ABSTRACT


Overall, the goal of this dissertation is to use three manuscripts to demonstrate why we need to use multi-channel data to measure self-regulated learning, and specifically metacognitive processes, during learning with advanced learning technologies, and why we need to move beyond using only self-report measures and traditional inferential statistics to measure how students use metacognitive self-regulated learning processes during learning with advanced learning technologies. I will provide specific empirical evidence revealing how using multi-level modeling, sequential pattern mining, and differential sequence mining can be used to explain the temporal dynamics of self-regulated learning unfolding over time, and how we can use this to develop advanced learning technologies that are adaptive to individual learning needs.

The first manuscript is a theoretical chapter where Azevedo, Taub, & Mudrick (in press) present the challenges faced by researchers when measuring self-regulated learning during learning with different types of advanced learning technologies, including relying on self-report measures, as well as new challenges to consider when measuring self-regulated learning with multi-channel data, such as temporal alignment of data. In the second manuscript, Taub, Mudrick, Azevedo, Millar, Rowe, & Lester (in press) used multi-level modeling with log files and eye-tracking data to determine how students’ use of metacognitive monitoring strategies impacted their performance on an in-game assessment when reading content during learning while playing CRYSTAL ISLAND, a game-based advanced learning technology that fosters scientific reasoning and literacy skills. In the third
manuscript, Taub, Azevedo, Bradbury, Millar, & Lester (revise and resubmit) used sequential pattern mining and differential sequence mining to investigate metacognitive monitoring during students’ hypothesis testing activities to determine if the sequences in which students tested their hypotheses different based on their efficiency in gameplay during learning with CRYSTAL ISLAND.

Overall, results indicated that by directly addressing the methodological and analytical issues discussed in the chapter about the need for using multi-channel data and non-traditional statistical techniques, the authors were able to investigate instances of metacognitive monitoring during learning while playing CRYSTAL ISLAND. These results reveal that metacognitive monitoring can be defined differently based on the activity students are doing (e.g., knowledge acquisition vs. hypothesis testing), and that using multi-channel data and non-traditional statistical approaches were helpful methodological tools and analytical techniques for investigating metacognitive monitoring, compared to traditional approaches of using self-report measures and traditional inferential statistics.

Broader implications from these manuscripts focus predominantly on the importance of context. Theoretically, defining metacognitive self-regulated learning processes will depend on the specific task students are doing. Methodologically, determining which data channels are the most appropriate variables to use to investigate metacognitive self-regulated learning processes will also depend on the specific activity students are engaging in. Analytically, the most appropriate statistical analysis techniques will depend on the research questions being asked for the analysis. Therefore, context will largely impact decisions for conducting research studies, with results aiming towards the design and development of adaptive advanced learning technologies that foster individualized learning for all student
© Copyright 2017 Michelle Taub

All Rights Reserved
Using Multi-Channel Data to Assess, Understand, and Support Metacognition with Advanced Learning Technologies

by
Michelle Taub

A dissertation submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

Psychology

Raleigh, North Carolina

2017

APPROVED BY:

Dr. Roger Azevedo
Chair of Advisory Committee

Dr. Jing Feng

Dr. Jeffrey A. Greene
Inter-Institutional Member

Dr. James C. Lester
DEDICATION

To Joel—You mean everything to me. You have supported me since day 1, and I feel like the luckiest person in the world to have you as my partner in crime. There is no one else I would rather share my life with.
I remember the day when I got a voicemail from a man named Roger Azevedo, who congratulated me on getting into the Master’s program at McGill, and wanted to meet me and discuss his research. I had graduated a few months before with a Bachelor of Arts in Psychology from McGill, and wasn’t certain about what the next journey would have in store for me. I always had an interest in helping learners of all ages—I worked for an after-school homework program for elementary school students for 2 years, I worked as a substitute teacher at a nursery school for 5 years, and I worked as a camp counselor for over 10 years for kids aging from 7 to 11. When I began volunteering at a school specialized for students with special needs, I developed even more of an interest in working with students with different types of abilities, whether they had special needs, different levels of motivation to learn, or different interests. I applied to the Learning Sciences Master’s Program where I felt I could explore this interest. When I received the call from Roger, I had no idea that it would change my life forever.

Roger shared his research with me: self-regulated learning, metacognition, advanced learning technologies. I was fascinated with everything he showed me, and couldn’t wait to learn more. I began reading and learning the literature about self-regulated learning. I became familiar with MetaTutor. I met and became friends with other lab members and collaborated with them, and finally began working on research of my own. I got to attend my first conference where I learned more than I could imagine. Then I began writing my own conference papers, and started traveling around the world to present my research, and to network with other researchers in the field. Before I knew it, 2 years had gone by and it was time to write my Master’s Thesis. I had already known since the start of my Master’s that I
wanted to continue to work with Roger to get my PhD, and couldn’t wait to begin that next chapter.

I will never forget the day when Roger announced to us that he was leaving for North Carolina State University in Raleigh, North Carolina, and that if we wanted to, we could go with him. He met with each of us individually to see how we were doing and what we were considering, and I told him I had to think about it, although I already knew what I wanted to do. I think I surprised him when I told him I was coming, but to this day, I don’t have any regrets.

Moving to Raleigh started a new chapter. I left my friends and family in Montreal, where I grew up, and moved to a different country only knowing 2 people, I began a PhD program, and I started a new life, where I made new friends, and met the most important and special person I have ever met. My interest in metacognitive monitoring grew as I worked to investigate different metacognitive processes among college and middle school students using eye tracking and facial expression of emotion data. I learned multi-level modeling and never though I would enjoy running statistics so much. I have worked so hard to publish papers in top-tier journals, submit and present at conferences all over the world, work with collaborators in interdisciplinary fields, and mentor new graduate students coming to work in our lab. Over the past 4 years, I have grown so much as a person, and I owe it all to Roger. Nothing would have happened for me if it wasn’t for Roger, and I don’t think I can ever thank him enough for everything he has given me for the past 6 years. I always try to think about what my life would be like if I hadn’t come to NC State, and even though I haven’t gotten to examine self-regulated learning and metacognition with children with special needs yet (there is still a long future of research ahead of me), I can’t imagine it any other way.
ACKNOWLEDGMENTS

I am so thankful for all the people who have supported me and worked with me throughout these past 4 years. Roger, I can’t thank you enough for everything you have done for me since I started my Master’s 6 years ago. You have been nothing but supportive—no matter what. You sacrifice everything for your families, both personal and academic, and I am so appreciative of that. Whenever I send you a draft, you have it read as soon as you can. You never keep your students waiting, and are always available whenever we need your help. You have mentored me to become a better researcher, a better student, and a better mentor to everyone else in the lab. I have learned so much from you, and I can’t imagine starting my academic career with anything besides working as your Post Doc. I can’t wait to begin the next chapter, and even though I look forward to becoming a faculty member soon, I will never forget the fun we had in the lab and at conferences, despite long trips and delayed flights! Thank you from the bottom of my heart for every opportunity you have given me.

Thank you to my committee members: Jing Feng, Jeffrey Greene, Thomas Hess, and James Lester. You have all been very supportive of all my hard work, and I really appreciate all the time you have spent reading drafts and providing me with feedback for my prelims, proposal, and defense. I am so lucky to have such a strong PhD committee, and I thank you all so much for everything.

No graduate student can get through it all on their own, and I am lucky to have gone through it all with my academic brother and friend, Nicholas Mudrick. Nick, we have been through it all together, and despite our bickering, you have been a great brother to me. I am so glad you tried so hard to convince me to come with you. Thank you also to my other academic siblings: Amanda Bradbury and Megan Price. Mandy, I have had a great time
working with you and mentoring you, and look forward to continuing. You are really the 
nicest person I know, and I hope you never lose your positivity. Megan, I am so glad you 
joined our lab. I have really enjoyed getting to know you, and I look forward to continuing to 
work with you as you transition to becoming a grad student. Thank you to all other lab 
volunteers: Dennis Hernandez, Carina Tudela, Alex Haikonen, Nicole Urh, Mitchell 
Moravec, and Pooja Ganatra, who have helped me code, do lit searches, take umpteen 
screenshots, or anything else I needed help with.

To all my friends and family—you may not all know exactly what I do, but I 
appreciate all your positive support. To my parents, sisters, brothers, niece, and nephew, I am 
so grateful for my family and even though I don’t get to see you too often, I know you are all 
always there for me. To all my friends back home and my new friends, I am so grateful to 
know you all, and you all mean so much to me.

Last but not least—Joel, you make me smile every single day. You are the most 
amazing person; you mean everything to me, and I wouldn’t be the same person I am today if 
I never met you. I am so lucky to share my life with you, have our adorable Iggy, and lean on 
you for support whenever I need you. Thank you for being you.
# TABLE OF CONTENTS

LIST OF TABLES .................................................................................................................. viii  
LIST OF FIGURES ................................................................................................................ ix  

CHAPTER 1: Introduction ..................................................................................................... 1  
   Aims of this Research .................................................................................................. 6  
   Addressing the Aims ................................................................................................. 7  
   Summary .................................................................................................................. 12  

CHAPTER 2: Understanding and Reasoning about Real-Time Cognitive, Affective,  
and Metacognitive Processes to Foster Self-Regulation with Advanced Learning  
Technologies ....................................................................................................................... 13  
   A Review of ALTs Detection of CAM Processes ...................................................... 14  
   Factors that Influence the Use of CAM Processes with ALTs .............................. 18  
   Measurement and Detection of CAM Processes During Learning with ALTs .... 19  
   Future Directions ................................................................................................... 23  
   Implications for Practice ......................................................................................... 24  
   Implications for Designing ALTs ........................................................................... 25  
   References ............................................................................................................. 26  

CHAPTER 3: Using Multi-Channel Data with Multi-Level Modeling to Assess In-  
Game Performance during Gameplay with CRYSTAL ISLAND .................................. 30  
   Introduction ........................................................................................................... 30  
   Methods ............................................................................................................... 34  
   Results ............................................................................................................... 38  
   Discussion ........................................................................................................... 40  
   References ........................................................................................................... 42  

CHAPTER 4: Using Sequence Mining to Reveal the Efficiency in Scientific  
Reasoning during STEM Learning with a Game-Based Learning Environment ....... 45  
   Introduction ........................................................................................................... 45  
   Methods ............................................................................................................... 53  
   Results ............................................................................................................... 61  
   Discussion ........................................................................................................... 71  
   References ........................................................................................................... 81  

CHAPTER 5: Integrated Discussion ................................................................................. 86  
   Defining Metacognitive Monitoring ........................................................................ 87  
   Broader Impacts .................................................................................................... 89  
   Future Directions ................................................................................................. 95  

REFERENCES .................................................................................................................. 102
LIST OF TABLES

CHAPTER 3
Table 1: Unstandardized Coefficients for Number of Books and Frequency by Title by Concept Matrix Attempts .............................................................. 38

Table 2: Unstandardized Coefficients for Proportions of Fixations on Book Content and Book Concept Matrices by Concept Matrix Attempts ............................................ 39

Table 3: Unstandardized Coefficients for Number of Books, Frequency by Title, and Proportions of Fixations on Book Content and Book Concept Matrices by Concept Matrix Attempts............................................................................ 40

CHAPTER 4
Table 1: Descriptions of Full Relevancy Codes................................................................. 60

Table 2: Correlations of Proportions of Time Testing, Reading, and Talking and DW Submissions ........................................................................................................ 64

Table 3: Pattern Mining Outcome Sequences by DW Group............................................. 67

Table 4: Differential Sequence Mining with 2- and 3-Coded Segments by DW Group .......... 70
LIST OF FIGURES

CHAPTER 2
Figure 1: Illustration of Instrumented Participant during Learning with an ALT ....... 20

CHAPTER 3
Figure 1: Overview of CRYSTAL ISLAND (top) and Book with Concept Matrix with Eye Tracking Overlay (bottom).................................................................................................................. 35

Figure 2: Cross-Level Interaction between Books and Frequency of Book by Title on Concept Matrix Submission Attempts.................................................................................. 39

Figure 3: Interaction between the Proportions of Fixations on Book Content and Book Concept Matrices on Concept Matrix Submission Attempts .............................................. 39

CHAPTER 4
Figure 1: Screenshots of CRYSTAL ISLAND Scanning Device (left) and Diagnosis Worksheet (right).......................................................................................................................... 54

Figure 2: 2-3-Coded Sequences with Support Values by DW Group........................................ 67

Figure 3: 4-Coded Sequences with Support Values by DW Group ........................................ 68
CHAPTER 1: Introduction

Self-regulated learning (SRL) is a psychological construct that implies students are playing an active role in their learning by deploying cognitive strategies and metacognitive monitoring processes, as opposed to being passive recipients of information (Pintrich, 2000; Winne & Azevedo, 2014; Winne & Hadwin, 1998, 2008; Zimmerman & Schunk, 2011; Schunk & Greene, in press). Research on SRL has revealed that when students engage in cognitive, affective, metacognitive, and motivational (CAMM; Azevedo, Johnson, Chauncey, & Graesser, 2011; Azevedo, Taub, & Mudrick, 2015; in press) processes, it can positively impact their learning, however, studies have also revealed that students do not typically deploy CAMM processes during learning, and therefore do not regulate their learning effectively and efficiently (Azevedo & Aleven, 2013). As such, researchers are developing advanced learning technologies (ALTs; e.g., intelligent tutoring systems, hypermedia, multimedia, serious games) that are designed to track, model, and foster the effective use of CAMM SRL processes by providing timely scaffolding to assist students with using particular types of cognitive and metacognitive SRL processes (Azevedo & Aleven, 2013; Sabourin & Lester, 2014). ALTs have been developed that address the use of SRL processes, with research using self-report measures and traditional inferential statistics to investigate SRL (Veenman, 2011; Greene, Robertson, & Costa, 2011), however there is a limited amount of empirical research using multi-channel data with advanced statistical techniques (e.g., multi-level modeling, educational data mining) to investigate the impact of how metacognition impacts students’ SRL during learning with these ALTs (Azevedo, 2015; Kinnebrew, Segedy, & Biswas, 2017). Multi-level modeling can be beneficial for this
research because it allows us to investigate instances of engaging in these SRL processes as opposed to assessing them as a product score, which aligns with the view of SRL as an event that unfolds over time (Greene & Azevedo, 2009; Winne & Azevedo, 2014; Winne & Hadwin, 1998, 2008). Specifically, the Information Processing Theory of SRL (Winne & Hadwin, 1998, 2008) views SRL as an event, and not a trait that cannot change, and so there can be multiple events (or instances) of engaging in SRL processes, which can be captured using this analytical technique. By investigating different events of engaging in SRL, we can therefore investigate each event separately (for example, each instance of reading), and how each individual event is associated with an outcome variable for that event, such as accuracy in completing the assessment of that book content. Educational data mining is another relevant analytical technique because it can also be used to investigate different SRL-related events including when events happen in relation to each other (Winne, 2017). For example, sequential pattern mining can be used to examine if there are sequences of engaging in SRL-related processes, such as testing different items, indicative of hypothesis testing, and in which order. Therefore, data mining techniques also investigate separate instances of engaging in SRL and SRL-related behaviors, as opposed to examining SRL as a composite variable. There is much research that needs to continue to be conducted in the area of metacognition with ALTs, however there are also theoretical, methodological, and analytical issues that need to be considered. My dissertation will address the methodological and analytical issues, by introducing research that has used different data channels with advanced statistical techniques to assess learning, SRL, and scientific inquiry with CRYSTAL ISLAND, a game-based ALT.
Methodological and Analytical Issues

Advances in metacognition are emerging, as interdisciplinary researchers converge on methodologies that focus on the real-time deployment of CAMM processes during complex learning with ALTs, such as MetaTutor and CRYSTAL ISLAND (see Azevedo & Aleven, 2013; Calvo, D’Mello, Gratch, & Kapas, 2015). There are several methodological challenges when collecting data, such as log files and eye tracking, in terms of temporal alignment, sampling rate, level of granularity, context, and inferences (Azevedo, Moos, Johnson, & Chauncey, 2010; Greene & Azevedo, 2010). In addition, there are also analytical challenges when using traditional methods (e.g., time series, inferential statistics, etc.) that impede our theoretical understanding of the complexity of these underlying processes, and therefore limit the implications for designing ALTs capable of providing students with real-time, adaptive scaffolding, modeling, and feedback (e.g., scaffolding use of sophisticated cognitive learning strategies).

Thus, in addition to using ALTs as learning tools for fostering the use of CAMM SRL processes, ALTs can be used as a research tool for examining how students use CAMM SRL processes during learning with these environments. For example, CRYSTAL ISLAND is a game-based ALT that can use multi-channel data to assess students’ accuracy in engaging in metacognitive monitoring processes during learning about science topics. CRYSTAL ISLAND, is a game-based learning environment that fosters students’ scientific reasoning and literacy skills as they learn about microbiology, and try to solve the mystery of what epidemic has impacted the inhabitants of the island (Rowe, Shores, Mott, & Lester, 2011). During learning and gameplay, students can gather clues by talking to non-player characters regarding their
symptoms, reading book content and completing embedded assessments about the content, and testing food items for contamination. Thus, we can examine how students are engaging in these activities, as indicators of metacognitive monitoring because we can examine how students are monitoring the questions they ask non-player characters and how they use their responses to impact which food items they choose to test; we can assess which books students read, and how they perform on the embedded assessments; and we can determine which food items students are testing and in which order. Using eye-tracking data, we can examine the durations of fixations on book content, and how the amount of time fixating on items influences metacognitive monitoring behavior related to scientific reasoning. Thus, we can use multi-channel data to investigate how students are engaging in gameplay strategies and metacognitive monitoring while they learn when playing the game.

Analytically, researchers predominantly use traditional inferential statistical techniques, such as repeated-measures ANOVAs, multiple linear regressions, etc., to assess metacognition during SRL. Such statistical methods require satisfying a set of assumptions, which are frequently violated due to the nature of these SRL processes. For example, one assumption is independence of observations, which assumes that all data cells are unrelated to each other. Typically in ALTs, students can deploy each SRL strategy more than one time, thus resulting in multiple data cells per person, per strategy. This would mean that the cells are dependent on each other (i.e., they come from the same person), violating the assumption of independence of cells. A repeated-measures ANOVA does allow for the inclusion of variables at multiple time points (there is no assumption of independence), however all time points must be equidistant, and there must be an equal number of time points for everyone,
such that if one person uses a total of 10 SRL processes every five minutes and another uses 20 every two minutes, this cannot be accounted for. In other words, we cannot have any missing cells, and so all participants need the exact same amount of data points, which occurred at the same time points. To address these analytical issues, we propose using multi-level modeling (MLM) (Raudenbush & Bryk, 2002), which combines the strengths of repeated-measures ANOVAs and multiple linear regressions to enable a researcher to determine if multiple instances of an event (which can differ across participants, therefore allowing missing cells) are predictive of some dependent variable. When using MLM, the assumptions required with inferential statistics are not required (however MLM does have its own statistical assumptions), thus when analyzing SRL data from ALTs, MLM is an ideal statistical technique used to address current challenges and lead to more sensitive and effective ALTs for complex learning. In addition, to MLM there are several other interdisciplinary techniques from educational data mining (EDM; Kinnebrew, Loretz, & Biswas, 2013) and machine learning (ML; Lallé, Taub, Mudrick, Conati, & Azevedo, 2017) that should be considered as they have the potential to augment our current understanding of the deployment of metacognitive processes during learning with ALTs. Sequential pattern mining and differential sequence mining can be used to determine if there are frequent patterns of a given behavior occurring, and whether or not there are statistical differences in the occurrences of these patterns between groups of interest (Kinnebrew et al., 2013). As such, MLM and EDM are both valuable techniques to be used for assessing multiple occurrences of a given behavior (i.e., a metacognitive monitoring strategy) during learning with ALTs.
Aims of this Research

Many studies investigating SRL with ALTs have predominantly used self-report measures as assessment tools to measure the impact of metacognition on SRL during learning with ALTs. In contrast, my dissertation will examine multi-channel process data to investigate the relationship between metacognition and learning, which is a less subjective method for examining SRL. We will be addressing questions, such as: (1) what are the individual and relative contributions of multi-channel data to understanding the influence of metacognition on measures of complex learning and scientific reasoning? (2) What are the key features of metacognitive processes during learning? For example, what is the duration, fluctuation, dynamics, sequence, etc. of individual metacognitive processes? In addition, more specific questions include: (1) Is there a relationship between concept matrix attempts and proportion of fixations on book content, and does this relationship depend on the proportion of fixations on book concept matrices during the scientific reasoning process in CRYSTAL ISLAND (Taub, Mudrick, Azevedo, Millar, Rowe, & Lester, in press)? (2) Are there unique sequences of hypothesis testing behavior based on the efficiency of solving the mystery during learning while playing CRYSTAL ISLAND (Taub, Azevedo, Bradbury, Millar, & Lester, revise and resubmit)? In sum, this research aims to capture and analyze multi-channel data using MLM and educational data mining (i.e., sequential pattern mining and differential sequence mining) with CRYSTAL ISLAND to build a blue print for designing ALTs capable of accurately detecting, tracking, modeling, and fostering metacognition.
Addressing the Aims

Overall, the goal of this dissertation is to demonstrate why we need to use multi-channel data to measure SRL, and specifically metacognitive processes, during learning with ALTs, and why we need to move beyond using only self-report measures and traditional inferential statistics to measure how students use metacognitive SRL processes during learning with ALTs. I will provide specific empirical evidence revealing how using multi-level modeling, sequential pattern mining, and differential sequence mining can be used to explain the temporal dynamics of SRL unfolding over time, and how we can use this to develop ALTs that are adaptive to individual learning needs.

This dissertation includes five chapters, which will each discuss and aim to address these methodological and analytical issues related to metacognitive monitoring and SRL in ALTs. Chapter 1 (this chapter) introduced the main issues to be addressed in this dissertation, and will finish by discussing the specific aims of this dissertation, including a brief introduction to each of the three manuscripts. Chapters 2, 3, and 4 are the full copies of each of the three manuscripts, where chapter 2 is a theoretical chapter, and chapters 3 and 4 are empirical journal articles. Chapter 5 will include an integrative discussion, which addresses key points from each manuscript, and how they each contribute, separately and together, to addressing the aims of this dissertation. We will also discuss the broader implications from these three manuscripts. Before presenting each manuscript, I will provide a brief overview of each manuscript.


In this handbook chapter, Azevedo, Taub, & Mudrick (in press) provide an overview of the challenges when measuring SRL during learning with different types of ALTs, including relying on self-report measures, as well as new challenges to consider when measuring SRL with multi-channel data, such as temporal alignment of data. The authors discuss different types of ALTs, some of which are grounded in theories of SRL, and how these systems can detect different CAM SRL processes, which demonstrates the shift towards using multi-channel data, and emphasizes that different ALTs can detect SRL in different ways. The authors note that it is important to consider factors (internal or external) that might influence CAM processes, such as levels of prior knowledge or context.

Additionally, Azevedo et al. (in press) discuss methods for measuring and detecting CAM SRL processes as students learn with ALTs, such as aligning data with different sampling (e.g., 120 frames per second for the eye tracker vs. 1 frame per millisecond for the log files). The goal of this chapter is to convey that all of these factors need to be considered towards designing adaptive ALTs, such that individual differences factors and aligning the data channels will be necessary for these adaptive systems to do to ensure they are fostering effective SRL so that students are learning the complex topics.


In this study, Taub, Mudrick, Azevedo, Millar, Rowe, & Lester (in press) used multi-
level modeling to examine undergraduate students’ SRL and scientific inquiry during
learning while playing CRYSTAL ISLAND. Specifically, they investigated book reading
behavior, such that the level of analysis was at the book instance level, where each instance
of opening a book generated one separate row of data. They used log-file data to assess how
many books students read, and how often they read each of those books; and used eye-
tracking data to determine the proportions of fixations on book content and their associated
in-game assessments (concept matrices), to see how these fixations impacted their
performance on the concept matrices. Based on these four variables, they ran three separate
models: (1) one using the log-file data variables (number of books, frequency of reading each
book) to determine their association with performance on the concept matrices, (2) one using
eye-tracking data variables (proportions of fixations on book content and proportions of
fixations on book concept matrices) to determine their association with performance on the
concept matrices; and (3) one model investigating the association between the interaction of
all four variables and performance on the concept matrices.

Their results revealed that not only were log-file data and eye-tracking data
individually significantly associated with performance on the concept matrices, but there was
also a 4-way interaction, revealing that participants who had the highest performance on the
concept matrices read fewer books, but read each of those books more frequently, and had
low proportions of fixations on the books and low proportions on the concept matrices,
suggesting that students were being strategic readers and were shifting back and forth
between the book and concept matrix, monitoring the content they were reading to ensure it
was relevant for answering the concept matrix questions, instead of spending time reading
the entire text and then completing the concept matrix (Taub et al., in press). This chapter, therefore, relates to the aims of this dissertation because it focuses on examining metacognitive monitoring during learning with a game-based ALT. To investigate metacognitive monitoring, the authors used online trace data (i.e., log files and eye tracking), as opposed to self-report measures, and used an advanced statistical technique (i.e., multi-level modeling) instead of traditional statistical techniques, that would not have allowed for investigation of each book instance. This manuscript, therefore, serves as one example in this dissertation revealing how we can move beyond traditional methods and analytical techniques in an attempt to address the methodological and analytical issues regarding research on ALTs.


In this study, Taub, Azevedo, Bradbury, Millar, & Lester (revise and resubmit) used sequential pattern mining and differential sequence mining to examine the differences in undergraduate students’ SRL and hypothesis testing behavior based on level of efficiency during learning while playing CRYSTAL ISLAND. Hypothesis testing was examined based on students’ use of a scanning device, which tested food items for being positive or negative for pathogenic or non-pathogenic substances (testing for positive for pathogenic viruses or bacteria would indicate that food item as the disease’s transmission source). This scanning behavior can be classified as hypothesis testing because it is assumed that students form
hypotheses based on the clues they gather during the game (e.g., based on food items non-player characters reported eating), such that they form hypotheses based on these clues, and then test their hypotheses by testing food items. This process involves not only hypothesis generating and testing, but also monitoring of food items that have already been tested, and what food items are being tested for (e.g., viruses or bacteria), requiring students to keep track of how much progress they have made towards solving the mystery. Taub et al. (revise and resubmit) examined the sequences of testing food items based on the relevancy of what they were testing (e.g., testing for a virus or mutagen when the clues indicate the disease is bacterial would not indicate a relevant test) and the relevancy of the food item itself (e.g., a sick patient reported eating eggs, and so eggs is a relevant food item to test). In addition, they examined the differences in these sequences based on students’ efficiency in solving the mystery, where efficiency was defined based on the number of attempts students made to solve the mystery (more efficient students made 1 attempt while less efficient students made more than 1 attempt).

Results revealed that the proportion of time students spent testing food items was significantly positively correlated with the number of attempts made to solve the mystery correctly, indicating that when students spent more time testing food items, they made more attempts at solving the mystery correctly. There were no significant correlations between time spent reading books and time spent talking to non-player characters, signifying the important relationship between hypothesis testing and solving the mystery. In addition, students who were more efficient at solving the mystery tested significantly fewer partially-relevant and irrelevant foods items than less efficient students, a result that was further
emphasized from the sequence mining results, which revealed that less efficient students generated more sequences, and tested significantly more sequences that contained partially-relevant food items. Therefore, these results demonstrated using log-files to investigate sequences of hypothesis testing behavior and how that differed based on levels of efficiency, demonstrating another method for using trace data with a non-traditional statistical technique to examine metacognitive monitoring during learning with ALTs.

**Summary**

Overall, these three manuscripts further emphasize the methodological and analytical issues faced as researchers who investigate how students engage in metacognitive SRL processes during learning with ALTs. There are many issues that need to be discussed to ensure that in future designs of ALTs, these issues are considered, which should help improve the ability for ALTs to foster both SRL and overall learning of complex topics. This is the goal of chapter 2, which addresses many of these concerns in a theoretical chapter. Chapters 3 and 4 aim to demonstrate potential solutions to the issues mentioned in the theoretical chapter, such as using multi-channel data to measure metacognitive monitoring during learning with ALTs, and analyzing these data using non-traditional statistics. Finally, chapter 5 will include an integrated discussion that is based on all the issues raised and the implications from the manuscripts, including posing questions that need to be considered towards developing adaptive ALTs that provide individualized tutoring for all students.
CHAPTER 2: Understanding and Reasoning about Real-Time Cognitive, Affective, and Metacognitive Processes to Foster Self-Regulation with Advanced Learning Technologies

Roger Azevedo, Michelle Taub, and Nicholas V. Mudrick

Self-regulated learning (SRL) involves learners’ ability to monitor and regulate their cognitive, affective, metacognitive, and motivational (CAMM) processes while using advanced learning technologies (ALTs). This alignment enables learners to direct their cognitive, affective, and metacognitive processing in real time while learning about challenging domains (e.g., science, mathematics) and while using advanced learning technologies (ALTs; e.g., intelligent tutoring systems, simulations, serious games, hypermedia, tangible computing, virtual reality). Additionally, emerging empirical evidence indicates that CAM processes play an important role in learning and problem solving as well as self-regulation with ALTs. However, capturing CAM processes during learning with ALTs poses several major conceptual, theoretical, methodological, and analytical challenges. For example, researchers currently measure CAM SRL processes using several online trace methodologies, such as concurrent think-alouds, eye tracking, log files, physiological sensors, and so forth. While these methods have the potential to advance current SRL frameworks, models, and theories, they still pose serious challenges (e.g., temporal alignment of data channels, lack of analytical techniques, and accuracy of inferences made from individual channels and across multiple analytical methods) that currently plague the field.

Another major challenge is related to the use of real-time CAM trace data to make ALTs adaptive. More specifically, using these trace data can provide support for learners’ CAM processes and domain learning in real time by allowing researchers to make inferences about the temporally unfolding deployment of the learners’ CAM processes. However, issues remain such as the lag time between capturing real-time deployment of CAM processes, inferences made by the ALTs to adapt to learners’ needs, and the ALTs’ ability to effectively monitor and regulate their own external regulation over time (e.g., a virtual agent self-regulates and modifies the timing, sequencing, and type of scaffolding of metacognitive judgments because it typically induces frustration in learners; Azevedo, Taub, Mudrick, Farnsworth, & Martin, 2016). In sum,
the educational effectiveness of ALTs hinges on researchers’ ability to collect real-time CAM trace data by converging a myriad of interdisciplinary methods (e.g., eye tracking) and analytical techniques (e.g., data mining, machine learning) to understand these processes, making accurate inferences regarding the underlying CAM processes, and modeling and embodying these processes to enhance learners’ ability to effectively monitor and regulate their own CAM SRL processes and overall learning (Taub, Azevedo, Bouchet, & Khorsavi, 2014).

As such, our chapter focuses on understanding and reasoning about real-time CAM processes to foster self-regulation with ALTs. More specifically, our chapter will focus on the following: (1) a critical review of various ALTs that use SRL models others that also focus on CAM processes but use different theoretical frameworks to analyze online trace data with ALTs (e.g., Winne and Hadwin’s information-processing theory); (2) analysis and discussion of key issues related to investigating CAM SRL processes during learning with ALTs with online trace methodologies (e.g., the role of contextual factors, representation of CAM processes, temporal dynamics); (3) a critical review of the factors influencing the use of CAM processes during learning with ALTs; (4) strengths and weaknesses of several interdisciplinary online trace methodologies used in ALTs to detect, track, model, and foster SRL and domain learning; and (5) future directions that will significantly augment our understanding of the role of CAM SRL processes during learning with ALTs and enhance the instructional effectiveness of these systems based on their ability to detect, track, model, and foster learners’ CAM SRL effectively. Lastly, we propose implications for using multichannel data to foster CAM SRL processes with ALTs, followed by implications for designing ALTs capable of detecting and fostering CAM processes.

A REVIEW OF ALTS’ DETECTION OF CAM SRL PROCESSES

Emerging empirical evidence indicates that CAM SRL processes play a critical role in learning and problem solving in different domains (e.g., science, mathematics, computer literacy) with ALTs (Azevedo, 2015). One key component of ALTs is that they have been shown to play an important role in facilitating students’ SRL by scaffolding, fostering, and supporting CAM processes (Azevedo et al., 2013). Many of these ALTs are unique not only in design, but also in their theoretical frameworks and intervention methods. Research on CAM SRL processes and ALTs is varied, and a succinct synthesis of the empirical research on their capability to detect, track, model, and foster SRL and domain learning is needed (Azevedo & Alevan, 2013a, 2013b). Currently, several unique ALTs are being designed to function as both a research and learning tool and therefore facilitate the collection, detection, tracking, modeling, and fostering of self-regulation of CAM processes through various methods (e.g., interface designed to afford learners opportunities to engage in metacognitive monitoring and regulate their learning by using sophisticated strategies and affective reactions by providing a wealth of nonlinear multimedia materials and conversations with artificial pedagogical agents). As such, the purpose of this section is to synthesize the current literature on ALTs that target SRL processes as well as ALTs that foster specific CAM processes and to provide an overview of the strengths and shortcomings of these systems.

ALTs have been developed to foster students’ SRL in response to research that has shown students do not typically deploy SRL strategies effectively or efficiently (Azevedo, Taub, & Madrick, 2015; Moos, 2018; this volume; Potras & Lajoie, 2018; this volume; Winne & Azevedo, 2014). Many types of ALTs can foster different CAM SRL...
processes such as intelligent tutoring systems, hypermedia- and multimedia-learning environments, game-based learning environments, and simulations, all of which task students with engaging in different types of SRL processes. For example, MetaTutor, nStudy, and SimSelf foster the use of metacognitive monitoring and cognitive learning strategies during learning, whereas Betty’s Brain requires students to create causal concept maps to teach Betty (a computer agent) science material, and Crystal Island requires students to engage in SRL and scientific inquiry to solve a mystery. In addition, many ALTs involve interactions between students and pedagogical agents so the ALT can provide scaffolding to students to teach them how to use SRL strategies effectively. MathSpring, AutoTutor, Affective AutoTutor, Gaze Tutor, Guru, and START are all intelligent tutoring systems that engage students in dialogue with the system to discuss their levels of affect and understanding of the material they are learning. As such, many different types of ALTs each have specific components to foster learning using CAM processes in different ways.

ALTs Theoretically Grounded in SRL Theories, Models, and Frameworks

For this chapter, we use the information-processing theory of SRL (Winne & Hadwin, 1998, 2008) as our leading theoretical framework. According to this model, SRL is viewed as an event that temporally unfolds over time, and occurs through a series of four cyclical stages, where information processing occurs via the use of cognitive and metacognitive strategies. We focus our attention on the second and third phases, setting goals and plans and use of learning strategies, because it is during these phases that students engage in planning, monitoring, and strategy use as they learn with ALTs (see Azvedo, Moos, Johnson, & Chauncey, 2010; Greene & Azvedo, 2009, 2010; Johnson, Azvedo, & D’Mello, 2011; Winne, 2018; this volume). Regarding ALTs with SRL as a guiding framework (e.g., Winne & Hadwin, 1998), there are similarities in how cognitive processes are detected and scaffolded, and differences in how they are modeled for learners and how adaptive they are to learner actions. More specifically, most of these ALTs (e.g., MetaTutor, nStudy) detect and track learners’ cognitive processes based on user-interface interactions such as selecting among various multimedia content, collecting scientific evidence and making hypotheses about a particular biological agent, taking notes on vast amounts of content, building new knowledge representation from existing system-provided information, and so forth. The real-time behavioral enactment of these cognitive processes is captured through eye tracking, log files, and pre- to post-test learning gains. Additionally, apart from Betty’s Brain, all of these systems require learners to take notes in some manner (i.e., through the note-taking feature of MetaTutor and nStudy vs. the concept matrix in Crystal Island; Azvedo et al., 2013; Lester, Mott, Robison, Rowe, & Shores, 2013). Another commonality among these ALTs is the use of different learning strategies such as coordinating informational sources (e.g., creating causal maps in Betty’s Brain vs. tagging information in nStudy; Beaudoin & Winne, 2009; Biswas, Segedy, & Kinnebrew, 2013).

Despite their similarities, these systems differ in their ability to scaffold and model cognitive processes in real time as well as cognitive learning strategies. For example, neither Crystal Island nor nStudy provides explicit scaffolding, but rather both rely on their interface or environment to promote the enactment of learning strategies (Beaudoin & Winne, 2009; Rowe, Shores, Mott, & Lester, 2011). In contrast, both MetaTutor and Betty’s Brain employ pedagogical agents to prompt and scaffold learners to enact specific learning strategies (e.g., summarizing in MetaTutor vs. crafting
inquiries in Betty's Brain). Furthermore, these systems are predominantly adaptive with the exception of nStudy. For example, MetaTutor makes inferences based on real-time analyses of trace data, using time thresholds and user behaviors to trigger production rules that, when met, adaptively scaffold the learner's cognitive strategy use (e.g., time threshold on page triggers a page quiz) and metacognitive monitoring. Furthermore, MathSpring adaptively regulates its level of difficulty through criteria such as math-problem selection based on specific learner actions (i.e., number of attempts on a problem), the amount of time spent on a problem, and whether help is requested (Arroyo et al., 2014). Overall, ALTs' ability to detect, track, model, and foster cognitive processes has produced significant results in learning, problem solving, reasoning, and scientific reasoning (see Aleven, 2013; Azevedo & Aleven, 2013b; Azevedo et al., 2013; Bùi & et al., 2013).

Most ALTs that are theoretically grounded in SRL ignore the critical role of affect, as few attempt to detect, track, and scaffold learners' emotions in real time while interacting with these systems (D'Mello & Graesser, 2015). The majority of current systems use self-report measures of emotions prior to, during, and following learners' interactions with specific components of ALTs, and therefore address learner emotions in a post-hoc fashion through analyses of self-report measures and post-hoc analyses of facial expressions of emotion data (Azevedo et al., 2016; Hartley, Bouchet, Hussain, Azevedo, & Calvo, 2015). However, MathSpring is one of the only ALTs situated within SRL that facilitates and scaffolds learners' affective processes in real time. For example, MathSpring monitors learners' affective states with physiological sensors, facial expressions of emotions, and self-reports, and it also scaffolds learners' affective processes with animated affective learning companions (Arroyo et al., 2014). These agents are designed as peer learners who offer support and guidance as learners encounter obstacles throughout their interactions with the system, and have been found to increase help-seeking behavior (Woolf et al., 2010). While this method of addressing learner affect is one that has found some success, affect detection is notoriously difficult and has continued to be a problem for ALT researchers for some time. Further, even when affect can be detected, most systems fail to employ system interventions or models to assist learners in regulating their affect (e.g., AutoTutor, Crystal Island, MetaTutor). Along with affect, metacognitive processes have also proved difficult to detect, facilitate, and scaffold as learners interact with these ALTs.

There are similarities and differences in the methods used to facilitate SRL strategies and metacognitive monitoring within these different ALTs. More specifically, these systems (1) promote similar macro-metacognitive processes (i.e., planning, setting goals, and monitoring progress towards goals within nStudy, MetaTutor, SimSelf, and MathSpring; (2) model efficient deployment of monitoring behaviors (e.g., a pedagogical agent in MetaTutor modeling how to appropriately set a subgoal vs. a pedagogical agent in SimSelf providing information on what metacognitive monitoring is: Azevedo, 2014; Traft, Mudrick, Azevedo, Markhelyik, & Powell, 2016); and (3) have specific components within their user interfaces that foster the self-initiation of specific micro-level metacognitive monitoring processes and SRL behaviors (i.e., self-initiating a feeling of knowing judgment by clicking the SRL palette of the MetaTutor interface vs. linking a self-made learning strategy with specific content in nStudy).

Although these systems model some of the same metacognitive processes, they still differ in the manner in which they scaffold SRL behaviors. For example, Betty's Brain is distinct from these other systems as it relies on the learning-by-teaching paradigm, which requires learners to use three important teaching principles that serve to support
SRL (Biswas et al., 2013); as the learners prepare to teach, they must interact with the agent, monitor, and reflect upon what has taken place. On the other hand, MathSpring facilitates self-reflection through an open learner model (Arroyo et al., 2014). Another distinct difference among these systems is the level of granularity of the metacognitive processes examined (i.e., micro-level SRL behaviors in MetaTutor; feelings of knowing, judgments of learning on a specific component of the human circulatory system vs. macro-level metacognitive processes in nStudy [planning, monitoring throughout the overall web-based interaction]; Greene & Azevedo, 2009; Winne & Hadwin, 2013).

Adaptively, many of these systems behave in much the same manner from a metacognitive standpoint as they do from a cognitive one. For example, Betty's Brain, nStudy, and SimSelf lack any form of adaptive scaffolding and rely on post-hoc assessments to detect and track learners' metacognitive behavior. This is a common theme within these ALTs, as detection is generally left to data derived from log files and self-report measures during post-hoc analyses.

**ALTs Theoretically Grounded in Theories Other Than SRL**

Our review of the literature has also uncovered another group of ALTs that have proved efficient in detecting, tracking, and modeling CAM without being grounded in theories of SRL. These systems generally use some version of two distinct theoretical frameworks: explanation-based constructivism with human tutoring models or the ACT-R theory of learning and performance. Explanation-based constructivist theories suggest learners must actively construct explanation-based meanings and knowledge through interaction, and progress is achieved through telling and doing (Aleven & Koedinger, 2002). In contrast, the ACT-R theory of learning and performance proposes that learning is achieved through the development of simple components that become complex in summation. Progress is achieved through the mastering of simple components that make up larger and more complex components (Anderson & Schunn, 2000). Although both of these frameworks lend themselves to supporting CAM processes within ALTs, they do so in different ways. Systems that use explanation-based constructivism (AutoTutor, Affective AutoTutor, Gaze Tutor, Guru, and eSTART) resemble human tutor-like interactions, and systems that use ACT-R (ASSISTments, PSLC Cognitive Tutor) attempt to build learner knowledge incrementally (Anderson & Lebiere, 1998; D'Mello & Graesser, 2012a, 2012b; D'Mello, Olney, Williams, & Hays, 2012; Jackson, Boonthum-Deneck, & McNamara, 2013; Mendicino, Razaq, & Heffernan, 2009; Olney et al., 2012; Singh et al., 2011). These different ALTs produce varied methods of detecting, tracking, and scaffolding learner CAM processes. However, similarities can still be found between these two groups.

While these systems do not scaffold, detect, or foster explicit SRL processes, they still address CAM components related to SRL; however, they are different in terms of their real-time adaptivity. For example, AutoTutor, Affective AutoTutor, and Guru address learners' cognitive processes in real time with natural language processing to adaptively model appropriate cognitive strategies and adapt content to learners' individualized abilities. Furthermore, Affective AutoTutor detects students' current affective states by monitoring facial expressions and body posture, whereas Gaze Tutor uses learners' eye movements to monitor attentional patterns and assess levels of learner engagement (D'Mello et al., 2012). While these systems primarily focus on learners' cognitive and affective processes, Cognitive Tutor and ASSISTments adaptively monitor and foster learners' cognitive and metacognitive processes through...
production rules and example-tracing (Anderson, Corbett, Koedinger, & Pelletier, 1995; Mendicino et al., 2009). Although dissimilar in their guiding theoretical frameworks, these systems still promote the effective use of CAM processes related to efficient SRL.

In sum, this section highlights the similarities and differences in the methods that contemporary ALTs use to detect, track, model, and foster CAM SRL processes. ALTs that use SRL as a guiding framework emphasize cognitive and metacognitive processes at the expense of affect, whereas ALTs that are grounded in explanation constructivism predominantly focus on cognitive and affective processes. Furthermore, all of these ALTs are distinct in the means by which they promote and monitor CAM SRL processes. Specifically, even though these ALTs attempt to foster similar CAM SRL processes, they do so in different ways. Although many of the ALTs discussed here emphasize adaptivity and converging multichannel data (e.g., log files, facial expressions of emotions), it is clear that no system addresses the detecting, scaffolding, and fostering of all components of CAM SRL processes. In the next section, we focus on factors that influence the use of CAM processes during learning with ALTs.

FACTORS THAT INFLUENCE THE USE OF CAM PROCESSES WITH ALTS

Context

One important factor to consider when investigating how students use CAM SRL processes during learning is the context in which they are learning. Learning can take place in a multitude of contexts that can differ based on the type of ALT being used (e.g., hypermedia-learning environment, game-based learning environment, intelligent tutoring system) or the topic the student is learning about (e.g., the circulatory system, microbiology, math, physics). No matter the distinction, contextual factors can impact CAM processes in many ways. Additionally, an ongoing issue regarding learning with ALTs is how learning can transfer to different domains over time. In the following section, we discuss the importance of context for near and far transfer of CAM processes.

The Impact of Context on Transfer

The context in which students are learning can impact their ability to transfer what they learn to different contexts. For example, if students successfully complete game-based learning using effective CAM processes, then these strategies can be applied to learning with a different environment, which would aid completion of the learning task with another ALT. The ability to transfer from one context to another would therefore depend on how well students use CAM processes in general. For example, if students can use the metacognitive monitoring strategy of judging their understanding of the material (i.e., judgment of learning) effectively with one ALT, they should be able to use this monitoring strategy with a different ALT (Heidig & Clarebout, 2011; VanLehn et al., 2007). However, in addition to considering how different contexts can impact transfer of learning using CAM processes, students’ levels of knowledge and skills (e.g., declarative, procedural, and conditional) can also impact how well they can transfer the knowledge they learn and the CAM processes they use. We discuss the impact of knowledge on the use of CAM processes in the next section.
Knowledge

When assessing how students learn with ALTs we often consider individual differences, such as students’ knowledge. When examining differences in knowledge and its impact on using CAM processes, we can use multiple categories to distinguish them, such as knowledge type (content vs. SRL knowledge, or declarative vs. procedural or conditional knowledge), quality versus quantity of CAM processes, and prior versus acquired SRL knowledge. In this section, we discuss some of these influencing factors.

Knowledge of Using CAM Processes Effectively

When we investigate how students use CAM processes, we often examine the frequency of use of these processes (e.g., frequency of judgments of learning), with the assumption that higher frequency means better use. However, this also involves an implicit assumption that might not be indicative of what is actually occurring. For example, if Student A takes a lot of notes during learning, as opposed to Student B who takes fewer notes, we might assume that Student A has better note-taking skills than Student B. However, if we further investigate the quality of these notes, we might determine that Student A took less-efficient notes (e.g., copied the text verbatim onto the notepad) than Student B, who took fewer notes but summarized the text into his or her own words. Thus, Student B used the cognitive learning strategy of taking notes more efficiently than Student A, despite the fact that Student A had a higher frequency of note-taking. This demonstrates the importance of differentiating between quality and quantity as well as how the quality of CAM processes can influence learning more than the quantity.

These issues related to knowledge are important in designing adaptive ALTs because any different types of individual differences can influence how a student uses CAM processes for learning. A major issue, however, in considering how to detect these different student characteristics lies in measuring these CAM processes with multiple data channels and being able to detect these processes based on the data. In the following section, we address these issues by discussing the different types of data channels typically used as well as the issues pertaining to aligning these data channels to accurately detect CAM processes during learning.

MEASUREMENT AND DETECTION OF CAM PROCESSES DURING LEARNING WITH ALTs

The previous section described temporally unfolding CAM processes with ALTs based on the use of both obtrusive and unobtrusive trace methodologies (Aleven, 2013; Azevedo, 2014, 2015; Azevedo et al., 2010; Bernacki, 2018; this volume; Bernacki, Nokes-Malach, & Aleven, 2013; Greene & Azevedo, 2010; Greene, Dekaens, Copeland, & Yu, 2018; this volume; Molenaar & van Jirwitti, 2014). When students engage in CAM processes during learning, we can collect multichannel data to investigate how their use of these processes might improve or impede their learning (see Figure 17.1). Such data channels include: (1) log files, (2) videos of facial expressions, (3) eye tracking, and (4) physiological data. Log files can be used to capture student activity within an ALT, which can inform us of how they are using cognitive and metacognitive processes by frequency, duration, and quality of responses. Videos of facial expressions can be run through facial expression recognition software (e.g., Attention Tool) to identify the
affective states (e.g., confusion) or action units being activated during the learning session. Eye-tracking data can generate students’ gaze fixations, saccades, and regressions, informing us of where on the screen (i.e., areas of interest; AOsIs) the student was looking (SMI Experiment Center) as potential indicators of CAM processes (Taub et al., 2016a, 2016b, in press). Physiological data (Empatica E4) can capture many behaviors, such as galvanic skin response, blood volume pulse, movement, body temperature, interbeat interval, and heart rate.

We can detect all of these data during learning, which can be indicative of engaging in cognitive or metacognitive processes or a change (e.g., a spike in physiological
arousal or a sudden facial movement) of affective states (e.g., confusion or frustration). Therefore, different data channels can capture different types of CAM process data, making it ideal to align and merge these data to identify how they can produce behavioral signatures of CAM processes during learning. To create these behavioral signatures, we must consider and address constraints pertaining to measuring affect and aligning data channels, which are discussed next.

**Measuring Affect**

Many SRL researchers have investigated how to measure the use of cognitive, metacognitive, and motivational SRL processes during learning (see Bernacki, 2018; this volume; Reimann & Bannert, 2018; this volume); however, little attention has been focused on methodological considerations when measuring students’ emotions. When measuring students’ affective states during learning, a series of approaches can be taken to detect and analyze these data. First, the term *affect* is an umbrella term that can encompass both emotions and mood, where emotions are short-term behaviors and mood is longer lasting (Scherer, 2009). Typically, as students learn we want to measure their emotional reactions to particular events that occur frequently. Therefore, we are interested in measuring their changing emotions as opposed to their mood when they begin learning.

Second, many categories of emotions could be of interest, such as discrete or non-discrete emotions, basic emotions, learning-centered emotions, and compound emotions. If we were to assume that emotions are discrete, this would imply that students can only experience one emotion at a time and not multiple emotions simultaneously. Research in this area might therefore select the highest evidence score of an emotion and assume that to be the emotion the student is expressing at that time. It can be difficult to single out one discrete emotion during learning with ALIs, as the multiple elements in the environment and learning context can evoke many emotions simultaneously. Thus, in contrast to examining discrete emotions, if emotions are nondiscrete (i.e., co-occurring) students can feel multiple emotions simultaneously during learning, which is more likely the case. For example, when assessing students’ evidence scores of enjoyment and confusion, it is possible to detect scores indicating the presence of both emotions, suggesting the students can be enjoying themselves but might also be confused about the material they are reading.

Students can also express different categories of emotions during learning. Basic emotions such as enjoyment, anger, fear, disgust, sadness, and surprise (Ekman, 1973) pertain to emotions typically felt during daily activities. In addition to nondiscrete emotions, they can also be compositions of different aspects of a range of emotions, called compound emotions, which combine these basic emotions (Du, Tao, & Martínez, 2014). For example, a student can be surprised, but this can be further differentiated into happily surprised, sadly surprised, fearfully surprised, or angrily surprised. Thus, compound emotions combine different dimensions of basic emotions to elicit 21 categories of emotions instead of the six basic emotions (Du, Tao, & Martínez, 2014).

While students can exhibit basic emotions during learning with ALIs, researchers have also investigated the following learning-centered emotions that are specifically demonstrated during learning: confusion, frustration, boredom, and engagement (D’Mello & Graesser, 2012a, 2012b). These emotions are typically expressed during learning, as they are common influences on or responses to learning-related activities. Thus, investigating these types of emotions is more specific to learning, and learning
with ALIs, than basic emotions (e.g., anger, joy). As illustrated above, many different types of affective states can be studied during student learning with ALIs. Once researchers have specified which aspects of affect they wish to study, they must then decide how they will detect these emotions methodologically.

Therefore, a final consideration in measuring emotions focuses on what exactly is being measured and analyzed. For example, when running video data through facial expression recognition software (e.g., FACET, FaceReader, Affdex), the software can yield evidence scores for both emotions and facial action units (i.e., possible movements of facial muscles on the learner’s face, such as brow lowering and lip tightening). Evidence scores (i.e., “output estimates of facial expression presence” or “facial expression recognition output”) of emotions can be limited in that people express emotions in different ways, and the software might not detect an instance of frustration if it is not elicited in a particular manner (i.e., based on how the software was developed to detect frustration). Thus, it can be beneficial to examine students’ action unit evidence scores, which demonstrate the areas on the face that are changing and at what evidence level. From these data, we can infer the students’ emotions using the action units that are associated with emotions (e.g., brow lowering as an indicator of confusion). In addition, this can allow for differences in expressing emotions as we can combine the evidence scores from different action units, similar to the work previously done on compound emotions (Du et al., 2014); however, we can expand on the emotions by including action units indicative of learning-centered emotions as well. This is important because it stresses not only that emotions can be expressed differently by different people, but also that there are different types of the same emotion, such as effective confusion and ineffective confusion. Therefore, detecting action units can be useful for assessing students’ emotions during learning with ALIs. It is evident that many challenges need to be addressed when assessing affective states, all of which need to be considered prior to conducting analyses regarding students’ use of CAM processes. In addition, if we are using multichannel data, other issues need to be considered prior to aligning the data; we address these issues in the following section.

**Aligning Multichannel Data to Create Behavioral Signatures of CAM Processes**

When collecting multichannel data using different types of data, a number of challenges need to be overcome prior to analyzing the aligned data. Different data channels collect data at different frequencies; for example, the SMI eye tracker measures eye movements at a rate of 120 or 250 data samples per second, facial expression recognition can be performed at every frame of video (typically 30 frames per second), and the Empatica E4 bracelet collects electrodermal activity data at a rate of four samples per second. Thus, to combine these data across a specific time period, the varied sampling rates must be reconciled. Once the data are combined, we can apply them to create behavioral signatures of different CAM processes used during learning.

Although it might seem the most beneficial to include all data channels when creating behavioral signatures, there are issues to consider in selecting data channels as well. When sampling multiple data channels, it is possible to gather enough data (and contextual information) from one data channel that might indicate a student is engaging in a particular CAM process. However, additional questions need to be answered once the data signal a behavior, indicating that perhaps we need more than one data channel to be certain we have found a behavioral signature of that CAM
process. For example, eye-tracking data might indicate the student is making a content evaluation, a metacognitive judgment that assesses the relevancy of the content to the current subgoal. In response, we can be confident that it is a content evaluation or we can turn to additional information for support, which would require the use of other aligned data such as assessing the student's emotions (or facial action units) at the time the eye-tracking gaze pattern was found. This then raises another issue: what additional data should we turn to for extra support and to provide contextual information (e.g., screen recording) to increase researchers' inference accuracy regarding the presence of specific CAM processes? Therefore, it is important to align all the collected data; however, the next step involves determining which of those data to use as indicators of CAM processes.

It is evident that using multiple data channels can help us to be more certain about which CAM processes a student is engaging in; however, there can be a downside to using multiple data channels as well. When all the data channels align to indicate the same CAM process, this can be beneficial; however, when the data yield conflicting results, indicating different CAM processes, this can lead to challenges in interpreting the data, thus causing uncertainties about which data to use. In addition, if these data yield conflicting results and we want to return to the data to further investigate them, but the student had continued to do something else in the ALT, this also poses challenges for determining which data channel is the most accurate because the CAM process has already ended. Therefore, this poses issues regarding timeframes and windows for measuring CAM processes, what kind of data we can use to best indicate the use of CAM processes during learning, and how to measure the processes while they are still occurring. These issues raise several theoretical, methodological, and practical implications for designing ALTs, which are discussed in the next sections.

FUTURE DIRECTIONS

This chapter presented some of the major advancements in ALTs in terms of how their design and use are theoretically based on SRL and other theories. We presented arguments regarding some of the major conceptual, theoretical, methodological, and analytical issues still plaguing the field when considering the challenges in collecting, aligning, analyzing, and making inferences from multichannel data that differ along several key dimensions. Future research should address these serious challenges as researchers continue to explore interdisciplinary methods to collect multichannel data and use a myriad of analytical techniques, which have the potential to advance current SRL frameworks, models, and theories. For example, serious effort should be devoted to addressing the following issues: temporal alignment of data channels; lack of analytical techniques and debatable accuracy of inferences made from individual channels and across data channels; the role of contextual factors that might interfere with the effective use of CAM processes; mechanisms; level of granularity in coding and making inferences about CAM processes; distinguishing among macro-level processes, micro-level processes, and valence when considering theoretical augmentation and implications for adaptivity for both humans and machines; analytical challenges in determining the correct unit of analysis (e.g., frequency vs. quality of CAM SRL knowledge and skills); quantitative and qualitative changes in CAM SRL processes over extended periods of time; individual and combined contributions of CAM processes and their relation to domain and SRL knowledge and skills; examining the existence of robust multichannel behavioral signatures for specific CAM SRL processes;
differentiating among declarative, procedural, and conditional SRL knowledge and skills learned during learning and problem solving with ALTs and the potential of treating motivational processes (e.g., self-efficacy, interest, task value) as trace data that contribute to CAM SRL but which may fluctuate at a higher time scale (e.g., minute, hour, day, week).

Addressing these underlying issues will allow researchers to build CAM SRL-sensitive ALTs capable of providing intelligent individualized support in real time. As argued in the chapter, a major challenge is related to the use of real-time CAM trace data to make ALTs adaptive. More specifically, using these trace data can provide support for learners’ CAM processes and domain learning in real time by making inferences based on the temporally unfolding deployment of the learners’ CAM processes. A major concern remains regarding the lag time between capturing real-time deployment of CAM processes, inferences made by the ALTs to adapt to learners’ needs, and the ALT’s ability to effectively deploy the most effective intervention (e.g., adaptive scaffolding by a virtual human) to address the learners’ needs at that moment in time. This major area of research has already made considerable strides, as documented in this chapter. The next generation of adaptive intelligent ALTs will include embodied agents who are capable of monitoring and regulating their own and learners’ external regulation over time (e.g., a virtual human self-regulates and modifies the timing, sequencing, and type of scaffolding of metacognitive judgments because scaffolding typically induces frustration in learners). In sum, the educational effectiveness of ALTs centers on researchers’ ability to collect real-time CAM trace data by converging a myriad of interdisciplinary methods (e.g., eye tracking, videos of facial expressions) and analytical techniques (e.g., data mining, machine learning) to make accurate inferences regarding the underlying CAM processes, and modeling and embodying them to enhance learners’ ability to effectively monitor and regulate their own CAM SRL processes and overall learning.

IMPLICATIONS FOR PRACTICE

Using multichannel data to measure how students use SRL processes during learning with ALTs allows us to capture actual student behavior, as opposed to relying on students to report these behaviors themselves. Multichannel data integrate a wide variety of sensor technologies (e.g., electrodermal bracelets, eye trackers) with traditional approaches to representing student behavior in learning environments. This combination of rich data sources holds the promise of discovering observable events that correspond to underlying CAM processes (Azevedo et al., 2013, 2015, 2016; Harley et al., 2015, 2016). This section describes current multichannel data sources that have become widely available and relatively affordable in recent years. Typical approaches to combining data are also discussed, with important implications for how these representations aid adaptation to the learner. Additionally, several techniques have recently been applied to analyze multichannel data while leveraging event sequences, as learning interactions unfold from moment to moment.

One prominent external data channel comes from eye tracking. These sensors identify eyes through reflection of infrared light. Combined with information about the display screen (e.g., size, distance from learner), these devices can be calibrated to show where the learner is looking at any moment. When AOIs are defined, a learner’s eye movements can be quantified as fixations on these AOIs and are representative of attentional and cognitive processing. Eye fixations are particularly
useful in quantifying reading behaviors, attention, and studying of graphical content such as videos or diagrams (Bondareva, Conati, Feyzi-Behnagh, Harley, & Azevedo, 2013; Jaques, Conati, Harley, & Azevedo, 2014; Taub et al., 2016a, 2016b, in press). For example, heat maps are also often used to produce visualizations of where individual learners or groups of learners fixated during display of particular content. Different parameters of heat maps can be adjusted to provide different insights into gaze behavior (e.g., to highlight shorter or longer periods of fixation) (Azevedo et al., 2017).

As learning environments have become increasingly sophisticated, the nature of system interaction has also evolved. ALTs now have more concerns than simply correctness of answers, as modern implementations include detailed system logs that provide a picture of how the learner interacts with each interface element at any given point in time. Thus, these learner–system interaction (or clickstream) data can provide information on whether students are taking time to work on a task based on their effective use of learning strategies or are simply advancing rapidly without consideration (perhaps by using hint-giving features) because they are not capable of making the accurate metacognitive judgments necessary to determine the relevance of the presented multimedia information.

**IMPLICATIONS FOR DESIGNING ALTS**

All the above-mentioned issues have further important implications for designing ALTs with pedagogical agents or intelligent virtual humans that can detect, model, and foster students’ CAM processes during learning. When designing these agents we must address these same issues regarding timing, because the agent must be provided with the appropriate threshold to detect the student’s activity and provide helpful feedback based on student performance.

An additional challenge when designing agents deals with the type of intervention the agent provides. The agent can intervene by prompting the student to engage in a particular CAM process or by confirming it has correctly identified which process the student is engaging in. We consider this a challenge because to create agents that are capable of intervening by prompting, we need to understand student behavior; and if we have trouble doing so as humans, due to the complexity and unpredictability of human behavior, how do we expect to program agents to be capable of doing it using algorithms that are not responsive to individual human behavior? Thus, instead of making the inferences based solely on data, the agents can be programmed to intervene by engaging in dialogues with the students to confirm if what they are detecting in the data is correct. In addition, instead of waiting for the right amount of data, the agent can intervene to obtain the ground truth about the student’s use of CAM processes and establish a rapport with the student to gain this ground truth. In principle, the more information the agent can obtain, the greater the likelihood it will be able to make accurate inferences regarding student behavior. Ideally, artificial agents should have access to multichannel data and be able to understand and reason from these data while determining how to adapt their own behavior to support learners’ SRL. Recent advances in educational data mining and machine learning have tremendous potential to provide intelligent, adaptive, and individualized feedback and scaffolding to support learners’ CAM processes as well as learning, problem solving, and performance with ALTs (e.g., Biswas, Baker, & Paquette, 2018/this volume).
CONCLUSIONS

Recent technological advances have allowed researchers to use interdisciplinary methods to collect rich multichannel trace data of learners’ CAM processes during learning and problem solving with ALIs. As such, researchers from the fields of educational, learning, cognitive, affective, social, engineering, and computational sciences have made major strides in developing ALIs to play a dual role. First, ALIs are used strategically as research tools to collect rich multichannel trace CAM SRL data to enhance our current framework, models, and theories of SRL by providing evidence of the complex, temporally unfolding nature of CAM processes in real time. Second, as learning tools ALIs are theoretically and empirically designed to afford learners the ability to foster CAM processes with some constraints. In sum, advances in understanding and reasoning about real-time CAM processes to foster self-regulation with ALIs are necessary to augment conceptual and theoretical issues as well as design intelligent systems to detect, track, model, and foster learners’ SRL.

ACKNOWLEDGEMENTS

This chapter was supported by funding from the National Science Foundation (DRL#1315352) and the Social Sciences and Humanities Research Council of Canada (SSHRC 895–2011–1006). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or Social Sciences and Humanities Research Council of Canada.

The authors would also like to thank Joseph E. Grafsaard, Carina Tulela, Mitchell Morace, Alex Haikonen, Daniel Bausch, Pooja Ganatra, and Sarah Augustine from the SMART Lab at NCSU for their assistance.

NOTES

1 We excluded motivational processes due to space limitations and the fact that using trace methodologies to measure, understand, and reason about these processes remains a major challenge for researchers and designers.

2 We use the acronym CAM (instead of CMM) throughout the chapter to denote our emphasis on cognitive, affective, and metacognitive processes only.

REFERENCES


28


CHAPTER 3: Using Multi-Channel Data with Multi-Level Modeling to Assess In-Game Performance during Gameplay with CRYSTAL ISLAND

ARTICLE IN PRESS

Full length article

Using multi-channel data with multi-level modeling to assess in-game performance during gameplay with CRYSTAL ISLAND

Michelle Taub *, Nicholas V. Mudrick *, Roger Azevedo *, Garrett C. Millar *, Jonathan Rowe *, James Lester *

* Department of Psychology, North Carolina State University, 2310 Business Drive, Raleigh, NC 27695-7502, United States
* Department of Computer Science, North Carolina State University, 800 Old Drive, Raleigh, NC 27695-8206, United States

ARTICLE INFO

Article History:
Received 1 July 2016
Revised 12 January 2017
Accepted 30 January 2017
Available online 2017
An earlier version of this study was presented at the 13th International Conference on Intelligent Tutoring Systems (ITS 2016) in Zagreb, Croatia and published in A. Micari, J. Langer, A. Pannasch (Eds.), Proceedings of the 13th International Conference on Intelligent Tutoring Systems—Lecture Notes in Computer Science 9890 (pp. 200-246). The Netherlands: Springer.

ABSTRACT

Game-based learning environments (GBLEs) have been touted as the solution for failure in educational outcomes. In this study, we address some of these major issues by using multi-level modeling with data from eye movements and log files to examine the cognitive and metacognitive self-regulation processes used by 50 college students as they read books and completed the associated in-game assessments (concept matrices) while playing the CRYSTAL ISLAND game-based learning environment. Results revealed that participants who read fewer books in total, but read each of them more frequently, and who had low proportions of fixations on books and concept matrices exhibited the strongest performance. Results suggest the importance of assessing quality vs. quantity during gameplay, such that it is important to read books in-depth, and compared to reading books once, the quantity. Implications for these findings involve designing adaptive GBLEs that scaffold participants based on their trace data, such that we can model efficient behaviors that lead to successful performance.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

Game-based learning environments (GBLEs) have been touted as the solution for failure in educational outcomes across several domains. Learning with GBLEs can be particularly effective for learning. Learning with GBLEs can be particularly effective for learning because they are designed to foster engagement during learning (e.g., Sabourin, 2013; Sabourin & Lester, 2014). Additionally, many games require self-regulated learning, in addition to learning processes, such as scientific reasoning (Millar et al., 2017). Scientific reasoning involves generating and testing hypotheses, and therefore students use self-regulated processes to assist in generating and testing these hypotheses. For example, during gameplay with CRYSTAL ISLAND, students are required to gather clues, and create and test hypotheses, to solve a mystery. It can thus be beneficial to integrate theories of SRL with scientific reasoning to investigate learning with GBLEs.

Despite the widespread enthusiasm, many critics have raised serious issues regarding the effectiveness of GBLEs for learning and problem solving (Mayer, 2015; Shute & Venkata, 2013). Unfortunately, the majority of published studies suffer from conceptual, theoretical, methodological, and analytical issues, undermining the value of GBLEs for improving learning, problem solving, and transfer of knowledge and skills across domains and age groups. Recent calls have been made to improve the quality of GBLE research. GBLEs are designed by including the conceptual and methodological approaches and interdisciplinary methods and analytical techniques to comprehensively improve the quality of GBLE research.

Please cite this article in press as: Taub, M., et al., Using multi-channel data with multi-level modeling to assess in-game performance during gameplay with CRYSTAL ISLAND, Computers in Human Behavior (2017), http://dx.doi.org/10.1016/j.chb.2017.03.030

* Corresponding author. North Carolina State University, Department of Psychology, 2310 Business Drive, Room 660, Raleigh, NC 27695-7502, United States.
E-mail addresses: mtaub@ncsu.edu (M. Taub), nmudrick@ncsu.edu (N.V. Mudrick), razervedo@gmail.com (R. Azevedo), gmillar@ncsu.edu (G.C. Millar), jjones@ncsu.edu (J. Jones), j Lester@ncsu.edu (J. Lester).
http://dx.doi.org/10.1016/j.chb.2017.03.030
(09-30-2017 Elsevier Ltd. All rights reserved.)
cognitive, affective, metacognitive, and motivational processes simultaneously during gameplay to understand their roles and impact other than the typical approach of using pre- to post-test measures and self-reports of motivation and engagement (Mayer, 2011). In this study, we address some of these major issues by using eye movements and log files to examine the cognitive and metacognitive self-regulatory processes deployed by college students while playing Crystal Island, a game that incorporates microbiology content and scientific reasoning to solve the mystery of a disease that has spread through a fictional remote island.

Research on self-regulated learning (SRL) indicates that students are self-regulating when they adaptively respond to both internal (e.g., use cognitive strategies during scientific reasoning) and external conditions (e.g., navigate a game environment in search of evidence) as evidenced by accurate monitoring and effective regulation of their cognitive, affective, metacognitive, and motivational processes during learning, problem solving, and performance (Azevedo, Johnson, Claessens, & Graesser, 2011; Azevedo, Taub, & Mutzick, 2015; Wiene & Azevedo, 2014; Wiene & Hadwin, 1998, 2008; Zimmerman & Schunk, 2011). Although research has shown that engaging in cognitive, affective, metacognitive, and motivational self-regulated learning processes can be beneficial for learning (Azevedo, 2009, 2014; Pintrich, 2000; Schunk & Greene, in press), research has also revealed that students do not typically deploy these processes effectively and efficiently during learning with advanced learning technologies such as intelligent tutoring systems, hypermedia, multimedia (see Azevedo et al., 2011, 2015; Graesser, 2015; VanLehn, 2016). Recent work on GILs and self-regulated learning has been conducted by Lester and colleagues (e.g., Sabornie & Lester, 2014) to examine of gameplay behaviors are predictive of learning, performance, engagement, and motivation using traditional statistics, data mining and machine learning. The current study extends this work by converging eye movements and log files to examine the underlying cognitive and metacognitive processes used by college students to solve the mystery on Crystal Island.

1.3. Theoretical Frame

Wiene and Hadwin’s (1998, 2008) Information Processing Theory (IPT) was used as the theoretical framework for the current study, which posits that learning occurs through a series of four cyclical phases, and information processing can occur within each phase. In the first phase, task definition, students must develop task understanding that drives their planning, monitoring and regulatory processes. In Crystal Island, students must understand the overall goal for the task, which is to solve the science mystery. In the second phase (goals and plans), students select goals for how they will accomplish the task (e.g., gather clues in each building) and plan how they will accomplish those goals (e.g., read books, complete embedded assessments). The third phase, strategy use, is when the students enact the plans to accomplish the goals they set in the previous phase (e.g., when students actually read the books and complete the embedded assessments). Strategy use and metacognitive monitoring can be inferred by analyzing in-game behaviors collected through eye movements and log files. The fourth phase (adaptation) is not addressed in this study. It is important to note that these phases are not necessarily sequential, and students can engage in multiple phases simultaneously, and in any order.

Information processing includes students engaging in cognitive, affective, metacognitive, and motivational processes to effectively self-regulate their learning (Azevedo et al., 2011, 2015; Azevedo, Taub, & Mutzick, in press), and these processes are related to monitoring and control. For example, students can monitor their use of strategies based on making metacognitive judgments (i.e., completing the in-game assessment an efficient strategy if the student does not understand the material), and use the goals and plans phase to dynamically monitor and control their use of strategies (i.e., re-reading a book as a cognitive learning strategy). Therefore, throughout the phases of learning, self-regulation implies that students engage in monitoring and control of self-regulated learning strategies.

Our use of Wiene and Hadwin’s (1998, 2008) model is advantageous because even though it has yet to be empirically tested, it is the only model that assesses SRL as an event that temporally unfolds over time (Wiene & Azevedo, 2014). The temporality of SRL is especially important during learning with GILs because students are presented with complex material, and their use of SRL strategies can change depending on the context. For example, during learning of complex text within a hypermedia-learning environment, we operationalize judgments of learning (JOLs) as assessing one’s understanding of the text by having them make that judgment, followed by a content quiz (Cenere & Azevedo, 2005). When making these judgments, there is a valence associated with it, such that a high rating of understanding would be a JOL, and a JOL would be indicative of low understanding. Although this might seem specific to hypermedia-learning environments, this can be applied to learning with GILs as well. During gameplay with Crystal Island, one activity students can engage in is reading books, which are associated with embedded assessments called concept matrices (Bowen, Stroh, Mott, & Lester, 2011). In each concept matrix, there are questions that test students on their understanding of the material in the book. Therefore, this can be seen as a JOL because students can self-evaluate their understanding of the text and go on to complete the concept matrix to test if they did understand the text. Furthermore, we can investigate the valence of the JOL based on the correctness of the responses. Thus, as opposed to the valence being associated with the students’ judgment of their understanding, the valence can be associated with how well the student performed on the assessment, such that low performance has a negative valence, and high performance has a positive valence. Therefore, we can apply IPT, specifically metacognitive monitoring and control to gameplay and scientific reasoning with GILs.

2. Literature review: GILs, assessment, and eye tracking

2.1. Game-based learning environments

The effectiveness of GILs across domains (e.g., math, computer science, biology, psychology, etc.) has come into question as several meta-analyses (Clark, Tanner-Smith, & Killingsworth, 2016; Connolly, Boyle, MacArthur, Hainey, & Boyle, 2012; Gitomer, Eccles, & Magman, 2012; Mayer, 2014; Wouters, van Nimwegen, van Oostendorp, & van der Spek, 2013) have revealed different results. More specifically, they have revealed that learning with GILs results in small to medium effect sizes for knowledge acquisition ($d = 0.29$, $p < 0.01$; Wouters et al., 2013), yet moderate to large effect sizes for knowledge acquisition ($g = 0.33$, $95\% CI [0.19, 0.48]$, $k = 27$, $n = 209$; Clark et al., 2016) and retention ($d = 0.36$, $p < 0.01$; Wouters et al., 2013), compared to conventional instructional methods (e.g., PowerPoint, classroom-based learning). However, findings did not show significant effects for learning with GILs on motivation ($d = 0.11$, $p > 0.05$; Wouters et al., 2013). In addition, Mayer (2014) determined that learning is most effective with games when the topic is science or second language learning, however games for teaching math and language arts is no more effective than using traditional classroom approaches (however a notable limitation is that few studies were investigated). Moreover, adventure games were found to be the most effective ($d = 0.72$).
followed by simulations (d = 0.62), and puzzle games (d = 0.45). Additionally, games were found to be the most effective for adults or college students (d = 0.74), then secondary students (d = 0.58), and elementary school students (d = 0.34). As such, Mayer (2014) posits that games for learning are effective, however the strength of the effect depends on the context, game type, and population. Several contributing factors explain these effects including a lack of operational definitions, a age of learners, learning phenomena and instructional approach (e.g., learning, problem solving, engagement, motivation), number and type of assessment methods (e.g., learning outcomes, transfer measures, self-reports), etc. which is due to a lack of theoretical grounding (Plas, Heiner, & Kürzer, 2015; Qian & Clark, 2016; Tsai, Huang, Hou, Hsu, & Chien, 2016; Filetecker and Kerres (2014) argue that game research needs to move beyond motivational and cognitive processes. Other processes, such as those underlying metacognitive monitoring, affect, and SRL have been largely unexamined (e.g., Sabourin & Lester, 2014).

Research regarding the effectiveness of GILs depends on the results across studies, lacks theoretical bases, and rarely targets higher-level skills like scientific reasoning, problem solving and SRL. Due to the varied learning domains tested and assessments used, research should emphasize the effect size of GILs rather than their statistical significance. Something also lacking from most of the meta-analyses assessing their effectiveness (Connolly et al., 2012; Girard et al., 2012; Qian & Clark, 2016). As many GILs are cut across several genres (e.g., narrative-centered, multimedia, role-playing), it is difficult to generalize results found with one GIL (Plas et al., 2015).

2.2. Assessment of learning

The educational impact of GILs has traditionally been assessed with quasi-experimental designs assessing knowledge acquisition/ content understanding from pre-to-post-test (Connolly et al., 2012). These studies have been designed to assess individual design characteristics such as uncertainty, feedback, learner control, cooperation, and interactivity and their influence on learning and self-reported experiences of motivation, engagement, etc. using traditional scales (e.g., Caplak, Ozçelik & Ozçelik, 2015; Calderon & Ruiz, 2015). Recently, more sophisticated analytical techniques such as path analysis and dynamic systems approaches have been employed to assess the effectiveness of GILs. For example, Caplak et al. (2015) investigated the influence of competition on learners’ motivation and post-test scores. Students were assigned to either the competition group, where they had access to other players’ scores, or the control group, where they did not have access. Results from the path analysis indicated that those in the competition group significantly outperformed those in the control on self-reported motivation scales and post-test scores. Alternatively, Snow, Allen, Jacyoyna, and McNamara (2015) investigated the influence of agency during learning and how it related to choice patterns and self-explanation quality in START-2, a serious game for reading strategy training and comprehension. Results indicated that the less chaotic (e.g., evenly ordered on random) and the more controlled (strategic and systematic) players’ movements were, the higher their perceptions of agency and learning outcomes. These analytical techniques are in-line with stealth assessment methodology rather than traditional analytical approaches (Shute, 2011), providing opportunities for using unobtrusive methods (e.g., eye tracking and log files) to examine the underlying cognitive and metacognitive SRL processes during gameplay with GILs.

The goal of stealth assessment is to provide valid, reliable, and unobtrusive measurements of students’ interactions with the game itself. Furthermore, stealth assessment allows for the evaluation of unfolding content understanding, where the interrelatedness of in-game actions (e.g., reading science books, collecting evidence) provide evidence about content learning, scientific reasoning, problem solving, and use of cognitive and metacognitive SRL processes compared to traditional studies (see Shute & Moore, in press). As such, the current study includes in-game assessment outcomes, eye tracking, and log files to provide a more comprehensive understanding of the cognitive and metacognitive processes underlying successful learning with Crystal Island.

2.3. GILs and eye tracking to assess self-regulatory processes

Despite the potential of eye tracking for making inferences about the cognitive and metacognitive processes underlying successful learning (Mayer, 2010; van Gog, Jamieson, Scheiter, Gerjets, & Paas, 2009), limited research has examined eye movements during learning with GILs. Specifically, analyzing eye movements can reveal what students are attending to and for how long during learning (Mayer, 2010; Scheiter & van Gog, 2009; van Gog & Jamieson, 2013). This methodology has recently been used to study multimedia and hypermedia-based intelligent tutoring systems (ITSs, e.g., Bondarev et al., 2013; Jaques, Conati, Harley, & Azevedo, 2014; Taub & Azevedo, 2016) to examine learners’ knowledge acquisition by assessing the total number of fixations on, and transitions between, areas of interest (AOIs) to assess content understanding (O’Mello, 2016; Huys & Nurmio, 2006).

Lastly, the limited research published on eye movements and GILs suffers from serious conceptual, theoretical, methodological, and analytical issues (e.g., Plas et al., 2015).

2.4. Crystal Island: a synthesis of previous work

Over the past 10 years, Crystal Island has served as a research platform for a broad range of studies including work on science problem solving (Sabourin, Rowe, Mott, & Lester, 2012; Spires, Rowe, Mott, & Lester, 2011), student learning (Adams, Mayer, McNamara, Koenig, & Wames, 2012), embedded assessment (Min, Rowe, Mott, & Lester, 2013), student affect recognition (McQuiggan, Robinson, & Lester, 2008), Sabourin, Mott, & Lester (2011), tutorial planning (Lee, Rowe, Mott, & Lester, 2014; Mott & Lester, 2006; Rowe & Lester, 2015), student knowledge modeling (Rowe & Lester, 2010); student goal recognition (Ha, Rowe, Mott, & Lester, 2011; Min, Mott, Rowe, Liu, & Lester, 2016), and virtual agent behavior (McQuiggan, Rowe, & Lester, 2008; Min, Wiggins, Perazzolo, Vail, Boyer, Mott, Frankowsky, Werber, & Lester, 2016; Rowe, Ha, & Lester, 2008).

A primary thread of research with Crystal Island has been investigating the relationship between learning, problem solving, and engagement in GILs. For example, Rowe et al. (2011) examined whether learning effectiveness and engagement are synergistic or conflicting in GILs. Compleimentary work investigating students’ off-task behavior, a symptom of student disengagement, found that going off-task was associated with reduced student learning, despite only accounting for approximately 5% of student gameplay time (Sabourin et al., 2012). Individual differences (e.g., prior knowledge, self-efficacy, gender) and SRL skills (cognitive strategy use) have also been found to serve an important role in student learning and problem solving in Crystal Island (Rowe, Shores, Mott, & Lester, 2010; Sabourin, Mott, & Lester, 2013). Moreover, a study by Nierfeld, Shores, and Hoffmann (2014) investigated the impacts of SRL on science content learning and problem-solving performance, as well as the influence of gender on SRL. Results found that strategy use (i.e., use of an in-game cognitive tool for diagnostic problem solving), monitoring bias, science self-efficacy, and situational interest were independently predictive...
of in-game performance, as well as gender. These prior studies are distributed from the current work by focusing primarily on datasets comprised of game trace logs and student self-reports, rather than multimodal data streams.

More recently, Min et al. (2016) investigated the application of multimodal learning analytics to devise models of companion agent behavior during game-based learning. By using two sequence-labeling techniques, long short-term memory networks (LSTMs) and conditional random fields, with multimodal data, including game trace logs, electrodermal activity, and facial action units (FAUs), they modeled companion agent behavior recorded during a Wizard-of-Oz study with middle school students. They found that utilizing FAUs and game trace logs in concert with LSTM models yields the best performing model of companion-agent behavior, yielding a 4.26% marginal improvement in predicting the wizard’s dialogue-act behaviors compared to a baseline approach. However, Min et al. did not collect eye gaze data, nor did they focus on student strategy use in their work. In MLM, the studies conducted by Lester and colleagues with Crystal Island are theoretically-based and converge learning outcomes, self-report measures, and trace data to examine certain SRL processes with adolescents.

3. Current study: assessing and converging multi-channel data with Crystal Island

The current study extends previous published research on GBLIS and the work on Crystal Island by Lester and colleagues by using eye tracking and log files to assess college students’ cognitive and metacognitive gameplay while using Crystal Island. More specifically, we converged specific in-game behaviors with eye tracking to assess how well students performed using in-game embedded assessments during learning and gameplay with GBLIS. As such, the goal of the current study was to use multi-level modeling (Raedensbush & Bryk, 2002) with log files and eye tracking to examine how students were reading books and completing the associated concept matrices as they played Crystal Island. The game has several in-game activity features, however for this study we focused on reading books and completing concept matrices because these students have low prior knowledge of microbiology, and these activities can help participants engage in scientific reasoning and use cognitive and metacognitive processes. Specifically, reading can foster knowledge acquisition, and the concept matrices indicate what is relevant in the adjacent text to answering the questions correctly (Lester, Matt, Robinson, & Rowi, 2013). Together, these features support the scientific reasoning process as well as self-regulated learning. Additionally, by investigating the in-game books and concept matrices, we are attempting to bridge cognitive and metacognitive processes, which is related to monitoring and control of learning. By investigating reading with its associated assessment, we can infer that when the student assesses they had read enough to complete the assessment, the student engaged in that control process, and switched from making that metacognitive monitoring judgment (i.e., the student decided they have read enough material) to using a cognitive learning strategy (i.e., began completing the assessment). Thus, in this study, we can single out making metacognitive judgments from using cognitive learning strategies, but continue to investigate how they are closely linked.

Furthermore, the foundation of measuring learning or learning-related behaviors with advanced learning technologies is being able to measure these behaviors overtly. For example, we do not need to infer if a student is reading a content page—we know when the student is doing so because log files indicating the student has clicked on that page, and eye tracking providing evidence of fixations on areas of content (i.e., kayaking the text). Alternatively, when measuring book reading and concept matrix completion, we can measure and track these behaviors, which is adhering to the foundational intent of doing research with advanced learning technologies. In this study, we know that students are reading in-game books because log files capture students’ mouse clicks to open and close the books, and we know they are reading because the eye-tracking data captures fixation duration on the books, and indicates that the students are reading particular books. In addition, we know when students are clicking from the book to the concept matrix, and when they are clicking to select their responses on the matrices. Additionally, we know from the eye tracking that they are reading the concept matrices. Therefore, we are able to track these behaviors during gameplay, and are thus not making inferences about student behavior, but instead, we are overtly capturing behavior during learning with Crystal Island.

When assessing in-game behavior, it is beneficial to use multi-level modeling (MLM; Raedensbush & Bryk, 2002) because we can account for between- and within-subject variance, along with a repeated-measures component, without violating statistical assumptions required for inferential statistics. Specifically, with Crystal Island, students can read as many books as they want, and MLM allows us to investigate reading books at multiple time points without having to collapse the data to an overall book reading instance. Furthermore, as we know SRL fluctuates over time, we want to investigate these fluctuations, both between and within participants. For example, one student might read a different number of books than another student, which would give us the between-subjects variability in book reading behavior. Additionally, a student might have longer fixation durations on books at the beginning, compared to the end of the game, demonstrating within-person fluctuations. Therefore, in this study, by using MLM we could investigate in-game assessment performance at multiple book instances (i.e., repeated-measures), and how fluctuations between and within-subjects (i.e., log files and eye tracking) were predictive of in-game performance.

3.1. Research questions and hypotheses

We address three sets of hypotheses to examine the number of concept matrix submission attempts. The number of attempts is a measure of in-game assessment as it measures how many times the students had to attempt to answer the concept matrix correctly (with a maximum of 3, as responses were auto-filled after three attempts). Our first research question focused on log-file data: Is there an association between the number of books read and the frequency of book opens by title with the number of concept matrix submission attempts? The second focused on eye-tracking data: Is there an association between the proportion of fixations on book content and on book concept matrices with the number of concept matrix submission attempts? The third research question combined the two: Is there a cross-level interaction between the number of books read, the frequency of book opens by title, and the proportion of fixations on the book content and concept matrices in the number of concept matrix submission attempts?

We proposed the following hypotheses: H1: the more books participants read, the more often they read each book, the less concept matrix submission attempts they made, resulting in better performance; H2: the longer fixation durations on the book content and concept matrices, the fewer concept matrix attempts, resulting in better performance; H3: there will be a significant interaction, such that log-file data (number of books and frequency of reading each book) and eye-tracking data (proportions of fixations on book content and concept matrices) will jointly impact concept
matrix submission attempts, with higher levels of all variables resulting in fewer attempts, and thus greater performance.

4. Methods

4.1. Participants

Fifty (N = 50) non-biology majors (56% female) from a large public university located in the southeast region of the US participated in a 1-session laboratory study. The participants’ mean age was 19.9 (SD = 1.69), 62% were Caucasians (n = 31), with the remaining racial and ethnic percentage including Native American, Asian/Asian American, African American, and Hispanic. 20% of participants reported level of proficiency and background with playing videogames to be skilled (from a range of not at all skilled to very skilled). The mean pre-test score was 55.8% (SD = 25.9), which indicates they had little prior knowledge in premedical fields, including premed at Creusa, Island, the game-based learning environment used in this study. Participants were monetarily compensated $10/hour and received up to $25 for completing the study.

This game-based learning environment was originally developed for middle school students; however, we adapted the game for college students so we could investigate gameplay among an older population, as according to Mayer (2014) games are the most effective for college students (see Section 2.1 above). We adapted the pre- and post-tests by making them more difficult, and we developed two separate versions so that participants did not complete the same test at pre and post.

4.2. Materials

Participants completed several self-report measures, including a demographics questionnaire, and others related to emotions and motivation, including the Emotion-Values Questionnaire (Azevedo, Harley, Trevers, Feyzi-Behnaagh, Duffy, Boucher, & Landis, 2013; Peckir, Elliott, & Malin, 2010), the Achievement Goal Questionnaire (Elliot & Murayama, 2008), the Intrinsic Motivation Inventory (Ryan, 1982), the Perceived Interest Scale (Schraw, 1997), and the Prensky Questionnaire (Witmer & Singer, 1998; Witmer, Jerome, & Singer, 2005). Participants’ prior knowledge about microbiology was assessed with a researcher-developed four-choice, 21 multiple-choice question pre-test and a similar post-test (adapted from Niefeld et al., 2014) immediately following game play. The questions contained 12 factual (e.g., declarative knowledge) and 9 procedural (e.g., You observe a biological agent and notice that it does not have a nucleus. What type of agent might you be looking at?) questions. These materials were chosen because we wanted to establish participants’ levels of prior content knowledge and prior gaming experience, as well as establish a baseline measure of levels of emotions and motivation. However, the questionnaires and content tests were not analyzed for this study.

4.3. Apparatus

4.3.1. Workstation

The equipment used in the study included a Dell Precison T7910 Workstation with 32 GB of memory and a 2.6 GHz Intel® Xeon®

CPU E5-2630 v3. This computer was connected to an external monitor with a resolution of 1920x1080 with the SMIVERSE® eyetracker attached below. Video recording of the session was captured using a Logitech 960 webcam positioned over the monitor and processed using the FACET module for Modern Attention Tool (2016). The participant was seated in front of the external monitor. The workstation also captured log-file data (i.e., information recorded about learner or system interactions, such as mouse clicks). This includes information regarding location, item, activity, and characters involved in any activity. For this study, we only analyzed eye-tracking and log-file data.

4.3.2. Eye tracking

Eye-tracking data is fed into Attention Tool in real time from the SMIVERSE® eyetracker (Sensomotoric Instruments, 2014). The infrared camera records with a sampling rate of 10 Hz, allowing for the smallest eye movement (0.03°) to be detected. The eyetracker collects participants’ eye gaze, fixations, and saccade movements, all used in tandem with MOTions to enable detailed post hoc analyses. For this study, we included fixations only, which is defined as looking at an area of interest (AOI) for a minimum of 250 ms, and no dispersion value because the AOIs are predefined and fixed, with a refresh rate of the monitor at 30 Hz.

The research team used the following algorithm to determine whether the participant was looking at an object in the 3D virtual world of the game: (1) using the screen coordinates of the participant’s eyes, an invisible ray is cast into the 3D world perpendicular to the screen, (2) if the ray collides with a 3D object (e.g., a book), the object is recorded and a timer is started, (3) if the timer exceeds the Attention Threshold setting (which is 250 ms), a fixation event is recorded in the trace data, and (4) the system continuously repeats these steps to determine how long an object is viewed and whether the participant shifts their attention to another object. Since CRISTAL ISLAND is a first-person GIBEL, it is important to note that the virtual camera the student is viewing the game world through can change position and orientation as the student navigates through books, areas or objects around in the virtual environment. In this situation, the student could be looking at the same position on the screen (e.g., the center of the screen), and the 3D objects being displayed at that screen location can change based on camera movement. To account for this, the CRISTAL ISLAND software ensures that a ray is cast at least as frequently as the camera is changing. This is configured using the Camera Movement Detection Frequency setting, which is typically set to 30 Hz. This means that the eye tracking logic is being executed 30 times per second, which matches the frame rate of CRISTAL ISLAND (30 frames per second). The frame rate indicates the number of times that the 3D virtual world is rendered as an image that is then displayed on the screen. This logic ensures that eye-tracking data is accurately mapped to 3D objects in CRISTAL ISLAND’s virtual environment.

4.4. CRISTAL ISLAND

CRISTAL ISLAND is a game-based learning environment designed (see Fig. 1) to teach students about microbiology scientific reasoning, and literacy, by having them gather clues to solve the mystery of what illness has impacted all the inhabitants of the island (Rowe et al., 2011). The game was developed using the Unity Engine (Unity, 2015) to be played on the computer. In the game, the participant played through the first-person view as the main character and having just arrived on the remote island where a mysterious illness has spread throughout a research camp located on the island.

Cristal Island is a narrative-centered learning environment that combines both inquiry-based learning as well as direct instruction.
(Lester, Ha, Lee, Mott, Rowe, & Sabourin, 2013). As the narrative plot of Cresta Iluso unfolds, players can naturally begin to develop inferences pertaining to possible future events. These inferences are then turned into hypotheses formed from acquired factual evidence, an essential component of inquiry-based learning (Lester et al., 2013). Furthermore, the exploratory model of inquiry-based learning is generally assisted by a facilitator and starts by posing problems or scenarios, rather than present a specific path to knowledge (e.g., all ten of the non-player characters within the learning environment).

Considering the narrative and exploratory nature of the environment that Cresta Iluso offers, explicit pedagogical instruction is needed to avoid decreasing the achievement of goals. This is accomplished by the direct instruction embedded in the exploratory learning architecture of Cresta Iluso, responsible for the planning of all the events throughout the narrative. This more structured and external regulated form of pedagogy is noticed in Kim the camp nurse’s instructions to the player. After informing the player that an illness has spread throughout the research camp and tasking the player to investigate the source of the illness, Kim, the camp nurse, lists the tools available by stating, “You can gather clues by talking to other team members, exploring the camp, and using the laboratory’s equipment”. As such, Cresta Iluso both supports inquiry-based learning by offering various ways to establish hypotheses throughout gameplay and direct instruction by programming fixed events (Lee, Mott, & Lester, 2011). As a result, the narrative-centered architecture successfully supports effective learning.

4.4.1. Buildings
In all buildings, there are different books and research papers, posters, food items, and non-player characters that participants can interact with. To gather clues, there are multiple buildings participating can visit, where they can engage in different activities (see below). In the infirmary, participants are given background information regarding the mystery. There are also sick patients who can help the participant gather clues. In the living quarters, participants can converse with experts on microbiology. In the dining hall, participants can collect food items that patients have been eating. Finally, in the laboratory, participants can test the food items they have collected in the other buildings.

4.4.2. In-game activities
To solve the mystery, the participant must identify the disease and its transmission source by exploring the environment and interacting with non-player characters (NPCs), reading posters, testing food items that are found in different buildings around the camp, completing the diagnosis worksheet, reading complex texts (e.g., books and lab articles), and completing concept matrices, which are assessments regarding the content found in those texts.

4.4.2.1. Conversations with non-player characters. The conversations throughout the game are multimodal representations of communication—they employ spoken language, facial expressions and text, as well as presenting the player with text dialogue options to choose from when communicating with the characters. The dialogue options are fixed dichotomies supplied by voice actors (Rowe et al., 2011). There are multiple non-player characters (NPCs)
that participants can converse during gameplay. Kim is the camp nurse and teaches the player how to successfully use the diagnosis worksheet and advises that speaking with the sick patients is a good start to establishing a hypothesis for a diagnosis. There are three sick patients: Teresa, Greg, and Sam, who report on their symptoms and previously eaten food. The player may also converse with Bryce, Ford, and Robert, who are experts in related topics and microbiology (such as viruses and bacteria), to gain a more sophisticated understanding of microbiology. The player can also talk to Quentin, the camp cook, who informs them of foods that inhabitants have been eating recently in pursuit of a solidified and coherent hypothesis regarding the source of the illness. Finally, Else works in the laboratory, and can assist with testing food items.

4.4.2. Testing food items. When participants collect food items, they can bring them to the laboratory to engage in hypothesis testing and test them for being potential transmission sources of the illness. The food items can be tested for contamination with viruses, bacteria, mutagens, or carcinogens. In addition, the scanner requires participants to input the reason for testing the items, for example, sick patient reported eating it. Reasons for testing food items include: “Sick members ate/drunk it”; “It wasn’t stored properly”; “It often carries disease”; “It looked dirty”; and “Sick members touched it”. The scanner will then identify if the test is positive or negative, and if the test is positive, the participant can confirm hypotheses that the illnesses is of a particular type (e.g., vi-

rus), and that the transmission source is a positively tested item (e.g., milk). When these data are collected, the participant can gather these clues by entering them in their diagnosis worksheet.

4.4.2.3. Diagnosis worksheet. During investigation, the participant uses a diagnosis worksheet to track and organize pertinent information (findings, hypotheses, and final diagnosis). The diagnosis worksheet serves as a scaffolding element within the game, such that it is designed to support problem-solving processes as well as a space for the participants to record and collect hypotheses and possible diagnoses (Lester, Spires, Netfield, Manogue, Mott, & Lobene, 2014). In this worksheet, participants can review their clues and reason about the hypotheses. The worksheet assists the players in solving the mystery, and allows them to travel back to the infirmary in order to find the completed diagnosis worksheet to Kim, the camp nurse. If the diagnosis proves to be incorrect, Kim will point out the error and recommend the player to reconsider either the transmission source or the illness or treatment plan. This feedback is valuable considering the player can use it to assess how close they are to solving the mystery and how to correct the original diagnosis. Once the player has listed the correct disease, source, and treatment plan, and submitted the diagnosis to Kim, the science mystery is solved.

4.4.4. Books and concept matrices. Participants could read books and research articles to learn about potential relevant material for solving the mystery (e.g., reading about viruses). In-game books were associated with embedded assessments, called concept matrices, to further facilitate students’ scientific reasoning and comprehension of complex scientific text. Concept matrices were associated with each book and article dispersed throughout the environment and were required to be completed any time a participant opened such an item. The questions within the matrices were presented in multiple-choice format and were directly relevant to the information being presented in the books and articles. For further usability, participants were able to switch between the matrix and scientific text so that they did not have to memorize the text. Furthermore, after three failed attempts at answering a question in the matrices, the correct response was then automatically filled in. As such, participants were always provided with the correct information (i.e., notes) on relevant content necessary to successfully solve the mystery.

The addition of the concept matrices helps ensure that the direct instruction embedded within the Coren Tinee environment is maintained. This feature is in addition to the participants’ ability to gather clues utilizing the backpack, converse with team members and sick patients (e.g., fixed dialogues associated with each game-based agent), explore the island (manually or through a fast-travel interface dependent on condition assigned), and use lab equipment to test food items (e.g., bread, egg, milk, etc.) for the source of the illness (e.g., salmonellosis, influenza, ebolavirus, etc.). This strategic instruction is also seen during the tutorial, when Else informs the player on how to travel within the environment and take notes when reading books or articles by way of a concept matrix. This explicit pedagogical scaffolding is directly indicative of direct instruction.

All of the abovementioned activities provide the participant the ability to engage in scientific reasoning, which involves hypothesis generation, followed by the formation of conclusions once the hypotheses have been successfully tested (Mills et al., 2011; Spies et al., 2011). For this study, we examined behaviors related to reading books and completed concept matrices as they support scientific reasoning and SRL.

4.5. Experimental procedure

The study took place during a single session. Depending on condition assigned, the session lasted anywhere from one to two and a half hours (M = 50.39 mins, SD = 29.50 mins). At the start of the session, the participants were greeted in the laboratory, and asked to have a seat in front of the workstation. Once seated, they were presented and asked to sign the consent form. After the consent form was signed, participants were instrumented, and then presented with an overview of the study. Following the overview, participants began assessing pre-session, including the demographics questionnaire, Emotion-Values Questionnaire, Achievement Goal Questionnaire, and the pre-test about microbiology.

After completion of the questionnaires, the researcher began calibration of the eye-tracking equipment. A 5-point calibration was accomplished by asking the participant to keep their head still while focusing on a white dot that moved around the screen. Calibration was repeated until the participant reached a satisfactory level (offset of eyemovements that are less than 0.05 mm) or made 5 attempts. Once the participant completed eye-tracking calibration in Coren Tinee, eye tracking was calibrated within Attention Tool. A baseline was then established for the facial recognition of emotion software as well as for the EDA bracket in Attention Tool.

After calibration, and before starting the game, the participant was presented with an overview of the experimental session. During the overview the participant learned about the setting of the game (i.e., remote island) as well as what their role was and how to solve the mystery. They were informed that their role throughout the game was to explore the camp to investigate the illness and solve the mystery of what has impacted all the inhabitants of the island. Additionally, the experimenter briefly mentioned what must need to be completed in order to successfully solve the mystery such as read books, articles and posters, interact with characters, gather clues and test food items.

Following completion of the game, the participants were asked to complete a series of questionnaires including the Emotion-
Values Questionnaire, Intrinsic Motivation Inventory, Perceived Interest Scale, Presence Questionnaire, followed by the microbiology content post-test. Upon completion of the questionnaires, participants were debriefed, thanked, and paid for their participation.

4.5.1. Experimental conditions

Prior to gameplay, participants were randomly assigned to one of three experimental conditions with varying levels of user agency. In the full agency condition, there were no constraints on which buildings to visit, or which order to visit them in. Additionally, participants could read whichever books, posters, and research papers they selected, and talk to whichever NPCs they chose to. The Partiel Agency condition required participants to travel through the buildings in a particular order, and only gave participants full autonomy once inside the buildings. In addition, each time they visited a new building, they were unable to leave until they read all of the resources, and conversed with all NPCs in that building. However, the order in which they interacted with the game elements within a building was up to them. In the No Agency condition, participants watched a prerecorded screen capture video of an expert researcher’s play-through of the game, along with a narration of his actions. The participants had no ability to interact with the game and as such, no trace data that captures user or system information was available. In total, the recording lasted 5462 s (91.5 min).

We did not examine the impact of experimental condition on concept matrix attempt submissions for this study. First, since we used online trace process data as our predictor variables, we could not include participants from the no agency condition, as they did not obtain online trace data during gameplay. Second, to increase our sample size at the individual level per group, we did not want to further categorize participants into two groups, which would result in fewer individuals per group. For example, we had a total of 50 participants, and did not want to further distinguish them into two groups of 25, as this is not enough per group for multilevel modeling (see below). Therefore, we did not include experimental condition as a predictor variable when coding and scoring our data.

4.6. Coding and scoring

For this study, we analyzed data from the log files and eye-tracking, both of which were automatically generated from our trace data pipeline, which was developed to calculate a preset list of variables related to gameplay, learning, problem solving, scientific reasoning, and self-regulated learning (e.g., session duration, fixation duration, number of books, book duration, etc.). These data were calculated at each instance level, for each activity. An activity is one of the in-game activities described above (e.g., book and concept matrix or testing lab items), and an instance is one occurrence of that activity (e.g., opening book number 1 and opening book number 12 are separate instances within the book activity). More specifically, we have instance-level data for books and concept matrices, conversations with non-player characters, game screen, workbook, open and edit, testing lab items, etc. Therefore, for each activity, we have a data set for each participant, at every instance of engaging in that activity. For example, if a student read 20 books, we would have a list of variables for each of those 20 instances, yielding 20 rows of data for that participant. As such, since each participant read a number of books, this yielded 1996 total rows of data. For this analysis, we analyzed log files and eye-tracking data from the data pipeline, at the book instance level.

4.6.1. Log files

We extracted two log-file variables for this study. First, we included the number of books participants opened. This value was automatically generated from the data pipeline. Additionally, we used book instances as our repeated-measures, level 2 variable (with the individual as the level 1 variable) for our multi-level modeling analyses. There were a total of 21 different books scattered throughout the buildings in Central I unaware, however participants could read books multiple times2 during gameplay (M = 24, SD = 9.36 for this study). We also extracted the frequency of book reads by title, which was also automatically detected from the log files. This variable defines how many times participants read each book, and so a book with a high frequency by title score reveals that the participant read that particular book many times. Finally, our dependent variable, concept matrix submission attempts, was coded from the log files, which indicated the number of times the concept matrix was attempted (M = 1, SD = 0.82), with a maximum score of 3 (see above), for each book instance.

4.6.2. Eye tracking

We also extracted eye-tracking data from the data pipeline, where we pre-determined eye-tracking variables we wished to analyze. These data were automatically extracted with the data pipeline, as the eye tracking was embedded in the game software, and thus eye-tracking data was captured and processed through the data pipeline. For this study, we analyzed the proportions of fixations on book content and the proportions of fixations on book concept matrices. The proportions were calculated based on the fixation duration on the books for each specific book instance, divided by the total fixations on all books. Proportions were calculated separately for the book content and the book concept matrices, yielding two proportions:

\[
\text{Book Content Fixation Duration per Book Instance} = \frac{\text{Total Book Content Fixation Duration}}{\text{Book Concept Matrix Fixation Duration per Book Instance}} = \frac{\text{Total Book Concept Matrix Fixation Duration}}{\text{Total Book Content Fixation Duration}}
\]

Therefore, we had two different proportion scores for each book instance. Once we coded and scored these log-file and eye-tracking data, we began to run our models to test our research hypotheses.

For this analysis, we used multi-level modeling (MLM), a statistical approach that combines the strengths of linear regression and repeated-measures ANOVA into one statistical test (Raudenbush & Bryk, 2002). This can be beneficial because during gameplay, students can engage in metacognitive strategies multiple times. For example, a student can read multiple books, and we can collect online trace data pertaining to student behavior during each book instance that may be indicative of metacognitive monitoring (e.g., return to previous book, prolonged fixation on relevant information supporting a particular hypothesis, etc.). However, when using traditional inferential statistics to examine process data during learning with GIBBS, we face the issue of violating statistical assumptions that must be satisfied when using these tests. For example, the assumption of independence of cells requires that each cell of data does not relate to one another (i.e., all cells are independent). However, if examining multiple instances of an event, each participant will occupy multiple rows of data (i.e., one row per instance). Thus, multiple cells will be dependent on

Please cite this article in press as: Taioli, M., et al., Using multi-channel data with multi-level modeling to assess in-game performance during gameplay with Core Business, Computers in Human Behavior (2017), http://dx.doi.org/10.1016/j.chb.2017.01.038
one another, as the data stems from the same participant. By using MLM, we can investigate multiple instances of events within students without having to satisfy this assumption, making MLM an ideal statistical technique when investigating learning and gameplay with CBILES, as the typical behavior is to engage in cognitive and metacognitive processes multiple times. In addition, MLM allows us to test for both between- and within-subject variance in the variable being investigated. For example, if we investigate the number of times a student reads a particular book, we can determine if this student reads books multiple times at the beginning of the learning session, compared to the end of the session, where they may read each book only once. This informs us of differences in reading behavior at different time points during the session, revealing within-subject variance in frequency of each book read.

In addition, to compare between-subjects, we can investigate how one student’s book reading behavior differs from another student. More specifically, we can determine if student A reads each book more frequently compared to student B, who only reads each book once. As such, we can investigate both of these types of subject variance simultaneously when using MLM analyses. Therefore, for this analysis, we used MLM to investigate how students read multiple books, and how students completed multiple associated concept matrices during gameplay with Creep house.

Although MLM is a powerful analytical technique, there are some limitations to using it, which led us to forego examining the impact of experimental conditions on concept matrix attempts. One requirement to using MLM is that we need a sufficient sample size, which is typically 30 participants per group. Although we did include data from 50 participants in this study, there were only 25 per condition, which was not sufficient for our models to be run. As such, we were not able to examine the impact of experimental condition on concept matrix attempts for this study, nor did we examine the effect of book-reading activities on post-test scores, but will address this limitation in future studies with larger samples per experimental condition.

5. Results

We used SAS software 9.4 (SAS Institute Inc., 2002) to run our analyses, with a restricted maximum likelihood (REML) estimation method, and a variance components covariance structure. The first step in using MLM requires a fully unconditional model, which allows us to determine if there is sufficient between- and within-subject variance in our dependent variable, and the intra-class correlation coefficient (i.e., the percentage of variance explained at the between- and within-subject levels), prior to running models with predictor variables (i.e., a null model). For this study, our fully unconditional model was run with number of concept matrix attempts as our dependent variable, using the following equation with the individual and book instance levels, where \( \beta_0 \) is the slope for the dependent variable, \( \tau_0 \) is the estimate of the grand mean of the number of concept matrix submission attempts, and \( \nu_{00} \) and \( \nu_{01} \) are the error terms at levels 1 and 2, respectively:

\[
\begin{align*}
\text{Level 1: Matrix Attempts} & = \beta_0 + \tau &
\text{Level 2: } \beta_0 &= \gamma_{00} + \nu_{00} \\
\end{align*}
\]

Results indicated that the grand mean (i.e., fixed effect) of the number of concept matrix submission attempts was significantly different from zero: \( \gamma_{00} = 1.01, t = 10.75, p < 0.0001 \). In addition, random effects results indicated that there was significant between- \( (\gamma_{00} = 0.36, z = 4.06, p < 0.0001) \) and within-subjects \( (0.56, z = 23.86, p < 0.0001) \) variance in the number of concept matrix attempts, informing us that participants had different numbers of concept matrix attempts compared to each other, and having varying numbers of attempts at different points during gameplay. In addition, this model revealed that 5% of the variance in the number of concept matrix submission attempts was between-subjects, and 83% of the variance in the number of concept matrix submission attempts was within-subjects. Therefore, based on these results, there was sufficient variance at both levels of the dependent variable to proceed to running models with predictor variables.

5.1. Research question 1: Is there an association between the number of books read and the frequency of book opens by title with the number of concept matrix submission attempts?

We addressed this research question by running a random intercept model with a fixed slope, with predictor variables at both levels 1 (frequency by title) and 2 (number of books). This model had the following equations, where \( \beta_0 \) is the slope for the frequency by title predictor, \( \gamma_{00} \) is the association between the number of books and concept matrix attempt submissions, \( \gamma_{01} \) is the association between the frequency of book by title, and \( \tau_0 \) is the cross-level interaction between the number of books and the frequency of book by title on concept matrix submission attempts:

\[
\begin{align*}
\text{Level 1: Matrix Attempts} & = \beta_0 + \beta_0 \text{[Frequency by Title]} + \tau_0 &
\text{Level 2: } \beta_0 &= \gamma_{00} + \nu_{00} \text{[Number of Books]} + \nu_{01} \\
\beta_0 &= \gamma_{00} + \gamma_{01} \text{[Number of Books]} \\
\end{align*}
\]

Results from this model (see Table 1 for an overview) revealed a significant association between number of books and concept matrix attempts: \( \gamma_{00} = -0.014, t = -2.20, p = 0.0329 \), such that an increase in books is associated with a decrease in concept matrix attempts, demonstrating that opening books leads to better performance on the concept matrices. Results also indicated a significant association between the frequency of books by title: \( \gamma_{01} = -0.44, t = -4.54, p < 0.0001 \), with an increase in books associated with a decrease in concept matrix attempts, revealing that if students read a book multiple times, they will perform better on the concept matrices. However, a significant cross-level interaction (see Fig. 2) revealed a significant association between the number of books and concept matrix attempts, especially strong for fewer books and higher frequencies of books by title, such that the fewest attempts, and thus the best performance, was for participants who read fewer books, but read each book more frequently. In contrast, participants with the most concept matrix submission attempts, and thus performed the worst, read fewer books overall, and read each book less frequently. This model accounted for 46.3% of the between-subjects variance, and 63.4% of the within-subjects variance in concept matrix submission attempts.

Table 1

<table>
<thead>
<tr>
<th>Unstandardized coefficients for number of books and frequency by title with concept matrix attempts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed effects</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>Matrix attempts, ( \beta_0 )</td>
</tr>
<tr>
<td>Intercept, ( \gamma_{00} )</td>
</tr>
<tr>
<td>No. books, ( \gamma_{01} )</td>
</tr>
<tr>
<td>Err by title usage, ( \gamma_{01} )</td>
</tr>
<tr>
<td>Books*Freq by title, ( \gamma_{01} )</td>
</tr>
<tr>
<td>Random effects</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>Within-person fluctuation (( \nu_{00} )</td>
</tr>
</tbody>
</table>

\( \nu_{00} < 0.05, \ p < 0.0001, \ \nu_{01} < 0.0001 \)
5.2. Research question 2: Is there an association between the proportion of fixations on book content and book concept matrices with the number of concept matrix submission attempts?

To address this research question, we ran a level 1 moderation model with constrained slopes, with the following equations, where $b_{0h}$ is the slope and $Y$ is the effect of proportion of fixations on book content, $b_{A}$ is the slope and $Z$ is the effect of proportion of fixations on book concept matrices, and $b_{BA}$ is the slope and $X$ is the effect of the interaction:

Level 1: $\text{Matrix Attempts}_i = b_{0h} + b_{A}Y_i + b_{BA}XY_i + r_{A}$
Level 2: $b_{0h} = \gamma_{00} + \gamma_{01}X$  
$b_{A} = \gamma_{10} + \gamma_{11}X$  
$b_{BA} = \gamma_{20} + \gamma_{21}X$

Results revealed (see Table 2) no significant association between the proportion of fixations on book content ($\gamma_{00} = -.080, t = -0.56, p = .57$) and no significant association between the proportion of fixations on book concept matrices ($\gamma_{01} = 0.023, t = 0.06, p = .95$). However, there was a significant interaction ($\gamma_{20} = -0.17, t = -14.31, p < 0.0001$), revealing that the effect between proportions of fixations on book content and book concept matrices is particularly detrimental for poorer performance, such that students with high proportions of fixations on both the book content and book concept matrices had the highest number of concept matrix submission attempts (see Fig. 3). Overall, this model accounted for 1.59% of the between-subjects variance and 27.5% of the within-subjects variance in concept matrix submission attempts.

5.3. Research question 3: Is there a cross-level interaction between the number of books read, the frequency of book opens by title, and the proportion of fixations on the book content and concept matrices on the number of concept matrix submission attempts?

To test for the cross-level interaction, we first examined the unique associations between each predictor variable and the number of concept matrix submission attempts, and then tested for the interaction between all four predictor variables and the number of concept matrix submission attempts. We used the following equation, with the cross-level interaction slope as $b_{AX}$ and the interaction effect as $\gamma_{AX}$:

Level 1: $\text{Matrix Attempts}_i = b_{0h} + b_{A}Y_i + b_{AX}XY_i + r_{A}$
Level 2: $b_{0h} = \gamma_{00} + \gamma_{01}X$  
$b_{A} = \gamma_{10} + \gamma_{11}X$  
$b_{AX} = \gamma_{20} + \gamma_{21}X$

For this model, we only investigated the unique associations between each predictor and the IV and the cross-level interaction, and not all other combinations of interactions (e.g., frequency by title, content by title, number of books, etc.), thereby not obtaining all possible gamma values (i.e., $Y_{A}$, $Y_{AX}$, $X_{A}$, and $X_{AX}$ were not included). Results revealed a significant association between number of books ($\gamma_{00} = -0.011, t = -2.17, p = 0.035$), frequency of books by title ($\gamma_{10} = -0.35, t = -12.01, p < 0.0001$), proportion of fixations on book content ($\gamma_{20} = 0.50, t = 4.06, p < 0.0001$), and proportion of fixations on book concept matrices ($\gamma_{30} = 1.04, t = 5.50, p < 0.0001$) and concept matrix submission attempts (see Table 2), revealing all significant unique associations between each predictor variable and the dependent variable. Additionally, we found a significant interaction; $\gamma_{AX} = 0.077, t = 10.41, p < 0.0001$, demonstrating the importance of including all of these variables in

Please cite this article in press as: Taat, M., et al., Using multi-channel data with multi-level modeling to assess in-game performance during gameplay with CoM, in Press. Computers in Human Behavior (2017), http://dx.doi.org/10.1016/j.chb.2017.01.038

---

Table 2

<table>
<thead>
<tr>
<th>Fixed effects</th>
<th>Estimate (std. error)</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matrix attempts, $b_{0}$</td>
<td>0.52 (0.082)</td>
<td>6.25***</td>
</tr>
<tr>
<td>Content fixations slope, $b_{A}$</td>
<td>-0.083 (0.14)</td>
<td>-0.56</td>
</tr>
<tr>
<td>Interaction slope, $b_{BA}$</td>
<td>0.013 (0.01)</td>
<td>0.11</td>
</tr>
<tr>
<td>Interaction effect, $b_{AX}$</td>
<td>0.17 (0.01)</td>
<td>14.11***</td>
</tr>
</tbody>
</table>

Random effects Estimate (std. error) t

| "Within-person fluctuations (σ²)" | 0.41 (0.018) | 23.99*** |

*p < 0.05, **p < 0.001, ***p < 0.0001*

---

*This was done to simplify our analyses, as we had examined all combinations (e.g., interactions between a combination of some level 1 predictors and number of books), this would have yielded far too many results.*
Table 3: Unstandardized coefficients for number of books, frequency by title, and proportions of fixations on book content and book concept matrices by concept matrix attempts.

<table>
<thead>
<tr>
<th>Fixed effects</th>
<th>Estimate (std. error)</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matrices, v</td>
<td>0.15 (0.05)</td>
<td>2.7</td>
</tr>
<tr>
<td>Intercept, b0</td>
<td>0.03 (0.02)</td>
<td>1.5</td>
</tr>
<tr>
<td>No. books, b1</td>
<td>0.04 (0.02)</td>
<td>1.9</td>
</tr>
<tr>
<td>Freq. by title, b2</td>
<td>0.02 (0.01)</td>
<td>1.1</td>
</tr>
<tr>
<td>Intercept, b3</td>
<td>0.04 (0.02)</td>
<td>2.1</td>
</tr>
<tr>
<td>Local fixations slope, b4</td>
<td>0.05 (0.03)</td>
<td>1.6</td>
</tr>
<tr>
<td>Matrices fixation slope, b5</td>
<td>1.04 (0.19)</td>
<td>5.1</td>
</tr>
<tr>
<td>Interaction slope, b6</td>
<td>0.077 (0.007)</td>
<td>10.1</td>
</tr>
<tr>
<td>Random effects</td>
<td>Estimate (std. error)</td>
<td>2</td>
</tr>
<tr>
<td>Matrices (u)</td>
<td>0.070 (0.02)</td>
<td>3.5</td>
</tr>
<tr>
<td>Within-person variance (w^2)</td>
<td>0.04 (0.02)</td>
<td>2.1</td>
</tr>
</tbody>
</table>
positive association between the variables, and we predicted a negative association. The lack of significant main effects, but a significant interaction effect demonstrates the importance of investigating cognitive and metacognitive processes together, and how they jointly predict performance, as opposed to singling them out. In addition, the low proportions of fixations predicting better performance may result from participants strategizing by examining the questions in the concept matrix and going back to the text to search for the correct responses instead of reading through the entire text and then answering the concept matrix, however we cannot confirm this because we did not sequentially analyze behaviors within each book instance, which we will aim to do in future studies.

These findings are similar to research conducted by Tai et al. (2016), who examined fixations between different components in the learning environment, as well as transitions between those different components, and found that participants with high comprehension of the material had more strategic fixations and transitions. Thus, these findings are similar to ours regarding strategic proportions of fixation durations, such that participants in our study who had fewer concept matrix submission attempts, and thus better overall performance on the concept matrices also seemed to have more strategic fixations, as opposed to just reading through the content without using any strategies.

In addition, this strategizing behavior seen in our study could relate to how participants were processing the information in the books and concept matrices. Specifically, instead of reading a large range of material, participants spent more of their time processing smaller amounts of information more thoroughly. Thus, participants seemed to be strategizing and selecting the most relevant material for completing the concept matrices, and then spending valuable time reading and processing that specific content. This relates back to our findings from the first research question regarding quality vs. quantity, such that participants who performed better on the concept matrices were not reading larger quantities of information, rather, the quality of what they were reading showed more thorough investigation of the content. This once again supports Wiese and Halvín’s IPT model because participants were taking more time and potentially using operations, referred to as SMART (searching, monitoring, assembling, rehearsing, and translating) during those shorter durations. Therefore, participants are selecting specific elements within the text, which are relevant to completing the concept matrices correctly, to apply these operations to. Searching for relevant material is known as the metacognitive strategy of content evaluation (CE; Arcevedo, 2009; Greene & Arcevedo, 2009); therefore, these participants could be using CE’s as well, which is resulting in better performance. Overall, it seems that they are monitoring the material and controlling what they choose to read depending on what is relevant for them to perform well on the assessments. As such, this behavior demonstrates that participants are engaging in monitoring and control strategies, which is also supported by the IPT model and other studies assessing the role of cognitive and metacognitive processes with other advanced learning technologies (e.g., Arcevedo et al., 2011, 2015).

Lastly, although we speculated that students are strategizing, monitoring, and controlling their reading behaviors, having to make inferences about this finding reveals that we cannot solely rely on using only one type of trace data to indicate what exactly participants are doing during gameplay with GameOn books. For example, we assumed that participants were strategizing based on examining their eye-tracking data, but if we were to add log files to our eye-tracking data, we can assess participants’ clicking behavior to examine when they clicked on the book when they clicked on the matrix, back to the book, etc., and then we can combine data from both channels to identify if participants were truly strategizing by clicking back and forth frequently during each book instance. Furthermore, we can add other data channels to differentiate between strategizing or engaging in a different behavior that resulted in higher performance (Arcevedo, 2015). For example, if we use video data of facial expressions, we can examine if students became confused while reading, which caused them to stop reading. However, they might have been able to resolve that confusion (D’Mello & Graesser, 2012) by completing the concept matrix, which led to better performance on the matrices. Therefore, by including more data channels, we do not have to make inferences regarding our results and we can analyze these data together to determine what exactly led participants to perform better on the concept matrices.

6.3. Research question 3: Is there a cross-level interaction between the number of books read, the frequency of book opens by title, and the proportion of fixations on the book content and concept matrices on the number of concept matrix submission attempts?

For this final research question, we hypothesized that when including data from the log files and eye tracking, we would find a significant interaction, with higher values of all variables resulting in the fewest number of concept matrix attempts (i.e., negative associations). Our results were partially supported, such that there were significant unique associations between each of the four predictor variables and concept matrix submission attempts, as well as a significant interaction. However, the number of books and proportions of time spent on the books and concept matrices yielded results in the opposite direction as we predicted. Our results revealed that the fewest number of concept matrix submission attempts were for participants with fewer books (positive association), high frequency of books by title (negative association), as we predicted, and low book and concept matrix proportions of fixations (positive associations). These results demonstrate that it is important to read books to perform better on the concept matrices, however in doing so, we must still consider quality vs. quantity (Crosley & Averett, 2009; McShane & Shapiro, 2005). Specifically, it is still important to read each book more frequently, as opposed to less frequently, because reading a book only one time will not guarantee that all of the relevant information will be processed and retained for the concept matrices, and for the post-test, which is taken at the end of the session. However, when reading books more frequently, strategizing works better; for example reading for longer proportions of time. As such, the quantity of books does matter; but the quality, i.e., how the books are read, can impact performance as well.

Additionally, these results emphasize the importance of using multi-channel data because it gives us the full picture of how participants were using SRL and scientific reasoning strategies as they were playing the game. Specifically, when we look at interactions that include many data channels, we find that with the interactions, we are seeing different results from the main effects. For example, the number of books had a negative unique association with concept matrix submission attempts (i.e., more books was associated with fewer attempts), but in the interaction, fewer books (positive association) lead to better performance. It can be more representative of learning when we examine all the data together because we cannot single out data and ignore the effects of other data, which is why the interactions are important to consider. Therefore, all of the results indicate that when we are analyzing performance on embedded assessments, we should use multi-channel data, and examine how these data interact with each other to predict performance, including cognitive and metacognitive processes.
6.4. Limitations

Despite our informative and significant findings, we must acknowledge the limitations of this study. First, we did not assess sequences of eye-tracking behavior within each book instance, and so we cannot confirm the order in which participants read books and completed concept matrices. Specifically, we summed all fixations for each book instance, so if a participant switched back and forth five times during one book instance, the data looked no different from another participant who only went from book content to the concept matrix once, in terms of calculating the proportions of fixations.

Additionally, we conceptualized JOLs with CReST using a valence that is based on the correctness of performance on the concept matrices, as opposed to asking them to judge their understanding of the material for that book. As such, we are inferring correctness based on performance, as well as inferring that participants are ready to complete the assessment (and have thus judged that they have read enough material to complete the assessment) when they switch from the book to the concept matrices. This is in contrast to assessing JOLs with hypermedia-learning environments where participants either select or are prompted to make a JOL. (e.g., Azevedo, 2004; Greene & Azevedo, 2009). Thus, these participants do not choose to make the assessment, rather they choose to make the judgment, which is followed by the assessment. Our measure of JOLs within CReST issuing did not actually allow for participants to overtly make a judgment that we could measure without inferring it based on their behavior.

Finally, for this analysis, we only investigated some SRL processes, such as cognitive learning strategies (e.g., reading), and metacognitive monitoring (e.g., JOL). In addition, we also only investigated some in-game activities, such as reading books and completing concept matrices. Therefore, the current trace data used for this study may have only captured some of the many cognitive and metacognitive self-regulatory processes that can be used during gameplay, SRL, and scientific reasoning with CReST.

6.5. Implications and future directions

The results from this study, as well as the addressed limitations, have important implications for conducting future studies, and designing GRLIs that are adaptive based on participant activity. For example, more fine-grained analysis of eye gaze, such as by using XY-coordinate data from eye tracking sensors, will enable identification of specific sub-portions of a text or an embedded assessment that a student has fixated upon, as well as investigation of how students transition between them. Further, the application of sequence mining techniques holds considerable promise, enabling the creation of models that characterize student gaze and problem-solving behavior over time. For example, differential sequence mining techniques can be used to distill common patterns of student problem-solving behavior in game-based learning environments that distinguish different groups of students, such as those who demonstrate high self-regulatory skills and low self-regulatory skills (Saloum et al., 2013). These computational techniques can be used to induce models that map student gameplay behaviors and physiological states to descriptions of high-level sequential problem-solving strategies that unfold over time.

Overall, results confirmed that when playing CReST, some participants are, in fact, reading books and completing concept matrices, which emphasizes that we can provide them with adaptive scaffolding that can walk them through this activity, to ensure they are doing it successfully. Specifically, we can use multi-channel data to demonstrate how an expert participant would read books and complete concept matrices, based on the results from our study. For example, the system can allow an expert selecting many books, and reading each of those books a few times. In addition, the expert can demonstrate how strategicizing will result in better performance, such that they monitor the feedback content is relevant for the matrix, and choose to read that content only, which results in better performance on the matrices. Therefore, we can use these data to model to participants how they can play the game and solve the mystery by engaging in effective cognitive, metacognitive, and scientific reasoning strategies.

Finally, these results can inform the design of GRLIs to adaptively support individual learner’s cognitive and metacognitive processes during scientific reasoning. For example, if the participant is not reading a lot, and is not completing the matrices, the system can prompt them to open relevant books and complete the associated assessment. In addition, we can use multi-channel data to drive real-time scaffolding for participants. For example, if the eye-tracking data indicates that a participant is spending a significant amount of time reading in a non-strategic fashion (e.g., no evidence of metacognitive judgments and use of cognitive strategies), the system can model a more efficient way to read books and complete concept matrices, which can result in better performance on the matrices. Specifically, we can model efficient eye movements (e.g., D’Mello, 2010; Jarodzka, van Gog, Dene, Schenker, & Geerjets, 2013) that will scaffold participants to use more effective self-regulated learning strategies. The ultimate goal for designing these GRLIs is to foster the most effective learning for students, which can be demonstrated through high overall performance and accurate metacognitive monitoring and effective cognitive strategy use (Azevedo et al., in press) and, if adding an adaptive component to these environments can improve in-game and overall performance, which we can detect by examining participants’ performance on in-game and post-test components at the end of gameplay, this confirms that these environments should continue to be used for all different types of learners, in different educational settings to foster and scaffold effective learning, gameplay, SRL, and scientific reasoning.

Acknowledgment

The research presented in this paper has been supported by funding from the Social Sciences and Humanities Research Council of Canada (SSHRC 985—2011—1006) awarded to the third and last authors. The authors would like to thank Robert Taylor and Andrew Smith for assistance with system development, and Megan Price for assisting with running participants.

References


SIM Experiment Center 1.8.65 (Apparatus and Software), (2004). Boston, Massachusetts, USA: SensoryMotor Instruments.


CHAPTER 4: Using Sequence Mining to Reveal the Efficiency in Scientific Reasoning during STEM Learning with a Game-Based Learning Environment

Abstract: The goal of this study was to assess how metacognitive monitoring and scientific reasoning impacted the efficiency of game completion during learning with CYSTAL 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.

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

Introduction

Self-regulated learning (SRL) is an effective way of learning for learners 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 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.

**Theoretical Framework**

As our theoretical framework, we use the information processing theory (IPT) model 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 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, Johnson, Burkett, Fike, Linteanu, Cai, & Rus, 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, which 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.

**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 ($d = .72$), followed by simulations ($d = .62$), and quiz or puzzle games ($d = .45$); and games had the highest positive effects for adults or college students ($d = .74$), followed by secondary students ($d =$...
.58), and elementary students \((d = .45)\). 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, Rodrigo, Baker, Paquette, Shute, & Ventura, 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 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 in 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.

**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 CRYSFAL 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.

**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 different 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 gain. (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 are 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.
Methods

Participants and Materials

641 undergraduate students (59% female) from a large public North American university participated in this study. Participants’ ages ranged from 18-26 years old ($M = 20$, $SD = 1.64$). They were randomly assigned to one of three experimental conditions (see section, 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.

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 Figure 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).

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

**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 source. Participants were able to test these hypotheses by collecting and scanning gathered food items (see Figure 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.

**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 Figure 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 present 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.

**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 ($M = 87.39$ minutes, $SD = 20.8$ 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.

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 for example. 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.
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: $\frac{\text{PostTestRatio} - \text{PreTestRatio}}{1 - \text{PreTestRatio}}$. 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 playing the game, and so calculating proportions accounted for this, and allowed us to control for longer durations.

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 worksheets 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.

**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, bacteria, carcinogen, or mutagen, however the correct response could only be a virus or bacteria (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.

Table 1

Descriptions of Full Relevancy Codes.

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>RELEVANT--RELEVANT</td>
</tr>
<tr>
<td>2</td>
<td>RELEVANT--PARTIALLY-RELEVANT</td>
</tr>
<tr>
<td>3</td>
<td>RELEVANT--IRRELEVANT</td>
</tr>
<tr>
<td>4</td>
<td>PARTIALLY-RELEVANT--RELEVANT</td>
</tr>
<tr>
<td>5</td>
<td>PARTIALLY-RELEVANT--PARTIALLY-RELEVANT</td>
</tr>
<tr>
<td>6</td>
<td>PARTIALLY-RELEVANT--IRRELEVANT</td>
</tr>
<tr>
<td>7</td>
<td>IRRELEVANT--RELEVANT</td>
</tr>
<tr>
<td>8</td>
<td>IRRELEVANT--PARTIALLY-RELEVANT</td>
</tr>
<tr>
<td>9</td>
<td>IRRELEVANT--IRRELEVANT</td>
</tr>
</tbody>
</table>

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 bacteria spread by eggs, and a participant first tested the eggs for a pathogenic virus, and then tested the eggs for a pathogenic bacteria, this would be coded as two testing events. The sequential pattern would be: **PARTIALLY 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.

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: 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.

**Results**

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* = .25, *d* = .30, revealing there were no significant differences in proportional learning gain between participants who were more efficient at solving the mystery (*M* = .22, *SD* = .25) and participants who were less efficient at solving the mystery (*M* = .30, *SD* = .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* = .078, *d* = .45, the proportion of time reading books (*t*(62) = .48, *p* = .63, *d* = .12), or the proportion of time talking to Teresa and Quentin (*t*(62) = 1.28, *p* = .21, *d* = .32). Specifically, there were no significant differences in the proportion of time spent testing food items between more (*M* = .020, *SD* = .0086) and less efficient (*M* = .025, *SD* = .013) participants, no significant differences in the proportion of time spent reading books between more (*M* = .33, *SD* = .073) and less efficient (*M* = .32, *SD* = .090) participants, and no significant differences in the proportion of time spent talking to the NPCs Quentin and Teresa between more (*M* = .023, *SD* = .0061) and less efficient (*M* = .021, *SD*= .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 \( r(62) = -.33, p < .01 \), 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; \( r(62) = -.48, p < .01 \), 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; \( r(62) = .32, p = .011 \), such that the more time participants spent testing food items, the greater number of diagnosis worksheet submission 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.
Table 2.

*Correlations of Proportions of Time Testing, Reading, and Talking and DW Submissions.*

<table>
<thead>
<tr>
<th>DW Submissions</th>
<th>Prop. Testing</th>
<th>Prop. Reading</th>
<th>Prop. Talking</th>
</tr>
</thead>
<tbody>
<tr>
<td>DW Submissions</td>
<td>-</td>
<td>.32*</td>
<td>- .23</td>
</tr>
<tr>
<td>Prop. Testing</td>
<td>-</td>
<td>-.48**</td>
<td>.034</td>
</tr>
<tr>
<td>Prop. Talking</td>
<td>-</td>
<td>-</td>
<td>-.33**</td>
</tr>
</tbody>
</table>

**p < .01, *p < .05.**

Note. DW submissions = number of times submitted diagnosis worksheet, prop. Testing = proportion of time 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.

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; \( t(62) = -1.38, p = .17, d = .35 \); however there were significant effects for number of partially-relevant items; \( t(54.331) = -2.54, p = .014, d = .63 \) (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; \( t(62) = -2.28, p = .026, d = .57 \). 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 was not significantly associated with number of times submitting the diagnosis worksheet, we focused solely on food testing behavior for our subsequent research questions.

**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 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 out of 32 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, bacteria, 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 Figure 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 Figure 3).
Table 3.

Pattern Mining Outcome Sequences by DW Group

<table>
<thead>
<tr>
<th></th>
<th>2-3-coded sequences</th>
<th></th>
<th>4-coded sequences</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&gt;50 %</td>
<td>&gt;90 %</td>
<td>Same</td>
<td>&gt;50 %</td>
</tr>
<tr>
<td>Once</td>
<td>109</td>
<td>5</td>
<td>94</td>
<td>56</td>
</tr>
<tr>
<td>More</td>
<td>256</td>
<td>27</td>
<td>441</td>
<td>3</td>
</tr>
</tbody>
</table>

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.

Figure 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.
Figure 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.

Overall, these results suggest that we were able to find common sequences of food testing item 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.

**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; $t(62) = -2.19, p = .032, d = .55$, and one for the 5 → 2 code; $t(62) = -2.28, p = .026, d = .57$. Specifically, participants who submitted their diagnosis worksheet correctly after one attempt (more efficient participants) had significantly lower frequencies of the PARTIALLY RELEVANT → RELEVANT → RELEVANT → PARTIALLY RELEVANT sequence and PARTIALLY RELEVANT → RELEVANT → PARTIALLY RELEVANT 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.
Table 4.

_Differential Sequence Mining with 2- and 3-coded Segments by DW Group_

<table>
<thead>
<tr>
<th></th>
<th>Once</th>
<th></th>
<th></th>
<th>More</th>
<th></th>
<th>t-test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1→4</td>
<td>1.03</td>
<td>.98</td>
<td>1.24</td>
<td>1.12</td>
<td></td>
<td>-.80</td>
</tr>
<tr>
<td>4→1</td>
<td>.58</td>
<td>.67</td>
<td>1.06</td>
<td>1.029</td>
<td></td>
<td>-2.19*</td>
</tr>
<tr>
<td>2→5</td>
<td>1.97</td>
<td>2.16</td>
<td>2.33</td>
<td>2.56</td>
<td></td>
<td>-.62</td>
</tr>
<tr>
<td>5→2</td>
<td>1.42</td>
<td>1.77</td>
<td>2.61</td>
<td>2.34</td>
<td></td>
<td>2.28*</td>
</tr>
<tr>
<td>1→2</td>
<td>.74</td>
<td>.73</td>
<td>.70</td>
<td>.85</td>
<td></td>
<td>.23</td>
</tr>
<tr>
<td>2→1</td>
<td>.68</td>
<td>.75</td>
<td>.58</td>
<td>.71</td>
<td></td>
<td>.56</td>
</tr>
<tr>
<td>1→4→1</td>
<td>.16</td>
<td>.37</td>
<td>.27</td>
<td>.45</td>
<td></td>
<td>-1.077*</td>
</tr>
<tr>
<td>2→5→2</td>
<td>.68</td>
<td>1.22</td>
<td>.64</td>
<td>.86</td>
<td></td>
<td>.16</td>
</tr>
<tr>
<td>2→5→8</td>
<td>.58</td>
<td>1.025</td>
<td>.55</td>
<td>1.12</td>
<td></td>
<td>.13</td>
</tr>
<tr>
<td>5→2→5</td>
<td>.68</td>
<td>1.35</td>
<td>.73</td>
<td>1.008</td>
<td></td>
<td>.17</td>
</tr>
</tbody>
</table>

**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.

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.
Discussion

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

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 larger proportional learning gains, 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, 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 multichannel 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, Biswas, Sengupta, Dickes, Kinnebrew, & Clark, 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 irrelevant food items, and rarely tested for carcinogens or mutagens (irrelevant testing options), which 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 specifically 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 students spending more time in phase 3), however the overlap of sequences across more and less efficient students reveals that all students 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 \textsc{PartiallyRelevant}--\textsc{Relevant} \rightarrow \textsc{Relevant}--\textsc{Relevant} (i.e., 4\rightarrow1) and \textsc{PartiallyRelevant}--\textsc{PartiallyRelevant} \rightarrow \textsc{Relevant}--\textsc{PartiallyRelevant} (i.e., 5\rightarrow2). 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\rightarrow1 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 bacteria). The 5\rightarrow2 sequence suggests that participants were testing the same pattern as the 4\rightarrow1 sequence, however in this case, they
were testing for a partially-relevant food item for both a virus and bacteria, 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 successful 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 students, and how important it is for students 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.
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 multimodal multichannel 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).

**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- \( \rightarrow \) re-reading\), 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.

**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.

**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.
References


CHAPTER 5: Integrated Discussion

The goal of this dissertation was not only to discuss the methodological and analytical issues plaguing interdisciplinary research on psychological and educational topics (such as SRL and learning with ALTs), but to also offer examples demonstrating how we can address some of these issues using multi-channel data and non-traditional statistical techniques. Specifically, issues pertaining to methodology include the overreliance on using self-report measures to determine how students engage in SRL processes. Additionally, when using trace data, issues arise regarding how to temporally align data from different channels, which have different sampling rates. All three chapters demonstrated (theoretically and empirically) how using multi-channel data does allow us to capture overt, behavioral data, instead of using data based on participants’ perceptions, which might not be an accurate measure, as research has revealed that students do not typically engage in reporting accurate metacognitive monitoring (Dunlosky & Lipko, 2007). Issues regarding analytical techniques are associated with violating statistical assumptions when assessing SRL as an event. Students can and should engage in multiple SRL processes during learning with ALTs, and when analyzing these SRL processes using traditional inferential statistics, we cannot truly examine how these processes unfold over time. However, there are statistical techniques that are better suited to examine the temporality of SRL. Multi-level modeling (Raudenbush & Bryk, 2002) and educational data mining (Kinnebrew et al., 2013) allow for the analysis of multiple instances of events, which is more representative of engaging in SRL.

Chapter 2 directly addresses these methodological and analytical issues that arise when investigating SRL with ALTs, and suggests ways in which we can overcome these
issues (e.g., using multi-channel data and non-traditional statistical techniques). Chapters 3 and 4 provide empirical examples of how we can use different types of multi-channel data and different statistical techniques, which revealed that we could find evidence of using metacognitive monitoring SRL processes from eye tracking and log-file data, collected at the instance level. Therefore, this dissertation demonstrated that there are novel and multiple approaches for investigating metacognitive monitoring during learning with ALTs, which provides greater insight into how students deploy metacognitive SRL processes during learning with ALTs.

**Defining Metacognitive Monitoring**

Metacognitive monitoring is widely studied in terms of an SRL process (Bjork, Dunlosky, & Kornell, 2013; Flavell, 1979; Veenman, Van Hout-Wolters, & Afflerbach, 2006), which can be classified into many micro-level strategies, such as judgment of learning and content evaluation (Greene & Azevedo, 2009). Therefore, metacognitive monitoring at the macro-level can mean something different across various contexts and different types of SRL-related activities, making it difficult to define. For example, metacognitive monitoring while engaging in scientific inquiry will require different processes than metacognitive monitoring during knowledge acquisition. In chapters 3 and 4 of this dissertation, Taub et al. (in press) and Taub et al. (revise and resubmit) demonstrated how we can define and investigate metacognitive monitoring differently based on different behaviors, using different data channels.

In chapter 3, metacognitive monitoring is defined in relation to knowledge acquisition, such that Taub et al. (in press) investigated how students monitored their reading
between the text and concept matrix within book instances during learning while playing *Crystal Island*. They defined monitoring in terms of judgments of learning, such that they defined shifting from the text to the concept matrix as an indication that students understood enough content to answer a particular question. To do this, they used eye-tracking data with log files and found evidence of strategic monitoring behaviors, such that students ensured they understood content for one question at a time, and when they responded to that question, they returned to the text to seek content to answer the next question correctly, and continued this process to answer each question in the concept matrix. Chapter 4, in contrast, involved defining metacognitive monitoring in relation to hypothesis testing while playing *Crystal Island*, where monitoring occurred via tracking which items were tested, and why. They were able to track metacognitive monitoring using log-file data, where monitoring was defined as monitoring progress towards goals, and the goal was to define the disease’s transmission source. After each instance of testing, it was assumed that students were monitoring if the result of that test moved them closer or further in determining the solution to the mystery. Thus, if they spent more time testing, this would indicate that they were further from solving the mystery, which was reflected in the positive correlation with attempts at solving the mystery, since being further away was associated with longer testing time. As such, based on the means by which the researchers were defining metacognitive monitoring, different data channels will be more or less appropriate for detecting metacognitive monitoring. This demonstrates how a construct within the same ALT can be defined differently, but can still provide evidence for each of the different definitions, demonstrating how context plays a large role when investigating CAMM processes during learning with ALTs.
Based on these results, many questions arise based on how we can define metacognitive monitoring across different ALTs, and whether we can define the same micro-level processes that students can deploy. For example, a judgment of learning was defined in chapter 3 based on students’ implicit behaviors. Would these implicit behaviors be shown during reading activities within different ALTs? Furthermore, does the type of ALT play a role here, such that we can define judgments of learning similarly across different game-based ALTs, but not for intelligent tutoring systems or hypermedia-learning environments? When hypothesis testing is fostered within an ALT, if the overarching goal is different for another ALT, would it be appropriate to assume the same metacognitive monitoring strategies would enhance performance? Answering these questions might help resolve the broader impacts of designing adaptive ALTs and conducting research regarding their effectiveness in fostering SRL, and subsequent learning.

**Broader Impacts**

Not only have these results emphasized how we can investigate metacognitive monitoring across different types of activities, these empirical results, as well as topics discussed in the theoretical manuscript provide insight into the broader implications for conducting research investigating metacognitive SRL processes using multi-channel data and non-traditional statistical techniques. These implications address theoretical, methodological, analytical, and practical considerations to be made with this research.

In addressing the methodological and analytical issues faced when assessing SRL during learning with ALTs, this has implications for addressing the theoretical issues as well. Specifically, there is a need to associate SRL, with the context the student is learning in. We
differentiated between metacognitive monitoring and different activities within one ALT (see above), but determining which specific strategies to foster in other ALTs will largely depend on the context. To address this issue, researchers need to consider what each ALT requires students to do. For example, when learning with a hypermedia-based ITS, such as MetaTutor (Azevedo et al., in press), students are required to read content so they can complete their sub-goals, and overall learning goal. To do so, they can use different cognitive (e.g., take notes, summarize) and metacognitive (e.g., judgment of learning, content evaluation) SRL processes to help them monitor their progress and ensure they understand the material. In a game-based ALT, such as CRYSTAL ISLAND, the goal is to solve the mystery, and students can choose which activities to engage in to help them gather clues. Specifically, they can and should read books and talk to non-player characters, but it is not required to achieve the goal of solving the mystery. In fact, based on results from chapter 4, there is no association between reading books and talking to non-player characters and solving the mystery. As such, reading is an important component to complete the task during learning with MetaTutor, however it is not a required component to complete the task during learning while playing CRYSTAL ISLAND. Based on these examples, it is evident that when designing ALTs, the type of SRL processes students need to engage in will depend on the particular activities required.

Methodologically, broader impacts from these manuscripts focus on the selection of the appropriate data channels for what is being investigated. In other words, not all data channels are appropriate for every analysis and the research must carefully consider which of these types of data will help address the specific issue that is being addressed. For example,
in chapter 3, the authors investigated how students metacognitively monitored their knowledge acquisition while reading books. In this case, using eye-tracking data was necessary and appropriate because this informed the authors of the amount of time students spent fixating on the book content and on the concept matrices. Furthermore, when the authors ran the models, they discovered that both the log-file data and eye-tracking data were significantly associated with performance on the concept matrices, revealing empirically that both of these data channels were appropriate for the analysis. In contrast the authors only used log-file data for their analysis in chapter 4 because that was the only necessary data channel needed to conduct their analysis. Hypothesis testing behavior was fully generated from the log files, such that the logs revealed each instance of testing, along with what items were tested. Log files also revealed the amount of times students attempted so submit their solution. In direct comparison to chapter 3, eye-tracking data was not appropriate because examining fixation duration during a scanning instance would not reveal any additional information because whether they were looking at the scanner or not during a scanning instance had no impact on what they were testing. Data revealing students’ emotions could have been an additional variable (which was noted for future directions in the chapter), however it again was not essential for determining exactly what food items students were testing, and in what order. Therefore, methodological implications from these manuscripts reveal that context is again important for extracting data, as depending on what students were doing, this will impact which data channels are the most appropriate for conducting analyses.

Analytically, and similarly to methodological implications, these manuscripts have implications for selecting the most appropriate analytical technique, such that depending on
the specific research questions that have been generated, different statistical tests will be most appropriate. In chapter 2, the authors discuss that when studying SRL as an event, and more specifically, when investigating different SRL events, traditional inferential statistics are not appropriate because they require creating composite scores, or investigating participants at common times points, without having any missing data. Therefore, the analyses conducted in chapters 3 and 4 (multi-level modeling, sequential pattern mining and different sequence mining) were more appropriate because they allowed for the investigation of SRL as a series of events that could change at each instance. Multi-level modeling allows for the examination of the predictive association between a variable at multiple time points, where Taub et al. (in press) used a series of log file and eye-tracking variables at each book instance to predict performance on concept matrices. Taub et al. (revise and resubmit) used educational data mining techniques to examine the sequences of hypothesis testing behavior, and whether the frequency of occurrence of these sequences differed based on levels of efficiency. This technique does not allow for investigating the association between a sequence and level of efficiency, but the research questions did not aim for this. Therefore, there are many advantages to using these techniques, but the specific one chosen will still depend on the research questions themselves. Furthermore, if research questions were to examine for significant differences between learning gains between efficiency groups, a \( t \)-test would be most appropriate, and not the other methods. Additionally, using MLM and EDM may not always be appropriate, as there are assumptions to be satisfied as well. For example, there is a required sample size for MLM, and models will not converge if there is not a sufficient sample. Therefore, basic inferential statistics can be the appropriate analysis, again,
depending on the context in which the analysis is being conducted. For investigating instances of SRL, these chapters reveal that using statistical techniques that do not require assumptions, such as the assumption of independence of cells, are the most appropriate technique.

In addition to the theoretical, methodological, and analytical implications presented from these three manuscripts, there are also practical implications from conducting these studies. One caveat to being able to conduct this type of cutting-edge research is being proficient in using different types of software (e.g., SAS, Python, spmf), which can be lesser known in certain fields. In addition, chapters 3 and 4 used data that were aligned by a computer scientist who developed a tool to align the data. Therefore, it is important to engage in interdisciplinary research, where psychologists and computer scientists work together to build and design ALTs, including capturing and aligning trace data to measure SRL processes, and methods for temporally aligning data channels. Therefore, if interdisciplinary teams work together, this can advance the fields of psychology, education, computer science, artificial intelligence, engineering, statistics, etc., and can allow for future studies and analyses to move even further beyond what has been accomplished so far in terms of investigating SRL during learning with ALTs using multi-channel data. This dissertation addressed issues related to conducting research, however there are practical issues that need to be considered as well, especially when it comes to acquiring the knowledge for using these methodological and analytical techniques, and for the actual design of these ALTs.

These broader impacts seem to have a unifying theme: the importance of context, whether it is the activity students are required to do to complete a given task, or the type of
ALT being assessed (e.g., Greene, Bolick, Jackson, Caprino, Oswald, & McVea, 2015). When defining SRL, it is necessary to include the task because this will impact how SRL is assessed. For example, SRL will be different when engaging in scientific reasoning compared to knowledge acquisition, which leads to questioning about the domain specificity of SRL. If all SRL processes can be investigated in different domains, can we assume SRL is not domain specific, or because we must adapt the operational definition of each specific SRL strategy based on the domain, do we assume it is not domain specific? Therefore, context plays a very important theoretical role when investigating SRL. In addition, based on the context of research, this will impact which methodological choices a researcher will make. To investigate how students test hypotheses; i.e., by knowing the order in which they rest each item, and what they test each item for in that same order, log-file data will be appropriate. However, if researchers are seeking to investigate how students feel about their ability to perform a task, it will be more appropriate to administer a self-report questionnaire about self-efficacy for them to complete. Context impacts analytical choices as well, such that certain research questions will be better answered given one statistical technique over another. Investigating data where students have multiple rows of data would better suit conducting multi-level modeling or sequential pattern mining techniques, with research questions asking about the predictive value of a given variable at multiple time points. However, using that same data, composite or mean scores can be computed, allowing researchers to use an ANOVA investigating the differences in mean duration spent using metacognitive monitoring processes across three different experimental conditions. Given the context of the research questions, some analytical approaches are more appropriate than
others. Therefore, given theory, method, or analysis, evidence from chapters 2-4 reveals that in making selections for research based on any of these three aspects, the context in which the study is conducted will play a major role in any decision.

Based on the specific results demonstrated by these two chapters, and the benefits of using these specific data channels and analysis techniques, these manuscripts also have broader impacts towards the design of adaptive ALTs. Adaptive ALTs can be beneficial for student learning because it guarantees that students with varying levels of abilities will benefit from using these systems, where these ALTs can therefore act as an additional aid for teachers.

**Future Directions**

Based on the abovementioned similarities and differences that have been observed across these manuscripts, and the broader impacts based on the reported findings, there are several questions that can be addressed in future studies, with the aim of developing adaptive ALTs. First, with this goal of developing adaptive ALTs, a first consideration must involve what the ALT will be adaptive based on. There have been many analyses that determined the impact of individual differences on SRL, such as prior content knowledge (Taub, Azevedo, Bouchet, & Khosravifar, 2014), achievement goals (Lallé et al., 2017), emotions (Sabourin & Lester, 2014), agency (Bradbury, Taub, & Azevedo, 2017; Sawyer, Smith, Rowe, Lester, & Azevedo, 2017), and personality (Lallê, Mudrick, Taub, Grafsgaard, Conati, & Azevedo, 2016). These studies have found that there are different uses of SRL processes and different levels of performance are associated with these factors, demonstrating how students with different characteristics learn differently, thereby implying that in order for all students to
receive the appropriate scaffolding and feedback intelligent, adaptive systems need to account for these factors when providing prompts and feedback to students. In addition to these factors, by collecting data on student behaviors using multi-channel data, as demonstrated in chapters 3 and 4, these results can be used as factors to include when developing adaptive ALTs. For example, if the ALT knows that students who fixate longer on the text and matrix will have lower performance, the system can suggest more strategic ways of reading the content to help improve their performance. As another example, based on the sequences of testing behavior students are engaging in, and based on the amount of time they are spending testing food items, the system can prompt them to try and change approaches because this might lead to inefficient gameplay behavior. Future studies should aim to incorporate more factors together to get an even better idea of how individual differences interact with trace data to impact SRL, learning gains, effectiveness of gameplay, etc., which can inform researchers of which variables should be included for the ALT to adapt to.

One issue, however, that arises when planning studies to investigate the effectiveness of a newly designed ALT, relates to the experimental design itself. When testing these new ALTs, a valid and standard experimental approach requires a comparison between the new treatment and a control condition that does not include the new treatment. If the newly designed ALT has been generated because the previous ALT is not sufficient for fostering learning, this indicates that even if the control condition includes learning with the old ALT, this will put the students in this control condition at a disadvantage because they do not get to benefit from this new technology. Therefore, in future studies, researchers must consider the
ethical dilemma of not providing all students the opportunity to learn with an ALT that caters exactly to their own learning needs. This poses a pressing question to experimental researchers—do we need to alter the gold standard for conducting research to allow for all participants to receive this opportunity? Should we always conduct within-subjects analysis to account for this? Given that there are statistical techniques that can investigate within-subjects designs (RM-ANOVA to investigate overall learning gains and multi-level modeling to investigate instances of SRL behaviors at the between- and within-subjects levels)? Based on advances using cutting-edge data channels and analysis techniques, future researchers should consider the benefits of providing all students equal opportunities for learning.

**Generalizability of findings.** Relatedly, many studies examining the effectiveness of ALTs are conducted in a controlled laboratory setting. This is important to control for internal validity, such that we guarantee the ALT is the only device the student is interacting with. Furthermore, when collecting data using equipment that requires precise measurement, such as an eye tracker, the surrounding environment must be controlled for, such as not having sunlight block the eye tracker’s ability to accurately and reliably collect data. However, this poses a threat to being able to generalize our findings, i.e., ecological validity, and we want to provide students with the same learning experience they would normally have in the classroom, which a laboratory setting does not represent. Therefore, we pose a threat to ecological validity when conducting laboratory studies, which are required when collecting multi-channel data that is as accurate as possible.

There is, however, a second aspect of generalizability, which not only involves generalizing our findings across physical learning environments (e.g., classroom vs.
laboratory), but also across different ALT environments. Researchers are constantly trying to attempt to generalize all findings, such that if they examine one SRL process during learning with one ALT, they seek to yield those same results during learning with a different ALT. Given the importance of context, however, this does not seem like a feasible goal. The broader impacts discussed the necessity to acknowledge the context in which learning occurs in order to conclude which SRL processes are the most appropriate. For example, if a learning environment requires students to solve math equations, it would not be feasible to foster engaging in content evaluation, which would require students to rate the relevancy of the content to their learning goal, because all content is relevant when solving a math equation. In contrast, if the environment includes math word problems, a content evaluation is an appropriate strategy because students often have to seek out relevant information from a word problem they will need to answer the question correctly. Therefore, it would not be feasible to compare content evaluations between these two ALTs because they are not an appropriate process to use in one of these environments. Additionally, if researchers were to compare the same process that is appropriate for two different ALTs, this may also not be appropriate because the meaning of these processes may differ from one ALT to the other. Chapter 3 investigated judgments of learning based on the assumption that students understood the material enough to proceed to the concept matrix. A judgment of learning in MetaTutor directly asks students to rate how well they understand the content they read on the current page, which is then followed by a 3-item multiple choice quiz. Based on the different definitions of these metacognitive monitoring strategies, it would not be feasible to seek out generalizing these findings, because the meaning of the strategy is so different in
these two ALTs, even though the activity of reading is the same. As a third example, it would not be feasible to generalize findings between reading books and hypothesis testing because even though the ALT is the same, the activity itself requires different types of monitoring skills. This challenges conventional research goals that aim to generalize findings across different environments, however if the activities or the operational definitions of the process being investigated are different from each other, obtaining generalizable results is not only nearly impossible, but it is not meaningful. Therefore, when conducting studies that investigate SRL processes in a new way (i.e., at the instance level, instead of using an aggregated score), this questions gold standards of research, and so future research should seek to propose new gold standards that align more accurately with new types of cutting-edge research that seeks to provide all students with the optimal educational experience.

Finally, the empirical chapters (Taub et al., in press; Taub et al., revise and resubmit) provided examples of using log files and eye-tracking data to investigate metacognitive monitoring during learning with ALTs, which are only a few samples of multi-channel data that could be used to analyze metacognitive monitoring and SRL. There has been a plethora of research investigating the impact of emotions on learning with ALTs (Calvo et al., 2015; D’Mello & Graesser, 2012). This research has revealed that learner-centered emotions, such as confusion and frustration, are frequently experienced during learning with ALTs, and depending on whether or not they can be resolved by engaging in problem-solving strategies, this will positively or negatively impact learning (D’Mello, Lehman, Pekrun, & Graesser, 2014). Therefore, although these emotions have a negative connotation to them, they can help improve students’ learning, as they require students to engage in processes to help them
resolve this confusion or frustration. Using video data of facial expressions of emotion can also be investigated at the instance level, and can therefore be used to predict performance on certain tasks, in a similar way to using eye-tracking data. These data can be combined with other important contextual factors for the development of adaptive ALTs, such that levels of emotions can be beneficial or not, depending on other factors, such as prior knowledge. For example, in investigating different facial areas as markers of emotions (i.e., action units), research has shown that action unit 4, which involves furrowing the eyebrows, can be indicative of both confusion and mental effort (D’Mello, Craig, & Graesser, 2009). Therefore, it can be the case that depending on students’ levels of prior knowledge, this can differentiate between different meanings of action unit 4. Thus, future studies should seek to include other multi-channel data, such as facial expressions of emotion, as another factor to include in designing adaptive ALTs to ensure all factors are considered when providing students adaptive feedback and support during learning with ALTs.

CONCLUSION

The ultimate goal for research on SRL and the use of CAMM processes with ALTs is to ensure that we are providing all students with the equal opportunity of learning complex topics to the best of their abilities. By continuing to investigate how students are learning with our existing technologies, we can use these data to help design ALTs that adapt to help them use SRL strategies appropriately, given the context they are learning in, and the activities they are required to do.

Based on the research that has been conducted, both in chapters 2-4 and in research all over the world on SRL with ALTs, a clear common goal for this research is towards the
development of adaptive ALTs that adapt, based on numerous factors, to the particular needs of each student who is learning complex topics. Although it is clear that these systems are the future for ALTs, there are still many future considerations for designing these ALTs that need to be resolved. Specifically, we know these systems need to be adaptive, but future research needs to pinpoint the most crucial criteria for them to adapt to. For example, many studies have shown that prior domain knowledge does impact learning (e.g., Taub et al., 2014), but what about procedural or conditional knowledge? What about prior knowledge of SRL, and using SRL processes? Does prior SRL knowledge play even more of a role than prior content knowledge? As such, future studies should seek to investigate not only the important variables that ALTs should adapt to, but also the levels of these variables. In order to conduct these investigations, it will be beneficial to use multi-channel data; whichever data channel will be the most informative for assessing individual learning processes, such as knowledge acquisition and scientific inquiry. Researchers need to truly understand the usefulness of each data channel, and to only use the applicable ones for their investigations. In addition, the analysis technique must be carefully chosen to ensure that the research questions are appropriately addressed. In doing so, we can make large advances toward developing these adaptive ALTs, which should greatly benefit students with different learning characteristics and different ability levels, to guarantee that all students are learning effectively and efficiently.
REFERENCES


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


