gameplay (Emotions and Values, Achievement Goals) and following gameplay (Emotions and Values, Perceived Interest, Presence).

Experimental procedure

This study took place over a single session that lasted about 1.5 hours ($M_{duration}$ = 69.4 min, SD = 21.7 min). Students began by signing an informed consent form, followed by getting an overview of the study and completing a demographics questionnaire. They then completed self-report questionnaires and the pre-test. Next, the experimenter calibrated the eye tracker, video (to collect facial expressions), and electrodermal activity bracelet. To calibrate the video, students were required to sit still for 6 seconds to establish a neutral baseline. Once calibrated, students began playing the game until they solved the mystery (see Crystal Island section, above). Once solved, students completed more self-report questionnaires and the post-test.

During gameplay, we collected a series of multichannel data: log files, videos of facial expressions of emotions, eye tracking, and electrodermal activity. For this study, we used log files and facial expressions of emotions only. Log files captured all gameplay behavior, including mouse clicks (e.g., selecting a book, using the scanner, and making changes to and submitting the diagnosis worksheet). Videos of facial expressions of emotions were run through FACET, a facial detection software run through Attention Tool 6.3 (iMotions, 2016). FACET has been empirically tested and shown to be an effective tool for detecting facial expressions (Dente, Küster, Skota, & Krumhuber, 2017). In addition, a recent study used FACET and found that the occurrence of emotions had a significant impact on students' metacognitive judgments (Sawyer, Mudrick, Azevedo, & Lester, 2018).

Data coding and scoring

Overall game score

We used overall game score as a proxy for learning instead of proportional learning gain (which measures gain in post-test score in relation to the pre-test score) because it takes into account all game activities (and their importance to completing the game based on an expert-defined model). For example, students get points for completing the in-game book assessment correct on the first attempt, fewer pointa on rhe second attempt, and they lose points if they have to try a third attempt. The overall game score is a previously expert-defined and validated measure for evaluating in-game performance in CRYSTAL ISLAND (Rowe et al., 2011). The measure was shown to be positively correlated with post-test score measured from a microbiology content post-test and a similar relationship is observed in this work ($r(59) = 0.32$, $p = 0.012$).

Overall game score rewards students for demonstrating positive problem-solving behaviors in an efficient manner and penalizes students for using poor problem-solving strategies such as "guess-and-check." The specific in-game actions that correspond to increases and decreases in overall game score are given in Table 1.

Table 1
Breakdown of points for in-game actions used to calculate overall game score

Action	Points (pts)
Overall Mystery Solution	
Correct Solution	500 pts
Solution Efficiency	(7500 / elapsed min) pts
Incorrect Solution Attempt	-100 pts
In-game Quiz Questions	
First Attempt Correct	25 pts
Second Attempt Correct	10 pts
Second Attempt Incorrect	-10 pts
Object Contaminant Testing	
Correct Object and Correct Contaminant	200 pts
Incorrect Object and Correct Contaminant	15 pts
Correct Object and Incorrect Contaminant	-15 pts
Incorrect Object and Incorrect Contaminant	-35 pts
Character Interactions	
Talk to Kim	(25 / elapsed min) pts
Talk to Teresa	(50 / elapsed min) pts
Talk to Ford	(125 / elapsed min) pts
Talk to Robert	(125 / elapsed min) pts
Talk to Quentin	(125 / elapsed min) pts
Total Maximum Points	**1665 pts**

Student scores can vary widely depending on how they approach solving the mystery, with less negative scans and incorrect worksheet submissions having a large impact on overall game score. This wide range of scores was observed, with scores ranging from -991 to 1502 with an average score of 705. A histogram of the overall game score is given in Figure 2, where a negative game score indicates the student engaged in more incorrect in-game activities (see Table 1). This figure includes a rug plot, where each student's overall game score is plotted as a dash along the x-axis. This figure also includes a probability density function estimate based on Gaussian kernels of each student's game score value given in the rug plot to display the empirical distribution of the game score without assigning the values to discrete bins. The density estimate shows the values are approximately normally distributed with some negative skew due to some very negative scoring students. These students likely exhibited "guess-and-check" type problem-solving strategies throughout gameplay (i.e., performed many negative scans or incorrect worksheet submissions trying to guess the correct answer).

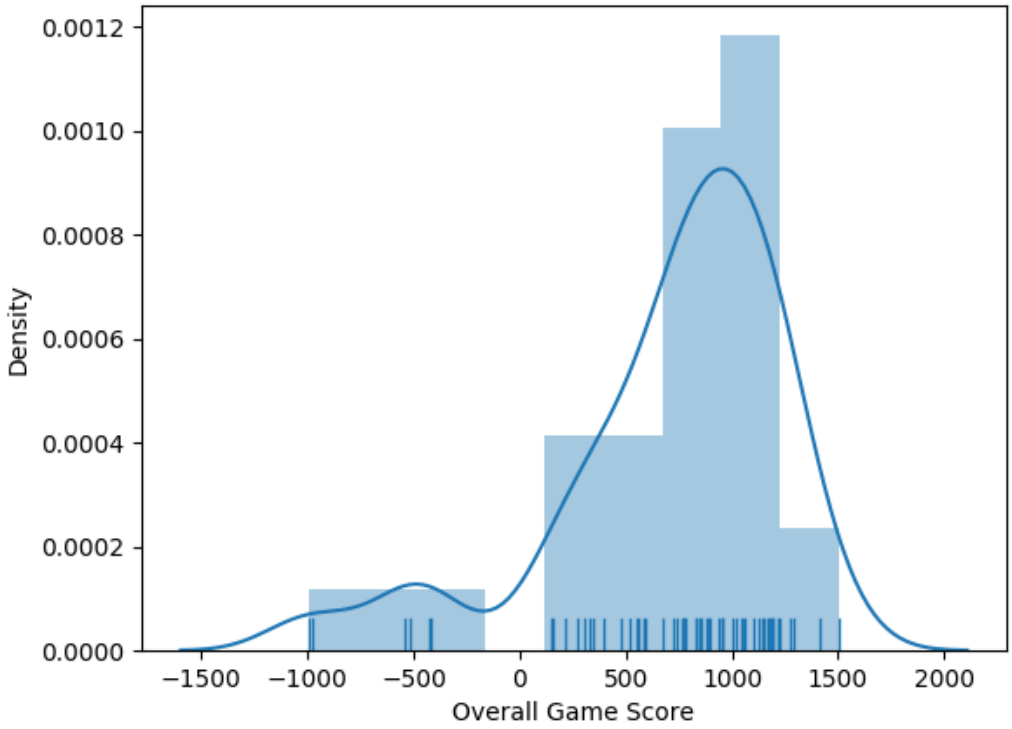

Fig. 2. Distribution of overall game score.

Context: Positive vs. negative activities

In CRYSTAL ISLAND, students perform many problem-solving behaviors in order to successfully solve the mystery. As demonstrated by the overall game score, some of these have been determined by an expert to be desirable within CRYSTAL ISLAND (as shown in Table 1). When students perform these actions, they are often informed of their action's outcome, such as a successful scan revealing the object has tested positive for the hypothesized contagion. Students are typically rewarded by overall game score for positive outcomes of these actions and penalized by negative outcomes of these actions. In this study, we wanted to examine how the positive or negative outcome of an action impacts a student's emotions during or immediately following the action. This can help to identify students who demonstrate effective self-regulation of their emotional and scientific reasoning processes when receiving negative feedback. Using these differences in emotion from positive and negative outcomes can reveal important information regarding the impact of the feedback on students and their overall performance measured by overall game score.

In this study, we focused on three specific actions that provide context for emotions experienced during positive and negative outcomes of these actions. The actions examined in this work are scanning an object, reading a book or article, and submitting a diagnosis worksheet. Scanning involves testing hypotheses (experimental space) and metacognitively monitoring the obtained results. Reading involves metacognitively monitoring one's understanding of the text to complete the assessments, identifying the relevancy of said text in relation to reported symptoms, and then using that information to generate hypotheses to be tested. Submitting the worksheet involves monitoring progress of obtained information and coordinating clues to form a final hypothesis of the solution. Thus, each action involves both self-regulatory and scientific reasoning processes. When students scan an object, they receive immediate

feedback as to whether the object scanned *Positive* for the specified contagion, or *Negative* for the specified contagion. When students read a book or article within the game, the book can be *Relevant* to the solution, containing microbiology content that helps students determine the likely cause of the mysterious outbreak, or *Irrelevant* to the solution, containing scientific content not necessary to solving the current mystery. When students submit their diagnosis worksheet, their submission can either be *Correct*, indicating a student has successfully solved the mystery, or *Incorrect*, indicating the student has answered with either an incorrect contaminated object, incorrect illness, or incorrect treatment plan.

Given that not all students perform actions, which result in both a positive and negative outcome, an analysis of the in-game actions was performed to determine valid students for each type of comparison. For example, if a student correctly submits their worksheet on the first try, they did not experience a negative worksheet submission, and thus a comparison between emotions when that student positively submitted and negatively submitted would not be possible. As another example, if a student was unable to solve the mystery correctly, they would not have a positive outcome for submitting the worksheet. Table 2 shows the overall number of positive and negative outcome actions performed by each student and the number of students that could be used in comparing emotions. Note that there is no variance in *Submission Correct* because students successfully complete the game when they correctly submit their worksheet, and therefore can only occur once.

Table 2
Number of positive and negative outcome actions

Action Outcome	**Students without Action Outcome**	**Valid Students**	**Mean of Valid**	**Std of Valid**
Scan Positive	2	59	1.27	0.52
Scan Negative	0	59	24.5	16.6
Book Relevant	0	61	15.7	5.12
Book Irrelevant	0	61	7.20	3.71
Submission Correct	2	26	1.00	0.00
Submission Incorrect	33	26	2.27	2.01

Note. Valid students indicates the number of students who experienced both a positive and a negative action outcome for that action.

Mean evidence scores: Duration proportions

Facial expression features were extracted automatically through a video-based facial expression tracking system, iMotions (2016).[1] The iMotions system extracts facial features at a frequency of 30 Hz that correspond to the Facial Action Coding System (FACS) (Ekman, Friesen, & Hager, 2002). Example action units would be eyebrow lowerer (action unit 4) or lip tightener (action unit 23). It uses an objective three-phase framework of facial detection from image input, feature detection (i.e., locating position of facial landmarks), and feature classification to report an evidence score representing the likelihood of a particular affective measure being present. For example, to detect action unit 4 (eyebrow lowerer), the software captures each video frame of the entire face, identifies the eyebrows in that frame, and classifies the likelihood that a human coder would identify the eyebrow as being lowered. A separate classifier is used for each affective measure, ranging from low-level facial expression features such as Action Units (AUs)

[1] The iMotions software was previously commercially available as FACET and the research-focused toolbox CERT (Littlewort et al., 2011).

to more complex, composite affective measures, such as Surprise and Joy. We used the facial expression tracking system to monitor students during gameplay to provide real-time evidence scores for 20 AUs and 9 emotion measures. The 9 emotion measures include Anger, Surprise, Frustration, Joy, Confusion, Fear, Disgust, Sadness, and Contempt.

We used a method combining relative thresholding and absolute thresholding of amplitude to calculate the durations students are in an elevated state of the emotion. This was done (1) to allow for comparison across students, and (2) to account for possible noise in the data. For example, if a student coughs, the software may erroneously detect surprise or confusion, but if the emotion has to reach a threshold, the emotion would have to be intended by the student. First. the evidence scores were standardized for each student by subtracting the mean evidence and dividing by the standard deviation evidence over their entire episode. While iMotions calibrates the evidence scores over the first few observations, it does not account for the potential variability of evidence for students. Thus, dividing by the standard deviation of a student's evidence score over the episode accounts for potential variability in expressiveness of individuals. Events were added to the affect log for the duration that the standardized evidence scores rose above a threshold of 1.65. These events represent moments when evidence scores rose 1.65 standard deviations (theoretical top 5% of standardized observations since these have been transformed to a normal distribution with mean 0 and standard deviation 1) above the mean evidence score for a particular student. This general process is shown in Figure 3, where the durations above threshold were used in contextualizing emotions around action outcomes.

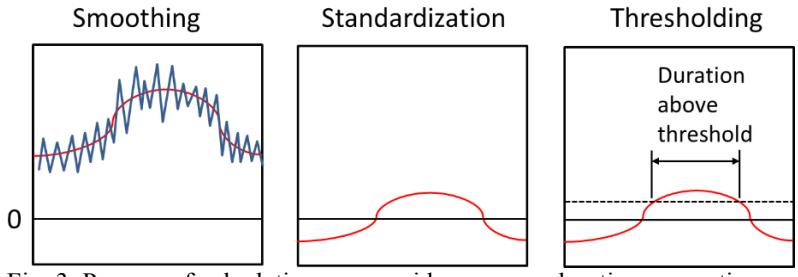

Fig. 3. Process of calculating mean evidence score duration proportions.

Since the scan and worksheet outcomes (positive/negative feedback) are given directly to the student, the emotions experienced in an interval immediately following the actions was conducted. The five seconds after receiving feedback regarding the action outcome were examined, which is a reasonable window to determine a student's emotional response to feedback, as emotions are short in duration and constantly changing when an individual engages in multiple behaviors (Scherer, 2005). More specifically, the proportion of the five-second window in which a student is in an elevated state (above the 1.65 standardized threshold) was calculated for each emotion. These were averaged over each similar outcome over a student's gameplay session to give one value per emotion-action outcome pairing representing the average proportion of the five second window after the action outcome the student is in an elevated state of the emotion. For example, one student had a measured value of 0.15 for *Frustration* experienced *After Scan Negative*, meaning on average of all of the student's scans, which resulted in a negative result, they were in an elevated state of frustration for 15% of the five seconds after the negative feedback was given. These values can range from 0, meaning they never experienced that emotion after the feedback, up to 1, meaning they always experienced the emotion during the five seconds succeeding the feedback.

The relevancy and irrelevancy of a book or article to the solution is never directly given to the student, so instead of taking the interval after reading, we used the full time the student spent reading the book or article as the interval to examine emotions. The final value reported for this contextual action represents the same proportion but is calculated using the proportion of time in an elevated emotional state over the full duration of reading a book (or article). For example, one student had a value of 0.042 for *Confusion* experienced *During Book Relevant*, meaning on average, while reading relevant books, the student was in an elevated state of confusion 4.2% of the time.

RESULTS

RQ1: Which emotions do students experience differently based on the context (i.e., positive vs. negative valences) of in-game action outcomes?

To answer this research question, we calculated the difference between the positive outcomes and negative outcomes of emotions that a student experienced either during or after the action (during book reading, after scanning, and after submitting the diagnosis worksheet). This was done because the first step was to determine if we could detect differences between different activity outcomes (i.e., a positive vs. a negative result). The differences greater than 0 indicate a student experienced a higher duration of that emotion after positive action outcomes than negative action outcomes. Similarly, a difference lower than 0 indicates a student experienced a lower duration of that emotion after positive action outcomes than negative action outcomes. This created three comparisons to investigate: one for each contextual action pairing, in which we compared emotions after positive action outcomes versus negative action outcomes. More specifically, this revealed whether students expressed different durations of joy, confusion, and frustration (i.e., 3 tests, with 1 for each emotion) while reading a relevant vs. irrelevant book, after obtaining a positive vs. negative scan result, or after making a correct vs. incorrect final diagnosis.

In this case, the null hypothesis is that there is no difference between emotions experienced after positive action outcomes against negative action outcomes. Thus, we began with three multivariate tests of paired differences to determine if there were differences in emotions between positive action outcomes and negative action outcomes. A Hotelling T^2 test was performed to test each action type for the multivariate paired differences against 0 since this is a within-subjects measure. To be more explicit, three tests were performed, one for scans, one for books, and one for diagnosis worksheet submissions, where each test was multivariate in the three emotions (joy, confusion, frustration) and compared the paired (by student) differences of the emotions against 0. Thus, a significant test statistic indicates that there is evidence that there was a difference between emotions experienced after positive action outcomes compared to negative action outcomes for that action type. We chose to use the difference score and compare the differences to 0 (as opposed to using each evidence score separately) because we wanted to keep our variables to a minimum due to a smaller sample size. This is especially important for our third research question where we used predictive models. Using the difference score limits the amount of predictors we have, especially since we are investigating three different emotions. Once we made this decision, we wanted to renain consistent and use the same variables for each research question. As discussed in Section: Data Coding and Scoring, the number of differences calculated varies by action type, since some students are invalid as they did not experience positive and negative outcomes of some actions. The results of these three tests are reported in Table 3, with an indication for how many paired differences (valid students) are in the test.

Table 3
Number of students who experienced positive and negative outcomes for each action

Action	**Valid Students**	**T^2(p)**
After Scan	59	9.61 (0.028)*
During Book	61	5.95 (0.13)
After Submission	26	6.40 (0.12)

*p<.05

The test results reported in Table 3 indicate that students experienced significantly different durations of emotions after *Scan Positive* (scanning a food item that tested positive) compared with *Scan Negative* (scanning a food item that tested negative). Since this is a multivariate test, it is not possible to tell which of these emotions is different after the scan outcomes. We thus conducted one sample paired difference *t*-

tests for each emotion using the student differences in emotion after *Scan Positive* and *Scan Negative*. The results from these tests are reported in Table 4, including the mean difference (difference of average emotion proportion after *Scan Positive* and average emotion proportion after *Scan Negative*) and effect size (Cohen's *d*). The results presented in this table indicate a significant difference in *Confusion* experienced after *Scan Positive* against *Scan Negative* where students tended to experience less proportions of confusion after *Scan Positive* compared with after *Scan Negative*.

Table 4
Results from one sample paired difference t-tests for after scan interval.

Emotion	t-stat (p-value)	Mean Difference (Std)	Effect Size (*d*)
Frustration	1.07 (0.29)	0.020 (0.15)	0.13
Confusion	-2.70 (< 0.01)**	-0.0076 (0.022)	-0.35
Joy	1.37 (0.18)	0.024 (0.14)	0.17

*$p < 0.05$, **$p < 0.01$
Note. Mean difference = positive outcome minus negative outcome.

RQ2: How are contextual emotions (i.e., the difference between positive vs. negative outcomes) related to overall game score?

While in RQ1 the emotions were compared to determine if they were significantly different between positive action outcomes and negative action outcomes, this section aimed to determine if the emotion differences have a relationship with overall game score. In this section, the Pearson correlation coefficient between differences in three learner-centered emotions (confusion, joy, and frustration) within the three contextual actions (after-scan, during-book, after-submission) and overall game score were calculated, for a total of 9 tests of correlation. The results of the highest five correlation tests by absolute magnitude are presented in Table 5. It should also be noted that while some *p*-values are below common significant thresholds, the Holm-Bonferroni method for family-wise error rates indicates that no correlations were significant. This could be due to the small sample size, noise contained within the proportions, or non-linear relationships between emotion difference and overall game score.

Table 5
Correlation results between context emotions and overall game score

Context Emotion Difference	Correlation with Overall Game Score	p-value	Sample Size (N)
After Submission Joy	0.305	0.13	26
During Book Joy	0.253	0.049	61
After Scan Frustration	0.193	0.144	59
After Scan Joy	0.173	0.189	59
During Book Confusion	-0.152	0.243	61

RQ3: What features of contextualized emotions and in-game behaviors are included in a more predictive model of overall game score?

While RQ2 analyzed the pairwise relationship between emotion differences after actions and learning outcomes, we were also interested in the additive effect of the emotion differences in predicting these learning outcomes, and what features were included in the best predictive model of overall game score. We

created a linear regression model for overall game score, using the differences in emotion by contextual action as covariates. Theoretically, this would use nine features, the same used in the pairwise correlation tests, but only 26 students had valid differences for emotions after diagnosis worksheet submissions, so we instead replaced the three emotion differences contextualized by submission outcome with the total number of worksheet submissions. This still allowed us to include components of self- and emotion-regulation, as well as scientific reasoning processes in our model, as submitting a diagnosis can be categorized as both a monitoring strategy (monitoring if a student has sufficient information to complete a goal) and scientific reasoning process (testing the hypothesis of the final diagnosis), in addition to emotion regulation. The total worksheet submissions should be a strong predictor of overall game score since students are penalized for incorrect submissions and rewarded for a correct submission, so a comparison against a model using only the total worksheet submissions against one that included the emotion features was conducted to assess the additive effect of the emotion differences with this strong baseline predictor of overall game score.

The full linear model for overall game score did have evidence that at least one coefficient was significantly different from 0; $F(8, 53) = 6.64$, $p < 0.001$. The maximum likelihood estimates of the coefficients from ordinary least squares regression are reported in Table 6. The model was tested under a leave-one-student out cross-validation procedure resulting in a cross validation R^2 of 0.285, indicating a 28.5% reduction (improvement) in mean squared error using predictions on held out data from the full model against predictions using the mean. This measure suggests the results presented in Table 6 are generalizable to future data. The reduced model using only the total worksheet submissions achieved a cross-validation R^2 of 0.191, meaning the full model achieved a 9.4% reduction (improvement) in mean squared error on held out test data against the reduced model. A nested F-test using the full model specified in Table 6 and the reduced model using the strong baseline predictor of total worksheet submits indicates a significant improvement in sum of squared error for the number of parameters introduced $F(6, 53) = 3.54$, $p < 0.01$.

The full linear model resulted in three significant predictors: total worksheet submits, during book joy, and after scan confusion. The negative coefficient for total worksheet submits is intuitive, as a higher number of submissions indicates more incorrect submissions and a lower game score. The positive coefficient on *During Book Joy* indicates that students with higher proportions of joy while reading relevant texts compared to reading irrelevant texts had higher overall game scores. The positive coefficient on *After Scan Confusion* suggests that students with higher proportions of confusion after a positive scan compared to confusion after a negative scan had higher overall game scores, all other things considered equal.

Table 6
Multiple linear regression model for final game score

	Coefficient	**Std Error**	**Std Coefficient**	**t-stat**
Intercept	1130	97.5	0	11.6
After Scan Frustration	631	415	0.165	1.52
During Book Frustration	-992	2850	-0.0530	-0.35
After Scan Confusion	4070	1970	0.218	2.07*
During Book Confusion	459	3240	0.0223	0.142
After Scan Joy	-335	458	-0.0811	-0.730
During Book Joy	5620	1310	0.403	3.71**

Total Submits	Worksheet	-184	34.3	-0.580	-5.37**
		Adj $R^2 = 0.397$		LOO-CV $R^2 = 0.285$	

*$p < 0.05$, ** $p < 0.01$.

DISCUSSION

For this study, we sought to determine how contextualized emotions were associated with overall learning, assessed by game score, during gameplay with CRYSTAL ISLAND. Overall, results revealed that emotions differed based on if the occurrence had a positive or negative outcome related to solving the mystery. Our findings shed light not only on how students' emotions differed from positive vs. negative in-game outcomes from engaging in self-regulatory and scientific reasoning processes, but also on how we can design intelligent systems that are adaptive to students' in-game behaviors.

Results from our first research question revealed that there was a significant difference in emotions for positive, compared to negative outcomes after scanning, but not during book reading or after submitting the diagnosis worksheet. This could be because during scanning, students are in the experimental phase of scientific discovery (Klahr & Dunbar, 1988), so they are obtaining results to prove or disprove a hypothesis, which can induce different types of emotions based on the result. Conversely, when students are reading books they are in the hypothesis space (Klahr & Dunbar, 1988), so they are not aware if their hypothesis will be supported or not at this point. It is possible that we did not find an effect for submitting the diagnosis worksheet because it does not occur as frequently as scanning or reading books. Our *t*-tests revealed that there was a significant difference in confusion after positive, compared to negative occurrences, but not for joy or frustration. Specifically, confusion was higher after a scan with a negative result than a positive result. This negative scan result might have caused cognitive disequilibrium, resulting in confusion (D'Mello & Graesser, 2012), as their hypothesis was not supported because they thought they found the transmission source of the illness since a sick patient reported eating it. For example, if a sick patient reported eating bread, milk, and eggs, the student might hypothesize that all three of these food items would be transmitters of the illness. However, the bread tested negative, which might have confused the student because they were not aware that only one item was the transmission source. This means that to resolve this confusion, students will have to return to the hypothesis space (Klahr & Dunbar, 1988) and make an adaptation (Winne, 2018) to their original hypothesis, and then return to the experimental space to test the new hypothesis. This partially supported H1 because we did predict confusion would be higher for negative, compared to positive outcomes, however we only found this result for scanning (not reading books or submitting the diagnosis worksheet), and only for confusion (not joy or frustration)

Results from our second research question revealed there were no significant correlations between contextual emotions and overall game score. This did not support H2 because we predicted there to be significant correlations between contextualized emotions and overall game score. This might be because we did not have a large enough sample to obtain significance, however the direction of these correlations yielded some interesting findings. For example, the *After Scan Confusion* value is positive when a student experiences more confusion after positive scans than after negative scans. Thus, a negative correlation with overall game score indicates that students who were more confused after positive scans tended to have lower overall game scores. Conversely, students who experienced less confusion after positive scans than after negative scans tended to have higher overall game scores, though at the current sample size this linear relationship is not significant at the 0.05 level. The positive correlation between *After Submission Joy* and overall game score suggests that students who had a higher proportion of joy after a correct submission compared with their incorrect submissions had a higher game score, potentially indicating they were joyful to finish the mystery early, which is reflected in the efficiency measures of the game score.

Results from our third research question revealed that final diagnosis worksheet submission significantly negatively predicted overall game score, such that making more submissions results in a lower game score. This can be indicating that these students were not engaging in effective self-regulated learning (Winne, 2018) or scientific reasoning strategies (Klahr & Dunbar, 1988) during gameplay, and might have simply been guessing for the solution. We also found that joy during book reading of a relevant book significantly positively predicted overall game score, indicating that students were happy to have found a relevant clue while reading, possibly because they were able to match the content with other information (e.g., matching symptoms from a sick patient with an illness they were reading about). This would result from an effective metacognitive judgment (content evaluation), so they could use this information to create a hypothesis. Lastly, we found that confusion expressed after scanning with a positive outcome significantly positively predicted overall game score, revealing that perhaps this confusion was beneficial (D'Mello, Lehman, Pekrun, & Graesser, 2014), allowing students to resolve their cognitive disequilibrium (D'Mello & Graesser, 2012) by concluding some food items were not the transmission source of the illness. This partially supported H3 because we did predict joy during a positive occurrence to predict overall game score, however this result was only for reading books. We also predicted confusion during a negative occurrence to be negatively predictive of overall game score, however we found confusion during a positive occurrence to be positively predictive of overall game score, not supporting H3. Additionally, we did not find any significant associations between frustration after any occurrence with overall game score, also not supporting H3.

Overall, our results partially align with prior studies that showed a negative relationship between confusion and learning (Andres et al., 2014, 2015; Ocumpaugh et al., 2017; Sabourin & Lester, 2014), as we did find that more diagnosis worksheet submissions correlated negatively with overall game score, which can be indicative of students being confused about submitting an incorrect diagnosis. However, we also found levels of confusion after a positive scan result to be positively predictive of overall game score, demonstrating that higher levels of confusion predicted a higher game score, which is the opposite of previous findings. It is possible that in this case, confusion was beneficial for student learning, whereas it was not in other cases. Other recent studies have investigated the impact of confusion and frustration on learning, and have found that expressing both confusion and frustration can lead to better learning outcomes (D'Mello et al., 2014; Liu et al., 2013; Richey et al., 2019) or use of SRL processes (Taub et al., in press). This demonstrates the importance of analyzing emotions within specific contexts so we can understand when being confused can help students learn effectively.

Limitations

This study examined the impact of contextual emotions, which provided interesting results for understanding how emotions differ in different contexts, however we must address the limitations. First, we only examined three emotions (joy, confusion, frustration), and students could have expressed other emotions that have not previously been theorized to relate to learning outcomes when engaging in these activities. Second, although FACET has been tested and validated as a facial expression detection software (see Dente et al., 2017), its generalizability to a setting of diverse college students has not been explicitly validated by the developers. Additionally, we collected multichannel data for this study, however we only used video data of facial expressions to measure emotions. Thus, in future studies, we should also use physiological data (i.e., electrodermal activity), as research has shown that this data can be indicative of emotions (Harley, Bouchet, Hussain, Azevedo, & Calvo, 2015). Finally, as previously mentioned, for this study, we aggregated all instances of joy, confusion, and frustration for all positive vs. negative outcomes for each in-game action instead of assessing individual instances of emotions and contexts, thus not including the temporal dimension of context. Now that we were able to detect these contextual differences in emotions, future studies should investigate emotion fluctuations over time and different in-game actions.

Implications and future directions

Results from this study have implications for understanding different levels of emotions students express during learning with all advanced learning technologies. Specifically, when these technologies include many different activities that require the use of different learning processes (e.g., knowledge acquisition, self-regulated learning, scientific reasoning), different emotions can be experienced, and in different ways, based on the outcome of that activity. Research has demonstrated that confusion can positively impact learning (D'Mello et al., 2014; Liu et al., 2013; Richey et al., 2019), but this may only be the case for some activities over others. For example, if students are reading multimedia material and are confused because they do not understand the content, perhaps they will not be able to resolve their confusion, which can lead them to be frustrated and ultimately bored. However, if a student is confused because they tested a hypothesis and determines it is disproved, this will provide them with the information they need to engage in scientific reasoning or self-regulation, leading to effective learning and problem solving, even if they were confused at first. Thus, it can be beneficial to understand the contexts in which different emotions can more beneficially impact overall learning with games.

Furthermore, based on the limitation mentioned above, future analysis should investigate the temporal nature of SRL, scientific reasoning, and emotions during game-based learning. These analyses can inform us of how students' emotions fluctuate over time, how they change use of SRL and scientific reasoning strategies over time, which can be indicative of making adaptations to less effective strategy use. If we do not detect changes, perhaps the student does not know how to resolve confusion or how to use different SRL and scientific reasoning strategies. Analyses can inform us of how early we can make performance predictions as well, which can also be informative for future game design, such that we can provide students with adaptive feedback as soon as we detect they are experiencing difficulties. Therefore, investigating the temporal aspects can be informative for understanding the changing nature of emotions, SRL, and scientific reasoning, and for future game design.

Lastly, future directions can aim to develop advanced learning technologies that are adaptive to students' emotions in different contexts, such as after receiving results from a scan, or while they are reading multimedia content. These adaptive environments can also ensure that students are engaging in effective learning strategies to ensure they are not engaging in maladaptive behaviors. In this study, results revealed that perhaps students were not self-regulating their learning or creating and testing hypothesis, but rather were formulating guesses to the solution, which resulted in lower overall game scores. Thus adaptive GBLEs can be designed to train students how to use self-regulatory strategies and create and test hypotheses so they can play the game effectively. As the long term objective of designing GBLEs is to motivate students to learn complex topics, the more help and guidance they can provide, the more effective these learning experiences will be.

ACKNOWLEDGMENTS

The research presented in this paper has been supported by funding from the Social Sciences and Humanities Research Council of Canada (SSHRC 895–2011–1006). The authors would like to thank Robert Taylor for assisting with system development.

REFERENCES

Andres, J.M.L., Rodrigo, M.M.T., Baker, R.S., Paquette, L., Shute, V.J., & Ventura, M. (2015). *Analyzing student action sequences and affect while playing Physics Playground.* Paper presented at the International Workshop on Affect, Meta-Affect, Data and Learning (AMADL 2015) at the 17th International Conference on Artificial Intelligence in Education (AIED 2015), Madrid, Spain.

Andres, J.M.L., Rodrigo, M.M.T., Sugay, J.O., Baker, R.S., Paquette, L., Shute, V.J., Ventura, M., & Small, M. (2014). An exploratory analysis of confusion among students using Newton's playground. In C.-C. Liu et al. (Eds.), *Proceedings of the 22nd International Conference on Computers in Education* (pp. 65-70). Ishikawa, Japan: Asia-Pacific Society for Computers in Education.

Azevedo, R., Mudrick, N. V., Taub, M., & Bradbury, A. E. (2019). Self-regulation in computer-assisted learning systems. In J. Dunlosky & K. Rawson (Eds.), *The Cambridge handbook of cognition and education* (pp. 587-618). Cambridge, MA: Cambridge Press.

Azevedo, R., Taub, M., & Mudrick, N. V. (2018). Using multi-channel trace data to infer and foster self-regulated learning between humans and advanced learning technologies. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance (2nd ed.).* (pp. 254-270). New York, NY: Routledge.

Clark, D.B., Tanner-Smith, E.E., & Killingsworth, S.S. (2016). Digital games, design, and learning: A systematic review and meta-analysis. *Review of Educational Research, 86,* 79–122.

Dente, P., Küster, D., Skora, L., & Krumhuber, E.G. (2017). Measures and metrics for automatic emotion classification via FACET. In J. Bryson, M. De Vos, & J. Padget (Eds.), *Proceedings of the Conference on the Study of Artificial Intelligence and Simulation of Behaviour (AISB)* (pp. 160–163). Red Hook, NY: Curran Associates.

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

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

D'Mello, S., Lehman, B., Pekrun, R., & Graesser, A. (2014). Confusion can be beneficial for learning. *Learning and Instruction, 29,* 153-170.

Ekman, P., Friesen, W.V, & Hager, J.C. (2002). *Facial action coding system*. Salt Lake City, UT: Netwprk Information Research Corporation.

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.

iMotions Attention Tool (Version 6.0) [Computer software] (2016). Boston, MA: iMotions Inc.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12,* 1-48.

Littlewort, G., Wu, T., Whitehill, J., Fasel, I., Movellan, J., & Bartlett, M. . (2011). CERT Computer Expression Recognition Tool. In *Automatic Face and Gesture Recognition* (pp. 298–305). New York, NY: IEEE.

Liu, Z., Pataranutaporn, V., Ocumpaugh, J., & Baker, R. S. J. d. (2013). Sequences of frustration and confusion, and learning. In S. K. D'Mello, R. A. Calvo, & A. Olney (Eds.), *Proceedings of the 6th International Conference on Educational Data Mining (EDM 2013)* (pp. 114-120). Educational Data Mining Society.

Mayer, R.E. (Ed.) (2014). *Computer games for learning: An evidence-based approach.* Cambridge, MA: MIT Press.

Munshi, A., Rajendran, R., Ocumpaugh, J., Biswas, G., Baker, R. S., & Paquette, L. (2018). Modeling learners' cognitive and affective states to scaffold SRL in open-ended learning environments. In T. Mitrovic, J. Zhang, L. Chen, & D. Chin (Eds.), *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization (UMAP '18)* (pp. 131-138). New York, MY: ACM.

Ocumpaugh, J., Andres, J.M., Baker, R., DeFalco, J., Paquette, L., Rowe, J., Mott, B., Lester, J., Georgoulas, V., Brawner, K., & Sottilare, S. (2017). Affect dynamics in military trainees using vMedic: From engaged concentration to boredom to confusion. In E. André, R. Baker, M. Rodrigo, & B. du Boulay (Eds.), *Proceedings of the 18th International Conference on Artificial Intelligence in Education* (pp. 238-249). Amsterdam, The Netherlands: Springer.

Ocumpaugh, J., Baker, R. S., & Rodrigo, M. M. T. (2015) *Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual* (Technical Report). New York, NY: Teachers College, Columbia University.

Plass, J.L., Homer, B.D., & Kinzer, C.K. (2015). Foundations of game-based learning. *Educational Psychologist, 50,* 258-283.

Pekrun, R., Lichtenfeld, S., Marsh, H. W., Murayama, K., & Goetz, T. (2017). Achievement emotions and academic performance: A longitudinal model of reciprocal effects. *Child Development, 88*, 1653–1670.

Poryaska-Pomsta, K., Mavrikis, M., D'Mello, S., Conati, C., & Baker, R. S. J. d. (2013). Knowledge elicitation models for affect modelling in education. *Journal of Artificial Intelligence In Education, 22,* 107-140.

Richey, J. E., Andres-Bray, J. M. L., Mogessie, M., Scruggs, R., Andres, J. M. A. L., Star, J. R., Baker, R. S., & McLaren, B. M. (2019). More confusion and frustration, better learning: The impact of erroneous examples. *Computers & Education, 139*, 173-190.

Rowe, J., Shores, L., Mott, B., & Lester, J. (2011). Integrating learning, problem solving, and engagement in narrative-centered learning environments. *International Journal of Artificial Intelligence in Education, 21*, 115–133.

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

Sabourin, J., Rowe, J., Mott, B., & Lester, J. (2012). Exploring inquiry-based problem-solving strategies in game-based learning environments. In S. A. Cerri, W. J. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Proceedings of the 11th International Conference on Intelligent Tutoring Systems (ITS 2012)* (pp. 470-475. Berlin, Heidelberg: Springer-Verlag.

Sabourin, J. L., Shores, L. R., Mott, B. W., & Lester, J. C. (2013). Understanding and predicting student self-regulated learning strategies in game-based learning environments. *Journal of Artificial Intelligence in Education, 23,* 94-114.

Sawyer, R., Mudrick, N. V., Azevedo, R., & Lester, J. (2018). Impact of Learner-Centered Affective Dynamics on Metacognitive Judgements and Performance in Advanced Learning Technologies. In C. P. Rosé, R. Martínez-Maldonado, H. U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren, & B. du Boulay (Eds.), *Proceedings of the 19th International Conference on Artificial Intelligence in Education* (pp. 312-316). Amsterdam, The Netherlands: Springer.

Scherer, K. R. (2005). *What are emotions? And how can they be measured?".* Social Science Information, 44, 693-727.

Taub, M., & Azevedo, R. (2018). Using sequence mining to analyze metacognitive monitoring and scientific inquiry based on levels of efficiency and emotional expressivity during game-based learning. *Journal of Educational Data Mining, 10,* 1-26.

Taub, M., Azevedo, R., Bradbury, A.E., & Mudrick., N. (in press). Self-regulation and reflection in game-based learning. In J. L. Plass, R. E. Mayer, & B. Horner (Eds.), *Handbook of game-based learning*. Boston, MA: MIT Press.

Taub, M., Azevedo, R., Bradbury, A. E., Millar, G. C., & Lester. J. (2018). Using sequence mining to reveal the efficiency in scientific reasoning during STEM learning with a game-based learning environment. *Learning and Instruction, 54,* 93-103.

Taub, M., Azevedo, R., Rajendran, R., Cloude, E. B., Biswas, G., & Price, M. J., (in press-b/online first 2019). How are students' emotions related to the accuracy of their use of cognitive and metacognitive processes during learning with an Intelligent Tutoring System? *Learning and Instruction.*

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, 76*, 641-655.

Winne, P.H. (2018). Cognition and metacognition within self-regulated learning. In D.H. Schunk & J.A. Greene (Eds.), *Handbook of self-regulation of learning and performance (2nd ed.)* (pp. 36-48). New York, NY: Routledge.

Winne, P., & Hadwin, A. (1998). Studying as self-regulated learning. In D. Hacker, J. Dunlosky, and A. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 227–304). Mahwah, NJ: Erlbaum.

International Journal of Artificial Intelligence in Education Volume# (YYYY) Number
Springer