Learning and Instruction 54 (2018) 93–103

Contents lists available at ScienceDirect

Learning and Instruction

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

Using sequence mining to reveal the efficiency in scientific reasoning during STEM learning with a game-based learning environment

Michelle Taub [a, *], Roger Azevedo [a], Amanda E. Bradbury [a], Garrett C. Millar [a], James Lester [b]

[a] Department of Psychology, North Carolina State University, 2310 Stinson Drive, Raleigh, NC 27695-7650, United States
[b] Department of Computer Science, North Carolina State University, 890 Oval Drive, Raleigh, NC 27695-8206, United States

ARTICLE INFO

Article history:
Received 28 December 2016
Received in revised form
12 June 2017
Accepted 23 August 2017
Available online 1 September 2017

Keywords:
Metacognition
Self-regulated learning
Scientific reasoning
Game-based learning
Sequence mining
Process data
Log files

ABSTRACT

The goal of this study was to assess how metacognitive monitoring and scientific reasoning impacted the efficiency of game completion during learning with Crystal Island, a game-based learning environment that fosters self-regulated learning and scientific reasoning by having participants solve the mystery of what illness impacted inhabitants of the island. We conducted sequential pattern mining and differential sequence mining on 64 undergraduate participants' hypothesis testing behavior. Patterns were coded based on the relevancy of what items were being tested for, and the items themselves. Results revealed that participants who were more efficient at solving the mystery tested significantly fewer partially-relevant and irrelevant items than less efficient participants. Additionally, more efficient participants had fewer sequences of testing items overall, and significantly lower instance support values of the PartiallyRelevant–Relevant to Relevant–Relevant and PartiallyRelevant–PartiallyRelevant to Relevant–Partially Relevant sequences compared to less efficient participants. These findings have implications for designing adaptive GBLEs that scaffold participants based on in-game behaviors.

1. Introduction

Self-regulated learning (SRL) is an effective way of learning for students of all ages (Azevedo, 2014; Winne & Azevedo, 2014). When students self regulate their learning, they are playing an active role in the learning process by engaging in different cognitive, affective, metacognitive, and motivational (CAMM) processes (Azevedo, Taub, & Mudrick, 2015). Research has shown that using different self-regulatory skills can enhance learning (Winne & Azevedo, 2014), however upon investigating how students use these skills in the classroom, research has revealed that students are often unsuccessful at self regulating their learning effectively and efficiently (Azevedo et al., 2015; Lester, Rowe, & Mott, 2013). As such, researchers have developed different types of advanced learning technologies (ALTs) designed to foster effective SRL (Azevedo et al., 2015; Biswas, Segedy, & Bunchongchit, 2016; Graesser, 2013; Lester et al., 2013).

One specific category of ALTs is game-based learning environments (GBLEs), which were designed to foster engagement and enjoyment during gameplay and learning (e.g., Crystal Island, Alien Rescue, Cache 17, iSTART). For this study, we investigated how participants used SRL and scientific reasoning processes (i.e., hypothesis testing) during learning with Crystal Island. We assessed hypothesis testing within the game, which we contextualized as testing food items for the possible transmission source of the mysterious illness that impacted the inhabitants of the island. In addition, we made the assumption that what was occurring between testing events involved SRL, specifically metacognitive monitoring processes and knowledge acquisition. This was contextualized within the game as time spent reading virtual books and posters for knowledge acquisition, time spent talking to non-player characters (NPCs) that could help further narrow down the transmission source, and frequency of tracking and monitoring food items being tested into a diagnosis worksheet. A major concern when assessing learning with GBLEs for scientific reasoning is the issue of efficiency in terms of participants choosing the relevant evidence, making appropriate inferences and hypotheses, and testing relevant evidence, while still enjoying the game. SRL researchers have not addressed this concern, nor have

* Corresponding author.
E-mail addresses: mtaub@ncsu.edu (M. Taub), razeved@ncsu.edu (R. Azevedo), aebradbu@ncsu.edu (A.E. Bradbury), gcmillar@ncsu.edu (G.C. Millar), lester@ncsu.edu (J. Lester).

http://dx.doi.org/10.1016/j.learninstruc.2017.08.005
0959-4752/

they investigated efficiency during gameplay. Therefore, for our analyses, we investigated the efficiency of scientific reasoning via hypothesis testing, and potential influences of SRL strategies (i.e., knowledge acquisition and monitoring processes) on the efficiency of hypothesis testing.

1.1. Theoretical framework

As our theoretical framework, we use the information processing theory (IPT) of self-regulated learning (Winne & Hadwin, 1998, 2008), which states that learning occurs through four cyclical stages: definition of the task, setting goals and plans, using learning strategies, and making adaptations to those goals, plans, and strategies, and, information processing, via the use of cognitive, metacognitive, and motivational SRL strategies, occurs during each stage. We use this model because each of these phases can relate to our study. In the first phase (definition of the task), students ensure that they are aware of what the task is asking them to do (e.g., solve the mystery of what disease has impacted all inhabitants by gathering clues from testing for the transmission source). In phase 2 (setting goals and plans), students set goals for how they plan to accomplish the task, as well as their plans for achieving those goals. For example, a student can set goals for gathering clues to help them solve the mystery, and their plans to do so would be to test different food items for the disease's transmission source, reading books to read about different diseases, including the symptoms associated with them, and talking to non-player characters who are sick patients and can report their symptoms, allowing students to match these symptoms to the ones they read about in the books, experts in microbiology who can tell them more about microbiology so they can narrow down if the disease is viral or bacterial, and workers on the island (i.e., camp cook) who could give more information about what food items inhabitants had been eating. The third phase, using learning strategies, involves students using cognitive and metacognitive strategies to enact the plans they set in the second phase. For this study, students could engage in cognitive learning strategies by reading information from the books and having conversations with non-player characters for knowledge acquisition about microbiology, and use metacognitive monitoring strategies to monitor the food items they were testing and their likelihood of being the disease's transmission source. It is important to note that knowledge acquisition is not an SRL strategy, however if students self-regulate their learning during knowledge acquisition, this can be beneficial because it can enhance knowledge acquisition by allowing students to actively acquire the information they need to complete the task. The fourth phase, making adaptations, involves the students adapting their goals, plans, and use of cognitive and metacognitive strategies based on their progression through the task. For example, students could decide to test all food items they found on the island, but after testing a large amount of items, they decide to only test those that a sick patient reported eating. Therefore, according to this model, students engage in self-regulated learning by using strategies related to monitoring and control, which allows them to actively pursue their goals and plans for accomplishing the task they are given.

This model is also particularly applicable to this study because it is the only model that views SRL as an event that unfolds in real time (Azevedo et al., 2010; Winne & Perry, 2000). For this study, we applied a sequence mining approach to examine specific events of food testing behaviors during learning via gameplay with Crystal Island. As such, we defined each testing event as an activity involving SRL and scientific reasoning, and examined sequences of how participants tested hypotheses to solve the mystery within the game.

When we study how learners use SRL strategies during learning, we should always include the context; i.e., what is required to complete the task itself (e.g., problem solving, scientific reasoning, etc.). With some GBLEs, learners must engage in scientific reasoning (or scientific inquiry), which involves using both theoretical and empirical bases for forming hypotheses that test science-related phenomena (White, Frederiksen, & Collins, 2009). As such, we used theories of SRL and scientific inquiry to investigate gameplay behaviors while learning during gameplay with Crystal Island, a GBLE that fosters SRL and scientific reasoning during learning about microbiology. For this study, we classified effective SRL as a strategic behavior, and therefore throughout the article, we refer to participants who are strategic or not strategic, which relates to their effective use of SRL processes.

1.2. Related work: research on SRL and GBLEs

Research on GBLEs has revealed not only that games are effective for learning, but has also provided guidelines for when games are the most effective. Mayer (2014) conducted a meta-analysis to investigate research comparing using games for teaching with teaching using traditional media devices (e.g., PowerPoint), as he states that research has shown that using games for teaching can be more effective. In doing so, Mayer took a four-step approach, where he did literature searches for the relevant papers, selected which papers fit the criteria for the meta-analysis, coded the experiments, and interpreted the results. Therefore, Mayer (2014) conducted this meta-analysis where he investigated different aspects (e.g., age group, content or subject, and type of GBLE) of GBLEs, and how these different types of GBLEs were found to impact learning, compared to traditional methods using media. Specifically, Mayer found that learning with games had the highest effect sizes for science and second language learning, whereas learning with games in math and language arts were found to be no better than using traditional teaching approaches (Mayer, 2014). In addition, adventure games were found to have the highest positive effects ($d = 0.72$), followed by simulations ($d = 0.62$), and quiz or puzzle games ($d = 0.45$); and games had the highest positive effects for adults or college students ($d = 0.74$), followed by secondary students ($d = 0.58$), and elementary students ($d = 0.45$). Therefore, based on this meta-analysis there is much promise for implementing games in classrooms using different domains and age groups.

Sequence mining is becoming an increasingly valuable analytical tool for assessing how students learn with ALTs, as during learning students can engage in multiple SRL strategies, and we seek to determine how their SRL unfolds over time. Studies using this approach have investigated overall performance (e.g., Baker & Corbett, 2014; Kinnebrew, Loretz, & Biswas, 2013), affect (e.g., Andres et al., 2015), and overall use of SRL skills (e.g., Bannert, Reimann, & Sonnenberg, 2014; Bouchet, Harley, Trevors, & Azevedo, 2013) during learning with various types of ALTs. Although the abovementioned studies have revealed the effectiveness of GBLEs for learning and SRL, few studies have aimed to use sequence mining to integrate how participants' scientific reasoning and inquiry, along with their metacognitive monitoring of SRL processes impacts their effectiveness in completing the games they are playing.

In addition to examining the processes of how students use SRL strategies, we must also examine *how efficiently* these processes are being used for a given task. Specifically, if students are told the overall goal of the game is to solve the mystery correctly, they might not feel it necessary to read all book content, especially content that will not be helpful for solving the mystery. In this case, the post-test might reveal a low score, however if the student solved the mystery correctly after one attempt, this can be indicative of efficient

behavior. Moreover, it is important to use monitoring strategies during activities such as hypothesis testing, but students will be selective on what items they are testing based on solving the mystery efficiently and will stop testing once they find the correct solution, and not to test all possible hypotheses.

Therefore, conducting research on learning with GBLEs is quite challenging as it requires balancing between efficiency and learning, such that an efficient learner might not reveal the highest learning gains, but did play the game as they were instructed to do so. Little research has aimed at addressing this balance between having full control while learning to be efficient with GBLEs and using accurate cognitive and metacognitive SRL strategies to investigate how students can use the appropriate strategies to result in efficient gameplay and learning. Thus, more research is needed to address what is missing in GBLE research to try to understand how SRL and scientific reasoning work together to enhance efficient learning, which is the aim of this study.

1.3. Current study

The goal of the current study was to determine if we could differentiate between efficient and less efficient participants in terms of hypothesis testing and SRL during gameplay with Crystal Island. We investigated amount of food items tested and their relevancies to solving the mystery as indicative of hypothesis testing, a key element of scientific reasoning. The relevancies of these items were indicative of participants engaging in monitoring processes as they selected the food items they wanted to test. Additionally, we assessed the number of diagnosis attempts because fewer attempts was indicative of efficiency, such that participants submitting the diagnosis on the first attempt were selectively monitored how they engaged in in-game activities.

1.3.1. Research questions and hypotheses

We investigated three research questions for this study: (1a) Does proportional learning gain differ between participants who solve the mystery more or less efficiently? (1b) Does the proportion of time spent testing food items, reading books, and talking to non-player characters (NPCs) differ between participants who solve the mystery more or less efficiently? (1c) Do the number of relevant, partially-relevant, and irrelevant food items differ between participants who solve the mystery more or less efficiently? (2) Are there frequent sequential patterns of testing for the transmission source of the illness? (3) Are there differential patterns of testing food items that are associated with efficiency in solving the mystery?

Subsequently, we hypothesized the following: (H1a): Participants who are more efficient at solving the mystery will have significantly higher proportional learning gains compared to participants who are less efficient at solving the mystery. (H1b): Participants who are more efficient at solving the mystery will spend significantly less proportions of time testing food items, reading books, and talking to NPCs Teresa and Quentin because they are more efficient with their time, compared to participants who are less efficient at solving the mystery. (H1c) More efficient participants will test significantly more relevant food items, and significantly fewer partially-relevant or irrelevant food items than participants who are less efficient at solving the mystery. (H2): There will be distinct sequential patterns for more and less efficient participants (i.e., there are no overlapping sequences) as the sequences are distinguishable based on efficiency. (H3): More efficient participants will obtain significantly higher instance support values (i.e., frequency of that sequence present per person) when testing relevant items (i.e., sequences with relevant items tested), and significantly lower instance support values when testing partially-relevant or irrelevant items (i.e., sequences with partially-relevant or irrelevant items). Specifically, more efficient participants will have higher instance support values for sequences with relevant codes and less efficient participants will have higher instance support values for sequences with partially-relevant or irrelevant codes.

2. Methods

2.1. Participants and materials

64[1] undergraduate students (59% female) from a large public North American university participated in this study. Participants' ages ranged from 18 to 26 years old ($M = 20$, $SD = 1.64$). They were randomly assigned to one of three experimental conditions (see section 2.3, Experimental Procedure), and were compensated $10 per hour for participating.

Prior to gameplay, participants completed a demographics questionnaire, followed by a series of self-report questionnaires asking them to report on their emotions and motivation. They also completed self-report questionnaires once they completed the game. Participants also completed a pre-test ($M = 55.6\%$, $SD = 2.77$ and post-test ($M = 68\%$, $SD = 2.69$, which were 21-item, four-choice multiple-choice tests on microbiology, with 12 factual and 9 procedural questions.

2.2. Crystal Island

Crystal Island is a narrative-centered game-based learning environment (GBLE) designed to foster self-regulated learning (SRL), scientific reasoning, and problem-solving skills (Rowe, Shores, Mott, & Lester, 2011). Participants experienced Crystal Island from a first-person perspective, where they arrived on a tropical island and were tasked with solving the mystery of what illness has spread and impacted the inhabitants of the island. Crystal Island (see Fig. 1) combines both inquiry learning and direct instruction, which allowed participants to gather clues and make inferences as they attempted to solve the mystery and discover its transmission source (e.g., pathogenic virus transmitted by eggs).

Participants explored multiple buildings where different books, research papers, posters, food items, and non-player characters (NPCs) are embedded to provide instruction and clues. In the infirmary, participants interviewed sick patients and interacted with Kim (the camp nurse), who provided pertinent information such as overall goals, background information, and indications of possible illness types and transmission sources. Specifically, Teresa (patient) informed participants what she had recently eaten. In the living quarters, participants conversed with microbiology experts and another patient and read more books and posters. In the dining hall, participants could collect more food items and speak with Quentin the cook, who also informed participants of food items inhabitants had been eating. There were food items in all buildings that participants could collect. Then, using information from the books, research papers, and posters, participants could make hypotheses of which items were most likely the transmission source of the illness, and then test these hypotheses in the laboratory.

2.2.1. Testing food items

Through interactions with game elements (e.g., reading books and posters, talking to NPCs), participants could create hypotheses regarding what food items were the most likely transmission

[1] This was a subsample of 94 participants. Only participants from two conditions were used due to limitations of one condition.

Fig. 1. Screenshots of Crystal Island scanning device (left) and diagnosis worksheet (right).

source. Participants were able to test these hypotheses by collecting and scanning gathered food items (see Fig. 1, left). Prior to scanning, participants had to specify why they selected that food item (i.e., sick members ate/drank it). The scanner then indicated whether the item was positive or negative for the selected illness type. Based on the results, the participant could confirm the transmission source and add their findings to the diagnosis worksheet

2.2.2. Diagnosis worksheet

Throughout their investigation, participants could track and organize pertinent information (e.g., symptoms, test results, and final diagnosis) via a diagnosis worksheet (see Fig. 1, right). The diagnosis worksheet supports problem-solving processes by providing a location for participants to offload gathered information (i.e., gathered clues as evidence), and later use this information to glean a final diagnosis, transmission source and treatment plan. Once participants felt they had correctly identified a diagnosis, transmission source, and treatment plan, they made their way back to the infirmary and presented their findings to Kim. If any part of the diagnosis was incorrect, Kim would provide specific feedback allowing them to reevaluate the incorrect portion or portions of their diagnosis. Once the participant correctly identified the illness type, transmission source, and treatment plan, the mystery was solved and the game ended.

2.3. Experimental procedure

Prior to gameplay, participants were randomly assigned to one of three experimental conditions, which varied based on the amount of agency they experienced. In the *full agency* condition, participants were free to play with no restrictions (i.e., could navigate to any building in any order, could read whichever books they wanted). In the *partial agency* condition, participants were required to follow a predetermined pathway (i.e., a set order of visiting the buildings via fast-track portal, which brought them from building to building without them having to navigate through the island themselves), and were required to interact with all artifacts in each location (i.e., had to read all books and posters, talk to all NPCs). In the *no agency* condition, participants did not play the game and instead watched someone play the game and narrate while he played. The *no agency* condition did not provide us with data on how participants interacted with the game (since they did not actually play and thus did not have log-file trace data), and therefore we did not include participants from this condition for this study. Furthermore, there were no restrictions on testing food items for participants in the other conditions, and we therefore included participants in both the *full* and *partial agency* conditions.

The study was conducted over a single session and lasted anywhere from one to two and a half hours depending on condition ($M = 87.39$ min, $SD = 20.8$ for the *full* and *partial agency* conditions). Participants were presented an informed consent form at the start of the experimental session. After signing the informed consent form, they received an overview of the study. They then put on electrodermal activity [EDA] bracelets, and were asked to complete pre-test measures including a demographics questionnaire, self-report measures about their perceptions of emotions and motivation, and the microbiology pre-test. Following the pre-test, the SMI EYERED 250 eye tracker was calibrated using a 9-point calibration. Following successful calibration, a baseline for the facial recognition of emotion software and the physiological bracelet were established using Attention Tool 6.3. Next, participants were given an overview, covering the game scenario, their role, and the importance of reading, interacting with NPCs, and scanning food items. Upon the game's completion, participants completed several self-report measures about emotions and motivation, and the microbiology content post-test. Participants were then debriefed, thanked and paid for their time.

2.4. Coding and scoring

When participants played Crystal Island, we collected the following multi-channel data: (1) log files, (2) eye tracking, (3) video of facial expressions, and (4) EDA. For this analysis, we only included log-file trace data, which captured all participant input into the game, such as selecting a food item to test. We focused on log files because this was our first attempt at using sequence mining to investigate efficiency in hypothesis testing, and log files are the most accurate data source for investigating efficiency (compared to eye tracking and videos of facial expressions of emotions) because they provide overt measures of student activity, and can be coded based on this overt behavior, without requiring inferences to be made that this behavior was observed. For example, we know the amount of attempts students made submitting their diagnosis worksheet based on their activity captured in the log files, whereas we would have to infer that a facial expression of frustration is a result of not getting the diagnosis correct. In addition, using additional data channels would require the use of multiple theoretical frameworks, and before doing so, we

wanted to first test sequence mining based on one theoretical framework only. Therefore, for this study, once all the log trace data were gathered, we coded and scored the data appropriately for our first attempt at conducting analyses using sequence mining to distinguish hypothesis testing behaviors based on level of efficiency.

2.4.1. Proportional learning gain and proportions of time testing, reading, and talking

To assess students' learning of microbiology, we used a proportional learning gain score using the following formula: $\frac{PostTestRatio - PreTestRatio}{1 - PreTestRatio}$. This formula allowed us to account for the amount of points gained in their post-test score in relation to their pre-test score. Therefore, our learning measure investigated learning of microbiology.

To calculate the proportions of time testing lab items, reading books, and talking to Teresa and Quentin (the NPCs that described food that had been eaten on the island), we extracted log-file trace data, which indicated the amount of time spent engaging in these activities, as well as the total session duration. We then calculated the proportions by dividing each activity's total duration by the total session duration, yielding three proportions. Calculating proportions allowed us to control for session duration, such that we could account for participants in the partial agency condition having longer durations. Based on the nature of the partial agency condition, participants had to spend longer time playing the game, and so calculating proportions accounted for this, and allowed us to control for longer durations.

2.4.2. Efficiency of solving the mystery

We grouped participants by their efficiency of solving the mystery in terms of the number of attempts they made to submit their diagnosis worksheet correctly. When submitting the worksheet, participants needed a correct diagnosis, transmission source, and treatment to complete the game. If any of these dimensions were incorrect, they were told the diagnosis was not correct, and they should try again. The log files recorded the number of diagnosis worksheet submission attempts ($M = 2.64$, $SD = 2.83$), and we then did a median split dividing participants into groups depending on the number of submissions they made. The median number of diagnosis worksheet submission attempts was 2, and so participants with two attempts or higher were in the 'more' group, and participants with 1 attempt were in the 'once' group. 31 participants submitted their diagnosis worksheet correctly on the first attempt ('once' group) and 33 participants submitted their diagnosis worksheet correctly after more than one attempt ('more' group), with number of attempts ranging from 2 to 16. As such, participants in the 'once' group were classified as solving the mystery more efficiently than participants in the 'more' group.

2.4.3. Full relevancy code

We created a full relevancy code based on the relevancies of two components of the food items being tested in the lab. Specifically, we coded the relevancy of what the item was being tested for, and the relevancy of the item itself. When testing an item, participants could choose to test for a virus, bacterium, carcinogen, or mutagen, however the correct response could only be a virus or bacterium (the specific correct test was randomly assigned at the beginning of gameplay). Although there was only one solution, other lab items could test positive for nonpathogenic viruses or bacteria. As such, the correct response according to the solution was assigned a relevant code, whereas the response that was not correct, but could still test positive for nonpathogenic substances was coded as partially-relevant. Carcinogens and mutagens were coded as irrelevant as they were never possible solutions. Additionally, we coded for the relevancy of the item being tested, which could also be relevant, partially-relevant, or irrelevant. There were many items participants could test, however the only relevant items were those reported by Teresa the patient that she had eaten (eggs, milk, or bread). Partially-relevant items were those that could be tested positive for a nonpathogenic virus or bacterium (e.g., apple, water, banana, orange, etc.), and irrelevant items were not eaten by anyone on the island, nor were they potential sources of nonpathogenic substances (e.g., peanuts or jelly). For example, if the solution was a pathogenic virus spread by milk, testing the milk for a virus would yield a Relevant–Relevant code, testing for bacteria would yield a Partially-Relevant–Relevant code, and testing for carcinogens or mutagens would yield an Irrelevant–Relevant code. All possible combinations of the codes yielded nine unique codes (see Table 1), and we used these codes to determine if there were patterns of testing food items based on these codes using sequential pattern mining.

2.4.3.1. Sequential pattern mining. Sequential pattern mining is a technique that examines if there are distinct sequences of a given event, which can be defined as a pre-determined behavior or activity (e.g., testing lab items, fixating on different areas on the screen during reading). We used the full relevancy codes to detect sequential patterns of how participants tested food items during gameplay with Crystal Island, where each event corresponded to testing one food item. Thus, we examined sequences corresponding to multiple events of testing food items. For example, if the solution was a pathogenic bacterium spread by eggs, and a participant first tested the eggs for a pathogenic virus, and then tested the eggs for a pathogenic bacterium, this would be coded as two testing events. The sequential pattern would be: PartiallyRelevant–Relevant → Relevant–Relevant (or 2 → 1). Therefore, based on all of the food items that participants tested, we were able to examine for common testing events across participants in both the efficient and non-efficient mystery solving groups.

2.4.3.2. Differential sequence mining. Differential sequence mining can be applied to test if the sequences obtained from sequential pattern mining have higher frequencies of occurrences in one group compared to another (Kinnebrew et al., 2013). For our analysis, we compared the sequences of testing food items by their relevancy codes (see section 2.4.3: Full relevancy code) between participants who solved the mystery more (i.e., 'once' group) or less (i.e., 'more' group) efficiently. To do so, we calculated an instance support value using brute force (Grafsgaard, 2014). An instance support value is the frequency that the sequence occurred within each individual. Thus, we calculated the instance support value for each participant within the 'once' group and the 'more' group, so we could determine if there were significant differences in these instance support values between the two groups.

Table 1
Descriptions of full relevancy codes.

Code	Description
1	Relevant–Relevant
2	Relevant–PartiallyRelevant
3	Relevant–Irrelevant
4	PartiallyRelevant–Relevant
5	PartiallyRelevant–PartiallyRelevant
6	PartiallyRelevant–Irrelevant
7	Irrelevant–Relevant
8	Irrelevant–PartiallyRelevant
9	Irrelevant–Irrelevant

3. Results

3.1. Research question 1: Does proportional learning gain(a), the proportion of time spent testing food items, reading books, and talking to NPCs (b), and the number of relevant, partially-relevant, and irrelevant food items tested (c) differ between participants who solve the mystery more or less efficiently?

For research question 1a, we ran an independent samples *t*-test with proportional learning gain as the dependent variable and DW group (diagnosis worksheet group; submitting the diagnosis worksheet once or more than once) as the independent variable. Results revealed a non-significant effect; $t(62) = -1.18$, $p = 0.25$, $d = 0.30$, revealing there were no significant differences in proportional learning gain between participants who were more efficient at solving the mystery ($M = 0.22$, $SD = 0.25$) and participants who were less efficient at solving the mystery ($M = 0.30$, $SD = 0.33$).

For research question 1b, we ran an independent samples *t*-test with proportions of time spent testing food items, reading books, and talking to NPCs who reported what inhabitants were eating (Teresa and Quentin) as the three dependent variables, and DW group as the independent variable. Results did not reveal a significant effect for the proportion of time spent testing lab items; $t(62) = -1.78$, $p = 0.078$, $d = 0.45$, the proportion of time reading books; $t(62) = 0.48$, $p = 0.63$, $d = 0.12$, or the proportion of time talking to Teresa and Quentin; $t(62) = 1.28$, $p = 0.21$, $d = 0.32$. Specifically, there were no significant differences in the proportion of time spent testing food items between more ($M = 0.020$, $SD = 0.0086$) and less efficient ($M = 0.025$, $SD = 0.013$) participants, no significant differences in the proportion of time spent reading books between more ($M = 0.33$, $SD = 0.073$) and less efficient ($M = 0.32$, $SD = 0.090$) participants, and no significant differences in the proportion of time spent talking to the NPCs Quentin and Teresa between more ($M = 0.023$, $SD = 0.0061$) and less efficient ($M = 0.021$, SD = 0.0052) participants.

In addition, we conducted correlations between the proportions of time testing food items, reading books, and talking to non-player characters Teresa and Quentin, and the number of diagnosis worksheet attempts. Results revealed a significant negative association between proportion of time talking to NPCs Quentin and Teresa and proportion of time spent reading books; $r(62) = -0.33$, $p < .01$, such that the more time participants spent talking to Quentin and Teresa, the less time they spent reading books. Results also revealed a significant negative association between proportion of time testing food items and proportion of time spent reading books; $r(62) = -0.48$, $p < .01$, such that the more time participants spent testing food items, the less amount of time they spent reading books. Finally, results revealed a significant positive association between the proportion of time spent testing food items and the number of diagnosis worksheet submissions; $r(62) = 0.32$, $p = 0.011$, such that the more time participants spent testing food items, the greater number of diagnosis worksheet submission attempts they made. Table 2 displays all the results from the correlation. In sum, these results reveal that the only variable significantly correlated with the number of diagnosis worksheet submissions was the proportion of time testing food items. All other proportion variables (testing, reading, and talking) were correlated with each other, however the proportion of time spent reading books and talking to non-player characters Quentin and Teresa were not significantly associated with the number of diagnosis worksheet submission attempts.

Finally, for research question 1c, we conducted an independent samples *t*-test with the number of relevant items tested, number of partially-relevant items tested, and number of irrelevant items tested as our dependent variables, and diagnosis worksheet group as the independent variable. Results revealed a non-significant effect for number of relevant items; $t(62) = -1.38$, $p = 0.17$, $d = 0.35$; however there were significant effects for number of partially-relevant items; $t(54.331) = -2.54$, $p = .014$, $d = 0.63$ (Levene's test for equality of variance was violated, and so we report the corrected degrees of freedom and *t*-test results), and number of irrelevant items; $t(62) = -2.28$, $p = 0.026$, $d = 0.57$. Specifically, there were no significant differences in the number of relevant food items tested between more ($M = 7.68$, $SD = 3.68$) and less efficient ($M = 9.09$, $SD = 4.43$) participants. However there were significant differences for partially-relevant food items tested, such that participants who were more efficient tested significantly fewer partially-relevant food items ($M = 10.9$, $SD = 7.28$) than less efficient participants ($M = 17$, $SD = 11.58$), and significant differences for irrelevant food items tested, such that participants who were more efficient tested significantly fewer irrelevant food items ($M = 2.65$, $SD = 3.20$) than less efficient participants ($M = 4.64$, $SD = 3.75$).

Overall, these results suggest that although there were no significant differences in proportional learning gain, the proportion of time spent testing, reading, or talking, or in the amount of relevant food items tested between groups, participants who solved the mystery less efficiently tested more partially-relevant and irrelevant food items, and correlations revealed that only the proportion of time testing food items was positively associated with submitting the diagnosis worksheet, which might explain why they tested more partially-relevant and irrelevant food items than participants who solved the mystery more efficiently. In addition, as proportions of time spent reading books and talking to non-player characters were not significantly associated with number of times submitting the diagnosis worksheet, we focused solely on food testing behavior for our subsequent research questions.

3.2. Research question 2: Are there frequent sequential patterns of testing for the transmission source of the illness for efficiency groups?

We ran the sequential pattern mining algorithm SPAM (Ayres, Flannick, Gehrke, & Yiu, 2002; Fournier-Viger, Gomariz, Campos, & Thomas, 2014) to detect sequential patterns of food items tested by their relevancy (see 2.4.3.1 in Coding and scoring), along with their support values, for each group. The SPAM algorithm generates

Table 2
Correlations of proportions of time testing, reading, and talking and DW submissions.

	DW Submissions	Prop. Testing	Prop. Reading	Prop. Talking
DW Submissions	–	0.32*	-0.23	0.18
Prop. Testing		–	-0.48**	0.034
Prop. Reading			–	-0.33**
Prop. Talking				–

**$p < .01$, *$p < .05$.
Note. DW submissions = number of times submitted diagnosis worksheet, prop. Testing = proportion of time spent testing food items, prop. Reading = proportion of time spent reading books, prop. Talking = proportion of time spent talking to non-player characters Teresa and Quentin.

sequential patterns of activities at a minimum support level of 50% (i.e., pattern must be found in at least 50% of participants). We examined sequential patterns, with segments of two, three, and four codes, for food items tested for more and less efficient participants separately, at a support value of both 50% and 90% (i.e., pattern occurring in at least 16 or 28 participants). We selected up to four-code segments due to the nature of testing food items, where participants had four options for what they were testing for (i.e., virus, bacterium, carcinogen, or mutagen).

Overall (see Table 3), more efficient participants (i.e., who submitted their diagnosis worksheet correctly on the first attempt) appeared to have fewer sequential patterns (at the 50% and 90% thresholds) of testing food items (regardless of relevancy) compared to less efficient participants (i.e., who submitted their diagnosis worksheet correctly after more than one attempt). Additionally, out of all sequential patterns, 94 of the 2- or 3-coded sequences occurred 50% of the time for participants in both groups (see Table 3), and 4 of the 4-coded sequences occurred 50% of the time for participants in both groups (see Table 3). Additionally, more efficient participants showed lower support values for all codes compared to less efficient participants, however both groups showed similar top five sequences (see Fig. 2). It is important to note that when investigating the 4-coded segments, we examined unique sequences because a repeated code within a sequence would be identified in a 2- or 3-coded sequence as well, thus not being a unique pattern. As such, results revealed lower support values overall for the 4-coded sequences, but higher support values for less efficient participants (see Fig. 3).

Overall, these results suggest that we were able to find common sequences of food testing behavior across groups, however there were far more sequential patterns that were not found in both groups, suggesting that some sequential patterns of food testing behavior might contribute to more effectively and efficiently solving the mystery than other sequential patterns.

3.3. Research question 3: Are there differential patterns of testing food items that are associated with efficiency in solving the mystery?

Differential sequence mining (see 2.4.3.2 in Coding and scoring) was applied to compare instance support values (i.e., frequencies of discovered patterns) for sequential patterns of testing food items (as determined in research question 2) between participants more or less efficient at solving the mystery. We used instance support values as our dependent variables, and conducted *t*-tests to compare these instance support values between the two groups. We selected 2- and 3-coded sequences as our dependent variables based on the highest mean instance support values. Results (see Table 4) revealed that for the six 2-coded sequences, there were two significant effects; one for the 4→1 code; $t(62) = -2.19$, $p = 0.032$, $d = 0.55$, and one for the 5→2 code; $t(62) = -2.28$, $p = 0.026$, $d = 0.57$. Specifically, participants who submitted their diagnosis worksheet correctly after one attempt (more efficient participants) had significantly lower frequencies of the PartiallyRelevant–Relevant → Relevant–Relevant sequence and PartiallyRelevant--PartiallyRelevant → Relevant–PartiallyRelevant sequence than less efficient participants. Additionally, results comparing the four 3-coded sequences (see Table 3) revealed that there were no significant differences in instance support values between participants who solved the mystery more or less efficiently.

Overall, these results suggest that there are differences in how participants who solved the mystery more efficiently tested lab items compared to participants who were less efficient. Although there were not significant differences between all patterns, the significant differences that we did find can be used to develop different types of scaffolding for different types of participants.

Table 3
Pattern mining outcome sequences by DW group.

	2-3-coded sequences			4-coded sequences		
	>50%	>90%	Same	>50%	>90%	Same
Once	109	5	94	56	0	4
More	256	27		441	3	

Note. >50% = frequency of sequences occurring in more than 50% of participants, >90% = frequency of sequences occurring in more than 90% of participants, Same = frequency of common sequences occurring in more than 50% of participants in both groups.

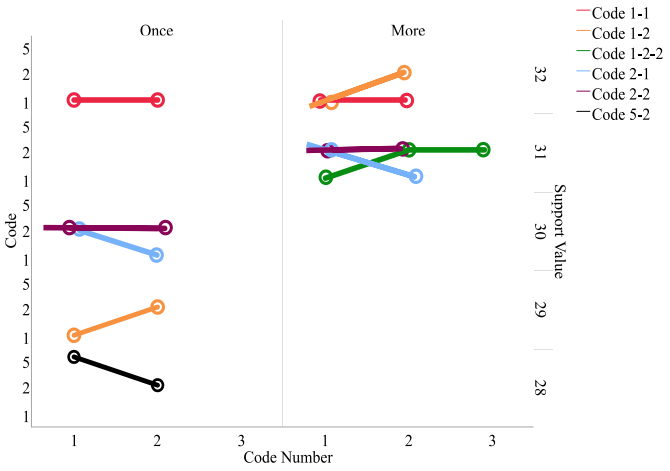

Fig. 2. 2-3-coded sequences with support values by DW group.
Note. Code Number = the location of the code within the sequence (1 = first code for that segment), Code = assigned code, which does not have a weighted value (i.e., code 5 is not a higher ranked code than code 1 or 2), Support Value is the support value for that code.

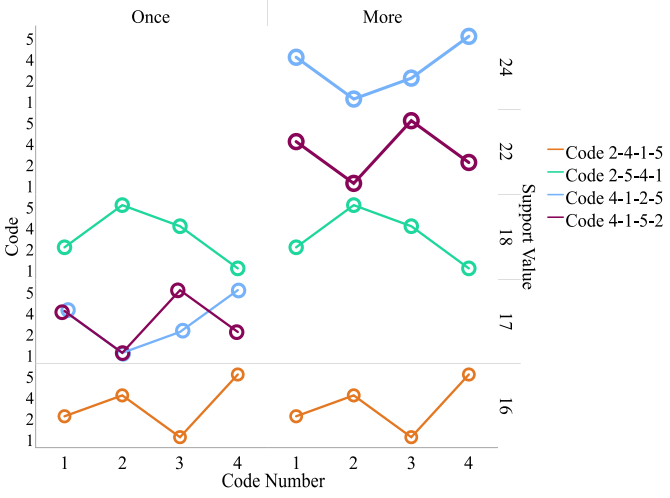

Fig. 3. 4-coded sequences with support values by DW group.
Note. Code Number = the location of the code within the sequence (1 = first code for that segment), Code = assigned code, which does not have a weighted value (i.e., code 5 is not a higher ranked code than code 1, 2, or 4), Support Value is the support value for that code.

4. Discussion

In this study, we addressed a major concern in SRL research by

Table 4
Differential sequence mining with 2- and 3-coded segments by DW group.

	Once		More		*t*-test
	M	*SD*	*M*	*SD*	
1→4	1.03	0.98	1.24	1.12	-0.80
4→1	0.58	0.67	1.06	1.029	-2.19*
2→5	1.97	2.16	2.33	2.56	-0.62
5→2	1.42	1.77	2.61	2.34	2.28*
1→2	0.74	0.73	0.70	0.85	0.23
2→1	0.68	0.75	0.58	0.71	0.56
1→4→1	0.16	0.37	0.27	0.45	-1.077+
2→5→2	0.68	1.22	0.64	0.86	0.16
2→5→8	0.58	1.025	0.55	1.12	0.13
5→2→5	0.68	1.35	0.73	1.008	0.17

**$p < .01$, *$p < .05$.
+ = Levene's test of equality of variances violated, corrected values reported.
Note. 1 = Relevant–Relevant, 2 = Relevant–PartiallyRelevant, 4 = PartiallyRelevant–Relevant, 5 = PartiallyRelevant–PartiallyRelevant, 8 = Irrelevant–PartiallyRelevant. Each sequence is a combination of two unique codes.

investigating the efficiency of SRL and scientific reasoning during gameplay with GBLEs using unobtrusive online trace methods and both traditional statistics along with data mining techniques. Results from our analyses indicated that we can investigate SRL and scientific reasoning, via hypothesis testing during gameplay with GBLEs to determine how efficiently participants completed the game and solved the mystery of what illness impacted the inhabitants of Crystal Island. Specifically, our research questions gave insight into the specific differences between efficient and non-efficient participants in terms of how they tested relevant, partially-relevant, and irrelevant food items to determine the transmission source of the illness. Below we discuss the results from each research question in greater detail.

4.1. Discussion of findings

Our first research question revealed that while there were no significant differences in proportional learning gain, or the proportions of time spent testing food items, reading books, or talking to the non-player characters Teresa and Quentin, less efficient participants tested more partially-relevant and irrelevant food items, but not relevant food items, compared to more efficient participants. In addition, proportion of time spent testing was significantly positively correlated with the number of diagnosis worksheet submissions, but proportions of time spent reading books and talking to Teresa and Quentin were not. This partially supports H1 since we predicted that less efficient participants would test more partially-relevant and irrelevant food items, however we also predicted smaller proportional learning gains, and longer proportions of time spent testing food items, reading books and talking to NPCs for less efficient participants, which we did not find, therefore not supporting H1 in its entirety. These results suggest that testing significantly more partially-relevant and irrelevant food items might be what is causing these participants to be less efficient. Furthermore, there were no significant differences in testing relevant food items between groups, which reveals that less efficient participants are not testing fewer relevant items, they are just not spending their time efficiently and in addition to testing relevant food items, they are also testing irrelevant food items. As such, they are testing items that do not need to be tested, which leads us to believe these participants are less efficient game players. Furthermore, not finding a significant difference in proportional learning gain further emphasizes the balance between learning and efficiency (see Section 1.2), When investigating learning with a GBLE, there is a balance between learning and efficiency, where solving the mystery correctly and quickly does not necessarily equate to learning everything about microbiology. This means that participants who did solve the mystery after one attempt might not have read about all the content in the post-test, meaning they did play the game efficiently, however they did not read all the content, resulting in no significant differences in proportional learning gain based on levels of efficiency of solving the mystery.

These findings align with the IPT model of SRL (Winne & Hadwin, 1998, 2008) for we can assume that less efficient students are engaging more in phase 3 (using learning strategies) and less in phases 2 (setting goals and plans) and 4 (making adaptations), thus not navigating through the entire SRL cycle to engage in effective SRL. Specifically, it appears as though less efficient students are not planning and monitoring their testing behavior, and are simply testing more items without a strategic approach. Furthermore, there was a significant positive correlation between number of times submitting the diagnosis worksheet and the proportion of time spent testing food items (and not reading books or talking to NPCs), which demonstrates that the more time participants spent testing food items, the more submissions they made, perhaps revealing that more time testing food items is indicative of guessing behaviors, and therefore less efficient gameplay. As such, more efficient students are, in fact, more efficient because they are monitoring their hypothesis testing behavior by ensuring they are testing plausible hypotheses, and are not testing all available food items. However, further research is needed to investigate if students' prior knowledge of microbiology and scientific reasoning might have impacted their monitoring, and how other multi-channel data can provide evidence of accurate monitoring of hypothesis testing behavior. These results relate to previous work investigating SRL during learning with ALTs as previous work done (Basu et al., 2016; Sabourin, Mott, & Lester, 2013) have revealed that the use of more SRL processes leads to better gameplay and problem solving, and our study revealed that using more monitoring processes leads to more efficient gameplay behavior.

Results from our second research question revealed that there were sequential patterns of testing food items for both efficient and less efficient participants, demonstrating that we can more thoroughly investigate the process of how participants hypothesis test by testing food items during gameplay. Specifically, these sequences revealed that no participants in the efficient group tested sequences of irrelevant food items, and rarely tested for carcinogens or mutagens (irrelevant testing options), which may demonstrate why these participants were more efficient than the less efficient group. These results partially support H2 as we did find distinct sequences of food testing, which we predicted, but we also found some overlap between efficiency groups (i.e., sequences occurring in both groups), which we did not predict, thus partially supporting our hypothesis. From this result we can infer that more efficient participants were more strategic in what food items they tested, such that they had fewer sequences of food testing behavior overall, which suggests that they were trying to strategically play the game and were not trying to guess or game the system. In contrast, less efficient participants seemed to have been testing all food items for all testing options, as evidenced by a larger number of sequences. As such, these participants were less efficient because they were not strategically testing items, but were guessing and testing all of the options. This might be especially true for 4-coded sequences, as this might have been indicative of testing the same food items for each of the four options in the scanner, implying they were guessing for the cause of the illness, as well as its transmission source. Finally, we did find higher support values for less efficient participants, which we might be able to attribute to the fact that there were so many sequences there was bound to be some overlap

between participants, however this remains an unanswered interpretation, which requires further investigation.

Results from the sequential pattern mining align with the IPT model in a similar way as the first research question (i.e., less efficient participants spending more time in phase 3), however the overlap of sequences across more and less efficient participants reveals that all participants can monitor to some extent. This demonstrates the importance of engaging in SRL strategies, but while also being efficient and not using SRL processes without knowing how efficiently to use them. Additionally, these results once again align with previous work demonstrating the important use of SRL strategies during learning (Basu et al., 2016; Sabourin et al., 2013), and also the benefits of using sequence mining to measure SRL (Azevedo, 2014, 2015; Bannert et al., 2014; Bouchet et al., 2013; Winne & Baker, 2013), for using sequential pattern mining revealed some different sequences, but some similar sequences across groups. Specifically, using traditional statistical techniques revealed that there were significant differences between efficiency groups, but these results do not inform us that there are some similarities as well.

Finally, our third research question revealed that there were some significant differences in the instance support values between efficient and less efficient participants during gameplay. Specifically, less efficient participants had significantly higher instance support values for the sequences PartiallyRelevant–Relevant → Relevant–Relevant (i.e., 4→1) and PartiallyRelevant–PartiallyRelevant → Relevant–PartiallyRelevant (i.e., 5→2). However, there were no other significant differences between groups, which may be reflected in the fact that participants tested equal numbers of relevant food items, and all participants did eventually solve the mystery, thus there are some patterns that are similar, but it might be the ones that are different that are differentiating participants. As such, this partially supports H3 because we did find significant differences between efficiency groups, however only for two sequences. The 4→1 sequence suggests that participants were testing the same relevant food item for both a virus and a bacterium, revealing that participants might have had a harder time selecting what to test the food item for (i.e., selecting between a virus and a bacterium). The 5→2 sequence suggests that participants were testing the same pattern as the 4→1 sequence, however in this case, they were testing for a partially-relevant food item for both a virus and bacterium, suggesting again that they were not able to discern what exactly they should have been testing for. This partially supports H3 because we expected less efficient participants to test the 5→2 sequence more frequently as it was not a relevant food item, however we expected more efficient participants to test relevant food items, and the 4→1 sequence is a more efficient pattern than the 5→2 sequence since it includes relevant food items, but was still found more frequently for less efficient participants.

Once more, these results align with the IPT model in terms of differentiating between the amount of time spent in the third phase of the SRL cycle leading to more or less efficient participants, and how important it is for participants to monitor their hypothesis testing behaviors during learning. Additionally, results again demonstrate the importance of efficient hypothesis testing in terms of knowing which hypotheses to test, and not testing all possible hypotheses, such that we need that balance between using SRL processes while engaging in efficient testing behaviors. These results provide the same alignment with previous research as the second research question however not only can we identify differences and similarities in testing behaviors between efficiency groups, using differential sequence mining allows us to examine for statistical significance of these results, revealing which specific sequential patterns are enacted more often in one group compared to another (e.g., Kinnebrew et al., 2013).

Overall, these results revealed that there are different types of participants who play Crystal Island, and by being able to differentiate and identify these types of participants, we can move toward developing adaptive GBLEs that scaffold participants based on their gameplay behaviors.

4.2. Limitations

Although our results revealed promising advances for investigating scientific reasoning through hypothesis testing within GBLEs, there are several limitations that we must acknowledge. First, when differentiating between efficient and less efficient participants, we categorized the amount of diagnosis worksheet submission attempts by conducting a median split, where the median was 2 attempts, categorizing participants in either the 'once' group or the 'more' group. Therefore, participants in the less efficient group had a large range of submission attempts, whereas the more efficient group only had one number of attempts, resulting in a larger range for the less efficient participants. Furthermore, to assess learning, we examined participants' proportional learning gains from pre-test to post-test of their scores on the microbiology content test, and therefore did not investigate participants' proficiency and learning of scientific inquiry and hypothesis testing. Therefore, future studies should include not only domain knowledge tests, but also knowledge and skills regarding their self-regulated learning and scientific inquiry to examine if they learned about using scientific inquiry processes effectively. In addition, as this was our first attempt at using sequence mining and classifying participants by efficiency, we only used log-file trace data in our analyses and did not include data from other channels, such as eye tracking, videos of facial expressions, and physiological data, which could have revealed other differences between efficient and non-efficient participants. Using log files was the most reliable data source as it allowed us to examine overt behavior (e.g., submitting the diagnosis worksheet), however participants' levels of emotions could have impacted their hypothesis testing behavior as well. As such, converging multi-modal multi-channel data are likely to address these limitations to gain a fuller understanding of how participants are efficient or not during gameplay with Crystal Island (see Azevedo, Taub, & Mudrick, in press).

4.3. Implications and future directions

Overall, the findings from this study have important implications for learning in many different types of environments and adapting to different types of learners who use these environments. In addition, these findings reveal many applications of sequence mining within one domain, and in many other domains. For example, we can investigate sequences of in-game activities in terms of relevancy (as was done in this study), or patterns of time spent engaging in different activities. Moreover, we can investigate attention allocation by examining participants' eye-tracking sequential patterns and determine which areas on the screen they fixate on and whether they display patterns of fixating on these areas in sequential order.

Finally, this research can be applied to all educational research to investigate how students at all age levels are learning using different types of ALTs, such as GBLEs, intelligent tutoring systems, hypermedia, simulations, etc., and how we can foster effective SRL for these students. Effective SRL requires the use of complex cognitive, affective, metacognitive, and motivational (CAMM) processes (Azevedo et al., 2015), and the accuracy of how students use these processes (e.g., Gutierrez, Schraw, Kuch, & Richmond, 2016). Specifically, we can examine how students read, plan by setting

sub-goals, assess their understanding of content, feel confusion or frustration, or lack task interest, etc. as they complete a task. We can code them as one of the CAMM processes and then examine the sequences of processes they engage in, such as: [feel confused → judge understanding (JOL) as not understanding (i.e., JOL-) → re-read], which is an effective sequence of strategies because they are aware that they do not understand, and then go back to try and read the material again. Therefore, using sequence mining to determine the most efficient sequences of using these processes can be beneficial for teaching students how to accurately self-regulate their learning.

4.3.1. Future directions

There are many future directions for conducting this research, which involve using sequence mining with multi-channel data to investigate cognitive, affective, metacognitive, and motivational SRL processes using additional theoretical frameworks including emotions and motivation. For example, to investigate emotions, this research might be relevant for science, technology, engineering, and mathematics (STEM) education for younger students who are likely to lack the self-regulatory knowledge and skills as well as lack the proficiency in scientific reasoning and hypothesis generation while using games for learning. Such learning will elicit negative emotions that could interfere with their performance (e.g., experiencing extreme confusion when considering alternative evidence). As such, determining the sequences of emotions (or increases in the expressivity of these emotions) that lead to these undesirable emotions can be used to train students how to regulate these emotions using effective regulation strategies, such as cognitive change.

4.3.2. Challenges for future research

In conducting future studies, there are many challenges we face that must be addressed in order to advance research in SRL with ALTs. For example, to conduct our differential sequence mining in this study, we selected instance support values based on the highest averages of sequences, however there could have been differences between other sequences that we did not investigate. Furthermore, when selecting the lengths of the sequences, the shorter sequences (i.e., 2-coded sequences) yielded the significant results instead of the longer sequences (i.e., 3-coded sequences). As such, this lead us to wonder about the value of including longer sequences in our analyses, and whether this is context specific. As such, future studies should also investigate which of these sequences are predictive (i.e., using linear regressions, logistic regressions, or multi-level modeling) of the efficiency of testing food items and solving the mystery efficiently.

Our goal in educational research is to ensure students are learning in the most effective way, and are learning to use the appropriate learning strategies. One approach in doing so is to develop adaptive learning environments, such as GBLEs, which cater to each student's individual learning needs, as teachers are not always accurate at judging their students' performance (Gabriele, Joram, & Park, 2016). For example, if we can predict gameplay behaviors early on, we can provide adaptive scaffolding based on these behaviors, and as such, predict whether participants will be successful or not at completing the game. As such, there are many challenges in designing adaptive learning environments, such as GBLEs, and future research should seek to investigate how we can address these issues to provide the most effective learning experiences for participants learning with these adaptive environments.

Acknowledgements

This study was supported by funding from the Social Sciences and Humanities Research Council of Canada (SSHRC 895-2011-1006). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of SSHRC.

The authors would also like to thank Nicholas Mudrick and Megan Price from the SMART lab at NCSU, and Robert Taylor, Andy Smith, and Robert Sawyer from the Intellimedia Group at NCSU for their assistance.

References

Andres, J. M. L., Rodrigo, M. M. T., Baker, R. S., Paquette, L., Shute, V. J., & Ventura, M. (2015, June). Analyzing student action sequences and affect while playing Physics Playground. In *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).

Azevedo, R. (2014). Multimedia learning of metacognitive strategies. In R. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 647–672). Cambridge, MA: Cambridge University Press.

Azevedo, R. (2015). Defining and measuring engagement and learning in science: Conceptual, theoretical, methodological, and analytical issues. *Educational Psychologist, 50*, 84–94.

Azevedo, R., Johnson, A., Burkett, C., Fike, A., Lintean, M., Cai, Z., & Rus, V. (2010). The role of prompting and feedback in facilitating students' learning about science with MetaTutor. In R. Pirrone, R. Azevedo, & G. Biswas (Eds.), *Proceedings of the AAAI Fall Symposium on Cognitive and Metacognitive Educational Systems* (pp. 11–16). Menlo Park, CA: AAAI Press.

Azevedo, R., Taub, M., & Mudrick, N. (2015). Technologies supporting self-regulated learning. In M. Spector, C. Kim, T. Johnson, W. Savenye, D. Ifenthaler, & G. Del Rio (Eds.), *The SAGE Encyclopedia of educational technology* (pp. 731–734). Thousand Oaks, CA: SAGE.

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

Ayres, J., Flannick, J., Gehrke, J., & Yiu, T. (2002). Sequential pattern mining using a bitmap representation. In D. Hand, D. Keim, & R. Ng (Eds.), *Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 429–435). New York, NY: ACM.

Baker, R. S., & Corbett, A. T. (2014). Assessment of robust learning with educational data mining. *Research & Practice in Assessment, 9*, 38–50.

Bannert, M., Reimann, P., & Sonnenberg, C. (2014). Process mining techniques for analysing patterns and strategies in students; self-regulated learning. *Metacognition and Learning, 9*, 161–185.

Basu, S., Biswas, G., Sengupta, P., Dickes, A., Kinnebrew, J. S., & Clark, D. (2016). Identifying middle school students' challenges in computational thinking-based science learning. *Research and Practice in Technology Enhanced Learning, 11*, 1–35.

Biswas, G., Segedy, J. R., & Bunchongchit, K. (2016). From design to implementation to practice—a learning by teaching system: Betty's brain. *International Journal of Artificial Intelligence in Education, 26*, 350–364.

Bouchet, F., Harley, J., Trevors, G., & 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.

Fournier-Viger, P., Gomariz, A., Campos, M., & 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, & H. Kao (Eds.), *Proceedings of the 18th Pacific-asia conference on knowledge discovery and data mining* (pp. 40–52). Cham, Switzerland: Springer.

Gabriele, A. J., Joram, E., & Park, K. H. (2016). Elementary mathematics teachers' judgment accuracy and calibration accuracy: Do they predict students' mathematics achievement outcomes? *Learning and Instruction, 45*, 49–60.

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).

Gutierrez, A. P., Schraw, G., Kuch, F., & Richmond, A. S. (2016). A two-process model of metacognitive monitoring: Evidence for general accuracy and error factors. *Learning and Instruction, 44*, 1–10.

Kinnebrew, J. S., Loretz, K. M., & 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., & Mott, B. W. (2013). Narrative-centered learning environments: A story-centric approach to educational games. In C. Mouza, & N. Lavigne (Eds.), *Emerging technologies for the classroom: A learning sciences perspective* (pp. 223–238). Amsterdam, The Netherlands: Springer.

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

Cambridge, MA: MIT Press.
Sabourin, J., Mott, B., & Lester, J. (2013). Discovering behavior patterns of self-regulated learners in an inquiry-based learning environment. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Proceedings of the 16th International Conference on Artificial Intelligence in Education—Lecture Notes in Artificial Intelligence 7926* (pp. 209–218). Berlin, Heidelberg: Springer-Verlag.
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.
White, B., Frederiksen, J., & 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., & 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., & 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., & Hadwin, A. (1998). Studying as self-regulated learning. In D. Hacker, J. Dunlosky, & A. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 227–304). Mahwah, NJ: Erlbaum.
Winne, P., & Hadwin, A. (2008). The weave of motivation and self-regulated learning. In D. Schunk, & B. Zimmerman (Eds.), *Motivation and self-regulated learning: Theory, research, and applications* (pp. 297–314). Mahwah, NJ: Erlbaum.
Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In M. Boekaerts, P. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 531–566). Orlando, FL: Academic Press.