Using Sequence Mining to Analyze Metacognitive Monitoring and Scientific Inquiry based on Levels of Efficiency and Emotions during Game-Based Learning

Michelle Taub and Roger Azevedo
University of Central Florida
michelle.taub@ucf.edu; roger.azevedo@ucf.edu

Self-regulated learning conducted through metacognitive monitoring and scientific inquiry can be influenced by many factors, such as emotions and motivation, and are necessary skills needed to engage in efficient hypothesis testing during game-based learning. Although many studies have investigated metacognitive monitoring and scientific inquiry skills during game-based learning, few studies have investigated how the sequence of behaviors involved during hypothesis testing with game-based learning differ based on both efficiency level and emotions during gameplay. For this study, we analyzed 59 undergraduate students' (59% female) metacognitive monitoring and hypothesis testing behavior during learning and gameplay with CRYSTAL ISLAND, a game-based learning environment that teaches students about microbiology. Specifically, we used sequential pattern mining and differential sequence mining to determine if there were sequences of hypothesis testing behaviors and to determine if the frequencies of occurrence of these sequences differed between high or low levels of efficiency at finishing the game and high or low levels of facial expressions of emotions during gameplay. Results revealed that students with low levels of efficiency and high levels of facial expressions of emotions had the most sequences of testing behaviors overall, specifically engaging in more sequences that were indicative of less strategic hypothesis testing behavior than the other students, where students who were more efficient with both levels of emotions demonstrated strategic testing behavior. These results have implications for the strengths of using educational data mining techniques for determining the processes underlying patterns of engaging in self-regulated learning conducted through hypothesis testing as they unfold over time; for training students on how to engage in the self-regulation, scientific inquiry, and emotion regulation processes that can result in efficient gameplay; and for developing adaptive game-based learning environments that foster effective and efficient self-regulation and scientific inquiry during learning.

Keywords: efficiency, emotions, game-based learning, scientific inquiry, self-regulated learning, sequence mining

1. INTRODUCTION

Self-regulated learning (SRL) is a highly studied educational construct that views students as active learners as opposed to being passive recipients of information (Pintrich, 2000; Winne, 2018; Winne & Azevedo, 2014; Winne & Hadwin, 1998, 2008; Zimmerman & Schunk, 2011). To engage in successful SRL, students must use the appropriate processes to ensure they are planning, monitoring, and strategizing, including cognitive, affective, metacognitive, and motivational (CAMM; Azevedo, Taub, & Mudrick, 2018) processes, to ensure they are learning the material to complete the task they are given. The specific types of strategies that are the most appropriate will depend on the task itself. For example, while engaging in accurate metacognitive monitoring processes are always beneficial for learning, the specific types of monitoring processes will be more effective for one task compared to another. When engaging in scientific inquiry (White, Frederiksen, & Collins, 2009), a student engages in hypothesis testing, which requires them to monitor the results they are obtaining to determine if their hypotheses are correct or incorrect. So while all SRL processes are important for learning, context will impact which specific SRL strategies will be the most beneficial for the given task. Although research has shown that successful SRL can lead to greater learning outcomes (Azevedo et al., 2018), research has also shown that students do not typically deploy the appropriate strategies effectively (Azevedo, 2009; Winne & Azevedo, 2014). As such, researchers have developed different types of advanced learning technologies (ALTs) to foster the use of effective SRL, whether it is by including pedagogical agents who externally regulate students' SRL by providing them prompts and feedback or by including tools within the ALT that students can use for regulating their learning (e.g., Azevedo et al., 2013; Graesser, 2013). There are many different types of ALTs, such as intelligent tutoring systems, simulations, hypermedia, multimedia, and game-based learning environments. For this study, we assessed how students used metacognitive monitoring and scientific inquiry processes during learning while playing CRYSTAL ISLAND, a game-based learning environment (GBLE) that teaches students about using SRL strategies to learn about microbiology.

GBLEs were specifically designed to foster learning while ensuring students stay engaged and motivated in the task (Mayer, 2014; Shute & Ventura, 2013). Although research has shown that games have the potential to maintain high levels of motivation during learning (Lester, Rowe, & Mott, 2013), all games differ in topic (e.g., physics, math, microbiology, critical thinking skills), target different populations (e.g., elementary school, middle school, high school, college, professional), and include different game narratives (e.g., science researcher, secret agent). Research has shown that students who learn via playing GBLEs typically show higher learning gains. However, these results are often for specific topics and specific populations (see Literature Review section for more details). Given that different games require the use of different processes (e.g., scientific reasoning and inquiry, problem solving), more research is needed on investigating how these specific processes unfold during learning. Additionally, there are many factors that have been found to influence learning with GBLEs, such as students' levels of emotions (Sabourin & Lester, 2014), and their efficiency in scientific inquiry (Eslinger, White, Frederiksen, & Brobst, 2008). However, there is limited research investigating how these factors have together impacted hypothesis testing during learning with GBLEs. As such, for this study, we investigated how students' levels of efficiency and emotions impacted their use of metacognitive monitoring processes while they engaged in scientific inquiry (via hypothesis testing) as they attempted to discover what mysterious illness has infected the inhabitants of CRYSTAL ISLAND (Rowe, Shores, Mott, & Lester, 2011).

1.1. THEORETICAL FRAMEWORKS

Since we investigated both self-regulated learning behaviors (i.e., metacognitive monitoring) and emotions during learning, we used one theoretical basis for each construct, as there are few theories that address both SRL as an event that unfolds over time and emotions as they unfold over time. Therefore, as our theoretical frameworks, we used but did not aim to directly assess, Winne and Hadwin's Information Processing Theory (IPT) of SRL (1998, 2008) and Scherer's Component Process Model (CPM) of emotions (2009).

1.1.1. Information processing theory

According to the information processing theory of SRL (Winne, 2018; Winne & Hadwin, 1998, 2008), learning occurs through a series of four cyclical phases where information processing occurs via monitoring and control processes within each phase. The first phase, *defining the task,* involves students understanding what they are being asked to do. The second phase is *setting goals and making plans* and is where students map out how they are going to accomplish the task. In the third phase, *use of learning strategies,* students use the learning strategies they planned to use in the second phase. The fourth phase, *making adaptations,* is where students can modify their goals and plans and engage in different learning strategies, if they were deemed not as useful as originally planned. Although these phases are named in order, it must be noted that they do not necessarily occur in sequential order, nor are they independent from one another. For example, a student can be engaging in learning strategies (phase 3) while making adaptations to their plans (phases 2 and 4). Specifically, this model posits that during each phase, students can engage in a set of cognitive and metacognitive strategies (i.e., conditions, operations, products, evaluations, and standards; COPES), and the specific set of operations can involve searching, monitoring, assembling, rehearsing, and translating (SMART). As such, according to this model, SRL is defined as an event that temporally unfolds over time through these phases of information processing.

In addition to the IPT model, we must also note that depending on the learning context and what is required of students to do (e.g., problem solving, hypothesis testing), there are additional factors that need to be considered within our theoretical framework. Specifically, for this study, students must engage in scientific reasoning and inquiry to complete the game. According to White, Frederiksen, and Collins (2009), the process of scientific reasoning and inquiry involves using both a theoretical basis and empirical evidence to form hypotheses that should be tested on different science topics. Additionally, to engage in effective scientific inquiry, students must have the appropriate metacognitive knowledge and skills to understand what the process of hypothesis formation and testing actually is, and why it contributes to effective scientific inquiry (White et al., 2009). For this study, students engaged in hypothesis testing to attempt to solve the mystery on the island, and so they must have been able to accurately monitor the items they were testing, along with the implications of the test results. Therefore, not only did students need to use the appropriate SRL strategies to select the relevant items to test, they had to do so by forming the correct hypotheses based on the clues they gathered throughout the game. Although including scientific inquiry with the IPT framework is relevant for examining hypothesis testing behaviors during gameplay, this model does not account for the impact of emotions during learning, which is why we also based our study on the component process model of emotions.

1.1.2. Component process model

The component process model of emotions (Scherer, 2009) is an appraisal model, which states that there are four different types of events (relevance, implication, coping, and normative significance) that can elicit different types appraisal processes or mechanisms, and emotions are the net result of those appraisal mechanisms. Specifically, during these events, people are performing evaluations (stimulus evaluation checks; SECs) that lead to forming different appraisals of the events, and it is these SECs that yield different emotions, and impact different components of an affective state. For example, if the event is hypothesis testing, the SEC might be to determine the relevancy of that item. If the student is unsure of the relevancy, this SEC would lead to an appraisal of uncertainty, which might lead to a state of confusion. In contrast, if the student makes the SEC that the item is relevant, this would lead to the appraisal of testing that item, leading to an affective state of certainty or confidence. Furthermore, the process is recursive, such that appraisals can occur multiple times during an emotional state or can reoccur during an event. Finally, this model is not limited to a set of emotions (i.e., Ekman, 1973), rather there is an unlimited set of emotional responses that people can elicit (e.g., confusion, certainty, confidence). Therefore, according to this model, all emotional states are the result of an appraisal of a particular event (e.g., use of an SRL strategy). We therefore based our study on these two models because they consider SRL and emotions as events that are cyclical in nature and can change over time, which is applicable to learning with GBLEs.

By assessing levels of efficiency and emotions, we attempted to integrate aspects from these two theoretical models, as both efficiency and emotions can impact self-regulated learning and appraisal processes. The IPT model focuses on using cognitive and metacognitive processes for effective self-regulation. This is related to efficiency, such that efficiency in self-regulation relates to accurately using these cognitive and metacognitive processes. Specifically, for this study efficiency in solving the mystery implies that students were accurately engaging in metacognitive monitoring processes during gameplay. For the CPM of emotions, in addition to levels of emotions being related to the amount of appraisals students make (i.e., more appraisals signifies high emotions), the quality of these appraisals might lead to efficient or inefficient hypothesis testing behavior. Therefore, both efficiency and emotions are closely linked to these theoretical models, signifying the importance of analyzing these variables for this study.

2. LITERATURE REVIEW

An increasing amount of research has been conducted to investigate the effectiveness of different types of GBLEs on student learning. For example, Physics Playground (originally called Newton’s Playground; Shute, Ventura, & Kim, 2013) has students learn about the laws of Newtonian physics by creating a simple machine to move a ball toward a given target. Operation ARIES! (Millis et al., 2011) is a game that teaches students about using the correct scientific reasoning skills by identifying incorrect ones. The narrative of this game involves aliens from the Aries constellation who are publishing work using the scientific method incorrectly, and gameplayers are secret agents who must identify these incorrect papers so the Federal Bureau of Science can stop them from getting published. Thus, this game requires students to be able to correctly identify the appropriate use of scientific reasoning and inquiry. CRYSTAL ISLAND, the game used in the current study (Rowe et al., 2011), also teaches students to use scientific reasoning and inquiry skills to solve the mystery of what epidemic outbreak has impacted all the island inhabitants (see CRYSTAL ISLAND section under Methods, below). As demonstrated from these few GBLEs, there are many different types of GBLEs, which teach a range of topics to a range of different populations of students. Given the large range in GBLEs, many meta-analyses have been conducted to determine the optimal use of GBLEs (Connolly, Boyle, MacArthur, Hainey, & Boyle, 2012; Girard, Escalle, & Magnan, 2012; Wouters, van Nimwegen, van Oostendorp, & van der Spek, 2013). The most recent meta-analyses (Clark, Tanner-Smith, & Killingsworth, 2016; Mayer, 2014) compared learning with games to learning with conventional methods, what types of games are the most effective, and for which populations. Specifically, Clark et al. (2016) found that learning outcomes were greater when learning with games compared to learning with non-games (*g* = .33) over a range of disciplines, spanning from ages 6 to 26. Similarly, Mayer (2014) found that games were better than traditional teaching methods; however this was only the case for teaching science and second language learning, and not for math or language arts. Additionally, Mayer (2014) found that learning with games was best for college students (*d* = .74), then secondary school students (*d* = .58), and then elementary school students (*d* = .34). With regards to type of game, Clark et al. (2016) found that an enhanced version of a game led to better learning outcomes than the standard version (*g* = .34), and specifically when scaffolding was provided (*g* = .41). More specifically, Mayer (2014) found that adventure games led to the best learning outcomes (*d* = .72), followed by simulation games (*d* = .62), and puzzle games (*d* = .45). In addition, Clark et al. (2016) revealed that there are many moderating factors that led to their results, such as number of sessions playing a game, where playing multiple times led to better learning gains (*g* = .44) than playing the game over a single session (*g* = .08); character goals and point of view, where single player, non-collaborative and non-competitive games yielded the best learning outcomes (*g* = .45); extra game elements, where adding points and badges led to better learning outcomes (*g* = .53) than providing additional scaffolding with points and badges (*g* = .25); etc. As such, it is evident from gaming research that overall, games can be beneficial for learning, but there are many additional factors that contribute to the game's effectiveness for learning.

Based on the increasing development of GBLEs, research on GBLEs has been increasing as well, and investigates a plethora of topics, such as the impact of gender (Edens, 2008; Nietfeld, Shores, & Hoffmann, 2014), motivation (Davies & Hemingway, 2014; Feng & Chen, 2014; Papastergiou, 2009; Vos, van der Meijden, & Denessen, 2011) engagement and immersion (Chen & Sun, 2016; Cheng, She, & Annetta, 2014; Spires, 2015; Tsai, Huang, Hou, Hsu, & Chiou, 2016), problem solving (Chang, Wu, Weng, & Sung, 2012; Marone, Staples, & Greenberg, 2016; Sabourin, Rowe, Mott, & Lester, 2012; Spires, Rowe, Mott, & Lester, 2011; Yang, 2012), and many others, and their roles in learning with GBLEs. As the focus of this paper is on SRL and emotions during learning with GBLEs, we will review specific research studies in those domains.

2.1. SRL AND GBLES

Many studies have investigated how students use different SRL processes during learning with GBLEs, with an increasing amount of research focusing on metacognitive strategies. For example, Sabourin, Shores, Mott, & Lester (2013) investigated how students used SRL processes during gameplay with CRYSTAL ISLAND. They asked students to frequently evaluate their current status and then classified students as low, medium, or high use of SRL processes based on the statements they made in their status updates. Based on these SRL groups, results revealed that high and medium SRL students had significantly higher learning gains than low SRL students, high SRL students used the tools necessary for learning content (i.e., read books and posters) more than medium and low SRL students, and high SRL students tested fewer items than medium and low SRL students (Sabourin et al. 2013), revealing that high SRL students were, in fact, better at self-regulating their learning, leading to better learning outcomes. A study conducted by Kim, Park, & Baek (2009) investigated how students used metacognitive strategies as they engaged in social problem solving with the game, *Gersang*, a Massively Multiple Online Role Playing Game where students embarked in a variety of quests with the goal of becoming a wealthy merchant in the Choseon Dynasty 200 years ago. In this game, there are scenarios of economic activity where students experience conditions such as inflation or international trade, as well as battle scenarios where students attempt to get better weapons and level upgrades. Thus, the goal is for students to understand a Market Economy and how to engage in problem solving and use metacognitive strategies to succeed in this Market Economy. The game provides each student with a gain credit score and a battle score, which were found to be correlated with each other (Kim et al., 2009). Based on correlations, their results revealed that thinking aloud and self-recording were significantly positively correlated with social problem solving, and self-recording and modeling were significantly positively correlated with achievement (Kim et al., 2009). Therefore, these results demonstrate the positive relationships between using SRL processes and metacognitive strategies and performance during learning with a GBLE.

In addition to investigating the use of SRL strategies, studies are focusing on how one's metacognitive awareness impacts their learning with GBLEs. In a series of studies (Snow, 2015; Snow, et al., 2015), students played iSTART-2 and were monitored on their performance on embedded quizzes. If they performed under a certain threshold level, the system informed them of their poor performance and transitioned them to a remedial task. Results revealed that after students (with both high and low levels of prior knowledge) were transitioned to this remedial task, they showed increased performance on the in-game assessments (Snow, 2015; Snow et al., 2015), revealing that when they are made aware of their low performance (i.e., metacognitive awareness), this increased their overall task performance. In contrast, in a classroom study (Ke, 2008), students played ASTRA EAGLE, a series of GBLEs developed by one of the school districts involved in the study. Results indicated that learning with games promoted students' levels of motivation; however there were no significant differences in performance on cognitive math assessments or metacognitive awareness (Ke, 2008). Therefore, it appears that more work is needed to investigate exactly how and when we can promote metacognitive awareness during learning with GBLEs.

2.2. EMOTIONS AND GBLES

In addition to investigating the use of metacognitive and SRL strategies during learning with GBLEs, more studies have begun to investigate the impact of students' emotions while learning with these environments. For example, Sabourin & Lester (2014) reported on a review of studies conducted with students playing CRYSTAL ISLAND, where they assessed the relationships between affect and overall learning, use of inquiry skills, problem solving abilities, and off-task behavior. Their findings revealed the benefits of positive affect on all of these factors during learning with GBLEs, including how these environments can be designed to increase positive affect and engagement to enhance learning.

Studies have also investigated specific emotions or emotional states that can either enhance or deter learning with GBLEs. A study conducted by Yeh, Lai, & Lin (2016) examined how students' emotions impacted their levels of creativity during learning with a game-based evaluation system. They categorized emotions based on the states of three factors: valence (positive or negative), activation (high or low), and focus (prevention or promotion). Results revealed that students who had a combination of positive valence, high activation, and promotion focused, indicative of being happy and elated, this emotional state facilitated their creativity during gameplay, whereas students with a combination of negative valence, high activation, and promotion focused, indicative of frustration and anger, led to decreased levels of creativity (Yeh et al., 2016). These results demonstrate the importance of designing systems that foster happiness and elation to induce creativity during learning with GBLEs. In another study, Andres et al. (2015) examined the relationship between students' action sequences and emotions during gameplay with Physics Playground. Specifically, they identified a group of sequences showing experimentation behaviors and another group of sequences representative of behaviors with unsolved activities. Results showed that out of the top seven experimentation sequences, two of them were correlated with confusion; these two sequences were those that showed less understanding of concepts. For the unresolved sequences, one of them was correlated with boredom; this sequence showed correct activities but demonstrated that students gave up on completing them. Therefore, these results revealed that there were distinct action sequences of student behavior during learning with GBLEs, and there was a significant connection to emotions during learning (Andres et al., 2015). It is clear from these studies that emotions play an important role during learning with GBLEs. One additional advantage to the study conducted by Andres et al. (2015) is their use of sequence mining to determine patterns of behaviors during learning.

2.3. SEQUENCE MINING

For the current study, we used sequence mining, a type of educational data mining technique, which allows us to examine the specific processes students use during learning, and not just outcome performance measures (Kinnebrew, Loretz, & Biswas, 2013; Winne & Baker, 2013). There are many different types of sequence mining techniques, such as sequential pattern mining and differential sequence mining (Kinnebrew et al., 2013), cluster analysis (Bouchet, Harley, Trevors, & Azevedo, 2014), and process mining (Sonnenberg & Bannert, 2015). Researchers have used educational data mining to investigate a range of phenomena, such as predicting overall learning outcomes (Baker & Corbett, 2014; Kinnebrew et al., 2013), use of cognitive and metacognitive SRL strategies (Bouchet et al., 2014; Sonnenberg & Bannert, 2015), using scientific inquiry skills (Gobert, Sao Pedro, Baker, Toto, & Montalvo, 2012; Gobert, Sao Pedro, Raziuddin, & Baker, 2013), and the impact of students' affect on learning (Andres et al., 2015; Ocumpaugh, Baker, Gowda, Heffernan, & Heffernan, 2014; Shute et al., 2015). Although these studies have all advanced our understanding of which GBLEs lead to successful learning and how SRL and emotions influence effective learning, less attention has been given to the process of how SRL behaviors and levels of emotion together lead to effective learning and gameplay with GBLEs, which is the goal of the current study.

3. CURRENT STUDY

The goal of the current study was to use sequential pattern mining and differential sequence mining to assess how students' metacognitive monitoring and hypothesis testing behavior differed for students with high or low levels of efficiency of completing the game and high or low levels of facial expressions of emotions during learning while playing CRYSTAL ISLAND, a GBLE where students learn about microbiology, where they had to collect clues to solve the mystery of what illness has infected inhabitants of the island.

We examined high and low levels of efficiency and emotions because our theoretical frameworks do not differentiate between different levels of these variables, and investigating both high and low levels can help us in the design of intelligent tutoring systems, such that different levels of scaffolding can be provided based on the specific characteristics of the learners. This leads to an imbalance in SRL research because we typically expect that students use all the SRL strategies they need to achieve higher learning gains. However, the narrative of gameplay is to solve the mystery and finish the game as quickly as possible. As such, the most efficient students might not be those who use all of the SRL strategies they need to learn but will instead limit the amount of strategies to ensure they are finishing the game quickly.

Additionally, we expect that students regulate their emotions during learning to avoid prolonged periods of frustration; however, these students might not aim to be aware of any level of emotions because they are focused on completing the game. As such, this results in an imbalance between effective SRL and efficient SRL, and we must keep in mind what the overarching goal of the study is for students participating, which will lead to forming different types of research questions and hypotheses.

We investigated metacognitive monitoring during learning and gameplay (i.e., hypothesis testing) with Crystal Island, a game-based learning environment that requires students to gather different clues to solve the mystery of what illness impacted inhabitants of the island. To gather these clues, students had to navigate to the five different buildings on the island where they could collect different types of clues, such as reading content about different illnesses or testing different food items for the transmission source. We focused specifically on clues that involved testing food items for the transmission source of the illness. This required students to collect food items and run them through a scanner, which determined if the food item tested positive or negative for pathogenic or non-pathogenic viruses or bacteria. When testing the item, students must select the substance they are testing for (bacteria, viruses, carcinogens, or mutagens), as well as why they chose to test it (e.g., sick inhabitant reported eating it). For our analyses, we sought to determine the sequences in which students tested food items in the scanner to determine the disease's transmission source. We investigated all instances of scanning behavior and coded each behavior based on two factors: (1) the relevance of the pathogenic substance they were testing, and (2) the relevance of the food item itself. We used the number of relevant, partially-relevant, and irrelevant food items, as well as the frequency of occurrences of sequential patterns of testing these food items, as our dependent variables.

In addition to gathering clues during gameplay, students could also monitor and track their progress by entering the information they gathered into a diagnosis worksheet. This worksheet was also where students recorded the final diagnosis, transmission source, and treatment plan. To solve the mystery correctly, students had to enter a correct final diagnosis, transmission source, and treatment plan, and submit it to the camp nurse. We therefore defined efficiency based on the number of correct submission attempts of the diagnosis worksheet, which we used towards our independent, or grouping variable, where one attempt was categorized as more efficient, and more than one attempt was less efficient. In addition to efficiency, we categorized students based on their levels of emotions, which we created based on evidence scores (i.e., the likelihood of a human coder coding for that emotion) of a composite of basic and learner-centered emotions. We conducted a median split on these scores, as determining different levels of emotions can be used for designing intelligent tutoring systems as well, and classified students with scores lower than the median as low emotions, or high emotions for students with scores above the median. Based on levels of efficiency and emotions, we created one grouping variable with four levels, which were the different combinations of these variables: (1) less efficient-low emotions, (2) less efficient-high emotions, (3) more efficient-low emotions, and (4) more efficient-high emotions. We used these groups for our independent variable for our analyses.

3.1. Research questions and hypotheses

We posed three research questions: (1) Are there differences in the number of food items tested by efficiency-emotion group? (2) Are there frequent sequences of testing for the transmission source for each efficiency-emotion group? (3) Are there differences in the sequences of testing for the transmission source by efficiency-emotion group? Providing answers to these questions will allow us to identify how students self-regulate their learning by engaging in metacognitive monitoring and hypothesis testing during gameplay and how this differs for students with different efficiency and emotion levels. The results have implications toward providing adaptive scaffolding to train students to self-regulate where if the system can identify sequences of hypothesis testing behavior that have been linked to less efficient gameplay and different levels of emotions, the system can provide the students scaffolding to help them engage in more effective behaviors to help them monitor their hypothesis testing during gameplay.

Based on our research questions, we generated three hypotheses, all of which presume that high levels of facial expressions of emotions can have different outcomes for metacognitive monitoring behaviors depending on level of efficiency, where (H1): there will be significant differences in the number of food items tested between efficiency-emotion groups, such that less efficient-high emotions students will test the most food items, and more efficient-high emotions students will test the fewest food items (H2) There will be frequent sequences of testing behavior for each efficiency-emotion group, where less efficient-high emotions students will show the greatest number of sequences, while more efficient-high emotions students will show the fewest sequences. (H3): There will be significant differences in the sequences of testing behavior between efficiency-emotion groups, such that less efficient-high emotions students will have the highest frequencies of sequences, while more efficient-high emotions students will have the lowest frequencies of these sequences,

4. METHODS

4.1. PARTICIPANTS AND MATERIALS

59[1] undergraduate students (59% female) from a large North American university participated in this study. Their ages ranged from 18 to 26 years old (*M* = 19.97, *SD* = 1.61). Students were randomly assigned to one of three experimental conditions (*full agency*, *partial agency, no agency*; see Experimental conditions section below), and they were compensated $10 per hour for participating.

Prior to gameplay, students completed a demographics questionnaire (asking them about their age, gender, amount of hours spent playing video games, and the types of video games they played), followed by a series of self-report questionnaires that asked them to report on their levels of emotions (i.e., Emotions and Values [EV]; adapted from Pekrun, Goetz, Frenzel, Barchfeld, & Perry, 2011) and motivation (i.e., Achievement Goals [AGQ]; Elliot & Murayama, 2008). After gameplay, students completed additional self-report questionnaires about their emotions and motivation (i.e., EV and AGQ), as well as the Perceived Interest Questionnaire [PIQ] (Schraw, Bruning, & Svoboda, 1995), which asked them to report on how interested they were in the task. Students also completed a pre- (*M* = 11.64 (or 55%), *SD* = 2.75) and post-test (*M* = 14.15 (or 67%), *SD* = 2.76), which were 21-item, four-choice multiple-choice tests containing 12 factual and 9 procedural questions on microbiology.

[1] This was a sub-sample of 89 participants, as we did not include participants who did not complete the game (i.e., solve the mystery), who had missing facial expression data, or did not generate any trace data.

4.2. Crystal Island

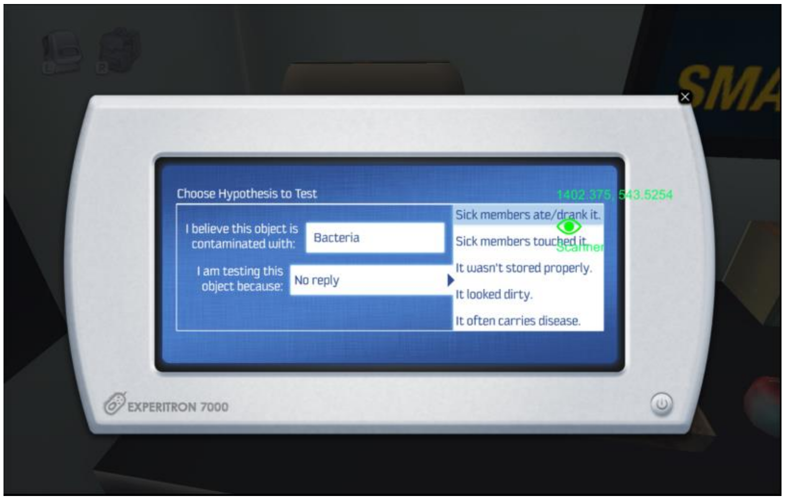
Crystal Island is a game-based learning environment that requires students to engage in self-regulated learning and scientific inquiry to learn about microbiology (see *Figure 1)*. Students were tasked with solving a mystery of what illness had spread and infected inhabitants of the island (Rowe et al., 2011), which required them to play a researcher (from a first-person perspective) and collect clues from different buildings by engaging in different in-game activities to determine the mysterious disease (salmonellosis or influenza), its transmission source (bread, eggs, or milk), and a suggested treatment plan (e.g., vaccine, rest).

Figure 1: Outside camp of Crystal Island.

There are five buildings on the island that students could navigate to and engage in different activities to collect information regarding the illness, such as reading books and research papers, talking to non-player characters, and collecting food items to be tested for the disease's transmission source. All elements in the game contain drop-down menus for students to input answers, and ask questions to sick patients, experts in microbiology, or employees on the island. Thus, to gather clues in the game, students navigated to the various locations on the island, and once in the location, selected books to read by clicking on them and responding to assessment questions by selecting from a list of answers, approached non-player characters by standing close to them and then selecting from a list of questions which ones to ask, and picked up food items by clicking on them and adding them to their backpack Books contained content about different diseases (e.g., Ebola, salmonellosis), which was coupled with an in-game assessment called a concept matrix. The concept matrix required students to answer questions about the book content. In the infirmary, there is Kim, the camp nurse who introduced the outbreak, as well as sick patients who reported on their symptoms. The living quarters (dorm room and Bryce's, a non-player character's, quarters) contained several non-player characters who are experts on microbiology topics (e.g., viruses, bacteria), along with books and research papers, posters, and other non-player character patients who were infected by the illness. In the dining hall, students could talk to Quentin, the camp chef, who reported on foods he had been cooking recently, and could gather more food items to take to the laboratory for testing.

4.2.1. Testing for the transmission source

In the laboratory, students could bring all food items they gathered in the other buildings in their backpack to be tested as the possible transmission source of the illness. When testing food items (see *Figure 2)*, students opened the scanner by clicking on it, opened their backpack and clicked on the food item they wanted to test to place it into the scanner, and then had to select (by clicking on) what they were testing the item for (i.e., bacteria, viruses, mutagens, or carcinogens) and the reason why they were testing it (i.e., sick patient ate it, it looked dirty). If the food item tested positive for pathogenic viruses or bacteria, this would indicate that item is the transmission source of the illness. However, many food items were potential nonpathogenic bacteria, and so if the food item tested positive for nonpathogenic bacteria, this is a possible source of that disease, but not the specific transmission source of the illness for that session. Once the students determined the pathogenic substance, they had discovered the transmission source and could mark down their results in the diagnosis worksheet by clicking on a menu of entry options on the worksheet.



Choose Hypothesis to Test

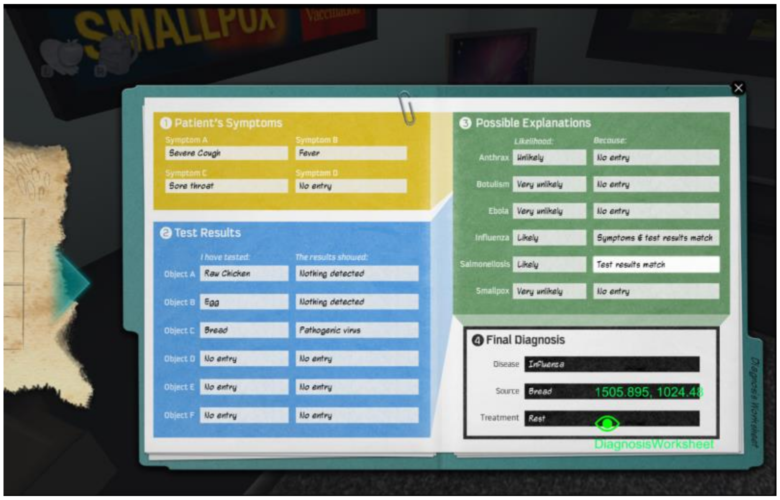
I believe this object is contaminated with: Bacteria

I am testing this object because: No reply

Figure 2: Scanning device used to test for the transmission source of the illness.

4.2.2. Diagnosis worksheet

The diagnosis worksheet is an in-game tool kept in the student's backpack that allowed students to monitor the clues they were gathering during learning (see *Figure 3)*. For example, when students read about different diseases in the books and spoke with non-player characters about the reported symptoms, they could mark down the likelihood of each of the illnesses being the solution to the mystery by selecting the likelihood as very, moderately, or not likely from the list of options. In addition, students could monitor the tested food items for pathogenic and nonpathogenic substances. To complete the game, students had to submit a correct final diagnosis. Once students determined the likely disease and transmission source along with the suggested treatment, they could insert that into the final diagnosis section and submit that to Kim, the camp nurse. If the diagnosis was correct, they had completed the game in one attempt, and then the game would be over. However, if it was not correct, Kim would direct them to the incorrect selection (e.g., check the transmission source). Once they changed it, they could reattempt to submit the worksheet until they had it correct. Thus, students who did not submit a correct final diagnosis had to make more than one attempt to solve the mystery correctly and complete the game. Therefore, the diagnosis worksheet allowed students to monitor their progress in gameplay to assist them in solving the mystery correctly and efficiently.



① Patient's Symptoms

Symptom A Severe Cough	Symptom B Fever
Symptom C Sore throat	Symptom D No entry

② Test Results

	I have tested:	The results showed:
Object A	Raw Chicken	Nothing detected
Object B	Egg	Nothing detected
Object C	Bread	Pathogenic virus
Object D	No entry	No entry
Object E	No entry	No entry
Object F	No entry	No entry

③ Possible Explanations

	Likelihood:	Because:
Anthrax	Unlikely	No entry
Botulism	Very unlikely	No entry
Ebola	Very unlikely	No entry
Influenza	Likely	Symptoms & test results match
Salmonellosis	Likely	Test results match
Smallpox	Very unlikely	No entry

④ Final Diagnosis

Disease	Influenza
Source	Bread
Treatment	Rest

Figure 3: Diagnosis worksheet used to monitor gathered information and submit the final diagnosis.

4.3. EXPERIMENTAL PROCEDURE

This study took place over one session lasting from 39.75 to 129.73 minutes (M = 87.24 minutes, SD = 21.22), depending on experimental condition (see below). When students began the session, they were first presented with a consent form, followed by an overview of the study, which involved the experimenter describing that they would play CRYSTAL ISLAND to solve the mystery of what illness has spread throughout the island and infected all the inhabitants. They were told that to solve the mystery, they would have to gather clues by exploring different buildings on the island and engaging in different activities, such as reading books and research papers, talking to non-player characters (e.g., the camp nurse, experts on microbiology topics), and test food items for the disease's transmission source. Finally, they were told that to complete the game, they would have to come up with a correct diagnosis, along with the correct transmission source and treatment plan. After the overview, students put on the electrodermal activity (EDA) bracelet so it could begin collecting a baseline. Next, students completed the pre-test measures, including a demographics questionnaire, several self-report questionnaires asking them to report on their emotions and motivation (see Participants and Materials section, above), and the 21-item pre-test on microbiology. Following the pre-test, students' eye tracking was calibrated with the SMI EYERED 250 (SMI, 2014) using a 9-point calibration, where calibration was acceptable when eye movements reached an offset of less than .05 mm or after five attempts at calibration. A baseline was then collected for the facial recognition of emotions and EDA bracelet using Attention Tool 6.2 software (iMotions, 2016). Once the equipment was set up, students were given a second overview of the game, where they were told specific details of the game scenario, including each of the activities they could engage in during the game (i.e., read books and research articles, talk to non-player characters, and test food items for the transmission source). After this overview, students began playing CRYSTAL ISLAND, and when they completed the game (i.e., solved the mystery), they completed post-test measures, including self-report measures about their emotions and motivation and the 21-item post-test on microbiology. Participants were then debriefed, paid for their participation, and completed the study.

4.4. CODING AND SCORING

During the session, we collected multi-channel data from each student, including: (1) log-file trace data, (2) videos of facial expressions of emotions, (3) eye-tracking data of fixations on different areas of interest (AOIs), and (4) electrodermal activity (EDA) of students' physiological responses. For this analysis, we only included log-file trace data and videos of facial expressions of emotions.

4.4.1. Efficiency-Emotion Groups

We grouped students based on how efficiently they solved the mystery and how emotionally expressive they were during learning. Efficiency in solving the mystery was based on the number of attempts students made at submitting their diagnosis worksheet (see above) correctly to solve the mystery. Log-file data generated the number of worksheet submission attempts (*M* = 2.51, *SD* = 2.58), and with this data, we created two groups of efficiency. More efficient students submitted their worksheet correctly after one attempt, and less efficient students submitted their worksheet correctly after more than one attempt (range: 2–16 attempts).

We also grouped students based on their levels of emotions during learning using their evidence scores from the videos of facial expression data. Evidence scores, as generated by iMotions Attention Tool 6.2 (2016) are described as the likelihood that human coders would code for that emotion. The numerical output uses a base 10 log scale, such that a score of 1 would indicate a likelihood of 10 human coders coding for that emotion, a score of 2 indicates a likelihood of 100 human coders coding for that emotion, a score of 3 indicates a likelihood of 1000 human coders coding for that emotion, etc. Scores could also be negative (i.e., the likelihood of human coders not coding for that emotion); however, we were only interested in the presence of emotions, and so we only used positive scores and transformed all negative scores to zero. Based on these values, we calculated a mean evidence score combining evidence scores of joy, anger, surprise, contempt, confusion, and frustration, as these are all emotions found to be expressed during learning with advanced learning technologies (D'Mello & Graesser, 2012; Ekman, 1973). We then performed a median split (median = .23) on these scores, where scores higher than the median were categorized as high emotions, and scores lower than the median were categorized as low emotions.

Finally, we created groups based on both efficiency and emotions, combining each type of classification (high vs. low of each). This generated four groups: (1) less efficient-low emotions (*n* = 15), (2) less efficient-high emotions (*n* = 14), (3) more efficient-low emotions (*n* = 15), (4) more efficient-high emotions (*n* = 15).

4.4.2. Relevancy codes

The log-file trace data generated instances of testing food items, including what items were being tested for. When students tested food items, they had to indicate what they were testing the food item for (viruses, bacteria, mutagens, or carcinogens; see above). We assigned relevant, partially-relevant, and irrelevant codes to what food item students were testing for. Students were not directly told if what they were testing for was relevant or not. However, they should have been able to make this inference based on the clues they gathered, such that there were books that explained these substances, and there were two microbiology expert non-player characters who provided information about these substances. The sick patients described their symptoms, and so students could deduce from this information that if a certain illness was associated with specific symptoms that the sick patient described, this should be the illness type.

When assigning these codes, half of them were specific to the solution. For example, the solution was selected at random as a pathogenic virus or bacterium, implying that both viruses and bacteria can be pathogenic substances. As such, if the type of transmission source was a virus, then testing for viruses was deemed relevant, but testing for bacteria was deemed partially-relevant, as bacteria could be a pathogenic substance, just not for this specific solution. In contrast, if the solution was a pathogenic bacterium, testing for bacteria was coded as relevant, and testing for viruses was coded as partially-relevant. Mutagens and carcinogens were never a possible type of pathogenic substance, and so these were always coded as irrelevant.

In addition, we coded the food items tested, and coded those items based on their relevance to solving the mystery as relevant, partially-relevant, or irrelevant. Relevant items were those reported as being eaten by the sick patient, Teresa, since those were the three potential transmission sources for the game (and one was chosen at random at the beginning of the game). The three relevant items were: bread, eggs, or milk, and the solution for the particular game would yield a result of a pathogenic virus or bacterium. Partially-relevant items were those that were not reported by Teresa as being eaten but were possible transmitters of nonpathogenic bacteria. This included fruit (e.g., orange, apple, banana, coconut), water, chicken, etc. Irrelevant items were therefore not eaten by island inhabitants, nor are they potential nonpathogenic bacteria (e.g., jelly, pie, peanuts). Therefore, again students were not given direct feedback about the relevancy of the food item. However, they were expected to infer that a sick patient had to have eaten the food in order for it to be the transmission source of the illness.

Based on the relevancies of what students were testing for and of the food items, we created a *full relevancy code*, which included two relevancy types. For example, if the solution to the mystery was a pathogenic virus spread by milk, and if the student tested the milk for viruses, this would yield a RELEVANT--RELEVANT code. We created codes with all possible combinations of relevancies, yielding nine unique codes (see Table 1). All codes are unique because they take order into consideration, such that the first code is always what is being tested for, and the second code is always what item is being tested. For example, a code of RELEVANT--PARTIALLYRELEVANT is different from a code of PARTIALLYRELEVANT--RELEVANT. Once we formed these codes, we used them to determine if there were sequential patterns of testing behavior between the efficiency-emotion groups.

Table 1: Description of nine full relevancy codes.

Code	Description
1: RR	RELEVANT -- RELEVANT
2: RP	RELEVANT -- PARTIALLYRELEVANT
3: RI	RELEVANT -- IRRELEVANT
4: PR	PARTIALLYRELEVANT – RELEVANT
5: PP	PARTIALLYRELEVANT -- PARTIALLYRELEVANT
6: PI	PARTIALLYRELEVANT – IRRELEVANT
7: IR	IRRELEVANT – RELEVANT
8: IP	IRRELEVANT – PARTIALLYRELEVANT
9: II	IRRELEVANT -- IRRELEVANT

4.4.3. Sequential pattern mining

Sequential pattern mining is an analytical technique (Baker & Yacef, 2009) that investigates if there are common sequences within an event, where an event can be defined as a given behavior or activity, such as testing food items or reading complex texts. For this analysis, we used testing food items as our event to determine if there were sequences of full relevancy codes students were using as they tested different food items for the disease's transmission source. For example, if students wanted to test the same food item for both viruses and bacteria, and the solution to the mystery was a pathogenic bacterium spread by eggs, this would generate a sequence of PartiallyRelevant--Relevant → Relevant--Relevant (PR → RR or 4→1). We used SPAM (Ayres, Flannick, Gehrke, & Yiu, 2002; Fournier-Viger, Gomariz, Campos, & Thomas, 2014) to examine 2-, 3-, and 4-coded sequences of food testing behavior for each efficiency-emotion group. The algorithm generates a support value for each sequence, which indicates the number of individuals per group that has enacted that sequence at least once. For example, a support value of 100% would indicate that all members of that group engaged in that sequence at least once during learning. We chose a maximum sequence of 4 items because there were four choices for selecting the transmitter of the pathogenic substance (bacterium, virus, pathogen, or mutagen), and so these sequences allowed us to account for students testing all possible pathogenic substances within one sequence As such, we wanted to determine if students were strategically selecting food items to test, or if they were simply testing all food items for all possible options, and which sequences were generated in each of the efficiency-emotion groups.

4.4.4. Differential sequence mining

Differential sequence mining can be used to detect if there are significant differences in the frequencies of sequential patterns between different groups (Kinnebrew et al., 2013). For our analysis, we used the sequential patterns of full relevancy codes to determine if there were significant differences in the number of occurrences of these sequences between efficiency-emotion groups. In doing so, we calculated an instance support value (i.e., the frequency of occurrence of that sequence per person) using a brute force technique (Grafsgaard, 2014), generating the instance support values for all combinations of 2-, 3-, and 4-coded sequences. Thus, we calculated instance support values for each sequence, for each student within efficiency-emotion groups.

5. Results

5.1. Research question 1: Are there differences in the number of food items tested by efficiency-emotion group?

A multiple analysis of variance (MANOVA) revealed a non-significant effect; Wilks' $\lambda = .78$, F (9,126.71) = 1.52, $p = .15$, $\eta_p^2 = .08$. This indicated that there were no significant differences in the number of relevant food items ($M = 8.09$, $SD = 3.61$), partially-relevant food items ($M = 13.09$, $SD = 9.47$), or irrelevant food items ($M = 3.22$, $SD = 3.36$) tested between efficiency-emotion groups (Table 2). This lack of significant results may be due to using traditional inferential statistics, and although we did not obtain significant differences from our MANOVA, perhaps using more advanced statistical techniques (i.e., educational data mining), we can determine that there are differences in these efficiency-emotion groups based on their patterns of using scientific inquiry to test for the transmission source of the illness.

Table 2: Number of food items tested by Efficiency-Emotion (E-E) group.

	Relevant		Partially-Relevant		Irrelevant	
	M	*SD*	*M*	*SD*	*M*	*SD*
Group LL	8.47	3.82	16.07	12.55	3.60	3.42
Group LH	8.57	3.30	15.14	8.85	4.50	3.78
Group HL	8.67	3.90	10.73	6.94	1.53	2.13
Group HH	6.57	3.30	10.36	8.04	3.36	3.54
	F		*F*		*F*	
MANOVA	1.10		1.44		2.13	

Note. Group LL = low efficiency-low emotion, Group LH = low efficiency-high emotion, Group HL = high efficiency-low emotion, Group HH = high efficiency-high emotion.

5.2. RESEARCH QUESTION 2: ARE THERE FREQUENT SEQUENCES OF TESTING FOR THE TRANSMISSION SOURCE FOR EACH EFFICIENCY-EMOTION GROUP?

We used the SPAM algorithm (Ayres et al., 2002, Fournier-Viger et al., 2014) to determine the sequential patterns of food item testing behavior for each efficiency-emotion group, using the full relevancy code for each behavior. Results revealed (see Table 3) that overall, there were many sequences. However, there were the fewest sequences for students in the more efficient-high emotion group and the most sequences for students in the less efficient-high emotion group. In addition, there were many more sequences for each group with a support value of 50%, with relatively few sequences (especially for 4-coded sequences) that all individuals per group used at least once (i.e., a support value of 100%).

Table 3: Sequential patterns using full relevancy codes for each E-E group.

	2-3 Coded Sequences		4-Coded Sequences	
E-E Group	SV 50%	SV 100%	SV 50%	SV 100%
Group LL	141	2	180	0
Group LH	344	24	915	5
Group HL	137	4	179	0
Group HH	53	1	11	0

Note. E-E Group = Efficiency-Emotion Group, SV = support value, 2-3 and 4-coded are the number of codes per sequence.

In addition, since a code of 1 (RR) indicates that this would be the correct transmission source for solving the mystery, we assessed the number of sequences that contained an RR code, and whether it was at the beginning or end of the sequence (see Table 4). We also assessed the number of unique patterns in the 4-coded sequences (see Table 5), as sequences with four codes would include 2- and 3- coded sequences if the codes were repeated (e.g., 1→2→3→2 could include sequences 1→2→3 and 2→3→2). Based on these results, for 2-3-coded sequences, groups had similar numbers of codes containing RR somewhere, with the less efficient-high emotion group containing the most sequences ending in RR, while more efficient-high emotions students had the fewest number of sequences ending in RR. We see a similar pattern for 4-coded sequences, such that less efficient-high emotions students had the most unique sequences, while more efficient-high emotions students had no unique 4-coded sequences. This may demonstrate that more efficient-high emotions students had more of a set of strategic sequences, as opposed to students who were less efficient-high emotions, who did not seem to have a strategic set of sequences to use. Overall, these results suggest that we were able to obtain frequent patterns of hypothesis testing behavior for each efficiency-emotion group.

Table 4: Sequential patterns include an RR code for 2-3 coded sequences.

E-E Group	Ends in RR	Ends in RR with other RRs	Has other RRs	Total	%
Group LL	19	10	37	66	46.81
Group LH	40	14	77	131	38.08
Group HL	20	11	32	63	45.99
Group HH	6	6	13	25	47.17

Note. RR = a RELEVANT--RELEVANT Code (code 1), E-E Group = Efficiency-Emotion Group. Ends in RR with other RRs = a pattern that has an RR code at the end of the sequence, but with other RRs in that sequence; Has other RRs = the sequence does not end in RR but does contain an RR code in that sequence.

Table 5: Sequential patterns with RR codes and unique codes for 4-coded sequences.

	Ends in RR	Has other RRs	No RRs	Total	%
Group LL	2	13	2	17	9.44
Group LH	64	114	99	277	30.27
Group HL	10	16	5	31	17.32
Group HH	0	0	0	0	0

Note. RR = a RELEVANT--RELEVANT Code (code 1), E-E Group = Efficiency-Emotion Group. Ends in RR = sequence ends with an RR code; Has RR = sequence contains an RR code but not at the end; and No RR = sequence is unique but does not contain an RR code.

5.3. RESEARCH QUESTION 3: ARE THERE DIFFERENCES IN THE SEQUENCES OF TESTING FOR THE TRANSMISSION SOURCE BY EFFICIENCY-EMOTION GROUP?

We used a brute force technique (Grafsgaard, 2014), which generated instance support values, defined as the frequency of occurrence of each sequence per person. We used the instance support values to determine if there were significant differences in the frequencies of sequences of hypothesis testing behavior between efficiency-emotion groups. To select the sequences (as we did not want to input over 1,000 sequences as dependent variables), we selected 2- or 3-coded sequences only because there were no unique 4-code sequences that had an instance support value of at least 1 across all 4 conditions. As such, we selected sequences that did occur at least once across all 4 conditions, leaving us with 9 dependent variables, and efficiency-emotion group as our independent variable, with four levels. We created 3 categories of sequences: (1) testing for the relevant pathogenic substance (i.e., sequences beginning with a 1, 2, or 3; with 4 sequences total), (2) testing for the partially-relevant pathogenic substance (i.e., sequences beginning with a 4, 5, or 6; with 3 sequences total), and (3) testing for irrelevant pathogenic substances (i.e., sequences beginning with a 7, 8, or 9; with 2 sequences total). We ran 3 separate chi-square analyses to test for each category of sequences.

Our first chi-square (see Table 6, Analysis 1) revealed a non-significant effect; $\chi^2(3) = 7.90$, $p = .54$, indicating there were no significant differences in sequences beginning with testing for the relevant pathogenic substance between efficiency-emotion groups.

Our second chi-square (see Table 6, Analysis 2) revealed a non-significant effect: $\chi^2(3) = 2.15$, $p = .91$, indicating there were no significant differences in sequences beginning with testing for the partially-relevant pathogenic substance between efficiency-emotion groups.

Our final chi-square (see Table 6, Analysis 3) revealed a significant effect; $\chi^2(3) = 8.87$, $p =.031$, indicating there were significant differences in sequences beginning with testing for irrelevant pathogenic substances (i.e., carcinogens or mutagens) between efficiency-emotion groups. Specifically, students who were more efficient with low emotions had the highest frequency of the IR → RR sequence (irrelevant pathogen, relevant food item → relevant pathogen, relevant food item), whereas students who were less efficient with low emotions had the lowest frequency of that sequence. Additionally, students who were more efficient with high emotions had the highest frequency of the IP → RP sequence (irrelevant pathogenic substance, partially-relevant food item → relevant pathogen, partially-relevant food item), while students who were less efficient with high emotions had the lowest frequency of that sequence.

Overall, these results suggest that based on the efficiency of gameplay and the levels of emotions students experienced, we can detect differences in behavior of testing for the transmission source of the illness.

Table 6: Frequencies of instance support values by E-E group.

Sequence	EE Group Frequency				χ^2
Analysis 1	LL (*n* = 15)	LH (*n* = 14)	HL (*n* = 15)	HH (*n* = 15)	
RR→PR	15	22	14	14	7.90
RP→RR	12	6	10	10	
RP→PP	32	36	27	27	
RP→IP	18	12	7	8	
Analysis 2					
PR→RR	14	15	13	6	2.15
PP→RR	5	7	2	2	
PP→RP	34	39	28	10	
Analysis 3					
IR→RR	5	7	11	8	8.87*
IP→RP	14	4	6	19	

*$p < .05$

Note. EE = Efficiency-Emotion. R = Relevant, P = PartiallyRelevant, I = Irrelevant.

6. Discussion

The goal of this study was to examine how students' levels of efficiency at solving the mystery and levels of emotions during gameplay related to their hypothesis testing behavior during gameplay with Crystal Island. Results revealed that when using a traditional statistical approach (MANOVA) using aggregate scores, no significant findings were found. However, when using sequential pattern mining and differential sequence mining, there were significant differences in the frequency of occurrence of sequential patterns of hypothesis testing behavior between groups.

Our first research question used a MANOVA to test for significant differences in the number of relevant, partially-relevant, and irrelevant food items tested between efficiency-emotion groups. Results revealed no significant effects, suggesting all students tested similar numbers of food items, regardless of their levels of efficiency or emotions during learning, thus not supporting H1, where we expected students with different levels of efficiency and emotions to test different numbers of food items. Theoretically, based on the IPT model, this suggests that students with both low and high levels of monitoring engaged in the same hypothesis testing behaviors, thus not suggesting different levels of metacognitive monitoring processes between groups. Furthermore, according to the CPM of emotions, students might not have had levels of emotion impeding their hypothesis testing behavior, such that they made similar numbers of appraisals and stimulus evaluation checks during learning, therefore not letting their emotions impact their hypothesis testing behavior. These non-significant MANOVA findings align with Ke's (2008) study who did not find a significant impact of GBLEs on the use of cognitive and metacognitive strategies, suggesting that perhaps Mayer's (2014) findings regarding GBLEs being the most beneficial for college students needs to be replicated. Additionally, perhaps Clark et al.'s (2016) findings regarding specific types of game elements, such as moderating and mediating factors, (i.e., if students had returned for more than one learning session) could be impacting the results.

For our second research question, we used sequential pattern mining to explore the sequential patterns of hypothesis testing behavior for each efficiency-emotion group. Results revealed that in contrast to results from research question one, we were able to detect distinct sequences of hypothesis testing behavior for each efficiency-emotion group. This supported H2; however, there were also sequences that were not unique to each group, thus support for our hypothesis was only partial. In comparison to using traditional statistical techniques, when using educational data mining techniques, we observed that students who were less efficient and highly emotionally expressive generated the most sequential patterns for 2-, 3-, and 4-coded sequences, revealing that perhaps it is these students who are exhibiting the least strategic behaviors, resulting in less monitoring of the relevancy of items they are testing. In line with the IPT model, perhaps these students have the least amount of control and monitoring behaviors, resulting in less strategic self-regulatory behaviors and scientific inquiry. Furthermore, as these students are exhibiting high levels of emotions, it seems as though making the stimulus evaluation checks and subsequent appraisals can be interfering with their ability to engage in efficient hypothesis testing behaviors, thus leading to a wider range of different patterns of behaviors compared to all other students. Therefore, in comparison to Andres et al. (2015), Sabourin & Lester (2014), and Yeh et al. (2016), whose studies found relationships between emotions and various factors, it is evident that high levels of emotions can impact performance, whether it can enhance it (e.g., high levels of happiness and elation leading to increased creativity; Yeh et al., 2016) or hinder it (e.g., high levels of confusion correlating with misunderstanding material; Andres et al., 2015).

Our final research question used differential sequence mining to detect significant differences in the frequency of occurrence of the sequential patterns we found in the second research question between efficiency-emotion groups. Results from chi-squares revealed significant effects for two sequences: IR→RR (7→1), and IP→RP (8→2), which contain relevant or irrelevant pathogenic substances, and relevant or partially-relevant food items. The frequencies of occurrence of both these sequences were higher for more efficient students, where the higher frequencies for the IR→RR (7→1) and the IP→RP (8→2) sequences were for students with low emotions and high emotions, respectively. In contrast, the lowest frequencies for these sequences were for less efficient students, where sequence IR→RR was the lowest for less efficient-low emotions students and sequence IP→RP was the lowest for less efficient-high emotions students. Thus, these results demonstrate that more efficient students with both levels of emotions showed higher frequencies of these behaviors and less efficient students with both levels of emotions showed lower frequencies of these behaviors, perhaps displaying that they were testing the same food item for multiple pathogens. Specifically, students who were more efficient and had low emotions enacted a sequence that ended with testing the correct transmission source for the correct illness type. Therefore, once testing that food item, they gathered enough information to solve the mystery. Students who were more efficient with high emotions also seemed to engage in the strategy of testing the same food item for different pathogens, however they were not testing for the correct transmission source, perhaps indicating an inability to regulate their emotions sufficiently to accurately monitor for the correct illness type *and* transmission source of the illness, but still demonstrating the ability to engage in strategic behaviors. These results do not support H3 because we expected the less efficient-high emotions students to have the highest frequencies of these sequences, and we found the highest frequencies to be for the more efficient students; we also expected more efficient-high emotions students to have the lowest frequencies of these sequences, and they had the highest for one of them. Our results tie into the IPT and CPM frameworks in a similar way to the second research question, such that it seems as though high levels of emotions impeded students' abilities to control and monitor their hypothesis testing behavior and engage in strategic testing behavior for some sequences, perhaps due to spending too much time engaging in appraisal mechanisms. As such, these results reveal the importance of how emotions relate to learning (e.g., Andres et al., 2015; Sabourin & Lester, 2014; Shute et al., 2015; Yeh et al., 2016), and how we need to ensure that students continue to use effective metacognitive strategies (e.g., Ke, 2008; Kim et al., 2009; Sabourin et al., 2013; Snow, 2015; Snow et al., 2015) alongside these emotions to ensure that GBLEs are providing the most efficient learning experience (Clark et al., 2016; Mayer, 2014).

6.1. LIMITATIONS

Despite our informative results on how high levels of emotions may impede the use of strategic metacognitive monitoring during scientific inquiry with low levels of efficiency, there are some limitations to our study that must be noted. First, our sample size was appropriate in sum; however, once we created four groups, each group only had 14 or 15 students, thus limiting our analyses of aggregated data (i.e., MANOVA for research question 1). Second, when conducting differential sequence mining, there is no theoretical justification for choosing which sequences to examine. Therefore, it is up to the researchers to theoretically or empirically justify why they chose the given sequences to analyze, as selecting all of the sequences can lead to at least 1,000 dependent variables. Thus, it is possible that there were more significant differences in the frequencies of sequences; we just did not choose them for our analysis. Additionally, our coding of efficiency and emotions was empirically based, such that we dichotomized efficiency as 1 vs. more than 1 and used a median split for emotions. It is possible, however, that students in the 'high' groups were more similar to those in the 'low' groups. For example, submitting the diagnosis worksheet twice might be more similar to submitting it once than submitting it 10 times; however, 2 times and 10 times were in the same group. Additionally, low levels of emotions, as classified in this study, might not mean low overall levels of emotions, it is just lower than the high emotions group. Additionally, as our analyses did not allow for us to infer causality, we cannot confirm that levels of emotions impacted hypothesis testing behavior, or if continuing to test items without determining the pathogenic substances is what caused high levels of emotions. Finally, dichotomizing the variables ignored their continuous nature, limiting our ability to address emotions as processes of events. As such, future studies should aim to address these limitations by increasing sample size, developing a clear set of standardized guidelines for sequence selection, and using data mining and machine learning algorithms that do allow us to infer causality and include all variables as continuous without dichotomizing them.

6.2. IMPLICATIONS AND FUTURE DIRECTIONS

Results from this study have important implications for training students on how to use metacognitive monitoring and scientific inquiry strategies while maintaining an appropriate level of emotions during learning with different types of advanced learning technologies. For example, theories of self-regulated learning often presume that students know *what* SRL is, and *how* to use SRL strategies effectively. However, most students are not trained in this, and therefore lack the prior SRL knowledge and skills needed when using these learning technologies to gain the most beneficial learning experience. Furthermore, these students must be taught that SRL should be applied differently depending on the task (e.g., hypothesis testing via scientific inquiry vs. solving a math problem), and so it is important for them to learn both the skills required for that task (e.g., hypothesis testing skills), and the appropriate SRL skills (e.g., metacognitive monitoring of testing behaviors) that complement task completion. In addition, students might also not be aware of how to control their emotions to prevent them impeding with their learning, and so they must also be taught the appropriate emotion regulation strategies (Gross, 2015) to avoid letting their emotions distract them from engaging in monitoring and control processes. Therefore, training sessions that include teaching students the declarative knowledge of SRL as well as the procedural and conditional knowledge of when and how to use SRL processes within specific contexts can greatly benefit students of all age levels, learning various topics with different types of learning technologies, and not just the specific sub-populations mentioned by Mayer (2014).

7. CONCLUSION

The goal of this study was to determine if we can identify how different types of students, based on their use of metacognitive processes and emotions, learn with GBLEs. Our results have implications for designing adaptive GBLEs that foster each individual student's learning needs. All students have different skills, abilities, and weaknesses (as demonstrated by investigating high and low levels of efficiency and emotions), and so a learning environment that does not address each student's challenges will not benefit each student. For example, one student might know how to use metacognitive monitoring and emotion regulation strategies but not know how to engage in scientific inquiry. In contrast, a student might know how to engage in scientific inquiry but not know how to use the appropriate metacognitive monitoring and emotion regulation strategies. In both cases, the students need scaffolding provided by the system to ensure they complete the task they are given. However, the type of scaffolding will differ because these students need assistance with different task elements. Furthermore, it is almost impossible to predict the combinations of strategies each student possesses before learning with computer technologies, and so predetermining scaffolding might not address all of the difficulties some students might have during learning. As such, developing environments, whether it is a GBLE, simulation, hypermedia, intelligent tutoring system, etc., that are adaptive in real-time will be able to scaffold students based on the particular challenges they are facing as they learn. In doing so, this can ensure that students of all ability levels can learn how to use

SRL, and other task-specific strategies the most effectively, which can ensure that they are all receiving the most optimal learning experience.

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 James Lester, Robert Taylor, Robert Sawyer, and Andy Smith for assisting with system development, and Megan Price for her assistance with data collection.

REFERENCES

ANDRES, J. M. L., RODRIGO, M. M. T., BAKER, R. S., PAQUETTE, L., SHUTE, V. J., AND 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.

AYRES, J., FLANNICK, J., GEHRKE, J., AND YIU, T. 2002. Sequential pattern mining using a bitmap representation. In D. Hand, D. Keim, AND R. Ng (Eds.), *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 429-435). New York, NY: ACM.

AZEVEDO, R. 2009. Theoretical, methodological, and analytical challenges in the research on metacognition and self-regulation: A commentary. *Metacognition & Learning, 4,* 87–95.

AZEVEDO, R., HARLEY, J., TREVORS, G., FEYZI-BEHNAGH, R., DUFFY, M., BOUCHET, F., AND LANDIS, R. S. 2013. Using trace data to examine the complex roles of cognitive, metacognitive, and emotional self-regulatory processes during learning with multi-agent systems. In R. Azevedo AND V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 427–449). Amsterdam, The Netherlands: Springer.

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

BAKER, R. S., AND CORBETT, A. T. 2014. Assessment of robust learning with educational data mining. *Research & Practice in Assessment, 9,* 38–50.

BAKER, R. S. J. D., AND YACEF, K. 2009. The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1,* 3–16.

BOUCHET, F., HARLEY, J., TREVORS, G., AND AZEVEDO, R. 2013. Clustering and profiling students according to their interactions with an intelligent tutoring system fostering self-regulated learning. *Journal of Educational Data Mining, 5,* 104–146.

CHANG, K. E., WU, L. J., WENG, S. E., AND SUNG, Y. T. 2012. Embedding game-based problem-solving phase into problem-posing system for mathematics learning. *Computers & Education, 58,* 775–786.

CHEN, L. X., AND SUN, C. T. 2016. Self-regulation influence on game play flow state. *Computers in Human Behavior, 54,* 341–350.

CHENG, M. T., SHE, H. C., AND ANNETTA, L. A. 2014. Game immersion experience: Its hierarchical structure and impact on game-based science learning. *Journal of Computer Assisted Learning, 31,* 232–253.

CLARK, D. B., TANNER-SMITH, E. E., AND KILLINGSWORTH, S. S. 2016. Digital games, design, and learning: A systematic review and meta-analysis. *Review of Educational Research, 86,* 79–122.

CONNOLLY, T. M., BOYLE, E. A., MACARTHUR, E., HAINEY, T., AND BOYLE, J. M. 2012. A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education, 59*, 661-686.

DAVIES, J. J., AND HEMINGWAY, T. J. 2014. Guitar hero or zero? Fantasy, self-esteem, and deficient self-regulation in rhythm-based music video games. *Journal of Media Psychology, 26,* 189–201.

D'MELLO, S. K., AND GRAESSER, A. C. 2012. Dynamics of affective states during complex learning. *Learning and Instruction, 22,* 145–157.

EDENS, K. M. 2008. The integration of pedagogical approach, gender, self-regulation, and goal orientation using student response system technology. *Journal of Research on Technology in Education, 41,* 161–177.

EKMAN, P. 1973. *Darwin and facial expression: A century of research in review.* New York, NY: Academic Press.

ELLIOT, A. J., AND MURAYAMA, K. 2008. On the measurement of achievement goals: Critique, illustration, and application. *Journal of Educational Psychology, 100,* 613–628.

ESLINGER, E., WHITE, B., FREDERIKSEN, J., AND BROBST, J. 2008. Supporting inquiry processes with an interactive learning environment: Inquiry Island. *Journal of Science Education and Technology, 17,* 610–617.

FENG, C. Y., AND CHEN, M. P. 2014. The effects of goal specificity on programming performance and self-regulation in game design. *British Journal of Educational Technology, 45,* 285–302.

FOURNIER-VIGER, P., GOMARIZ, A., CAMPOS, M., AND THOMAS, R. 2014. Fast vertical mining of sequential patterns using co-occurrence information. In V.S. Tseng, T.B. Ho, Z. Zhou, A.L.P. Chen, AND H. Kao (Eds). *Proceedings of the 18th Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 40-52). Cham, Switzerland: Springer.

GIRARD, C., ESCALLE, J., AND MAGNAN, A. 2012. Serious games as new educational tools: How effective are they? A meta-analysis of recent studies. *Journal of Computer Assisted Learning, 29,* 207–219.

GOBERT, J. D., SAO PEDRO, M. A., BAKER, R. S. J. D., TOTO, E., AND MONTALVO, O. 2012. Leveraging educational data mining for real-time performance assessment of scientific inquiry skills within Microworlds. *Journal of Educational Data Mining, 4,* 111–143.

GOBERT, J. D., SAN PEDRO, M., RAZIUDDIN, J., AND BAKER, R. 2013. From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences, 22,* 521–563.

GRAESSER, A. C. 2013. Evolution of advanced learning technologies in the 21st century. *Theory into Practice, 52,* 93–101.

GRAFSGAARD, J. F. 2014. *Multimodal affect modeling in task-oriented tutorial dialogue.* (Doctoral dissertation). Retrieved from ProQuest. (3690271).

GROSS, J. J. 2015. The extended process model of emotion regulation: Elaborations, applications, and future directions. *Psychological Inquiry, 26,* 130–137.

iMOTIONS ATTENTION TOOL (VERSION 6.0) [COMPUTER SOFTWARE] 2016. Boston, MA: iMotions Inc.

KE, F. 2008. Computer games application within alternative classroom goal structures: Cognitive, metacognitive, and affective evaluation. *Educational Technology Research and Development, 56,* 539–556.

KIM, B., PARK, H., **AND** BAEK, Y. 2009. Not just fun, but serious strategies: Using meta-cognitive strategies in game-based learning. *Computers & Education, 52,* 800–810.

KINNEBREW, J. S., LORETZ, K. M., AND BISWAS, G. 2013. A contextualized, differential sequence mining method to derive students' learning behavior patterns. *Journal of Educational Data Mining, 5,* 190–219.

LESTER, J. C., Rowe, J. P., and Mott, B. W. 2013. Narrative-centered learning environments: A story-centric approach to educational games. In C. Mouza AND N. Lavigne (Eds.), *Emerging Technologies for the Classroom: A Learning Sciences Perspective* (pp. 223-238). New York, NY: Springer US.

MARONE, V., STAPLES, C., AND GREENBERG, K. H. 2016. Learning how to learn by solving bizarre problems: A playful approach to developing creative and strategic thinking. *On the Horizon, 24,* 112–120.

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

MILLIS, K., FORSYTH, C., BUTLER, H., WALLACE, P., GRAESSER, A., AND HALPERN, D. 2011. Operation ARIES!: A serious game for teaching scientific inquiry. In M. Ma, A. Oikonomou, AND L. Jain (Eds.), *Serious games and edutainment applications* (pp. 169–196). London, UK: Springer-Verlag.

NIETFELD, J. L., SHORES, L. R., AND HOFFMANN, K. F. 2014. Learning environment self-regulation and gender within a game-based learning environment. *Journal of Educational Psychology, 106*, 961–973.

OCUMPAUGH, J., BAKER, R., GOWDA, S., HEFFERNAN, N., AND HEFFERNAN, C. 2014. Population validity for educational data mining models: A case study in affect detection. *British Journal of Educational Technology, 45,* 487–501.

PAPASTERGIOU, M. 2009. Digital game-based learning in high school computer science education: Impact on educational effectiveness and student motivation. *Computers & Education, 52,* 1–12.

PEKRUN, R., GOETZ, T., FRENZEL, A., BARCHFELD, P., AND PERRY, R. 2011. Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary Educational Psychology, 36,* 36–48.

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

ROWE, J., SHORES, L., MOTT, B., AND 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., AND 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., AND LESTER, J. 2012. Exploring inquiry-based problem-solving strategies in game-based learning environments. In S. A. Cerri, W. J. Clancey, G. Papadourakis, AND K. Panourgia (Eds.), *Proceedings of the 11th International Conference on Intelligent Tutoring Systems—Lecture Notes in Computer Science 7315* (pp. 59–64). Amsterdam, The Netherlands: Springer.

SABOURIN, J. L., SHORES, L. R., MOTT, B. W., AND LESTER, J. C. 2013. Understanding and predicting student self-regulated learning strategies in game-based learning environments. *International Journal of Artificial Intelligence in Education, 23,* 94–114.

SCHERER, K. 2009. Emotions are emergent processes: They require a dynamic computational architecture. *Philosophical Transactions of the Royal Society,* 364, 3459–3474.

SCHRAW, G., BRUNING, R., AND SVOBODA, C. 1995. Sources of situational interest. *Journal of Literacy Research, 27,* 1–17.

SHUTE, V. J., D'MELLO, S., BAKER, R., CHO, K., BOSCH, N., OCUMPAUGH, J., VENTURA, M., AND ALMEDA, V. 2015. Modeling how incoming knowledge, persistence, affective states, and game progress influence student learning from an educational game. *Computers & Education, 85,* 224–235.

SHUTE, V., AND VENTURA, M. 2013. *Measuring and supporting learning in video games: Stealth assessment.* Cambridge, MA: The MIT Press.

SHUTE, V. J., VENTURA, M., AND KIM, Y. J. 2013. Assessment and learning of qualitative physics in Newton's playground. *The Journal of Educational Research, 106,* 423–430.

SMI EXPERIMENT CENTER 3.4.165 [APPARATUS AND SOFTWARE]. 2014. Boston, Massachusetts, USA: SensoMotoric Instruments.

SNOW, E. L. 2016. *Promoting self-regulation and metacognition through the use of online trace data within a game-based environment.* (Doctoral dissertation). Retrieved from Arizona State University Libraries, ASU Electronic Dissertations and Theses.

SNOW, E. L., MCNAMARA, D. S., JACOVINA, M. E., ALLEN, L. K., JOHNSON, A. M., PERRET, C. A., DAI, J., JACKSON, G. T., LIKENS, A. D., RUSSELL, D. G., AND WESTON, J. L. 2015. In C. Conati, N. Heffernan, A. Mitrovic, AND M. F. Verdejo (Eds.), *Proceedings of the 17th International Conference on Artificial Intelligence in Education—Lecture Notes in Computer Science 9112* (pp. 786–789). Basel, Switzerland: Springer International Publishing.

SONNENBERG, C., AND BANNERT, M. 2015. Discovering the effects of metacognitive prompts on the sequential structure of SRL-processes using process mining techniques. *Journal of Learning Analytics, 2,* 72–100.

SPIRES, H. A. 2015. Digital game-based learning: What's literacy got to do with it? *Journal of Adolescent & Adult Literacy, 59,* 125–130.

SPIRES, H. A., ROWE, J. P., MOTT, B. W., AND LESTER, J. C. 2011. Problem solving and game-based learning: Effects of middle grade students' hypothesis testing strategies on learning outcomes. *Journal of Educational Computing Research, 44,* 453–472.

TSAI, M. J., HUANG, L. J., HOU, H. T., HSU, C. Y., AND CHIOU, G. L. 2016. Visual behaviour, flow and achievement in game-based learning. *Computers & Education, 98,* 115–129.

VOS, N., VAN DER MEIJDEN, H., AND DENESSEN, E. 2011. Effects of constructing versus playing an educational game on student motivation and deep learning strategy use. *Computers & Education, 56,* 127–137.

WHITE, B., FREDERIKSEN, J., AND COLLINS, A. 2009. The interplay of scientific inquiry and metacognition: More than a marriage of convenience. In D. Hacker, J. Dunlosky, & A. Graesser (Eds.), *Handbook of metacognition in education* (pp. 175–205). New York, NY: Routledge.

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

WINNE, P. H., AND AZEVEDO, R. 2014. Metacognition. In K. Sawyer (Ed.), *Cambridge handbook of the learning sciences (2nd ed.)* (pp. 63-87). Cambridge, MA: Cambridge University Press.

WINNE, P. H., AND BAKER, R. S. J. D. 2013. The potentials of educational data mining for researching metacognition, motivation, and self-regulated learning. *Journal of Educational Data Mining, 5,* 1–8.

WINNE, P., AND 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.

WINNE, P., AND HADWIN, A. 2008. The weave of motivation and self-regulated learning. In D. Schunk AND B. Zimmerman (Eds.), *Motivation and self-regulated learning: Theory, research, and applications* (pp. 297–314). Mahwah, NJ: Erlbaum.

WOUTERS, P., VAN NIMWEGEN, C., VAN OOSTENDORP, H., AND VAN DER SPEK, E. D. 2013. A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology, 105*, 249-265.

YANG, Y. T. C. 2012. Building virtual cities, inspiring intelligent citizens: Digital games for developing students' problem solving and learning motivation. *Computers & Education, 59,* 365–377.

YEH, Y. C., LAI, S. C., AND LIN, C. W. 2016. The dynamic influence of emotions on game-based creativity: An integrated analysis of emotional valence, activation strength, and regulation focus. *Computers in Human Behavior, 55,* 817–825.

ZIMMERMAN, B., AND SCHUNK, D. (EDS.) 2011. *Handbook of self-regulation of learning and performance.* New York, NY: Routledge.