2021

Leveraging Multimodal Learning Analytics to Understand How Humans Learn with Emerging Technologies

Elizabeth Cloude
University of Central Florida

LEVERAGING MULTIMODAL LEARNING ANALYTICS TO UNDERSTAND HOW HUMANS LEARN WITH EMERGING TECHNOLOGIES

by

ELIZABETH CLOUDE
M.A. University of Central Florida, 2020
B.S. Christopher Newport University, 2016

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Learning Sciences and Educational Research
in the College of Community Innovation and Education
at the University of Central Florida
Orlando, Florida

Fall Term
2021

Major Professor: Roger Azevedo

ABSTRACT

Major education and training challenges are plaguing the United States in preparing the next generation of the future workforce to meet the demands of the 21st Century. Several calls have been released to improve education programs to ensure learners are acquiring 21st century knowledge, skills, and abilities (KSAs). As we embark on the digital and automation ages of the 21st century, it is essential that we move away from traditional education programs that define and measure KSAs as static constructs (e.g., standardized assessments) with little consideration of the actual real-time deployment of these processes, missing critical information on the degree to which learners are *acquiring* and *applying* 21st century KSAs. The objective of this dissertation is to use 1 book chapter and 2 journal articles to illustrate the value in leveraging emerging technologies and multimodal trace data to define and measure scientific thinking, reflection, and self-regulated learning--core 21st century skills, across contexts, domains, tasks, and populations (e.g., medical versus undergraduates versus middle-school students). Chapters 2-4 of this dissertation provide evidence of ways to leverage multimodal trace data guided by theoretical perspectives in cognitive and learning sciences, with a special focus in self-regulated learning, to assess the extent to which learners engaged in scientific thinking, reflection, and self-regulated learning *during* learning activities with emerging technologies.

Overall, results from these chapters illustrate that it is necessary to utilize methods that capture learning processes as they unfold during learning activities that are guided by theoretical perspectives in self-regulated learning. Findings from this research hold significant broader impacts for addressing the education and training challenges in the United States by collecting multimodal trace data over the course of learning to not only detect and identify how learners are developing KSAs such as scientific thinking, reflection, and self-regulated learning, but where these data could be fed into an intelligent and adaptive system to repurpose it back to trainers, teachers, instructors, and learners for just-in-time interventions and individualized feedback. The intellectual merit of this dissertation focuses predominantly on the importance of utilizing rich streams of multimodal trace data that are mapped onto different theoretical perspectives on how humans self-regulate across tasks like clinical reasoning, scientific thinking, and reflection with emerging technologies such as a game-based learning environment called Crystal Island. Discussion is incorporated around ways to leverage multimodal trace data on undergraduate, middle-school, and medical student populations across a range of tasks including learning about microbiology to problem solving with a game-based learning environment called Crystal Island and clinically reasoning about diagnoses across emerging technologies.

I dedicate my work to education, mentorship, and future generations of women and underrepresented minorities in STEM and research.

ACKNOWLEDGMENTS

I recognize and honor the privilege of having the opportunity to pursue higher education while being surrounded by a community of supporters who challenge, inspire, and elevate not only my research but my thinking and growth as an individual throughout the last 4.5 years.

Roger, I would not be where I am today without your extensive mentorship and dedication to my graduate training. I am fortunate to receive high-quality mentorship and academic support from you since arriving in your lab in 2017. Thank you for taking a chance on me. I cannot thank you enough for the countless hours spent reading my papers, research ideas, proposals, and so many others. You never kept me waiting, and your transparency and willingness to communicate hard things is something to be acknowledged. You have inspired me to be a better researcher, scholar, student, mentor, and teacher, and I will take these values with me everywhere I go. Because of your commitment to my training and mentorship as a young scholar and scientist, it has instilled in me how deeply meaningful and important it is to create a close mentorship with graduate students and lab members that encompass authenticity, inclusivity, support, leadership, communication, and teamwork. I have grown so much since the first day in your lab, and it is with so much gratitude and joy to begin a new chapter as a research scientist and scholar. Thank you for everything you have given me.

Thank you to my committee members: Drs. Michelle Taub, James Lester, Mary Jean Amon, and Dario Torre. You have all been supportive of my ideas and work, and I cannot thank you enough for the time spent reading through my dissertation drafts and challenging me to think about a program of research built around intellectual merit and broader implications. I am so lucky to have a strong a dissertation committee, and I thank you all for everything.

Thank you to my wonderful lab mates, Megan and Daryn, for their tremendous support and encouragement over the years. I could not have done this without you either. From countless weekends spent studying linear regression models to white board sessions and debating theoretical models, you have expanded my perspectives in more ways than I can express.

Thank you to my family for supporting me on my educational journey throughout my entire life. Thank you to my wonderful friends and girlfriend for listening, understanding, and encouraging me to push forward when I doubted myself during highly stressful events. I would not be here with you, and I cannot begin to express my gratitude for believing in me when I could not believe in myself. Thank you.

Finally, I would like to thank the funding agencies, the National Science Foundation and Social Sciences and Humanities Research Council of Canada, for making this research possible.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACROYNMS

ACGME	Accreditation Council for Graduate Medical Education
AIC	Akaike Information Criterion
AOI	Area of Interest
CCSS	The Common Core State Standards
GBLE	Game-based Learning Environment
GLA	Game-learning Analytics
KSA	Knowledge, Skills, and Abilities
LOG	Logfile
MLA	Multimodal Learning Analytics
NASEM	National Academies of Sciences, Engineering, and Medicine
NCLB	No Child Left Behind Act
NDS	Nonlinear Dynamical Systems
NGSS	Next Generation Science Standards
NPC	Non-player Character
NRC	National Research Council
OECD	Organisation for Economic Co-operation and Development
SCT	Social-cognitive Theory
SDDS	Scientific Discovery as Dual Search
SRL	Self-regulated Learning
STEM	Science, Technology, Engineering, and Medicine
VIF	Variance Inflation Factor

CHAPTER ONE: INTRODUCTION

Evidence shows learners in the United States today do not have the knowledge, skills, or abilities (KSAs) to deal with the demands of the 21st century (National Academies of Sciences, Engineering, and Medicine [NASEM], 2018). For the U.S. to compete in the global economy, there have been several calls to transform education and training programs to ensure that learners develop the core 21st century KSAs including analytical thinking, problem solving, reflection, adaptability, critical and scientific thinking, effective communication and listening, as well as collaboration (National Research Council [NRC], 2015, 2012). As our nation is faced with persisting, and sometimes novel, educational, socioeconomic, environmental, and health-related challenges (e.g., COVID-19 pandemic, climate change), it is essential to equip future generations with the KSAs needed to solve grand society challenges. Since learners are not prepared for the current and future workforces, it raises questions about how KSAs have changed from the past to the present to plan for what is needed in the future. In this chapter, I will discuss (1) the U.S. educational system and how it has changed over the previous 20 years with a special focus on policy and decentralization to highlight its relation to (2) the current U.S. educational system and how its structure may be a result of the significant education and training challenges in preparing current and future workforces. Finally, I will review the role of emerging technologies and their capacity to mitigate education and training issues within public-school systems by proposing a program of research.

<u>U.S. Educational System and KSAs: The Past</u>

KSAs are determined by components of the U.S. education system. For a learner to illustrate they have mastered KSAs, they must receive a credential (e.g., high-school diploma), which requires the learner to successfully complete a curriculum of courses and oftentimes pass a series of examination and an annual standardized assessment. For example, to measure whether learners meet the minimum requirements in reading, writing, math, science, and history, Virginia Public Schools administer annual Standards of Learning assessments (Virginia Department of Education, Commonwealth of Virginia, 2021). To dictate whether a learner has mastered core KSAs, this model of education maintains underlying assumptions that a credential and series of passing standardized assessments hold the capacity to measure a learner's ability to apply KSAs. This model began after the 20th century when the U.S. educational system changed from a decentralized to a more centralized state, heavily influencing how the system operates today (DeBoer, 2012). One major step toward centralization occurred in 2001 when President George W. Bush signed the No Child Left Behind Act (NCLB) into law. The NCLB increased the role of the federal government in holding public schools accountable for the education of all learners by supporting a standards-based reform, shifting the education system toward standardization and national-level policy. One of the main reasons for the shift toward centralization was globalization and international competitiveness (West, 2012).

Based on the very nature of the NCLB, states were required to define what educational achievement meant, meanwhile the federal government expanded its role in education by implementing annual testing and academic reports, teacher qualifications, and significant changes to public-school funding. Students' achievement and scores were tracked across states, leading to standardized assessments in service of closing achievement gaps such that no child was left behind. By placing a scarcity in school funding, teachers were required to adopt a more active role in the classroom to ensure their students were prepared to pass standardized assessments, such that the rate at which a teachers' classroom passed determined the school accreditation and federal funding. As a result of this, curriculums became increasingly narrower and classrooms became teacher-centered, where students were forced to adopt a more passive role in choosing which instructional materials they engaged with (West, 2012). Because of these shifts in education over the last 20 years to a more standards-based approach, where students demonstrate achievement by passing assessments, it begs the question as to what KSAs the assessments have been measuring, and whether these KSAs are meeting the demands of the workforce?

In 2007, the NRC conducted a review to assess U.S. competitiveness in STEM and released a report suggesting the U.S. was in significant risk of losing its economic and STEM leadership (NRC, 2007). The report highlighted that test scores for science or mathematics were significantly lower in the U.S. compared to other international countries, and as a result, one of the main takeaways was that STEM education "inadequately prepares students to work outside of university" (p. 94) and that "primary and secondary schools do not seem able to produce enough students with the interest, motivation, knowledge, and skills they will need to compete and prosper in the emerging world" (NRC, 2007, p. 94). The Organisation for Economic Co-operation and Development reported similar findings (OECD, 2007). In 2010, the NRC updated the report and confirmed that the U.S. ranked 15th on the 2005 Programme for International Student assessment in mathematics and 21st in science (NRC, 2010). A more recent report highlighted similar findings, where the U.S. continued to rank low in mathematics and science relative to 40 other countries (OECD, 2016).

Based on these findings, many curricular changes have taken place with goals to enhance STEM education, such as The Common Core State Standards (CCSS) and Next Generation Science Standards (NGSS; NGSS Lead States, 2013; Suter & Camilli, 2019). There was also a shift in focus around connecting the educational system such as standardized assessments with research-based theories of how learning occurs across tasks, population, contexts, and domains to guide classroom design, instruction, and teacher training (Shepard, Penuel, & Pelligrino, 2013). However, progress is slow in moving these efforts forward and does not address the significant issue presented within the system where there continues to be a major disconnect between *acquiring* and *applying* KSAs in the 21st century. In formal education settings such as the classroom, learners are tasked with memorizing factual information that is often isolated from the real world in which the phenomena exist and take on greater meaning in preparation for passing an assessment. Learners need to pass said assessment to demonstrate the level of required mastery of KSAs. Yet, most learners are unable to apply the same KSAs they ‘mastered’ in the classroom to their everyday lives (NASEM, 2018). Could the educational crisis of the 21st century be a result of adopting a ‘one-size-fits-all’ paradigm, where all learners are required to follow the same curriculum and pass the same assessments in order to demonstrate they have acquired the necessary KSAs to compete in the workforce? Designing the education system as a ranking tool (e.g., pass versus fail) that is flooded with “standards” of successful learning and achievement continue to promote inequality and lack of inclusivity for different types of learners, prompting systemic bias across races, socio-economic status, gender, sex, and among many others (Blikstein & Worsley, 2017).

Some of the societal challenges we are experiencing today (e.g., climate change, systemic racism, anti-vaccinations for COVID-19 pandemic) may be a rippling effect from impacts of going through the U.S. educational system, where learners are *not* taught to acquire and apply 21st century KSAs, such as developing self-regulatory skills, scientific thinking skills, and reflection, since learners need to sit still, be quiet, and follow a specific regime such that they are adequately prepared to regurgitate information and answer a “standard” ratio of correct answers on a series of standardized items. We cannot expect the future workforce to master acquiring and applying effective self-regulatory skills, scientific thinking, or reflection, among many others if we continue to uphold the current educational system. We must collect information on how learners acquire and apply KSAs across a range of contexts, tasks, domains, and situations with emerging technologies that are synonymous to the real-world in which they will be used and provide learners with the agency to play an active role in their learning, practicing, making mistakes, and refining their KSAs. These issues draw me to contemplate the tools available today, and the degree to which emerging technologies could be leveraged to gauge *all* learners’ ability to acquire and apply 21st century KSAs such as scientific thinking, reflection, and self-regulated learning across contexts, tasks, domains, etc. guided a theoretical lens in learning sciences.

Emerging Technologies and KSAs: The Present

Several government reports have acknowledged the importance of integrating technology into the educational system (NRC, 2015, 2012; U.S. Department of Education, 2017) due to the compounding empirical evidence that technology could serve to not only address the educational and training challenges today, but augment 21st century skills for future generation. The essential role of technology in education could not be made clearer during the pandemic of COVID-19 which reduced location and time barriers around learning especially within higher education. Further, the digital and automation ages of the 21st century have led to promising techniques for discovering insights into the complex nature of learning, reasoning, problem solving, performing, and achieving in a variety of tasks, domains, and settings using emerging technologies (e.g., immersive virtual reality, game-based learning environments, intelligent tutoring systems). Research suggests that emerging technologies have assisted learners in developing 21st century KSAs across multiple domains (e.g., problem solving skills; Azevedo, Mudrick, Taub, & Bradbury, 2019; Carpenter, Cloude, Rowe, Azevedo, & Lester, 2021), and the rise of technology provides researchers with a platform to monitor, track, model, and foster 21st century KSAs using high-sampling resolution devices capable of capturing rich streams of multimodal data, such as data captured across multiple channels and modalities including eye movements from an eye tracker, facial expressions of emotions using facial recognition software, learner-system interactions via timestamped logfiles, physiological responses via wearable bracelets across learning tasks (Azevedo & Gašević, 2019; Noroozi, Alikhani, Järvelä, Kirschner, Juuso, & Seppänen, 2019). Current research efforts are being employed to investigate ways for teachers to leverage rich streams of multimodal trace data on their students' learning using emerging technology by building dashboards (Holstein, McLaren, & Aleven, 2019; Molenaar & Knoop-van Campen, 2019; Wiedbusch, Kit, Yang, Park, Chi, Taub, & Azevedo, 2021). A particularly relevant application of leveraging the rich information captured from multimodal trace data generated while students engage in learning activities with emerging technologies has been multimodal learning analytics (MLA).

For the last few years, researchers have been utilizing MLA within educational research, a term that, "at its essence… utilizes and triangulates among non-traditional as well as traditional forms of data in order to characterize or model student learning in complex learning environments" (p. 1346, Worsley, Abrahamson, Blikstein, Grover, Schneider, & Tissenbaum, 2016). It is important to note that researchers use MLA in a variety of different ways, but the key distinguishing characteristic of MLA, especially as it relates to its use in this dissertation and subsequent chapters, is that this technique acknowledges that learning occurs across multiple modalities, i.e., a particular mode in which something exists or is experienced, and these modalities encompass multiple data channels, variables, and analyses with the overarching goal of defining and assessing learning as it occurs in the real world (Worsely et al., 2016). For example, MLA can encompass multiple data channels such as gestures and speech to model learning, or simply one data channel with multiple variables such as eye movements where different events can be extracted from one dataset such as the number of seconds a learner fixates on an area of interest, or how often they engage in saccadic activity between two areas of interest on an interface. MLA can also be broken down into different modeling techniques, such that the decision to analyze specific streams of multiple data channel or key variables requires analytical decision making determined by a theoretical lens such as the information-processing theory self-regulated learning (Winne, 2019) to determine *how* the data should be modeled (e.g., dynamically and non-linear) to ensure operational definitions and modeling techniques align with assumptions of how the learning construct being studied unfolds according to theory. As such, MLA is rather complex, and for purposes of this dissertation, MLA is referred to as encompassing more than one data channel such as converged measures of both eye movements and learner-system interactions during learning activities, that are modeled in alignment with theoretical frameworks in cognitive and learning sciences, with a specialized focus on the role of self-regulated learning (Winne, 2019).

We focus on self-regulated learning (SRL) specifically in this dissertation due to a plethora of empirical evidence suggesting that SRL plays a key role in developing 21st century

KSAs using emerging technologies needed to address the demands of the future workforce (Schunk & Greene, 2018). In particular, decades of research show that SRL with emerging technologies demonstrates significant promise for enhancing KSAs using instructional design principles and theoretical perspectives situated in affective, educational, psychological, learning, cognitive, and computational sciences (Azevedo & Gašević, 2019; Biswas, Segedy, & Bunchongchit, 2016; Blikstein & Worsley, 2016; Clark, Tanner-Smith, & Killingsworth, 2016; Greene, Deekens, Copeland, & Yu, 2018; Qian & Clark, 2016; Taub, Sawyer et al., 2020; Winne, 2019). Theoretical frameworks in SRL also provide flexibility in studying how learning unfolds across time and contexts, tasks, populations, emerging technologies, etc., offering a researcher tool for scientists to study all types of learners in acquiring and applying 21st century KSAs to investigate solutions to the education and training demands in the U.S.

Specifically, SRL is a multi-faceted, non-linear process that fluctuates over time to meet task demands and evolving goals, and encompasses many components including cognitive, affective, metacognitive, social, and motivational processes which all play a crucial role in learners' ability to acquire and apply core 21st century skills such as effectively solving problems in a complex environment with continuously changing constraints (Winne, 2018). Most research studying SRL with emerging technologies have almost exclusively examined the role of cognitive and metacognitive processes, excluding the effect of motivational, social, and affective processes and their interdependent relationship with SRL, learning, and performance (Järvelä & Bannert, 2021). This focus likely results from the zeitgeist of the late 1980s, which had deep roots in understanding learning from an information-processing perspective which inherently ignores context, time, and motivational/social/affective components involved in human learning, performance, collaboration, communication, and many other core 21st century KSAs, especially in regard to cognitive sciences and its contributions to advancing our understanding of cognition and metacognition using emerging technologies and artificial intelligence such as ACT-R (Anderson, Betts, Bothell, & Lebiere, 2021). This movement offered a offered a great deal of scientific knowledge and empirical foundation for building intelligent systems within emerging technologies using design principles that scaffold cognitive strategies and metacognitive processes to augment KSAs (Cloude, Taub, Lester, & Azevedo, 2019; Mangaroska, Sharma, Gaševic, & Giannakos, 2020).

Several studies find that emerging technologies designed to scaffold SRL result in improved learning and performance outcomes (Azevedo, Taub, & Mudrick, 2018; Cloude, Taub, & Azevedo, 2018; Cloude et al., 2019; Taub, Sawyer, Smith, Rowe, Azevedo, & Lester, 2020). Traditional research paradigms for testing the effectiveness of intelligent systems in their ability to scaffold SRL has involved pre/post experimental designs that typically utilized self-report measures and knowledge assessments. Although, more recently there has been a noteworthy shift where researchers are beginning to examine learning processes using rich time series data generated across learning activities, emerging technologies, and contexts (Blikstein & Worsley, 2020). For this very reason, the underlying focus of this dissertation involves a heavy emphasis on studying the role of SRL in scientific thinking, reflection, and clinical reasoning (core 21st century KSAs), and we guide MLA to reflect this such that methodological and analytical decisions take special care to align with assumptions outlined by theoretical perspectives on self-regulation. In particular, the constructs selected in this dissertation involve some component of self-regulation, such as scientific thinking in chapter three or reflection in chapter 4, both core

21st century KSAs. Within each chapter, the construct of study (e.g., reflection) is cleared defined by a theoretical perspective (e.g., model of reflection by McAlpine and colleagues, 1999 in chapter four) such that there is little confusion about how MLA was leveraged to study learning constructs to ensure it is aligned with a theoretical perspective.

Summary of Research Program

We are evolving beyond a society that simply acquires knowledge, requiring future generations to master highly complex 21st century KSAs like reflection and scientific thinking. Multimodal data captured continuously throughout learning sessions with emerging technologies grounded in theoretical perspectives on SRL provide insight into the developing capabilities of the future workforce by providing fine-grain information on how learners are performing in real-time. In this dissertation, I provide evidence of how we can harness the power of theoretically based MLA grounded in SRL theory generated across learning activities, contexts, populations, and domains with emerging technologies to shed light on scientific thinking, reflection, and clinical reasoning. Specifically, chapter two introduces a theoretical discussion for using MLA guided by the socio-cognitive theory of self-regulation to transform medical education, where Cloude, Wiedbusch, Dever, Torre, and Azevedos' (accepted with minor revisions) chapter entitled, "the role of metacognition and self-regulation on clinical reasoning: leveraging multimodal learning analytics to transform medical education" to be published in the 2nd edition of the *Handbook of Multimodal Learning Analytics*, co-edited by Drs. Michail Giannakos, Xavier Ochoa, and Daniel Spikol. This chapter emphasizes the importance of moving away from traditional assessments to study the role of metacognition and self-regulation on clinical reasoning. In chapter three, Cloude, Dever, Wiedbusch, and Azevedos' (2020) journal article entitled, “quantifying scientific thinking using multichannel data: toward individualized game-learning analytics was published in *Frontiers in Education*. It discusses the significant challenges in pinpointing when learners are engaging in scientific thinking using unimodal data (i.e., logfiles alone) with game-based learning environments. They demonstrate the need to combine both eye movements and logfiles to operationalize when learners scientifically reason about information, such as gather information, make hypotheses, and experimentally test those hypotheses during game-based learning using multimodal data guided by a theory in scientific thinking (i.e., MLA). The fourth chapter by Cloude, Carpenter, Dever, Lester, and Azevedo (2021) title, “game-based learning analytics for supporting adolescents’ reflection” was published in the *Journal of Learning Analytics* and provides empirical evidence for leveraging multimodal data grounded in a model of reflection (i.e., MLA), a critical aspect of SRL, across learning activities to define not only the quantity of reflection but also the quality and account for the goals designed into a game-based learning environment to define successful performance and learning.

Overall, results from these chapters illustrate that it is critical to utilize methods and multimodal data that capture learning processes as they unfold during learning activities that are guided by theoretical perspectives in SRL (i.e., MLA) to enhance the development of 21st century KSAs such as self-regulatory skills with emerging technologies. Theoretically, my dissertation contributes toward advancing the science of learning by leveraging a range of theoretical frameworks based in how individual learn, reflect, self-regulate, and think scientifically using multimodal data that is mapped onto theoretical constructs. Methodologically, my dissertation contributes toward capturing learning processes as they unfold in real-time across contexts, populations, domains, and tasks. Analytically, my dissertation contributes toward modeling multimodal data that is based in theoretical frameworks and reveals the significance of learning constructs that would otherwise not be clear (e.g., modeling reflection as it relates to different goals within Crystal Island). Findings from this research hold significant implications for addressing the educational and training challenges in the United States by designing principles into systems that can capture MLA to not only detect and identify how learners are reflecting, self-regulating, or thinking scientifically, but also provide information back to teachers and instructors for just-in-time intervention and feedback to address deficiencies in learning with a particular emphasis on personalized and adaptive learning. This approach would eliminate the ‘one-size-fits-all’ paradigm and potentially more toward a non-standardized reform, where MLA generated on all learners across time, tasks, contexts, domains, and populations would inform one’s readiness to enter the workforce based on their ability to acquire and apply KSAs across learning activities, rather than solely relying on standardized assessments scores to reflect level of competency. Moving toward this approach would redirect the infrastructure of the educational system toward a more learner-centered model, such that the learner plays a more active role in their engagement with materials and applying KSAs across tasks rather than passively listening to instruction. The value in demonstrating the ubiquitous utility in applying theoretically based MLA in SRL to enhance 21st century KSAs across multiple dimensions (e.g., populations, contexts, time, etc.) highlights its capacity to provide solutions for educational challenges that can be expanded beyond the classroom, potentially transforming workforce training and addressing societal challenges we are all facing today such as the spread of misinformation via social media. For this reason alone, we emphasize the value of this research in advancing intellectual merit and broader impacts in the concluding section of this dissertation and offer future directions that would move this work forward.

CHAPTER TWO: THE ROLE OF METACOGNITION AND SELF-REGULATION ON CLINICAL REASONING: LEVERAGING MULTIMODAL LEARNING ANALYTICS TO TRANSFORM MEDICAL EDUCATION

This chapter titled, "the role of metacognition and self-regulation on clinical reasoning: Leveraging multimodal learning analytics to transform medical education" was led by Elizabeth Cloude with contributing co-authors Megan Wiedbusch, Daryn Dever, Dr. Dario Torre and Dr. Roger Azevedo. It has been accepted for publication in 2022 within the 2nd edition in the *Handbook of Multimodal Learning Analytics*.

Abstract

Medical errors are defined as preventable adverse effects of health care that often result from faulty clinical reasoning processes. Moreover, studies find that medical errors have been linked to failures in clinical-reasoning strategies and medical education, suggesting a need to transform curricula that prioritizes knowledge, skills, and abilities needed for effective clinical-reasoning practices. In this chapter, first we discuss the need to incorporate metacognition, the process of monitoring and evaluating one's own clinical reasoning, and self-regulation, the ability to adapt one's clinical-reasoning strategies to address these significant issues within medical education. However, challenges exist within medical education because most educational programs rely on 'snapshots' of students' performance (e.g., pre/post) to define competency using standardized assessments and self-report methodologies, missing information on what and when is occurring when applying knowledge, skills, and abilities *during* training activities. Next, we introduce multimodal learning analytics as a novel research approach for studying the role of metacognition and self-regulation on clinical-reasoning processes as they

unfold using multiple streams of time series data, such as eye movements, facial expressions, concurrent verbalizations, physiological data, and many others, during medical education and training with emerging technologies with a specific focus on the socio-cognitive cyclic model of self-regulated learning by Zimmerman and Moylan (2009). Finally, we discuss implications for utilizing multimodal learning analytics to transform the KSAs needed for the next generation of medical professionals.

Introduction

The Accreditation Council for Graduate Medical Education (ACGME) expects medical students to obtain competency in six areas to perform safely in practice: (1) patient care, (2) medical knowledge, (3) interpersonal and communication skills, (4) professionalism, (5) practice-based learning and improvement, (6) systems-based practice (ACGME, 2021), and recently, clinical reasoning (Connor, Durning, & Rencic, 2020). Metacognition and self-regulation are critical for clinically reasoning. In other words, metacognition is an ability to monitor and regulate one’s own thinking while engaging in the diagnostic-reasoning process. Once an individual identifies their reasoning process, they must evaluate if and when it may be influenced by biases or discrepancies in patient information, lab tests, etc., requiring medical professionals to step back and view their reasoning from an objective lens in pursuit of patient safety and quality of care (ACGME, 2021). If and when errors or biases are identified through metacognitive monitoring, a medical professional must then regulate their reasoning by engaging in self-regulation (Schunk & Greene, 2018).

Training medical professionals to be competent in metacognition and self-regulatory skills--i.e., to monitor, evaluate, and adapt their reasoning strategies, could be accomplished through building curricula and core competencies around metacognition and self-regulation (Dunlosky & Tauber, 2016; Nelson & Narens, 1994). Unfortunately, there is a sufficient lack of programs built to address these deficiencies in clinical reasoning within the United States, and the extent to which medical students develop metacognition and self-regulatory skills prior to graduation is currently unknown (Rencic, Trowbridge, Fagan, Szauter, & Durning, 2017). Developing a metacognitive and self-regulated learning research program could help address the extent to which medical students develop these skills, while also examining the role of metacognition and self-regulation on clinical reasoning is a possible solution for making informed decisions about medical education curricula (Batalden, Leach, Swing, Dreyfus, & Dreyfus, 2002; Cleary, Konopasky, LaRochelle, Neubauer, Durning, & Artino, 2019; Patel, Kaufman, & Kannampallil, 2018; Rencic et al., 2017).

Assessing and Evaluating Competency in Clinical Reasoning

The traditional research paradigm in medical education is permeated with standardized assessment tools to define competency, that is often supplemented with subjective self-report (questionnaire) ratings on learning processes like metacognition. For example, accredited medical institutions in the United States prepare medical professionals by first immersing them in a curriculum of lectures (e.g., lectures on anatomy and biology; Mourad, Jurjus, & Hussein, 2016 and then clinical experiences (e.g., shadowing physicians across specialities like surgery) including simulation-based training activities that utilize emerging technologies, such as the da Vinci system to train surgical skills (Fard, Ameri, Darin Ellis, Chinnam, Pandya, & Klein, 2018). or high-fidelity mannequins (Meyers, Mahoney, Schaffernocker, Way, Winfield, Uribe,... & Lipps, 2020; Petrizzo, Barilla-LaBarca, Lim, Jongco, Cassara, Anglim, & Stern, 2019). More recently, medical educators have also begun utilizing virtual and/or augmented reality systems

(Bric, Lumbard, Frelich, & Gould, 2016; Kuehn, 2018; Winkler-Schwartz, Bissonnette, Mirchi, Ponnudurai, Yilmaz, Ledwos, ... & Del Maestro, 2019).

Within clinical experiences, the traditional approach is to present medical students with common information about a disease. Students are required to identify critical findings associated with patient histories and symptomatology which should lead them to a particular illness script, disease, and treatment plan (Custers, 2015). Measurements used to assess and evaluate performance during clinical training, especially within medical education research, is defined by accuracy and errors scores, such as time to complete a procedure, number of omission and commission errors (Winner & Millwater, 2019) that are captured using self-reported ratings, gradesheets, and checklists completed by subject-matter experts (Wood & Pugh, 2020). We argue this method fails to capture how medical students might be clinically reasoning about information, and thus missing information on whether they are using faulty reasoning strategies (e.g., anchoring towards a particular disease based on common information), and more importantly, if medical students are monitoring and evaluating their reasoning process when they may be potentially vulnerable to deficiencies.

Moving toward a research program that collects data *during* learning/training activities could augment the assessment, evaluation, and ultimately, readiness of medical professionals. Collecting multiple streams of data could reveal when they have failures or vulnerabilities within their clinical-reasoning process. For instance, capturing data on developing clinical reasoning competency during simulation-based training might provide information to educators and instructors when failures in clinical reasoning might be occurring in real-time so they can intervene and provide immediate feedback to scaffold medical students' clinical reasoning (Dyre & Tolsgaard, 2018, Duffy, Azevedo, Sun, Griscom, Stead, Crelinsten,... & Lachapelle, 2015).

We argue that while standardized assessments within medical education present strengths (e.g., reliability), capturing data during learning/training activities could serve to augment medical education by measuring clinical-reasoning strategies *during* training activities, offering opportunities to gauge individual differences in readiness and potentially opens a platform for capturing transfer of knowledge, skills, and abilities in real-world clinical settings that current standardized assessment cannot capture address (e.g., Did the patient survive? Their symptoms reduced, etc.). This argument is supported by many throughout the medical field, such as Krupat (2018) who argues that research in medical education needs to shift toward capturing clinical reasoning as a process rather than taking a snapshot of an individual's readiness *before* and *after* medical-training scenarios. To capture clinical reasoning as a process, what kind of data might provide insight into a medical student's clinical-reasoning process, and more importantly, when there may be deficiencies in reasoning, what role could metacognition and self-regulation play in mitigating their harmful effects?

We argue that utilizing multiple data channels, i.e., multimodal time series data, to assess metacognitive and self-regulatory skills as they relate to clinical reasoning *during* training scenarios might provide insight into a medical student's reasoning strategy. No one data channel could provide enough context or information about the temporal unfolding and quality of understanding when there are failures in clinical reasoning with implications for instructional decision-making (Dyre & Tolsgaard, 2018). Multimodal data have recently been used to study metacognition and self-regulated learning with emerging technologies (Azevedo, 2020; Cloude, Dever, Wiedbusch, & Azevedo, 2020; Cloude, Carpenter, Dever, Lester, & Azevedo, 2021; Lajoie, Pekrun, Azevedo, & Leighton, 2020). Multimodal data, especially eye tracking and sensor-motion technologies, are currently being leveraged in some medical research subspecialties such as surgery and radiology to gauge clinical-reasoning strategies and approaches while medical students are engaged in training activities (Ashraf, Sodergren, Merali, Mylonas, Singh, & Darzi, 2018; Fox & Faulkner-Jones, 2017; Hermens, Flin, & Ahmed, 2013). A study by Ahmidi, Ishii, Fichtinger, Gallia, and Hager (2012) developed an objective and automated method for assessing competency using a combination of eye tracking and tool-motion data during surgical simulations. They argued that how a surgeon moves their tools in conjunction with their eye movements--i.e., where a surgeon looks when they operate, contains sufficient information to quantitatively and objectively evaluate surgical skill competencies. Other studies have found similar supporting evidence that eye gaze reflects cognitive strategies and predicts an individuals level of expertise in medicine (Ahmidi et al., 2012; Ashraf, Sodergren, Merali, Mylonas, Singh, & Darzi, 2018; Shinnick, 2016; Tien, Pucher, Sodergren, Sriskandarajah, Yang, & Darzi, 2014; Richstone, Schwartz, Seideman, Cadeddu, Marshall, & Kavoussi, 2010). Studies like these highlight the need to not only collect multimodal data in medical education and training, but also synthesize and provide that information back to educator in real-time to gain access to the developing competency of the medical student. Additionally, we would like to acknowledge that while eye movements provide rich, granular information on what medical professionals are perceiving within their field of view, eye movements alone cannot be the only data channel used to assess one's level of developing competency. Rather, combining multiple measures of process data such as think-alouds could reveal what the medical student is reasoning about and potentially help identify the judgments they are making about the medical case.

One area of promise in studying clinical reasoning as a process, and the role that metacognition and self-regulation might play in mitigating errors, could be by referring to Learning Analytics. Learning Analytics is a transdisciplinary research paradigm that aims to advance our understanding of learning by measuring, collecting, analyzing, and visualizing data about the learning process (Chan, Sebok-Syer, Thoma, Wise, Sherbino, & Pusic, 2018; Lang, Siemens, Wise, Gasevic, 2017). The overall objective of the Learning Analytics community is to optimize scaffolding and build effective interventions and programs to reveal learning processes as they unfold during problem solving or task activities, and ultimately assess relationships between these insights and training and performance outcomes (Joksimović, Kovanović, & Dawson, 2019). Within medicine, much attention on analytics has been for clinical purposes, such as quality improvement, with few efforts being directed within medical education. Since the age of data analytics is upon us, the time is now for medical education to leverage the potential of Learning Analytics (Chan et al., 2018).

Multimodal Learning Analytics in Medical Education

There is a longstanding tradition in the medical expertise and education literature regarding the use of individual data channels (e.g., eye movements, log-files, and concurrent think-alouds) to examine medical reasoning processes across sub-specialities (e.g., radiology, surgery; Ericsson, Hoffman, Kozbelt, & Williams, 2018; Feltovich, Prietula, & Ericsson, 2018; Norman, Grierson, Sherbino, Hamstra, Schmidt, & Mamede, 2018). However, converging these individual data challenges together has not been part of this research tradition. Multimodal learning analytics is an emerging interdisciplinary field led by computer scientists, engineers, cognitive and learning scientists, psychometricians, and data scientists who have focused on detecting, measuring, understanding, and predicting how various processes (e.g., metacognition) and data channels (e.g., concurrent think-alouds, eye movements, log-files, etc) are predictive of learning, reasoning, problem solving, etc. Azevedo & Gašević, 2019; Baltrušaitis, Ahuja, & Morency, 2018; Ochoa, 2017). We argue that multimodal learning analytics is extremely relevant and can significantly transform medical education to reveal developing clinical reasoning competency and the role of metacognition and self-regulation on improving medical students' ability to clinically reason effectively (e.g., Lajoie, Li, & Zheng, 2021).

Computational capabilities have grown exponentially over the last decade (Ochoa, 2017 and there has been a noteworthy shift in Learning Analytics toward utilizing multimodal data, i.e., multiple streams of data, within other domains. Multimodal learning analytics (MLA) operates to support advances in multimodal data capture, especially as it relates to signal processing, to address the challenges of studying a variety of complex learning-relevant constructs such as clinical reasoning as observed in complex learning environments such as medical settings (Worsley & Blikstein, 2015). Some examples of multimodal data include speech, video, electrocardiology, eye tracking, facial expressions of emotions, learner-system interactions, and psycho-physiological indicators (see examples by Blikstein, Worsley, Piech, Sahami, Cooper, & Koller, 2014; Sharma & Giannakos, 2020). Currently, however, the majority of studies utilize single data channels such as think-aloud protocols or human machine interactions to capture clinical reasoning during learning activities (Cloude, Ballelos, Azevedo, Castiglioni, LaRochelle, Andrews, & Hernandez, 2021; Forsberg, Ziegert, Hult, & Fors, 2014). Leveraging multimodal data has driven researchers to develop innovate ways for defining and measuring learning processes and medical education research could benefit from its expertise to define and capture metacognition and self-regulation and their role on clinical reasoning across a variety of contexts and interactions with patients (Azevedo & Gašević, 2019; Boulet & Durning, 2019; Lodge, Panadero, Broadbent, & de Barba, 2018).

However, leveraging MLA techniques to study metacognition and clinical reasoning within the field of medicine is still in its infancy. We argue MLA has the potential to address the significant challenges discussed by capturing metacognition, self-regulation, and clinical-reasoning processes both in real-time and over time to study when failures in reasoning might occur, and the role that metacognition and self-regulation could serve in mitigating these deficiencies for training purposes. For instance, what streams of data (e.g., concurrent verbalizations, eye movements, facial expressions of emotions, physiological sensors, gestures) could inform when a medical student may be failing to accurately reason about a diagnosis, and where do these failures stem? Why, and under what conditions, are these failures being made (e.g., cognitive bias, incomplete illness script, expertise level, metacognitive monitoring, fatigue, cognitive load, stress) and how might raising metacognitive awareness and self-regulation shed light on deficiencies in clinical reasoning to mitigate their detrimental effects?

To implement MLA into research within medical education aimed at improving clinical reasoning, it is essential to rely on a theoretical lens to guide selecting and defining key variables, methodologies, experimental design, and interpretation of findings since these data can be so large and at various time scales, granularities, and dimensions. Utilizing a theoretical perspective is required for making sense of big data and generating meaningful MLA for education and training as it provides a foundation with which to contextualize data signals that might indicate changes in clinical reasoning, metacognition, and/or self-regulation (Dawson, Joksimovic, Poquet, & Siemens, 2019). In order to accomplish this, we referred to the socio-cognitive cyclic (SCT) model of self-regulated learning by Zimmerman and Moylan (2009) in this chapter because it describes relations between metacognition and self-regulation as it relates to higher-order cognitive processes such as clinical reasoning. Specifically, we used this model to describe the current state of technology-supported (e.g., simulations, computer-assisted systems, high-fidelity mannequins, etc.) medical education literature targeting skills like surgical procedures and clinical reasoning. We further examine the degree to which these technologies capture and support medical students' metacognition and self-regulation, and their use of MLA. Additionally, we provide suggestions of ways that MLA might be applied, augmented, and improved from an SCT perspective for medical educators.

Socio-cognitive cyclic model of self-regulated learning

Not only do medical students need to learn the procedures and protocols for medical practices, they must also develop the self-regulatory processes (e.g., goal setting, metacognitive monitoring, reflecting, etc.) necessary for effectively applying and adapting clinical reasoning processes. Zimmerman's socio-cognitive cyclic model of self-regulated learning (SCT; Zimmerman & Moylan, 2009) highlights relations between various metacognitive processes that give rise to self-regulation and higher-order cognition such as clinically reasoning about a patient's disease.

According to Zimmerman and Moylan (2009), self-regulated learning is organized across three phases: (1) forethought, (2) performance, and (3) self-reflection. The Zimmerman model argues that during the forethought phase, individuals engage in setting goals by initiating metacognitive judgments about the problem space (e.g., patient presents with symptoms in conjunction with their history), and then they must make plans for ways to approach the task (e.g., run several lab tests). Within the performance phase, individuals act on their plans and metacognitive judgments by continuously monitoring and evaluating their progress. Finally, when individuals in the self-reflection phase, they continue to initiate metacognitive judgments based on their performance in addition to any changes occurring within their environment and context (e.g., the patient’s health begins declining at an accelerated rate). This new information informs the other phases, illustrating the cyclic nature of metacognition and self-regulation by pushing the individual to reconsider their goals in relation to their performance and the problem space in order to assess whether they need to adapt their plans and clinical-reasoning approach to meet the demands of the environment (Zimmerman & Moylan, 2009).

Previous works that including metacognitive activities throughout medical training, specifically simulation-based learning, can improve performance. (Josephsen, 2017) used observations and interviews of nursing students prior to and during simulation and debriefing to identify metacognitive strategies, revealing while some metacognitive strategies were enacted by students, they did not use enough to engage completely and self-regulate their own learning (Josephsen, 2017). This suggests that specific features throughout training should be developed to help students learn not only the necessary medical skills and knowledge for their professions but also the skills and processes needed to initiate metacognitive processes and self-regulate their learning for future training and education. Interventions for adapting and initiating these processes, however, require rich information about learning that can be captured using multimodal data. Compared to traditional assessment learning analytics focused on learning phenomenon as “event-focused” and “ontologically flat” (Noroozi, Alikhani, Järvelä, Kirschner, Juuso, & Seppänen, 2019), MLA can provide objective and contextualized information about one’s otherwise normally invisible metacognitive and self-regulated learning processes (i.e., planning, performance, and reflection) hidden within the black box of the mind. For example, if a medical student reveals poor judgment accuracy within the forethought phase which is displayed with MLA prior to performance, intervention can occur before the mistakes compound in later phases that would be otherwise operating from weak metacognitive skills and could be misattributed as a poor cognition event instead, missing opportunity to augment the student's reasoning and performance.

Forethought Phase

The forethought phase is essential for both orientating to the problem space (e.g., new patient within the hospital setting) and planning/adapting actions before and after performing (e.g., running lab tests; Lajoie, Poitras, Doleck, & Jarrell, 2015). Metacognitive processes continuously guide one's attention and efforts in future phases, such as developing hypotheses about a disease while clinically reasoning about a medical case. However, few studies capture MLA on medical students' forethought phase of metacognition and self-regulation prior to intervention, but instead rely on a "snapshot" of the previous state of a medical student as measured via pre-test assessments. It is important to capture processes within the forethought phase continuously across learning activities given the cyclic nature of SCT.

For example, BioWorld, an intelligent tutoring system designed to provide novices with opportunity to practice clinical reasoning using embedded tools and features such as a Hypothesis Manager, which allows students to plan and track their differential hypothesis and initiate metacognitive judgments about diseases over time, fostering the forethought phase (Lajoie et al., 2015). Specifically, they captured and converged timestamped learner-system interactions, concurrent verbalizations, and self-report questionnaires from 27 medical students. Utilizing these data and text mining analysis, they defined diagnostic performance by calculated the percentage of steps that medical students had taken in relation to subject-matter experts steps taken to reach a final diagnosis. Their results showed that the more often medical students initiated metacognitive feelings-of-knowing judgments (e.g., *"That's why I don't think it's just the anxiety, the panic attacks, and it must be something else."*), the higher their diagnostic efficiency. Yet, the more metacognitive judgments-of-learning initiated during clinical reasoning (e.g., *"I guess that would be urine, glucose, no, no, no, no, sodium."*), the lower their diagnostic efficiency (Lajoie et al., 2015).

While this approach uses MLA (i.e., logfiles, self-report measures, think-alouds) to study metacognition and self-regulation within the forethought phase during clinical reasoning with computer-based systems, we argue that supplementing other data channels beyond self-report and human-system interactions could offer a more complete and dynamic capturing of the forethought phase (see Cloude, Dever, Wiedbusch, & Azevedo, 2020). For example, studies find that eye movement metrics (e.g., measuring time spent fixating on relevant (or irrelevant) elements within an interface) contextualize log files and have been shown to reflect metacognitive judgments (Mudrick, Azevedo, & Taub, 2019; Wiedbusch & Azevedo, 2020). Eye movements have also been predictive of future actions and intentions (Deng & Gu, 2021; Koochaki & Najafizadeh, 2021), and have now been used to distinguish between intentional reasoning versus unintentional interaction with a system (e.g., Cloude et al., 2020). By aligning eye-tracking with these self-report and human-interaction measures, MLA become richer and more dynamic in a phase of self-regulation that is normally very internal.

BioWorld is one of the few medical education technologies that specifically targets the forethought phase of self-regulation. A large portion of medical education also involves simulation-based technologies, such as the da Vinci simulator, but currently there are no methodological approaches for capturing the forethought phase during learning with such technologies. Instead, most methods are utilized to capture data on developing procedural skills (e.g., catheter placement), and consequently studies using these technologies demonstrate gaps in literature because there is little known about the role of metacognition or self-regulated learning during the forethought phase on procedural skill performance with simulation-based technologies.

In a systematic review of simulation-based medical education studies, Rooney and others (2018) found that across 54 studies (covering screen-based simulators, simulated environments, virtual reality/haptic systems, simulated patients, part-task trainers, and computer-based systems with mannequins), clinical skill sets were assessed using pre/post-test assessments and contouring analysis in 26 of studies. This work highlights that, although studies show that simulations can be used as treatment-planning (i.e., future medical course of actions), a potential component of planning behaviors assumed within the forethought phase, there is still a disconnect between planning and practical application tasks (Rooney, Zhu, Gillespie, Gunther, McKillip, Lineberry,... & Golden, 2018). That is, surgical competency is not just the performance of mechanical tasks. Surgery also requires decision-making and the judgment of every action being within the best interest of the patient's safety and quality of care (Hall, Ellis, & Hamdorf, 2003). It is not unreasonable to assume that when considering these types of simulation-based technologies, measuring and analyzing metacognition and self-regulation using MLA could shed light on the abilities of medical students to help provide more individualized scaffolding when there are deficiencies in performance during training activities.

A study by Kahol, Vankipuram, and Smith (2009) proposed a multimodal architecture to be used in psychomotor cognitive simulators, or simulators that train clinical reasoning in addition to the mechanical movements requirement in medical activities. This architecture combines sensory-level cognitive mechanisms, captured using eye movements and skin conductance data analytics, with more (meta)cognitive processes (such as movement planning) captured on a higher level using haptic-based analytics about intermodal coordination and transfer (i.e., exchanges between perception from multiple sources). While developing the architecture and cognitive surgical simulator, ten residents completed a psychomotor skills task and several cognitive skills tasks using a series of pegboard tasks. Within the basic ring transfer task (testing psychomotor skills), residents were asked to pick up a ring with a haptic joystick and then place it on a highlighted peg on a simulated pegboard. The tracing tasks required residents to pick up the ring again, however it was moving within the environment. To test orientation, the residents had to account for different placement orientations. In a preparatory attention task, the highlighted peg was only highlighted briefly as compared to always being lit. Working memory was assessed by highlighting a sequence of pegs instead of just one. Finally, visio-haptic transfer was tested by displaying and hiding the entire pegboard throughout the task. Residents were evaluated on the number of cognitive errors made and hand/tool movement captured through a tracking glove and joystick. Hand and tool movement smoothness was used as a measurement of surgical psychomotor proficiency. Movement planning was measured during the 2D and 3D tracking tasks as well as the orientation and visio-haptic transfer tasks. Their work found that cognitive training prior to simulation-based training improved performance, and tasks such as movement planning were important for developing future surgical skills (Kahol et al., 2009) This highlights how medical education technology needs to collect and measure not just the learning analytics around the performance of actions, but also MLA of metacognition and self-regulation during the forethought phase.

Performance Phase

Most technology-enhanced simulation education directly measures the performance phase. They target developing and applying skills and medical knowledge before and after training activities. These educational devices utilize various technologies such as virtual computer patients that either enhance hands-on simulation practice (e.g., high-fidelity simulation mannequins) or completely virtual patients only requiring a computer (i.e., without any additional hands-on simulation). In a meta-analysis and systematic review by Cook and colleagues (2011), the authors report that technology-enhanced simulation (excluding studies with virtual patients that only require computers), resulted in both high effect sizes for acquiring medical knowledge (N = 118 studies, d_{pooled} = 1.20, p < 0.001), skills (N = 426 studies, d_{pooled} = 1.09, p < 0.001), and behaviors (N = 20 studies, d_{pooled} = 0.79, p < 0.001) of medical students as well as moderate effect sizes for patient-related outcomes (N = 32 studies, d_{pooled} = 0.50, p < 0.001). Additionally, they reported the majority of outcomes measured in the studies (N = 690, 75%) assessed performance based on time to complete a task, global subjective ratings (by faculty or computer) of process (number of errors made), and task products (e.g., detection of key abnormalities; Cook, Hatala, Brydges, Zendejas, Szostek, Wang,... & Hamstra, 2011). Comparatively, a meta-analysis examining computerized clinical case simulation found that while the simulation resulted in a large positive effect compared to no intervention, they were small when compared to non-computer instruction (Cook, Andriole, Durning, Roberts, & Triola, 2010). The authors highlight that most of the studies largely focused on short-term knowledge and failed to capture long-term effects (i.e., transfer), or process-based metrics, suggesting the field develop new measures for evaluation. No mention was made specifically about additional data streams or metacognition and self-regulation, which could shed light on developing expertise using MLA generated across activities to assess long-term effects of training (i.e., transfer to real-world settings).

Some work has been done to capture performance using multimodal data such as eye-tracking. For example, Krupinski and colleagues (2006) used eye movements to compare how medical students and residents studied virtual slides for diagnosing pathology as compared to practicing pathologists. They showed that across all levels of expertise, slide readers attended to similar locations that would contribute to correct diagnostic decisions, however the speed at which these locations were identified were faster for more experienced observers (Krupinski, Tillack, Richter, Henderson, Bhattacharyya, Scott... & Weinstein, 2006). This highlights two important implications- (1) there needs to be multiple sources of data to assess performance and (2) physiological and behavioral data alone is not enough to determine if/where there are failures during the clinical-reasoning process.

The advantage of incorporating measures of metacognition and self-regulation during the performance phase to generate MLA would be to pinpoint what and when medical students are evaluating during their clinical-reasoning process to assess its effect on performance. Specifically, to distinguish characteristics between the forethought and performance phases, we argue that the reference would be to the students' goals. For instance, metacognition and self-regulation that was related to the goal of the training session would be indicative of the forethought phase based on planning and adaptive behaviors such as adapting strategies based on their performance in relation to meeting a goal. Metacognition and self-regulation within the performance phase might be solely focused on completing the task rather than assessing planning and strategic behaviors for achieving their goals (based on the information received from the performance phase--e.g., *"how well did I do?"*).

Self-Reflection Phase

Some work has been done to capture process data on reflection-based metacognitive processes during clinical reasoning. For example, Lajoie et al (2015) used sequential pattern mining on log files with another study in BioWorld to identify self-reflection during clinical reasoning. Specifically, they examined the actions students took prior to and immediately after requesting help from the tutoring system, capturing temporally when students had reflected on their current state of reasoning and evaluated their need for more assistance using learner-system interactions. This work highlights how metacognitive strategies and processes can be captured within simulations and feedback into the adaptive aspects of these technologies. Instead of offering undirected feedback at generic decision making points in training, BioWorld gives direct feedback based on a medical students' learning in real time.

In another study by Wesiak and others (2014) students using an enhanced-training simulator (ImREAL) with (affective) metacognitive scaffolding were asked to provide feedback about the usefulness of the last part of the simulation. Additionally, students were able to use a notepad throughout learning to reflect or take notes on, and the authors found that despite the fact that the number of users using the notepad decreased over time, the number of reflections within those uses increased (Wesiak, Steiner, Moore, Dagger, Power, Berthold... & Conlan, 2014). It is important to note that this study only used self-report measures post-learning and the coding of the notepad to capture metacognition the self-reflection phase during learning activities. However, this work highlights how building in tools that directly support and engage students in active metacognitive processing can lead to deeper learning without necessarily requiring more work from students.

Birt, Moore, and Cowling (2017) explored the effectiveness of mobile mixed-reality medical and health sciences education of first and second year physiology and anatomy students. In their work, students were given narrated 3D models of the brain, spinal cord, and brainstem for knowledge acquisition. They were also provided a simulation focused on skill development that had students perform a laryngoscopy on a virtual patient while being provided step by step instructions. In addition to the traditional pre and post skills and knowledge assessments, students were retroactively interviewed about their experiences, and asked to reflect on the affordances of using mixed-reality for training. The authors found that fidelity of the simulation was not as important for learning, but rather it is the design and human affordances of the technology (Birt et al., 2017). The reflection of their experiences revealed that students also wanted the ability to do more than recall their experience, but also rewatch their simulation to further reflect. On a longer-term scale, many students are now being required to develop an "ePortfolio", or a digital collection of student assignments that encourages reflection of one's progress and provides a source of faculty assessment (Lewis & Baker, 2007; Norris & Gimber, 2013).

After-action reviews, also known as debriefing, are systematically structured retrospective procedures for capturing self-reflection about one's performance and previous experiences (Keiser & Arthur, 2021). The effectiveness of simulations is dependent on these debriefings to provide timely and accurate performance feedback, typically provided through verbal debriefing, or review of a videotaped performance with the trainee and instructor (Henneman, Cunningham, Fisher, Plotkin, Nathanson, Roche... & Henneman, 2014). Only recently have these reviews begun to include other, more objective, measures of performance. For example, Henneman and others (2014) examine how including eye-tracking during an after-action review could improve simulation performance. The authors found even though it did not improve knowledge of medication or treatments, including eye-tracking technology in an after-action review improved patient safety practices such as patient identification. Feedback on other objective measures, such as spatial and temporal location data (i.e., body posture) and communication have also been used in after-action reviews for simulation training. The

Interpersonal Scenario Visualizer (IPSViz) is a debriefing tool to help students reflect on their interactions with simulated patients (Raij & Lok, 2008). This tool was shown that providing students with visualizations of multiple perspectives of body positioning and conversation flow metrics (i.e., rapport-building, levels of friendliness, etc.) changed perceptions and awareness of the simulated interactions, with many students reporting they would change their future behavior after the review (Raij & Lok, 2008). These studies emphasize the impactful role that not only the collection of multimodal data can have on medical education, but also the effectiveness with the inclusion of MLA provided during the self-reflection phase of self-regulated learning.

Implications of multimodal learning analytics to improve medical education

In this section, we present the implications of leveraging multimodal learning analytics to improve medical education. We focus on the four major pillars of multimodal learning analytics: detecting, modeling, tracing, and fostering metacognition and self-regulated learning during clinical reasoning with emerging technologies. Based on extensive research on multimodal data channels (D'mello, 2017), we situate our examples of specific multimodal data, including but not limited to log-files, screen recordings of human-machine interactions, eye movements, and concurrent think-alouds according to each the three phases related to SCT theory of self-regulated learning (SRL; Zimmerman & Moylan, 2009). We highlight that our proposed implications do not include all possible streams of data (e.g., facial expressions of emotions, EEG, and other physiological devices and sensors), SRL processes (e.g., emotional, motivational and social processes), and learning analytics techniques and data visualizations, but offer guidelines for future research directions[1].

Detecting Metacognition and Self-regulated Learning Processes

Detecting is a critical first step in designing theoretically-based and empirically-driven multimodal learning analytics to understand clinical reasoning. Detecting metacognition and SRL involves using various devices and sensors including log-files to collect human-machine interactions while medical professionals learn or train using systems such as BioWorld (Lajoie et al., 2020); eye tracking systems to collect fine-grained eye metrics such as rate of fixations, saccades, regressions, among many others to reflect medical professional's attention allocation and cognitive processing of information—e.g., do they fixate on relevant parts of a mammogram that characterizes a tumor, and if so, how long do they spend fixating on it and do they return to it and why? Video and audio recordings of human-machine interactions while medical professionals interact with simulators capture contextual information, allowing researchers to make inferences based on verbal (e.g., think-alouds that can be coded for metacognitive and SRL processes and other variables associated with clinical reasoning; Artino, Cleary, Dong, Hemmer, & Durning, 2014; Azevedo & Lajoie, 1998; Cleary, Durning, & Artino, 2016) and non-verbal behaviors, such as tool-hand coordination during surgical procedures (Menekse Dalveren & Cagiltay, 2020), including how they navigated and interacted with the system (e.g., hypermedia system (Cloude et al., 2021), intelligent tutoring system (Feyzi-Behnagh, Azevedo, Legowski,

[1] Not all the same data sources will provide relevant information in all of the three phases of the SCT model, and the rate and way in which data are fused should change based on the research questions, experimental design, theoretical framework, nature of the task, etc.

Reitmeyer, Tseytlin, & Crowley, 2014), virtual reality system (Orlosky, Itoh, Ranchet, Kiyokawa, Morgan, & Devos, 2017), and their features (e.g., response to virtual patients' reactions), how they interacted with other students (e.g., team performance using high-fidelity mannequins; Duffy et al., 2015). In sum, these detection devices and sensors provide important data on various aspects of metacognition and SRL processes but two majors questions exist— (1) the accuracy of detecting metacognitive and SRL processes that are relevant to clinical reasoning but based on the three phases of SCT model of SRL (Zimmerman & Moylan, 2009), and (2) implications for using multimodal learning analytics to improve medical education. What is the accuracy of multimodal data detection in collecting metacognitive and SRL processes that are relevant to clinical reasoning but based on the three phases of SCT (Zimmerman & Moylan, 2009)?

During the *forethought phase*, detecting clinical reasoning-relevant metacognitive and SRL processes include setting goals and strategic planning which can be accomplished by using log-files, screen recordings of human-machine interactions, eye movements, and concurrent think-alouds. Each of these data channels are capable of providing data as to when, where, and how medical professionals engage in setting goals and strategic planning prior to clinical reasoning. For example, they may read a clinical history, review radiological evidence, interview a virtual patient, retrieve a relevant illness script from long-term memory, etc.

In the *performance phase*, there are a dozen relevant self-control task, attentional, and self-observation strategies that we would detect by using log-files, screen recordings of human-machine interactions, eye movements, and concurrent think-alouds. The performance phase is when medical professionals would use several metacognitive, SRL, and clinical-reasoning strategies in real-time and therefore the full suite of multimodal data channels would be ideal to capture the cognitive and metacognitive processes enact during clinical reasoning regardless of medical education context (e.g., virtual reality simulators, high-fidelity mannequins, hypermedia, intelligent tutoring system, etc.). This phase of metacognition and SRL is ideally suited for using multimodal data (Azevedo, Taub, & Mudrick, 2018; D'mello, 2017; Emara, Hutchins, Grover, Snyder, & Biswas, 2021; Järvelä, Malmberg, Haataja, Sobocinski, & Kirschner, 2019; Lajoie et al., 2020; Winne, 2019). However, there is a major lingering question—i.e., how many of these data channels are truly needed to detect, model, trace and foster metacognition and SRL during clinical reasoning? This question will be revisited in the next sections.

The *self-reflection phase* is quite interesting from a multimodal data perspective in supporting learning analytics to improve medical education. More specifically, this phase focuses on self-judgments and self-reactions following clinical reasoning. As such, the most important multimodal data to collect to understand these processes would likely be concurrent verbalizations and eye movements performed retrospectively while the medical professionals review their data such as a video recording of their own multimodal data collected during the performance phase. These data could be used to reflect their level of expertise and focused attention (Ericsson, 2007, 2015; Kok & Jarodska, 2017; Normal et al., 2018) to provide insight into why they missed, misused, etc. critical/relevant part(s) of the clinical scenario (e.g., relevant patient data, misinterpreted findings, emotions of the virtual patient, deterioration of the high-fidelity mannequin) including other students (e.g., team member in emergency room) to support self-reflection. In sum, decisions made about detecting metacognition and SRL processes in supporting clinical reasoning and multimodal learning analytics is critical and feeds the other issues covered in this section, modeling, and tracing, and fostering multimodal learning analytics to improve medical education.

Modeling metacognition and self-regulated learning processes

The multimodal data detected in the previous sections can be used to model these processes both in the students (i.e., medical professional in various specialties and levels of expertise) and emerging technologies for medical education (e.g., BioWorld; Lajoie et al., 2021). Ideally, both students and learning technologies should each be modeling the metacognitive, SRL processes, and clinical processes while solving medical cases based on the detection techniques described above.

In this section, we assume that all students are capable of accurately and dynamically model the metacognitive, SRL, and clinical reasoning processes and that the modeling of the processes is made "visible" or externalize from the multimodal data (Azevedo et al., 2018). More specifically, students have the potential to model these processes since their concurrent verbalizations can reveal metacognitive (e.g., I am trying to understand how this find in the chest x-ray fits with the diagnosis of pneumonia), SRL (e.g., I need to plan how to solve this case since it is rather complex and I do not have experience with solving atypical cases), and clinical-reasoning processes (e.g., *"I really think all evidence points to a diagnosis of ductal carcinoma in situ"*). In serial or parallel fashion, depending on the data channel and temporal dynamics of information, the processes being uttered may show their eye movements focusing their attention on relevant aspects of the clinical case (e.g., prolonged fixations on the tumor shown in the mammogram; Hermens et al., 2013), etc. We argue that while detection methods make the processes "visible" to other students (e.g., researchers, team members, peers) that may not all be modeled by the medical professionals for various reasons including accessibility to these processes, extraneous cognitive load, awareness of these processes, automaticity of these processes based on level of medical expertise, dynamics of the processes and the nature of their temporal unfolding in time as serial or parallel processes (Azevedo, 2020; Paas & van Merriënboer, 2020).

In summary, while students may reveal these processes by using sensors, devices, and other detection techniques and therefore make them visible, they may not be able to model them all internally (i.e., in the cognitive architecture). As such, we argue that intelligent learning technologies (e.g., intelligent tutoring system, intelligent immersive virtual reality environments with AI-based virtual trainer and virtual patients) capable of modeling these processes by having direct access to the medical professionals' multimodal data could be more accurate and objective in modeling these processes due to their computational power and AI-based inferencing and sense-making of these data. Despite the intelligent learning technologies' potential, we acknowledge that several constraints and limitations still exist since intelligent systems are not capable of fully modeling all processes accurately for a variety of reasons. One of the major reasons for intelligent learning systems not being capable of accurately modeling all metacognitive, SRL, and clinical-reasoning processes is that a major component of modeling is still not accessible to students.

For example, while intelligent learning systems collect log-files of human-machine interactions at very precise timescales (e.g., milliseconds) they need an inference engine to make sense of what to model, how to model, and when to model these processes utilizing a theoretical and empirical perspective. At a basic level, log-files can be shown to the student as the system collects them in real-time, but such data will not make sense to a student and needs to be computed, re-represented, and eventually shown to a student to support their clinical reasoning, depending on the objectives of log-files data for tracing and fostering metacognition, SRL, and clinical reasoning (e.g., time spent solving a case, frequency of errors, timing and sequencing of clinical reasoning steps, etc.).

This issue about log-files exists when considering modeling metacognition, SRL, and clinical reasoning processes with other multimodal data. For example, modeling these processes from concurrent think-alouds would require natural language capabilities that would allow the system to parse, infer, and model the processes for the student. Similarly, modeling physiological data would require several transformations from the pure signal to a representation that would make sense to a student and the time for this type of processing and inferencing may take time to model. By contrast, eye movements, i.e., gaze behaviors can be displayed in real-time as interface overlays and the student can accurately and simultaneously compare their eye movements with that of an expert to get immediate feedback on the learning, performance, etc. (Chetwood, Kwok, Sun, Mylonas, Clark, Darzi, & Yang, 2012). However, gaze behaviors alone are not sufficient to model metacognition, SRL, and clinical reasoning processes (Kok & Jarodzka, 2017).

In addition, other variables extracted from eye movement data (e.g., fixations, saccades, blink rate, dwell time, scanpaths, pupil dilation, etc.) need to be processed, inferred, and re-represented in a manner that makes sense to a student such as including contextual information and pre-defined Areas of Interest (AOIs) around relevant cues within the environment as they relate to different steps within task procedures (Erridge, Ashraf, Purkayastha, Darzi, & Sodergren, 2018). For example, what type of representation is ideal to model the student's metacognitive monitoring? Would the representation be an amalgamation of the frequency and duration of fixations, saccades, and regressions on relevant and irrelevant AOIs? In sum, we argue that to accurately and dynamically model these processes students and machines can each play their individual roles and strategically take advantage of their strengths (e.g., computational power of machines to collect, amalgamate, represent processes) while minimizing their weaknesses (e.g., student's lack of metacognitive awareness and accurate monitoring of their physiological, facial, and eye movements). These challenges are described below as they have implications for tracing and fostering metacognitive, SRL, and clinical processes.

Tracing Metacognition and Self-regulated Learning Processes

Tracing refers to a system's ability to identify students' externalized learning processes continuously over time (Ochoa, 2017). Specifically, within the field of learning analytics, it is essential to study how students' behaviors, captured through multimodal data such as eye-tracking, log-files, concurrent verbalizations, etc., change over time as an external representation of their dynamic learning processes, performance, and outcomes while using emerging technologies. These technologies often capture students' interactions with the system (i.e., log-files, word processing) over time to examine how learning processes occur and illuminate the internal processes of the student. However, to fully understand how medical students deploy learning processes (e.g., metacognitive strategies, self-regulated learning processes) over time while interacting with emerging technologies, researchers need to converge multimodal data using a theoretical perspective grounded in metacognition and self-regulated learning. This subsection highlights how emerging technologies might trace students' internal learning processes of metacognition and self-regulated learning to transform clinical-reasoning education.

Several emerging technologies mentioned previously (e.g., BioWorld, ImREAL) naturally capture data, such as log files, to trace how students use metacognitive and self-regulated learning strategies, i.e., timing and use of self-reflections, over time while using clinical reasoning to successfully complete objectives within the learning environments.

However, the largest limitations of tracing overt behaviors within these environments include: (1) relying on unimodal data (i.e., only using one mode of data to trace behaviors); (2) context (e.g., students' existing metacognitive skills; environmental influences) is integrated by the researcher, not intelligently supplied by the system; and (3) behavior and measures of performance were evaluated and used for interpreting students' metacognitive and self-regulated process use post-hoc. With these limitations, researchers must first establish how to theoretically build emerging technologies and collect and interpret multimodal data in tracing students' metacognitive and self-regulated learning processes over time. In using the SCT model of self-regulation (Zimmerman & Moylan, 2009), emerging technologies can accurately and intelligently trace, via multimodal data, how students deploy different metacognitive and self-regulated learning processes throughout learning activities, including which and in what order these processes (e.g., goal setting, help-seeking, self-evaluation) were used.

Through using several combinations of data streams to trace *forethought*, we can identify students' (in)accurately using metacognitive and SRL strategies within this phase as well as predict the transition to the *performance* and *self-reflection* phases of SCT. For example, goal setting and planning can be identified by both students' concurrent verbalizations and eye-tracking data. While concurrent verbalizations can externalize students' implicit processing, this modality alone is limited to students' awareness of their actions, internal processes, and use of metacognitive and self-regulated learning processes. Similarly, eye-tracking data can identify the context-specific information that is being attended to, but not explicitly the internal processes being utilized. For example, a student may verbalize that they want to correctly identify the source of a patient's symptoms (goal setting) by running a blood work test (planning). However, when examining where the student is looking using eye-gaze metrics, one might see longer dwells on symptoms that could not be tested through blood work, hinting at a more complex differential diagnosis that requires additional testing during the performance phase. In addition, these data can reveal anchoring biases through contradictory implications between data modalities. In utilizing multiple modalities, data reveal a more context-driven and comprehensive understanding of students' internal learning processes during forethought than the use of unimodal data interpretations.

Performance can be traced over time using a combination of log-file and eye-tracking data where the SCT model proposes performance is represented by students' deployment of task strategies (i.e., help-seeking, note-taking) that can be directly measured over time by log files, and students' metacognitive monitoring and attention allocation measured by eye-tracking data. For example, log files within an emerging technology focused on clinical reasoning can explicitly measure the use of note-taking behaviors as the student deploys task strategies to diagnose an illness, where eye-tracking would support conclusions as to the potential success of clinical reasoning as students allocate their attention to certain symptoms or diagnoses.

Students' *self-reflection* of their success in clinical reasoning and use of metacognitive and self-regulated learning process deployment can be identified using log-file (e.g., word processing) and facial expressions of emotion (e.g., frustration). For example, logfiles can identify self-reflections of a student's perceived clinical reasoning ability and performance as well as metacognitive and self-regulatory processes. Facial expressions of emotions, in conjunction with logfiles, can be used to trace students' self-satisfaction and reactions to their self-reflections, -judgments, -evaluations, etc. For example, after or during a reflection, students may emote disgust or frustration to a perceived inaccurate clinical diagnosis whereas joy would be emoted during a perceived accurate diagnosis. In combination with log-files that can also demonstrate students’ (in)accurate performance and strategy use, facial expressions of emotion can identify the accuracy of students' self-reflections during clinical reasoning.

This example of self-reflection demonstrates how tracing is essential to examine the transition between metacognitive and self-regulatory phases as well as using multiple data modalities to provide contextual interpretations of learning processes over time, using the SCT model as a theoretical foundation. Tracing learning processes throughout learning with an advanced learning technology can illuminate how students use metacognition and self-regulated learning processes and provide implications for how to foster students’ accurate use of these processes during clinical reasoning.

Fostering Metacognition and Self-regulated Learning Processes

It is essential for emerging technologies to foster metacognition and self-regulated learning processes during clinical reasoning to improve medical education, as most students are incapable of effectively and accurately using metacognitive and self-regulated learning skills and processed during complex tasks (Josephsen, 2017, Zheng & Zhang, 2020). Issues in deploying and monitoring the use of metacognitive and self-regulated learning processes can be demonstrated by students’ inability to identify when a specific process or strategy is required given students’ prior content and metacognitive knowledge or task requirements, constraints, and resources (Winne, 2019). As such, emerging technologies can foster students’ more frequent and accurate use of metacognitive and self-regulated learning strategies theoretically grounded in SCT of self-regulation (Zimmerman & Moylan, 2009).

The *forethought* phase is required for students to orient themselves within the environment of the emerging technology (e.g., identify tools and resources), identify or set goals to accomplish (e.g., correctly diagnose an illness), and construct plans around the strategies needed to achieve their set goals (e.g., take notes, monitor anchoring biases). Few emerging technologies have incorporated metacognitive and self-regulated learning fostering techniques. BioWorld, as previously explored, contains embedded tools that require students to make hypotheses and metacognitive judgments to achieve goals (Lajoie et al., 2015). From these features, essential processes are fostered for better clinical reasoning development. Similar to BioWorld, other emerging technologies can better foster the forethought phase by incorporating features that explicitly, or covertly, prompt students to engage in goal setting, planning, and evaluating potential outcomes, task values, and time and effort.

While most studies examine performance in terms of learning outcomes for medical knowledge both during and after a student has interacted with emerging technologies (Cook et al., 2010, 201; Josephsen, 2017, Zheng & Zhang, 2020), fostering performance requires researchers and instructional designers to question not the quantity of medical knowledge students obtain from the environment but *how* medical knowledge outcomes, clinical reasoning, and metacognitive and self-regulated learning skills can be increased using features within the learning technology. Features theoretically grounded in the SCT model of self-regulation (Zimmerman & Moylan, 2009) would encourage and foster students' attention on relevant information, deployment of effective and accurate learning strategies (e.g., note taking, summarizing), help-seeking behaviors, and monitoring their knowledge, skills, and progress towards goals. Future research and emerging technologies should prioritize students' ability to deploy performance-enhancing strategies that can apply across medical domains and within clinical reasoning tasks. These techniques and features would further foster *self-reflection* to support students' affective responses to self-evaluations and guide students in understanding how to then adapt plans, goals, and metacognitive and self-regulatory processes deployed during the performance phase. In fostering both performance and self-reflective strategies centered around the SCT model of self-regulation, emerging technologies can better support the development and execution of clinical reasoning skills.

However, the future of medical education is not only centered around how emerging technologies can implicitly include techniques and features to foster these metacognitive and self-regulatory processes but intelligently and adaptively support individualized clinical reasoning development across medical students. To do so requires the use of multimodal data in detecting, modeling, and tracing metacognition and self-regulated learning processes and strategies over time to better inform these technologies in how to foster the use of these processes across students. In this manner, these learning technologies can transform how medical knowledge is conveyed, the method in which clinical reasoning skills are diagnosed and developed by these technologies via multimodal learning analytics, and the ability for metacognition and self-regulation to be cornerstones of medical education.

CHAPTER THREE: QUANITFYING SCIENTIFIC THINKING USING MULTICHANNEL DATA WITH CRYSTAL ISLAND: IMPLICATIONS FOR INDIVIDUALIZED GAME-LEARNING ANALYTICS

This chapter, “quantifying scientific thinking using multichannel data with Crystal Island: Implications for individualized game-learning analytics” was originally published in an open-access peer-reviewed journal called *Frontiers in Education*, vol. 5, and led by first author Elizabeth B. Cloude and co-authors Daryn A. Dever, Megan D. Wiedbusch, and Dr. Roger Azevedo (2020).

Abstract

Quantifying scientific thinking using multichannel data to individualize game-based learning remains a significant challenge for researchers and educators. Not only do empirical studies find that learners do not possess sufficient scientific-thinking skills to deal with the demands of the 21st century, but there is little agreement in how researchers should accurately and dynamically capture scientific thinking with game-based learning environments (GBLEs). Traditionally, in-game actions, collected through log files, are used to define if, when, and for how long learners think scientifically about solving complex problems with GBLEs. But can in-game actions distinguish between learners who are thinking scientifically while solving problems versus those who are not? We argue that collecting multiple channels of data identifies if, when, and for how long learners think scientifically during game-based learning compared to only in-game actions.

In this study, we examined relationships between 68 undergraduates’ pre-test scores (i.e., prior knowledge), degree of agency, eye movements and in-game actions related to scientific-thinking actions during game-based learning, and performance outcomes after learning about microbiology with Crystal Island. Results showed significant predictive relationships between eye movements, prior knowledge, degree of agency, and in-game actions related to scientific thinking, suggesting that combining these data channels has the potential to capture when learners engage in scientific thinking and its relation to performance with GBLEs. Our findings provide implications for using multichannel data, e.g., eye-gaze and in-game actions, to capture scientific thinking and inform game-learning analytics to guide instructional decision making and enhance our understanding of scientific thinking within GBLEs. We discuss GBLEs designed to guide individualized and adaptive game-analytics using learners' multichannel data to optimize scientific thinking and performance.

Introduction

Scientific thinking has steered the crusade of discovery and changed the way in which we understand, interact, and exist in the world (Klahr, Zimmerman, & Matlen, 2019). As our communities are faced with persisting, and sometimes novel, socioeconomic, environmental, and health-related challenges (e.g., interrupting traditional classroom instruction and resorting to remote learning due to the COVID-19 pandemic), it is essential to equip future generations of learners with scientific-thinking skills (Kuhl, Lim, Guerriero, & Damme, 2019; NASEM, 2018; NRC, 2015, 2012). Recent advances in learning technologies, such as game-based learning environments (GBLEs), have guided design principles (e.g., role of agency in fostering cognitive engagement underlying scientific thinking) that enhance learners' skills and knowledge related to scientific thinking and offer solutions to problems that learners face in formal education settings (e.g., learners being tasked with memorizing factual information outside of real-world application; Deater-Deckard, Change, & Evans, 2013; Morris, Croker, Zimmerman, Gill, & Romig, 2013; National Research Council, 2011; Plass, Mayer, & Homer, 2020). After decades of studying cognitive processes underlying scientific thinking for training and educational purposes (Byrnes & Dunbar, 2014), it has become clear scientific thinking relies on factors that are personal to each learner (Dunbar & Klahr, 2012; Klahr & Dunbar, 1988).

For example, prior knowledge plays a key role in learners' ability to formulate hypotheses. However, studies with GBLEs use an approach that generalizes across learners, failing to account for individual characteristics that may impact learners' ability to think scientifically. These issues further compound current problems associated with providing individualized instructions in GBLEs. If we do not know how to capture if, when, and for how long learners engage in scientific thinking with GBLEs, how do we provide data-driven, individualized instruction to meet learners' needs? Researchers face additional challenges when capturing scientific thinking with GBLEs because most studies rely on in-game actions, measured solely through log files, to define if, when, and for how long a learner is thinking scientifically about solving problems, such as the amount of time using a scanner to test evidence (Smith, Shute, & Muenzenberger, 2019).

But we argue in-game actions do not provide enough information to identify whether learners are thinking scientifically about solving problems with GBLEs. For example, can researchers distinguish between a learner aimlessly engaging in an action (e.g., testing random food items) versus a learner engaging in an action because they are thinking scientifically about solving a problem (e.g., testing food items based on current hypotheses developed from information gathered)? Other data channels such as eye movements and pre-test scores could supplement information captured via in-game actions to inform if, when, and for how long learners are thinking scientifically during game-based learning. In this study, we investigated whether eye movements, pre-test scores, degree of agency, in-game actions, and post-test scores identified if, when, and for how long learners were thinking scientifically about solving problems with GBLEs. Our findings provide implications for designing GBLEs to guide individualized and adaptive interventions based on learners' individual needs using their multichannel data to optimize scientific-thinking skills and performance.

What is Scientific Thinking?

Scientific thinking defines two types of thinking (Dunbar & Klahr, 2012; Klahr et al., 2019). The first is literally thinking about the content of science-related topics, such as the characteristics of a virus. The second type of thinking encompasses a set of reasoning strategies or cognitive processes, such as inductive and deductive reasoning, problem solving, as well as hypothesis formulation and testing (Dunbar & Klahr, 2012; Zimmerman & Croker, 2014). These two types of scientific thinking have a codependent relationship with each other (Dunbar & Klahr, 2012; Klahr et al., 2019). For instance, a learner must think about the idiosyncrasies of a virus in relation to other infectious diseases to conceptually understand the content, which lays the foundation for reasoning about how a virus might spread in a research camp compared to bacteria. Because of this, learners' prior knowledge plays a crucial role in scientific thinking as related to reasoning strategies and other cognitive processes (Zimmerman & Croker, 2014). However, few studies account for the role of learners' prior knowledge in their ability to think scientifically during problem solving with GBLEs. Klahr and Dunbar (1988) developed a framework called scientific discovery as dual search (SDDS) which accounts for prior knowledge by describing hypothesis formulation as relying on long-term memory to identify gaps in knowledge to drive information-gathering actions (Dunbar & Klahr, 2012; Zimmerman

& Crocker, 2014). Because of this, SDDS has the potential to explain learners' individual characteristics which may influence their ability to think scientifically.

Scientific Discovery as Dual Search

Klahr and Dunbar (1988) conceptualize scientific thinking as searching within and between two problem spaces-- the hypothesis space and the experimental space. Each space is distinguished by its own set of operations and representations. In the hypothesis space, learners search their long-term memory (i.e., prior knowledge) and/or metacognitive knowledge and experiences to formulate hypotheses, while the experimental space is guided by planning and investigating current hypotheses. The learner must search through their hypothesis testing results to inform the generation of alternative hypotheses in this space, or make conclusions about the state of the phenomenon in which they are studying. The model also suggests that in order to initiate successful scientific thinking, learners must engage in three key components. First, learners must gather information before they can formulate a hypothesis. This requires searching through memory and the environment in which the learner is learning. Second, learners formulate a hypothesis for testing based on their understanding. Last, once a hypothesis is formed, the learner tests their current hypothesis and evaluates the results. Based on the results, they can either come to a decision about the phenomena they are studying (i.e., reject the hypothesis), or use the results to formulate alternative testable hypotheses (i.e., accept the hypothesis; Klahr & Dunbar, 1988).

As such, the amount of time spent gathering information is contingent on the learners' prior knowledge about the content, where the more time that learners spend gathering information, or searching through memory, the less time they have to allocate to other components of scientific thinking such as formulating and testing hypotheses. Previous research using SDDS to study scientific thinking found that a variety of individual characteristics contribute to scientific thinking and thus impact subsequent learning and performance outcomes (Lazonder, & Harmsen, 2016), such as prior knowledge about the domain (Dunbar & Klahr, 2012) and types of scaffolds used to guide learners (Lazonder, Hagemans, & De Jong, 2010; Mulder, Lazonder, & de Jong, 2011; Taub, Sawyer, Smith, Rowe, Azevedo, & Lester, 2020). However, major gaps persist as this work has not been extended to problem solving with GBLEs.

Game-based Learning Environments

GBLEs designed to foster scientific-thinking and problem-solving skills use a range of game features that influence learners' personally, such as enhancing their emotional engagement, perceived agency, and interest in learning (Plass, Mayer, & Homer, 2020; Plass, Homer, & Kinzer, 2015). A number of meta-analyses provide evidence that the design of GBLEs amplifies knowledge and skill acquisition (Clark, Tanner-Smith, & Killingsworth, 2016; Mayer, 2014; Wouters, Van Nimwegen, Van Oostendorp, & Van Der Spek, 2013) by (1) providing an incentive structure such as granting rewards for completing tasks, (2) visual and auditory aesthetics used to engage learners, and (3) a narrative with clearly-defined rules that limit agency by requiring learners to finish tasks to successfully complete the objective of the GBLE (e.g., talking to a non-player character (NPC) to gather clues to move forward with other tasks). However, these meta-analyses have not examined the role or impact of scientific thinking with GBLEs, nor have they provided actionable multichannel data collected during game-based learning to prescribe the most important features that can be used to individualize scientific thinking with GBLEs. Agency, or learners' ability to control their actions during learning (Bandura, 2001), is a critical design within GBLEs and has been shown to impact scientific thinking (Taub et al., 2020). For instance, research has found that while affording total agency--

i.e., no restrictions to learners' actions, results in increased engagement, motivation (Mayer, 2014; Shute, Rahimi, & Lu, 2019), and performance outcomes with GBLEs (Loderer, Pekrun, & Lester, 2020' Plass et al., 2020), there are mixed findings regarding agency and its impact on scientific thinking. GBLEs that provide total agency require learners to effectively engage in scientific thinking all on their own, where learners may become distracted by other activities or seductive but extraneous features that are unrelated to scientific thinking, such as exploring the environment or trying to game the system. Because of this, varying degrees of agency (e.g., partial versus total agency) serve as an implicit scaffolding technique for developing scientific-thinking skills with GBLEs.

Agency that Supports Scientific Reasoning as Inconspicuous Scaffolding

Implicit scaffolds, such as partial agency, support knowledge and skill acquisition by quietly changing the way learners interact with information, tools, and game elements built into GBLEs. For example, Crystal Island is a narrative-centered GBLE designed to teach learners about microbiology using varying degrees of agency (Taub et al., 2020). The total agency condition affords learners absolute control over their actions during game-based learning, while the partial agency condition requires learners to engage in a predefined and fixed sequence of actions (e.g., learners must read all books within each building before testing hypotheses) with the assumption that the required in-game actions are beneficial for effective scientific thinking (e.g., gathering information prior to generating hypotheses), learning, and performance outcomes. Learners in the partial agency condition, however, are still given some control by allowing them to use tools at any point during game-based learning, such as recording clues and formulating hypotheses using the scientific worksheet. The no agency condition does not afford learners any control over their actions as they watched a scientific-thinking expert in 3rd person complete the game. A study investigating the effect of agency on scientific thinking and performance using Crystal Island found that learners in the partial agency condition demonstrated more scientific thinking in-game actions and higher performance outcomes relative to learners in the total or no agency conditions (Taub et al., 2020). However, because scientific thinking was solely measured using in-game actions, challenges exist as the authors cannot ensure learners were actually thinking scientifically versus completing actions to move forward in the game. Other sources of data are critical to inform if, when, and for how long learners engage in scientific thinking during game-based learning.

Quantifying Scientific Thinking with GBLEs

A study by Shute, Wang, Greiff, Zhao, and Moore (2016) captured scientific thinking using the number of in-game actions learners initiated during problem solving and found that in-game actions were associated with successful problem solving and a strong predictor of performance compared to other methodological approaches such as problem-solving ability assessments (e.g., Raven’s Progressive Matrices; Raven, 2000, 1941). Yet, major methodological and analytical limitations still exist in this work. Leveraging in-game actions that suggest scientific thinking does not capture information on whether the learner was actually thinking scientifically about solving the problems. For example, scientific thinking was operationally defined as learners’ in-game actions that suggested analyzing resources. But can in-game actions capture when learners analyzed resources without information on where they were visually attending? Perhaps the learner was aimlessly playing the game and interacted with a resource by chance. A similar study by Taub and colleagues (2018) examined the sequence of in-game actions to quantify successful scientific thinking, where in-game actions were defined based on the relevance of learners’ actions toward solving the problem during game-based learning (e.g., testing food items and pathogens relevant to the problem solution). Their results revealed that learners were more effective in their scientific thinking if their in-game actions suggested they tested fewer hypotheses that were more relevant to the problem solution compared to learners who tested more hypotheses that were less relevant to the problem solution (Taub, Azevedo, Bradbury, Millar, & Lester, 2018). However, the authors used in-game actions alone to define scientific thinking, making the assumption that all learners, regardless of prior knowledge and other individual characteristics, are scientifically thinking about solving the problem during game-based learning. Still, can one be certain that when learners tested fewer hypotheses more relevant to the problem solution, they did not accidentally select a relevant hypothesis when using in-game actions data? It is critical to investigate whether other sources of data might supplement information captured via in-game actions suggesting scientific reasoning. Without capturing more information on what the learner is engaging with during game-based learning and accounting for individual characteristics such as prior knowledge, major gaps persist in research studying scientific thinking during problem solving with GBLEs.

Eye-tracking Methodology

Including other data channels to supplement in-game actions during game-based learning may better inform when learners are actively engaging in scientific thinking. Decades of literature suggest eye-gaze data could be a promising direction in revealing implicit cognitive processing like scientific thinking (Scheiter & Eitel, 2017). Adopting a cognitive psychology perspective, where learners visually attend is based on the size of the retina. The retina determines if and how much visual information is available for processing and discloses the learners' foci of attention and thus what information is being processed (van Zoest, Van der Stigchel, & Donk, 2017). A number of studies provide empirical evidence supporting this perspective, such that where participants look is indicative of their reasoning behaviors such as scientific thinking (Miller Singley & Bunge, 2018; Plummer, DeWolf, Bassok, Gordon, & Holyoak, 2017). For instance, Vendetti, Starr, Johnson, Modavi, and Bunge (2017) showed that participants' saccades and fixations data were indicative of reasoning. Specifically, they found that participants’ eye movements revealed optimal strategies for solving visual analogy problems when they were faced with distracting stimuli. Their findings showed that when participants attended to relevant relationships in stimuli, they were more likely to solve the analogy problems compared to participants who attended to distracting stimuli (Vendetti et al., 2017). Several studies have also found that where learners visually attend is related to their intention such that the eye-gaze data predicted future in-game actions (Hillaire, Lécuyer, Breton, & Corte, 2009; Huang, Andrist, Sauppé, & Mutlu, 2015; Park, Lee, Lee, Chang, & Kwak, 2016; Rajendran, Kumar, Carter, Levin, & Biswas, 2018; Rayner, 2009) and reasoning behaviors (Bondareva, Conati, Feyzi-Behnagh, Harley, Azevedo, & Bouchet, 2013).

A study by Munoz, Yannakakis, Mulvey, Hansen, Gutierrez, and Sanchis (2011) analyzed eye-gaze data using an artificial neural network to assess whether these data predicted future actions. Their model achieved an accuracy rate above 83\% for predicting future actions based on eye-gaze positions on the screen, such that the more often learners fixated on game elements within the game, the more likely they were to initiate in-game actions using those game elements in the future (Munoz et al., 2011). A similar study by Huang and colleagues (2015) examined whether eye-gaze data predicted future actions. Their model predicted actions 1.80 seconds before the learner initiated the action based on where they had previously fixated at an accuracy rate of 76% (Huang et al., 2015). As such, do these findings transfer to game elements in GBLEs which are intentionally designed to foster scientific thinking? Specifically, can eye-gaze data inform whether a learner is engaging in scientific thinking based on relationships between eye-gaze data and in-game actions related to scientific thinking with GBLEs? A study by Singh, Miller, Newn, Sonenberg, Velloso, and Vetere, (2018) investigated whether there were differences between a model with in-game actions versus a model with both eye-gaze data and in-game actions collected during a multiplayer game in predicting future in-game actions. Their results found that the model with both eye-gaze and in-game actions data achieved a better fit with 71\% accuracy for predicting future actions compared to the model with only in-game actions at 41\% accuracy. The aforementioned studies demonstrate how eye-gaze data might inform scientific thinking in-game actions (Singh et al., 2018). In sum, it is essential to use more than one data channel to capture and analyze scientific thinking during game-based learning. Eye-gaze data that supplements in-game actions may inform if, when, and how long learners are thinking scientifically during game-based learning.

Supporting Game-based Learning using Multichannel Data

GBLEs allow researchers to capture rich data about complex learning constructs (Taub et al., 2020). Game-learning analytics (GLA) are techniques developed for capturing, storing, analyzing, and detecting critical information about what, when, and how long learners interact with game elements such as tools and other resources while solving problems during game-based learning. GLA afford opportunities to guide instructional prescriptions and decision making based on what learners are doing to optimize and transform their learning experience based on their multichannel data (Freire, Serrano-Laguna, Manero, Martínez-Ortiz, Moreno-Ger, & Fernández-Manjón, 2016; Lang, Siemens, Wise, & Gasevic, 2017). As such, GLA opens a door for capturing scientific thinking with GBLEs. Since theoretical models and previous studies suggest individual characteristics of learners play a crucial role in their ability to think scientifically such as the degree of agency (Taub et al., 2020) and prior knowledge about the content being studied during game-based learning (Dunbar & Klahr, 2012; Klahr & Dunbar, 1988), it is crucial to capture and analyze GLA to advance the science of learning and guide instructional decision making.

A number of studies have used GLA to gain insight into the role of individual characteristics to analyze its relation to learning and performance outcomes (Giannakos, Sharma, Pappas, Kostakos, & Velloso, 2019) used learners' multichannel data to capture skill acquisition while they played a pac-man game. Skill acquisition was defined by motor-movement adaptation and decision making, where learners were required to control four buttons on a keyboard over the course of 3 sessions that increased in difficulty. Five sources of data were collected: eye-gaze--i.e., pupil diameter, fixation, saccades, and events (e.g., number of fixations, number of saccades, etc.), keystrokes, EEG, facial expressions of emotions, and wristband data--i.e., heart rate, blood pressure, temperature, and electrodermal activity. Their findings showed that keystrokes demonstrated the lowest amount of model accuracy in predicting skill acquisition at an error rate of 39\%. But, by fusing multichannel data, they achieved a 6\% error rate in predicting skill acquisition (Giannakos et al., 2019). Similarly, Alonso-Fernández, Cano, Calvo-Morata, Freire, Martínez-Ortiz, and Fernández-Manjón (2019) captured learners' interaction data during game-based learning tasks to examine its relation to learning outcomes. By building predictive models using data-mining techniques, they found that in-game actions capturing how learners interacted with game elements during game-based learning were 89.7% accurate in their predictive ability, leaving approximately 11% of error. However, similar to the aforementioned studies, they found that by combining both the learners' pre-test scores before game-based learning and their in-game actions during game-based learning, their predictive model demonstrated an accuracy rate of 92.6%, reducing its error rate to less than 8% (Alonso-Fernández et al., 2019). Similarly, Sharma and others (2020) captured and analyzed learners' in-game actions and physiological data--i.e., facial expressions of emotions, eye tracking, EEG, and wristband data to predict learners' effort based on patterns in their multichannel data while they were learning. Their results showed that Hidden Markov Models could detect learners' effort from their multichannel data generated during learning in a way that fostered individual instructional prescriptions in real-time that were informed by patterns in learning behaviors (Sharma et al., 2020).

In sum, these studies highlight how important it is to analyze multichannel data captured over a learning session to gain insight into skill and knowledge acquisition with GBLEs. GLA techniques are useful in tracking, modeling, and understanding what learning processes are initiated and how each contributes to an individual learner's conceptual understanding of the domain, or the degree at which a skill is acquired and mastered. The value of GLA is emphasized based on its ability to reveal behaviors which may be detrimental to a learner's performance that go beyond in-game actions captured via log files, offering opportunities to detect maladaptive behaviors and intervene during learning activities to redirect learners toward optimal learning, skill acquisition, and performance outcomes. For example, if a learner has less prior knowledge about game content, or is showing elevated time spent examining game elements, then it could signal the learner's scientific thinking skills. Further, this information could guide an instructional intervention tailored to the specific needs of the learner based on their individual characteristics, in addition to when, with what, how often and how long they are engaging with game elements related to the overall objective of the learning session. Yet, contrary to these promising directions, a number of studies continue to use in-game actions alone to capture scientific thinking with GBLEs (Shute et al., 2016).

Current Study

In this study, we examined college students' scientific thinking with GBLEs to determine optimal features for personalizing instruction during science learning. Current studies apply a generalized methodological and analytical approach by only accounting for performance and in-game actions logs to define if, when, and how long learners think scientifically about solving problems with GBLEs (Shute et al., 2016; Smith et al., 2019; Taub et al., 2018, 2020), ignoring individual characteristics of learners that have been shown to impact their ability to think scientifically during game-based learning. To address the challenges mentioned above, we investigated whether learners' multichannel data were related to scientific thinking during game-based learning and performance outcomes. Specifically, the objective of our study was to examine whether relationships existed between multichannel data--i.e., eye gaze, in-game actions, degree of agency, and pre-test and post-test scores. Our research questions were grounded in SDDS theory (Klahr & Dunbar, 1988) and previous empirical evidence related to modeling scientific thinking to examine whether the proportion of time fixating on game elements related to scientific reasoning, pre-test scores, and degree of agency predicted the proportion of time interacting with game elements related to scientific reasoning and post-test scores. Our research questions and hypotheses are provided below:

Research Question 1:

To what extent does the time fixating and interacting with scientific reasoning related game elements predict post-test scores after game-based learning, while controlling for prior knowledge and degree of agency? We hypothesize there will be predictive relationships between the time interacting with scientific reasoning related game elements and post-test scores after game-based learning, while controlling for pre-test scores and degree of agency. Our hypothesis is both grounded in Klahr and Dunbar's (1988) SDDS theory and previous empirical evidence that multichannel data generated over a learning session (e.g., Alonso-Fernández et al., 2019), prior knowledge (Dunbar & Klahr, 2012), and the degree of agency (Lazonder et al., 2010; Mulder et al., 2011; Taub et al., 2020) has been related to performance outcomes.

Research Question 2:

To what extent does time fixating on scientific reasoning related game elements predict time interacting with scientific reasoning related game elements during game-based learning, while controlling for pre-test scores and degree of agency? We hypothesize there will be predictive relationships between the time fixating on scientific reasoning related game elements and the time interacting with scientific reasoning related game elements during game-based learning, while controlling for pre-test scores and degree of agency. Our hypothesis is based on empirical evidence suggesting that eye movements are related to learners' cognitive processing and future in-game actions (Hillaire et al., 2009; Huang et al., 2015; Park et al., 2016; Singh et al., 2018).

Research Question 3:

To what extent does the time fixating on non-scientific reasoning related game elements predict the time interacting with non-scientific reasoning related game elements during game-based learning, while controlling for pre-test scores and degree of agency? We hypothesize there will be predictive relationships between the time fixating on non-scientific reasoning related game elements and time interacting with non-scientific reasoning related game elements unrelated to scientific reasoning during game-based learning, while controlling for pre-test scores and degree of agency. Our hypothesis is based on empirical evidence suggesting that eye movements are related to learners' cognitive processing and future in-game actions (Hillaire et al., 2009; Huang et al., 2015; Park et al., 2016; Singh et al., 2018).

Materials and Methods

Participants and Materials

A total of 138 learners were recruited from three large, public universities in North America. For this paper, a subset of 68 participants (*n* = 68, 68% female, age: *M* = 20.01, *SD* = 1.56) were included in our analyses because they met the inclusion criteria: complete eye-tracking, log file, and performance data and were randomly assigned to either the total agency (*n* = 41) or partial agency conditions (*n* = 27). Of the subsample, the majority identified as Caucasian (75%), while the remaining identified as Asian, Black, Hispanic or Latino, and Other. Additionally, the majority of the subsample reported rarely playing video games (37%), while the remaining reported occasionally (24%) or frequently (16%), or very frequently (9%) playing. Most participants reported an average level of video game playing skill (41%), whereas others reported none at all (13%), limited skills (21%), skilled or very skilled (25%). Sixty-three percent of participants reported playing video games 0-2 hours per week (63%), while the remaining reported playing 3-4 hours (16%), 5-10 hours (7%), 10-12 hours (12%), or more than 20 hours (1%) per week. This study was approved by the Institutional Review Board prior to recruitment and informed written consent was obtained prior to data collection.

To measure knowledge about microbiology, a 21-item, 4-option multiple choice assessment was administered before and after participants finished the game, regardless of whether they had solved the mysterious illness plaguing the island. Participants answered between 6 and 18 correct items (*Med* = 12, *M* = 57%, *SD* = 0.13)[2] on the pretest, and between 10 and 19 correct items (*Med* = 15, *M* = 72%, *SD* = 0.12) on the post test. The assessments contained 12 factual (e.g., *"What is the smallest type of living organism?"*) and 9 procedural questions (e.g., *"What is the difference between bacterial and viral reproduction?"*). Several self-report questionnaires were also administered to participants before and after game-based learning to gauge their emotions, motivation, self-efficacy, and cognitive load[3]. Scientific thinking was measured using a combination of log files and eye-gaze behaviors (see Coding and Scoring subsection for details). Game play duration ranged from 69 to 94 minutes (*M* = 83, *SD* = 17.5).

Experimental Design

Crystal Island was designed with three experimental conditions: (1) total agency where participants had the autonomy to make their own decisions during game-based learning without any restrictions, (2) partial agency where participants had limited autonomy to make their own decisions during game-based learning due to restrictions around the sequence of buildings they could go to and tools they could use, and (3) no agency conditions where participants watched in third-person as another player worked toward identifying the mysterious pathogen in 3rd person. Upon beginning the game, participants within the total agency condition could go to any of the

[2] The mean was calculated using a ratio score of total correct items over total items on the pre/post-test assessments. Median was reported for the total number of correct items.

[3] We do not provide information about the self-report questionnaires because they were not included in the analyses. Readers are encouraged to email the corresponding author to request information about the scales administered.

buildings and open whichever tool they deemed important toward identifying the unknown pathogen at any time they chose, while participants in the partial agency condition were required to visit the nurse first and discuss the symptoms the inhabitants were experiencing before being required to follow a "golden path" of steps to solve the mystery. The partial agency condition was designed to optimize scientific-reasoning actions by requiring participants to first gather information related to the unknown pathogen (e.g., symptoms), read specific books and research articles touching on types of pathogens, and then experimentally test their generated hypotheses. The no agency condition was designed as a means to model what scientific-reasoning actions should look like during game-based learning with Crystal Island.

Crystal Island

Crystal Island is a GBLE designed to foster and improve scientific-reasoning skills through an objective-based quest during complex problem solving. The science content in the game aligns with the Standard Course of Study Essential Standards for Eighth-Grade Microbiology (McQuiggan, Goth, Ha, Rowe, & Lester, 2008). The problem-solving activities emphasize the practice of scientific inquiry as called for by the Next Generation Science Standards, and they also align with standards from the Common Core State Standards for English Language Arts on Reading: Informational Text. Participants were tasked with identifying a mysterious pathogen infecting inhabitants in a research camp on an isolated island (see Figure 1), adopting the role of a Center for Disease Control and Prevention agent to identify an unknown pathogen that infected a team of researchers (Rowe, Shores, Mott, & Lester, 2011).

Participants were instructed to provide an accurate diagnosis and treatment solution by testing possible transmission sources (e.g., food items such as eggs and cheese), talking to NPCs on the island about their expertise or symptoms, and collecting information through resources (e.g., books and research articles) scattered around the island in order to complete the game. During the game, participants moved through a variety of buildings, each containing different game elements providing information related to pathology including NPCs, research articles, books, posters, food items, and devices to test hypotheses that foster scientific thinking. Specifically, Crystal Island was built with a (1) dining hall (where participants can collect food items and interact with the cook--i.e., NPC to assess which food has been served recently; see Figure 2) where an orange is illustrated as a food item); (2) laboratory (where participants can test food items picked up throughout the game to assess whether they have been infected with a pathogen; see Figures 1 & 3); (3) infirmary (where participants can interact with a nurse and sick patients who provide information about their symptoms (see Figures 1 & 4); and (4) dormitory (Figure 1) as well as (5) living quarters (where participants can interact with other team members in the research camp; (see Figure 1). Using the information gathered from the various island locations, participants can record clues and hypotheses for later testing using a range of tools provided during learning with Crystal Island.

Tools Designed to Foster Scientific Thinking

Six tools were built into Crystal Island to foster scientific thinking and help participants navigate the game challenges and mystery. First, *books, posters,* and *research articles* (see Figure 1) were scattered around the island and served as some of the main sources of information about microbiology and pathology. The books, articles, and posters were originally written by a former middle-school science teacher and ranged in topic, length, and text difficulty. Each of the sources explained distinct characteristics related to certain types of pathogens and the ways to treat (1) bacteria, (2) viruses, (3) parasites, (4) fungi, and (5) genetic diseases. Pathogens covered in books, research articles and posters included influenza, tapeworm, Anthrax, Salmonella, E.

Coli, Polio, Ebola, and sickle-cell anemia. Participants within the partial agency condition were required to “read” every book, poster and research article available to them. However, it is important to note that the system counted a book, poster, and research article as read if it was opened (for however brief). While log files also indicated how long the book remained open, participants could have not actually chosen to read, but instead looked elsewhere on the screen. Participants within the total agency condition, however, were not required to read any of the books, posters, or research articles if they chose not to. Within books and research articles, participants had an opportunity to test their understanding of the new information they had just read about using a *concept matrix* (see Figure 3).

Concept matrices were presented as a matrix where participants matched pathogen types (e.g., virus) to their associated and distinct characteristics (e.g., reproduce quickly in living host cells). Participants were given a total of three attempts to provide correct information in the mapping exercise. If participants failed to answer correctly within the first 3 trials, the correct answers were displayed to them (see Figure 3). Participants within the partial agency condition were required to successfully complete (or at least attempt 3 times) every concept matrix. This was done with hopes participants would be more likely to read all of the information available within the game and therefore increase their microbiology knowledge and make more informed deductions and hypotheses regarding the mystery. However, given that the correct answer was provided to participants after three failed attempts, participants could be incentivized to game the system for non-relevant (or uninteresting) books and research articles. Because participants within the total agency condition were not required to complete concept matrices if they chose not to, it is possible they were less likely to game the system as there was no reward for doing so or punishment for not completing them. Participants also had access to *scanner* in the laboratory building (see Figure 3), where they could test food items they hypothesized as transmitting the pathogen to the sick researchers.

During the game, participants were also provided with a *diagnosis worksheet* (see Figure 2) that allowed them to record information related to (1) different types of pathogens and their distinct characteristics; (2) symptoms the sick inhabitants were experiencing; and (3) hypotheses about the transmission source of the pathogen (e.g., orange, milk, eggs). For example, participants could keep track of food items they had previously tested to foster their scientific thinking (e.g., if cheese had negative results for bacteria, participants could deduce either the cheese was not the source of the illness or bacteria was not the pathogen). Items did not have to be scanned in order to correctly solve the mystery, however it was available to test hypotheses generated from other clues. Participants within the partial agency condition were required to scan items, while those in the total agency condition could choose if they wanted to. While participants could fill out and edit the diagnosis worksheet at any time during the game, they had to submit an accurate diagnosis, transmission source, and treatment solution to successfully solve the mystery and complete Crystal Island.

Procedure

When participants arrived at the laboratory, a researcher confirmed their identity and ensured no clothing, hair, or glasses (including eye conditions such as astigmatisms) would interfere with the eye-tracker calibration and data collection on eye movements. First, participants completed informed, written consent and then were instrumented with an electro-dermal bracelet and calibrated to the eye tracker and facial expressions of emotions software (see

Figure 5 for experimental setup)[4]. After successful calibration, participants were instructed to complete several questionnaires gauging motivation, emotions, and self-efficacy as well as a 21-item, multiple choice pre-test assessment on microbiology. Next, instrumented participants started learning with Crystal Island while we collected their multichannel data. On average, it took participants 81 minutes (*SD* = 23 minutes) to identify the unknown pathogen and a correct treatment solution. If participants did not solve the mysterious illness within 90 minutes, they were exited from the game. Immediately after, participants were instructed to complete several questionnaires gauging motivation, emotions, and presence in addition to a 21-item, multiple choice post-test assessment on microbiology. Upon completing the post-test session, participants were paid $10/hr (up to $30), debriefed, and thanked for their time and participation.

Apparatus

Eye-gaze behaviors

To record eye-gaze behaviors, an SMI EYERED 250 eye tracker (SensoMotoric Instruments, 2014) was used in this study and detected pupil and fovea location using infrared light. We used a 9-point calibration and configured the eye tracker to capture eye-gaze data at a sampling rate of 30 Hz, capturing relatively small eye movements at an offset of less than 0.05mm. Eye-gaze fixations were processed in iMotions software (iMotions, 2016), which provided granular post-hoc analysis for creating dynamic areas of interest (AOIs) around game

[4] For this study, we only analyzed data collected via the eye-tracker, performance assessments, and keyboard stroke/mouse clicks. As such, we do not provide information about other data channels, but readers are encouraged to email the corresponding author to request more information.

elements related to scientific reasoning (e.g., time spent fixating on complex text was defined as information gathering; Figure 4).

We defined two types of AOIs: (1) around game elements in which the learner was not interacting with (e.g., fixating on a book from a distance), and (2) around the content of the game elements in which the learner was interacting with (e.g., opening a book and fixating on the text). These two types of AOIs distinguished between the eye-gaze and interaction variables, where the AOIs capturing interaction content was used to define when learners were engaging with materials by combining the data with in-game actions.

In-game Actions

In-game actions were recorded and time-stamped in log files when participants used the mouse and/or keyboard for analyses. Specifically, this data channel provided event- and time-based actions over the course of game-based learning. Throughout this paper, we refer to this data channel as ‘interaction elements’ since it signaled when participants interacted with game elements.

Coding and Scoring

Performance measures

Prior knowledge and knowledge acquired during game-based learning was measured by creating a ratio of correct responses over total items on the pre- (*M* = 0.57, *SD* = 0.13) and post-test assessments (*M* = 0.72, *SD* = 0.12).

Fixations and Interactions with Game Elements

Game elements were categorized as either related to scientific reasoning (e.g., books, research articles, etc.) or non-scientific reasoning (e.g., doors, furniture, etc.) grounded within Klahr and Dunbars’ (1998) SDDS theory. Scientific reasoning game elements were further separated into 3 variables: (1) *gathering information*, (2) *generating hypotheses*, and (3) *experimentally testing hypotheses*. To define scientific reasoning based on learners' interacting with game elements, we aligned in-game actions with eye-tracking data in order to measure two behaviors: (1) when a participant fixated on game elements, but did not interact with same game elements, and (2) when a participant fixated and interacted with game elements at the same time (i.e., interacting and fixating on game elements to ensure their engagement with game elements). The time a participant spent interacting[5] with a game element required that the log file not only show that the game element was interacted with (e.g., a book opened, non-player character asked a question, or the diagnosis worksheet edited), but also that the participant was fixating on the content of the game element with which they were interacting with (see Figure 6). For example, if a participant fixated on a book on a table but did not touch it, this was considered a fixation. If they opened the book, but the eye-tracking data did not suggest they fixated on the book’s content (e.g., looked elsewhere, opened and closed the book quickly, etc.), this was not considered interacting with a game element. However, if participants opened the book and fixated on the content, this was counted as the participant interacting with the book. Interactions were defined in this manner to ensure that the participant was engaging with the game element and so that we did not bias our analysis to consider interactions where the participant was not fixating on the content (i.e., accidentally interacting with a game element). Additionally, it helped prevent making the assumption that log files were indicative of scientific reasoning--that

[5] Throughout this paper, we use interaction to signal that participants were not only interacting with game elements via in-game actions, but also fixating on game elements during their interaction with the game element via eye gaze.

is, we do not have to assume that all in-game actions captured within the log files are meaningful.

As such, the eye-tracking data helped augment and contextualize the log files to make a more meaningful interpretation and highlights the novelty of our approach compared to methods of only using log files. This is especially important for the partial condition participants who were required to complete certain actions before moving on. While the log file might alone indicate participants completed all actions in the partial agency condition, eye-tracking might illustrate other behaviors such as gaming the system. Additionally, for participants within the total agency condition, we would be able to determine if some actions were more exploratory search actions versus scientific reasoning actions. Further, fixations were defined using gaze points within 1-degree visual angle exceeding at least 250ms (Salvucci & Goldberg, 2000) on AOIs described in 2.5.1. Next, the fixations were aggregated across different element AOIs. Additionally, the time a participant spent fixating or interacting with game elements were aggregated across the categories described above (i.e., scientific reasoning elements [further grouped as either gathering information, generating hypothesis, or experimentally testing hypothesis] and non-scientific reasoning elements). We then used these categories to find the proportion of time either fixating or interacting with elements compared to total time spent trying to solve the mysterious illness with Crystal Island. This was done to help control for the time differences produced by condition requirements of the partial agency group. It is important to note that the AOIs of game elements used to define fixations were different than the AOIs used to define interactions (see Tables 1 & 2) which outline all of the various AOIs that were used to distinguish and categorize interactions and fixations with elements for our analysis). This is due to new visuals being shown when items were clicked on. For example, clicking on a book opened it up to content that was previously not visible. As such, fixating on a book was defined as a fixation on a closed book, while an interaction with a book was when a participant opened a book and fixated on the content in the book.

Statistical Analyses

We cleaned and processed our data in R [Version 3.6.2] (R Core Team, 2013) using ‘read_xl’ (Wickham & Bryan, 2017), ‘dyplr’ (Wickham, Francois, Henry, & Muller, 2018), and ‘reshape2’ (Wickham, 2007) packages for the data wrangling, manipulation, and melting features. Utilizing the ‘bestNormalize’ package (Peterson & Cavanaugh, 2019) and ‘plot’ function from the ‘base’ package (R Core Team, 2013), non-normally distributed variables were transformed into a normal distribution. Specifically, we used log, square root, and ordered quartile transformations, and in some cases, standardization. All variables were normalized or standardized with the exception of pre- and post-test ratio scores as these were normally distributed. Next, twelve participants were eliminated as they demonstrated significant outlying observations via Grubb's test (Grubbs, 1969).

Upon building the models, we used the ‘stepAIC’ function from the ‘MASS’ package to conduct stepwise model selection using Akaike information criterion (AIC) for our first research question (Venables & Ripley, 2002). The ‘summary’ function from the ‘base’ package was used to access object and model statistics, while the ‘ggplot2’ package was used to visualize models and corresponding relationships among variables (Wickham, 2016). To visualize interaction terms, packages ‘tidyverse’ (Wickham, Averick, Bryan, Chang, McGowan, François, ... & Yutani, 2019), ‘sjPlot’ (Lüdecke, 2018), and ‘sjmisc’ (Lüdecke, 2018) were used. To probe the relation between interactions, we used the package, ‘interactions’, by calling the ‘probe_interaction’ function (Long, 2019).

Results

Preliminary Analyses

To examine the potential effect of the experimental manipulation on the results (i.e., level of agency), we conducted independent two-sample *t*-tests on all variables to assess if there were differences between conditions. Analyses revealed no significant differences in both pre- and post-test scores between conditions (*ps* > 0.05) and so we did not include condition to maintain a parsimonious model for the first research question. For the second and third research questions, we found significant differences in time interacting with elements related to gathering information, $t(60) = -3.09$, $p = 0.003$, and generating hypotheses, $t(50) = 2.29$, $p = 0.026$, between experimental conditions and so we included condition in each equation (*ps* < 0.05; see Figure 7 and Table 3). For research questions 2-3, we did not use the AIC method to select the best fit model because the predictor variables were empirically (i.e., literature suggests prior knowledge and agency impacts scientific thinking), theoretically (i.e., the SDDS model suggests prior knowledge impacts scientific thinking), and statistically justified (e.g., significant differences in eye-gaze and interaction data between conditions; see Current Study subsection for more details).

To what extent does time fixating and interacting with scientific reasoning related game elements predict post-test scores after game-based learning, while controlling for pre-test scores?

To examine whether relationships existed between proportion of time fixating and interacting with game elements related to scientific reasoning during game-based learning and post-test scores, while controlling for pre-test scores, we used AIC stepwise selection. Separate models for fixations and interactions with game elements were built, because including these data in the same equation to model performance created multicollinearity issues (VIF > 10). The first model was built using eye-gaze data to assess its relation to post-test scores while controlling for pre-test scores. A baseline model was determined using stepwise selection via AIC (Akaike, 1974). The AIC method revealed that the best model was a simple linear regression equation, where pre-test scores were included as the only predictor variable (see Table 4). The fitted model estimated that average post-test scores increased by 0.43 points for each correct item on the pre-test assessment, where pre-test scores explained approximately 22% of the variance in post-test scores. As such, time spent fixating on game elements related to scientific and non-scientific reasoning during game-based learning were unrelated to post-test scores (*ps* > 0.05).

The second model was built using interaction with elements data to assess its relation to post-test scores, while controlling for pre-test scores. The AIC method indicated that the best model fit was a multiple linear regression equation, where pre-test scores and proportion of time interacting with game elements related to gathering information and generating hypotheses were included as predictor variables (see Table 5). The fitted model estimated that average post-test scores increased by 0.53 points for each correct answer on the pretest and increased by 0.24 points for each second increase based on a three-way interaction between proportion of time interacting with game elements related to gathering information and generating hypotheses, as well as pre-test scores correct item on the pre-test assessment. These predictors explained approximately 30% of the variance in post-test scores (see Figure 8). To probe the complexity of the three-way interaction relationship, we used a robust statistical technique known as the Johnson-Neyman interval (Johnson & Neyman, 1936). The Johnson-Neyman technique relies on the range of values of the moderators (i.e., time spent interacting with elements related to gathering information and generating hypotheses), where the slope of the predictor (i.e., pre-test scores) is significant compared to non-significant at an alpha level of 0.05. In our analysis, we considered pre-test scores as our focal predictor and proportion of time interacting with game elements related to gathering information and generating hypotheses as moderators hypothesized to affect the relationship between pre- and post-test scores (see Figure 8) for visualizations of pre- and post-test relationships as the moderators change from -1 standard deviation, to the mean, and +1 standard deviation.

Refer to Table 5 for a breakdown of how the relationships between post- and pre-test scores change based on the proportion of time interacting with game elements related to gathering information and generating hypotheses. In sum, the analysis suggested a significant positive relationship between pre- and post-test scores was present, except when participants spent less than 1 standard deviation (*SD* = 0.025) away from the mean (*M* = 0.083) generating hypotheses, but more than 1 standard deviation (*SD* = 0.061) away from the mean (*M* = 0.102) when gathering information. In other words, when participants spent a larger proportion of time during game-based learning gathering information compared to the relative average and spent a smaller proportion of time generating hypotheses compared to the relative average, the relationship between pre- and post-test scores was not significant ($p > 0.05$).

To what extent does time fixating on scientific reasoning related game elements predict time interacting with scientific reasoning related game elements during game-based learning, while controlling for pre-test scores and degree of agency?

To examine whether relationships existed between proportion of time fixating on game elements related to scientific reasoning and the proportion of time interacting with game elements related to scientific reasoning during game-based learning, while controlling for pre-test scores and experimental conditions, we built multiple linear regression equations.

Specifically, we built separate models for each scientific-reasoning process: information gathering, hypothesis generation, and experimental testing, since including these variables in the same equation created major multicollinearity issues (VIF > 10). For each model, fixation variables were included as predictors while interaction with elements were included as targets. First, we built a model to examine relationships between the proportion of time fixating on game elements related to gathering information and interacting with game elements related to gathering information, while controlling for pre-test scores and experimental conditions. A significant multiple linear regression equation was found, $F(7,60) = 3.243$, $p < 0.001$, for the proportion of time interacting with game elements related to gathering information between total and partial agency conditions and a two-way interaction between pre-test scores and proportion of time fixating on game elements related to gathering information (see Table 6 for beta coefficients and adjusted R^2). The fitted model estimated that the average proportion of time interacting with game elements related to gathering information decreased by 0.03 if participants were assigned to the total agency condition relative to the partial agency condition.

We also found a significant two-way interaction, where the relationship between interactions and fixations for elements related to gathering information was affected by pre-test scores, such that the lower pre-test score (e.g., 44% correct items on pre-test) revealed a positive relationship between interactions and fixations for elements related to gathering information. However, the higher pre-test scores revealed a negative relationship between interactions and fixations for elements related to gathering information. Together, the predictors explained approximately 19% of the variance in the proportion of time interacting with game elements related to gathering information during game-based learning.

Our second model was built to examine relationships between the proportion of time fixating on game elements related to gathering information via interactions and fixations for elements related to generating hypotheses, while controlling for pre-test scores and experimental conditions. A significant multiple linear regression equation was found, $F(7,60) = 2.225$, $p = 0.044$, suggesting relationships between interactions and fixations for elements related to generating hypotheses and experimental conditions (see Table 6 for beta coefficients and adjusted R^2). The fitted model estimated that the average proportion of time interacting with game elements related to generating hypotheses increased by 0.96 for each second increase in the proportion of time fixating on game elements related to generating hypotheses (see Figure 9), as well as the average proportion of time spent interacting with elements related to generating hypotheses decreased by 0.02 if participants were assignment to the partial agency condition compared to the total agency condition. Together, the predictors explained approximately 11\% of the variance in the proportion of time interacting with game elements related to generating hypotheses during game-based learning.

Finally, our last model was built to examine relationships between interactions and fixations for elements related to experimental testing, while controlling for pre-test scores and experimental conditions. We did not find a significant multiple linear regression equation, suggesting there were no relationships between interactions and fixations for elements related to generating hypotheses, while controlling for pre-test scores ($p > 0.05$). In sum, our analyses revealed a significant two-way interaction where pre-test scores moderated relationships between interactions and fixations for elements related to gathering information. We also found significant relationships between experimental conditions and interactions and fixations for elements related to generating hypotheses. We did not find relationships between interactions and fixations for elements related to experimental testing ($p > 0.05$).

To what extent does time fixating on non-scientific reasoning related game elements predict time interacting with non-scientific reasoning related game elements during game-based learning, while controlling for pre-test scores and degree of agency?

To examine whether relationships existed between interactions and fixations for elements related to non-scientific reasoning, while controlling for pre-test scores and experimental condition, we built a multiple linear regression equation. For this model, the fixations on elements was included as a predictor while interactions with elements was included as the target. Analyses suggested there were no significant relationships between interactions and fixations for elements related to non-scientific reasoning, while controlling for pre-test scores and experimental condition (*ps* > 0.05).

Discussion

In this study, we examined scientific thinking with GBLEs to determine optimal features for individualizing instruction during science learning. Specifically, the objective of our study was to examine whether learners' multichannel data generated during game-based learning with Crystal Island were related to scientific thinking and performance and could be used to guide individualized instruction with GBLEs.

Research Question 1

Our first research question examined relationships between fixations and interactions for game elements related to scientific reasoning and post-test scores, while controlling for pre-test scores. The predictive models suggested there were no relationships between fixation variables related to scientific reasoning and post-test scores (*p* > 0.05). However, the predictive models did reveal a significant three-way interaction between pre-test scores, interactions with game elements related to gathering information and generating hypotheses during game-based learning, and post-test scores. Specifically, the model suggested interactions with game elements related to gathering information and generating hypotheses moderated relationships between pre- and post-test scores. We found that when holding interactions for game elements related to gathering information and generating hypotheses constant at various increments--i.e., at mean - 1 SD, mean, mean + 1 SD; see Figure 8), the relationships between pre- and post-test scores changed. For example, when learners spent more time, or an average proportion of time generating hypotheses, the proportion of time they spent gathering information positively influenced relationships between pre- and post-test scores. However, when the learner spent more time gathering information and less time generating hypotheses, the relationships between pre- and post-test scores were no longer significant (*p* > 0.05).

This finding suggested that when learners spend a higher proportion of time gathering information and less time generating hypotheses during game-based learning captured by combining both eye-gaze and in-game actions, it was detrimental to performance regardless of how much prior knowledge learners had brought to the learning session. These findings are partially consistent with our hypothesis such that there were relationships between interactions for game elements related to scientific reasoning and post-test scores based on empirical studies (Alonso-Fernández et al., 2019; Giannakos et al., 2019; Lazonder & Harmsen, 2016; Sharma et al., 2020) and SDDS theory (Klahr & Dunbar, 1988); however, since no relationships between fixations on game elements related to scientific reasoning alone and post-test scores were found, this was inconsistent with previous literature (Alonso-Fernández et al., 2019). We would like to emphasize that interaction data were *only* included in our models if eye-gaze data suggested that while learners were interacting with elements via log files, they were also fixating on those same elements (see subsubsection 2.6.2). While we did not find relationships when only including eye-gaze data, a possible explanation could be that when assessing eye-gaze data alone (i.e., without considering interaction data), it is too granular to capture its relation to performance. Instead, other modalities of data generated during game-based learning may need to supplement statistical models in order to detect an effect or relationship, which has been a consistent finding across a number of GLA studies (Giannakos et al., 2019; Sharma et al., 2020). Future research should look toward including data channels such as concurrent verbalizations (Greene, Deekens, Copeland, & Yu, 2018). For instance, proportion of time individual learners fixate and interact with game elements related to different components of scientific thinking that are supplemented by think-alouds could be useful in informing GLA that is individualized to each learner's needs.

Research Question 2

Our second research question examined whether the proportion of time fixating on game elements related to scientific reasoning predicted the proportion of time interacting with game elements related to scientific reasoning, while controlling for pre-test scores and experimental condition. The analyses revealed significant relationships between pre-test scores as well as interactions and fixations with elements related to gathering information. Specifically, we found a two-way interaction where the predictive model revealed that learners' prior knowledge about microbiology moderated relationships between interactions and fixations with game elements related to gathering information, such that the less prior knowledge learners had upon entering the learning session (e.g., less than 50% items correct on the pre-test assessment), then there was a positive relationship between interactions and fixations for elements related to gathering information. However, if the learner had scored higher on the pre-test (e.g., more than 50% items correct on the pre-test assessment), there was a negative relationship between interactions and fixations on games elements related to gathering information.

We also found significant positive relationships between experimental conditions as well as interactions and fixations for game elements related to generating hypotheses. This finding suggests that learners in the partial agency condition spent significantly less time interacting with game elements related to generating hypotheses compared to the total agency condition. The predictive model also suggested that, while controlling for experimental condition, there was a significant positive relationship between interactions and fixations for game elements related to generating hypotheses during game-based learning. Unfortunately, we did not find significant relationships between interactions and fixations for elements related to experimental testing ($p > 0.05$). It is also important to note that we built separate models for each scientific reasoning variable (e.g., gathering information or generating hypotheses) due to multicollinearity issues. However, multicollinearity was not an issue for non-scientific reasoning variables, suggesting that combining eye-gaze and in-game actions data might provide insight into the extent to which a learner is engaging in scientific reasoning. In other words, if eye-gaze and in-game actions data are unrelated for certain game elements (e.g., non-scientific reasoning elements), it might suggest that the learner is not engaging in scientific reasoning.

These findings are partially consistent with our hypothesis. Specifically, significant relationships between pre-test scores as well as interactions and fixations for game elements related to gathering information was consistent with our hypothesis, empirical literature (Hillaire et al., 2009; Huang et al., 2015; Park et al., 2016; Singh et al., 2018), and SDDS framework (Klahr & Dunbar, 1988), where we expected eye-gaze and prior knowledge to predict scientific reasoning in-game actions with GBLEs. Additionally, significant relationships between experimental condition as well as interactions and fixations for elements related to generating hypotheses was consistent with our hypothesis and empirical literature (Hillaire et al., 2009; Huang et al., 2015; Park et al., 2016; Singh et al., 2018; Taub et al., 2020). However, the findings did not reveal significant relationships between interactions and fixations for elements related to experimental testing, which is inconsistent with our hypothesis and empirical evidence (Hillaire et al., 2009; Huang et al., 2015; Park et al., 2016; Singh et al., 2018). A possible explanation of this result could be that, while fixations and interactions for elements were related for gathering information and generating hypotheses game elements, the game elements defined as experimental testing such as the scanner for testing food items did not require the learners to interact with as much as the game elements defined for gathering information and generating hypotheses. For instance, learners may have spent the majority of their time during game-based learning reading about the content and finding clues for formulating their hypotheses, and so upon testing hypotheses, it only required a few seconds or minutes to test those food items, reducing the proportion of time learners could have fixated and interacted with those game elements. As such, it is critical for researchers, educational technologists, and instructional designers to critically examine and build game elements that are not biased toward one component of scientific thinking relative to another in order to capture the nuances of scientific thinking. For example, Crystal Island has a disproportionate amount of game elements that learners can fixate on and interact with that reflect gathering information, such as the learner gathering clues from posters, research articles, books, and non-player characters. Or providing multiple tools that learners can use to track their clues and generate hypotheses during problem solving, yet there are few tools available for testing those hypotheses. Because of this, it is imperative to design GBLEs as both training and research tools and critically evaluate the design of GBLEs to capture dynamic learning behaviors that not only support GLA but are also based on theoretical frameworks and empirical evidence.

Research Question 3

Our third research question examined whether relationships existed between interactions and fixations for game elements related to non-scientific reasoning, while controlling for pre-test scores. Our analyses revealed no significant relationships between interactions and fixations for game elements related to non-scientific reasoning. These findings were inconsistent with our hypotheses and previous empirical literature (Hillaire et al., 2009; Huang et al., 2015; Park et al., 2016; Singh et al., 2018), where we expected there to be relationships between interactions and fixations for game elements related to non-scientific reasoning. A possible explanation of this could be related to the objective of Crystal Island. Learners were required to solve a mysterious illness in order to complete the game and so they had no reason to fixate on or interact with game elements unrelated to scientific reasoning, or completing the objective of the game such as fixating and interacting with a plant or furniture during game-based learning. As such, a disproportionately low volume of data were present for fixations and interactions with elements related to non-scientific reasoning game elements, which may not have provided enough variance to detect a relationship in the predictive models. These findings highlight how critical it is to account for the objective of a learning environment when using GLA to guide instructional decision making and gain insight into complex learning behaviors.

Limitations

Limitations of this research include sampling bias such that the participants who completed the study were all undergraduate students and so the modeling results may not generalize to other learners who are not undergraduate students. Additionally, the sample was exclusively young adults, restricting our modeling interpretations to learners ranging from ages 18 to 25. Additionally, the shortcomings associated with using stepwise regression for model selection include biased parameter estimation, among others (Harrell, 2015). We also had significantly less volume of data for fixations and interactions with non-scientific reasoning related game elements. We also did not analyze information on learners' background such as academic major, race, etc. The findings from this study with Crystal Island may also not generalize to other game-based learning environments.

Implications and Future Directions

Future research should examine the effect of modifying the degree of agency during game play based on multichannel data generated from individual learners that depends on several issues such as prior knowledge, background such as major, race, age, etc., proportion of time fixating and interacting with game elements related to scientific thinking, efficaciousness in using scientific-thinking skills, developing competency etc. to guide individualized game-learning analytics. For instance, developing a temporal threshold that monitors how often and how long learners are fixating and interacting with game elements related to scientific reasoning to identify whether there is a need for intervention could be a novel approach for addressing questions related to how and what to adapt during game-based learning. If a learner demonstrates high prior knowledge about the content, then affording the learner more agency seems appropriate, but what if the learner is unable to apply their knowledge during game-based learning to think scientifically? Based on the learners' multichannel data gathered during game-based learning, GLA should be used in real-time to inform whether more or less agency is needed to scaffold scientific thinking and optimize individual learners' performance. Studies should also examine other data channels (e.g., concurrent verbalizations, physiology, facial expressions of emotions) to assess whether these sources of information could supplement eye-gaze, in-game actions, and pre-test scores in their relation to scientific thinking and performance with GBLEs (Plass et al., 2020).

Implications of our findings lay a foundation for using multichannel data to define and capture scientific thinking during game-based learning that is crucial for supporting and optimizing individual scientific thinking, learning, and performance. Specifically, our study provides suggestions for building GBLEs that leverage agency as scaffolding techniques intended to, not only foster scientific-thinking skills, but capture, adapt, and inform instructional decision making based on the needs of individuals' scientific thinking during game-based learning. Further, to implement GLA in the classroom to inform instructional decision making, it will require more support and technological resources than are currently available to most educators, such as a teacher dashboard, to illustrate the various data channels and likely aid in making sense of the GLA and determining the appropriate intervention for individual learners.

Conclusions

As our world and communities are faced with persisting, and sometimes novel, socioeconomic, environmental, and health-related issues (e.g., prolonged interruptions in classroom instruction due to the COVID-19 pandemic), it is essential to equip learners of future generations with an interdisciplinary set of skills which contribute to higher-order thinking (e.g., scientific thinking; National Research Council, 2018, 2012). The objective of our study was to examine relationships between scientific thinking and performance during game-based learning using learners' multichannel data--i.e., eye gaze, in-game actions, level of agency, and pre-test scores and post-test scores. Results showed significant predictive relationships between eye-gaze, pre-test scores, and interaction data related to scientific reasoning, suggesting that eye-gaze, prior knowledge, and agency play a crucial role in scientific thinking and performance with GBLEs. Overall, our findings highlight how critical it is to capture multichannel data, specifically a combination of in-game actions, eye gaze, degree of agency, and pre-test scores, when investigating scientific thinking with GBLEs. Using multiple data channels to define scientific thinking has the potential to support individualized instructional decision making that is guided by a combination of the learners' level of prior knowledge about the topic, as well as how long and what learners are fixating on and interacting with in terms of their relation to scientific reasoning.

For instance, since the models revealed that prior knowledge moderates relationships between fixations and interactions with game elements related to gathering information, then a learner's pre-test score should inform the researcher, instructor, or system about if and when they should intervene based on whether the individual learner is demonstrating GLA which may be detrimental to learning and performance. Perhaps, the learner has low prior knowledge about microbiology upon entering the learning session, and so this would inform the instructional decision being made. If the learner is not fixating on game elements related to gathering information, which is critical for their scientific thinking (i.e., hypothesis generation) when there is little prior knowledge about microbiology, then the researcher, instructor, or system would be aware that these data suggest the learner may not be engaging in scientific-thinking actions. Yet, if the learner has a relatively high prior knowledge about the content, then the researcher, instructor, or system need not worry about whether the learner is fixating on game elements related to gathering information as it is not critical for their scientific-thinking actions. To implement GLA in the classroom to assist in instructional decision making, educators will need more support and technological resources to help effectively monitor and understand learners' data. Recent studies have begun investigating dashboards as tools in the classroom to assist in data sense-making for guiding instruction across individual learners (Perez-Colado, Perez-Colado, Freire-Moran, Martínez-Ortiz, & Fernández-Manjón, 2017).

Additionally, our findings draw us to evaluate the role of agency designed into Crystal Island when referring to the three-way interaction. In other words, was one experimental condition designed to require learners to engage in one component of scientific reasoning more than another and why (i.e., why require learners in partial agency to interact with game elements related to gathering information compared to generating hypotheses)? Since our preliminary analyses suggested that learners assigned to the partial agency condition spent a significantly higher proportion of time interacting with game elements related to gathering information compared to the total agency condition, while the total agency condition spent a significantly higher proportion of time interacting with game elements related to generating hypotheses compared to the partial agency condition, it suggests that the degree of agency afforded to learners serves as a critical scaffolding technique that impacts scientific thinking and performance (Taub et al., 2020). As such, it is our responsibility as researchers and scientific thinkers to critically evaluate what each condition is requiring learners to do and why? Perhaps, designing levels of agency in GBLEs should be based on the prior knowledge the learner brings to the session to optimize their scientific thinking, learning, and performance. In the progressing realm of interconnected human and technology ecosystems, implications of our findings could enhance our understanding of scientific thinking during game-based learning to effectively address future GBLEs that are designed for individualized and adaptive instruction to optimize scientific thinking, learning, and performance.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

All authors contributed to the conception of the work and revised the final manuscript. RA designed and conducted the study. EC, DD, and MW wrote the first draft of the manuscript. DD, MW, and RA provided several rounds of edits on the manuscript.

Funding

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

Acknowledgments

The authors would like to thank team members of the SMART Lab and Hans Doderlein at the University of Central Florida as well as the Intellimedia group at NC State for their assistance and contributions.

Data Availability

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

CHAPTER FOUR: GAME-BASED LEARNING ANALYTICS FOR SUPPORTING ADOLESCENTS' REFLECTION

This chapter, "game-based learning analytics for supporting adolescents' reflection" was originally published in an open-access peer-reviewed journal called the *Journal of Learning Analytics*, vol. 8, issue 2, pages 51-71. The article was led by first author, Elizabeth B. Cloude, and co-authors Dan Carpenter, Daryn A. Dever, Dr. Roger Azevedo, and Dr. James Lester in 2021.

Abstract

Reflection is critical for adolescents' problem solving and learning with game-based learning environments. Yet, challenges exist in literature as most studies often lack a theoretical perspective and clear operational definition to inform how and when reflection should be scaffolded during game-based learning. In this paper, we address these issues by studying the quantity and quality of 120 adolescents' written reflections and its relation to their learning and problem solving with Crystal Island, a game-based learning environment. Specifically, we (1) define reflection and how it relates to skill and knowledge acquisition; (2) review studies examining reflection and its relation to problem solving and learning with emerging technologies; and (3) provide direction for building reflection prompts into game-based learning environments that are aligned with the learning goals built into the learning session (e.g., learn about microbiology versus successfully solve a problem) to maximize adolescents' reflection, learning, and performance. Overall, our findings emphasize how important it is to not only examine the quantity of reflection, but also the depth of written reflection as it relates to specific learning goals. We discuss the implications of using game-learning analytics to guide instructional decision making in the classroom.

Introduction

According to American philosopher John Dewey, meaningful experiences are not possible without some element of reflection (Dewey, 1923; 1933). Dewey (1933) suggests reflection is the process of actively and carefully appraising situations or experiences that requires learners to step back, observe, and meaningfully contemplate how they solve problems and whether a particular set of problem-solving strategies is effective for achieving their goals (Dewey, 1993; Tarricone, 2011). Thus, reflection is essential for effective problem solving because it promotes new knowledge and higher order thinking skills (Dewey, 1933), which has significant bearing on learning and performance (Alzaid & Hsiao, 2018; Izu, & Alexander, 2018; Luo & Baaki, 2019; Patel, Baker, & Scherer, 2019; Tarricone, 2011). Yet, studies find most learners in the United States do not have skills for successful problem solving (Azevedo, Taub, & Mudrick, 2018; Azevedo, Feyzi-Behnagh, Duffy, Harley, & Trevors, 2012; NRC, 2015, 2012; NASEM, 2018; OECD, 2016), begging the question as to whether their insufficiency stems from a lack of reflecting. Since reflection is not taught in U.S. curricula (NASEM, 2018), it is plausible learners are not reflecting to cultivate skills which contribute to solving real-world problems (Moshman, 2013; Wang, Chen, Lin, & Hong, 2017).

To deal with the educational challenges adolescents face (i.e., gaining the knowledge and skills needed to meet the demands of the 21st century), researchers have built game-based learning environments (GBLEs), such as Crystal Island, that incorporate features (e.g., narrative, pedagogical agents) designed to scaffold skill acquisition while simultaneously stimulating motivation and engagement (Taub, Sawyer, Smith, Rowe, Azevedo, & Lester, 2020). Clark and others (2016) found that, out of 57 studies, there were consistently higher learning outcomes for participants learning with GBLEs compared to those who learned in conventional settings without GBLEs ($\bar{g}$ = 0.33, 95% CI [0.19, 0.48], tau^2 = 0.28). Similar results were found by Ak and Kutlu (2017) compared differences in learning gains between traditional settings, 2D GBLEs, and 3D GBLEs and found no differences in learning gains across the three learning environments (*ps* > 0.05). A possible explanation for the mixed findings could be that, while GBLE features are designed to foster higher-order thinking, most game features fail to scaffold reflection, which is important for acquiring knowledge and cultivating effective problem-solving skills in dynamic environments (Tarricone, 2011). This could be in part due to reflection often being vaguely defined and lacking a theoretical perspective, a dearth of studies that use multimodal learning analytics (Mangaroska, Sharma, Gaševic, & Giannakos; Geden, Emerson, Carpenter, Rowe, Azevedo, & Lester, 2021), and the fact that game features (e.g., prompts) are not intentionally designed to scaffold reflection (Taub, Azevedo, Bradbury, & Mudrick, 2020).

In this paper, gaps were addressed using a theoretical perspective and multimodal data on reflection to examine whether the quantity and quality of reflections were related to two learning goals presented within a GBLE, Crystal Island. Specifically, we (1) describe the challenges associated with how reflection has been previously defined and provide a clear definition that is supported by theoretical and empirical work; (2) highlight previous studies that have captured and analyzed reflection with emerging technologies; and finally (3) describe a model of reflection (McAlpine, Weston, Beauchamp, Wiseman, & Beauchamp, 1999) that emphasizes reflection as a function of achieving learning goals, offering a unique perspective to studying reflection with game-based learning environments which provide implications to augment instructional decision-making in the classroom to scaffold adolescents' learning and problem solving.

What is Reflection?

Significant challenges exist in literature as reflection has often been defined interchangeably with introspection and metacognition (Brown, 1982; Flavell, 1979; Pintrich, 2002; Rosenthal, 2000). While reflection relies on introspection and serves as the foundation for metacognition, these terms are distinctive (Tarricone, 2011). Specifically, reflection is the process of purposeful contemplation or focused thinking, whereas introspection is the process of looking within such that one draws awareness to their thoughts, feelings, and reactions to stimuli (Tarricone, 2011). On the other hand, metacognition is the act of monitoring and controlling cognition (Flavell, 1979), while reflection is the act of observing, inspecting, and contemplating (1) a belief or supposed form of knowledge, (2) the evidence that supports said belief or knowledge, and (3) the conclusions with which one draws from observing and inspecting their belief or knowledge.

Moreover, literature refers to reflection as the building block of higher-order thinking (e.g., metacognitive monitoring; effective problem solving; Tarricone, 2011). In other words, self-knowledge that results from introspecting and reflecting fuels metacognitive knowledge because it encompasses the beliefs learners hold about themselves and how they exist in the world (Pintrich, 2002). Traditional theoretical perspectives offer support explaining self-knowledge as a vital ingredient in developing metacognitive knowledge (e.g., model of metamemory by Flavell (1979), leading to a number of empirical studies supporting that metacognitive knowledge is essential for metacognitive processes such as feelings of knowing, and that reflection contributes to enhanced metacognition and learning outcomes (Bannert &

Reimann, 2012). Further, the importance of developing intentional and conscious reflection is critical for developing higher order thinking skills and plays a role in learning (Azevedo, Taub, & Mudrick, 2018; Alzaid & Hsiao, 2018; Dewey, 1923; Izu, & Alexander, 2018; Luo & Baaki, 2019; Patel, Baker, & Scherer, 2019; Tarricone, 2011). Next, we discuss studies examining the role of reflection on skill/knowledge acquisition with emerging technologies.

Reflection with Emerging Technologies

Most studies capture reflection using direct (i.e., the learner is prompted to reflect on specific aspects of their learning; e.g., "*what is the most important information you learned about X?"*) or open prompts (i.e., the learner is simply prompted to reflect in general; e.g., *"how is your learning overall going so far?"*; ter Vrugte, de Jong, Wouters, Vandercruysse, Elen, & van Oostendorp, 2015) elicited using written statements or pedagogical agents with emerging technologies (Carpenter, Geden, Rowe, Azevedo, & Lester, 2020; Geden et al., 2021; Taub et al., 2020). For instance, a study by Wu and Looi (2012) prompted adolescents (ages 13-15 years) to reflect using pedagogical agents with Betty's Brain, an agent-enabled, learning-by-teaching environment (Biswas, Segedy, & Bunchongchit, 2016). Participants were required to study concepts and diagrams related to elementary economics and then teach the agents based on their understanding of the learning materials. Participants were assigned to one of three groups: generic prompt, specific prompt, and a no-prompt group. The generic prompts were designed to scaffold participants to examine their perspectives, beliefs, and experiences by reflecting on metacognitive strategies and beliefs used (e.g., *"What do you think about teaching and who is it for?"*; Wu & Looi, 2012, p. 342), while the specific prompts were designed to scaffold participants to achieve the learning objectives by reflecting on task- and domain-specific skills (e.g., *"Can you explain the concepts you just taught me?"*, Wu & Looi, 2012, p. 342).

Participants were prompted to reflect when they indicated something incorrect about a concept during teaching to the agent, where the agent was triggered to prompt the learner to reflect. Specifically, they examined relations between the quality of written reflections (i.e., did the written statement illustrate reactivity, contemplation, or elaboration?), immediate learning, and transfer knowledge using knowledge assessments across the three conditions.

Results showed participants assigned to either generic- or specific-prompt groups had higher immediate learning outcomes compared to the no-prompt group; however, there were no differences in immediate learning outcomes between the generic-prompt or specific-prompt groups. Their findings also suggested adolescents in the generic-prompt group performed better on the transfer test relative to the specific-prompt and no-prompt groups (Wu & Looi, 2012). Additionally, their results showed that participants in the specific-prompt group demonstrated more reactivity in their written statements, while participants in the generic-prompt group demonstrated more contemplation in their written statements. However, significant challenges exist in this study, particularly in the timing of the reflection prompts via agents, which were initiated only when participants demonstrated incorrect patterns in relation to the learning material. For instance, should learners *only* be prompted to reflect on their knowledge when they illustrate misconceptions? At which points should adolescents be prompted to reflect while learning and teaching complex materials, and how does this prompting relate to achieving learning goals presented within the environment?

Many different GBLEs have been created to foster specific skill sets (e.g., problem solving, scientific reasoning) or conceptual understanding for various domains (e.g., computer science or microbiology) using clearly structured learning goals built into the environment. Yet, few studies design game features to scaffold reflection using a theoretical perspective (Taub et al., 2020), presenting significant challenges in our understanding of reflection during game-based learning and its role on learning and problem solving as it relates to achieving learning goals. Moreno and Mayer (2005) conducted a study where participants designed plants capable of surviving in different climates with a GBLE. Some participants were assigned to a condition where they received corrective or explanatory feedback on problem solving from a pedagogical agent, while others were assigned to a condition in which they were asked to reflect by explaining their problem-solving solutions (e.g., *"Why do you think that particular type of root/stem/leaf will survive in this environment?"*; Moreno & Mayer, 2005, p. 123).

The results showed participants who were asked to reflect on their solutions demonstrated better retention and transfer compared to the corrective or explanatory feedback condition. However, this effect was only present if participants reflected on a solution that was correct. While these findings demonstrate the promise of pedagogical agents scaffolding reflection and its positive effect on learning, we highlight the study did not apply a theoretical model explaining or defining reflection, drawing us to question whether the prompts were eliciting reflection. This issue is especially emphasized when there is no analysis of the quality of the reflection responses as it relates to achieving learning goals. Additional gaps exist as there was no theoretical underpinning determining *when* reflection should be elicited during learning during game-based learning.

Similarly, Fiorella and Mayer (2012) used paper-based worksheets to scaffold reflection and examined whether reflection prompts impacted solving electric-circuit problems with GBLEs. The prompts were designed to draw learners' attention to specific information that was related to the problem solution (e.g., *"If you take away a resistor in serial, the flow rate increases."*; Fiorella & Mayer, 2012, p. 177). They found participants who were prompted to reflect performed higher on the transfer test relative to those who were not (Fiorella & Mayer, 2012). However, we would like to raise similar issues emphasized in Moreno and Mayer (2005). If prompts were designed to direct learners' attention to specific information for finding the solution, learners could be metacognitively monitoring instead of reflecting. More issues exist because drawing learners' attention to specific information does not mean the learner is practicing reflection. Without a clear definition of reflection, it is difficult to make conclusions that might inform relations between reflection and learning with GBLEs, especially when reflection responses are not evaluated based on the quality of the responses as they relate to learning goals. A study by Johnson and Mayer (2010) comparing the effects of two different types of prompts on performance: open and directed, against a no-prompt condition. Learners in the open-prompt condition were required to generate self-explanations after every action during game-based learning, while learners in the directed-prompt condition selected from a list of reasons after every action. Results showed learners in the directed-prompt condition performed significantly better on the transfer assessment compared to the no-prompt condition (*d* = 1.20). They also found learners in the open-prompt condition did not significantly differ from the no-prompt condition (Johnson & Mayer, 2010). It is surprising no effect of open-ended reflection on learning was found since previous studies have found that open-ended written reflection responses reveal reflection depth that relates to higher learning outcomes (Wu & Looi, 2012; Carpenter et al., 2020; Ullmann, 2019). Additional challenges exist as the simple act of selecting reasons for initiating an action does not necessary elicit reflection, drawing us to question whether reflection played a role in the differences between the conditions.

Conversely, ter Vrugte and others (2015) found different reflection prompt types did not impact performance. They examined written reflections elicited via prompts during learning about math with a GBLE to assess its relation to reasoning and learning. Specifically, four conditions were built into the system: reflection prompt condition, reflection prompt with procedural information condition, procedural information condition with no reflection prompts, and a control condition. Results showed reasoning skills improved across all conditions, suggesting reflection prompts or procedural information did not have an effect on learning (ter Vrugte, de Jong, Wouters, Vandercruysse, Elen, & van Oostendorp, 2015).

While the authors argued the reflection prompts were too demanding to have an impact on learning, we would like to raise the issue that examining the quality of the reflection responses (e.g., depth at which one reflects) may provide more insight into the findings of this study, particularly in why there was no effect of reflection prompts on performance. For instance, can we ensure the learners were reflecting solely based the prompt they received and the number of times they completed a written response, i.e., the quantity of written reflection? Carpenter and colleagues (2020) demonstrated a novel technique to gauge the quality of reflection by examining adolescents' written reflection responses using machine learning to automatically detect the quality at which learners reflected to assess its relation to learning with Crystal Island, a GBLE designed with two learning goals: learn about microbiology concepts and solve a mysterious illness on the island (Taub et al., 2020). Reflection quality was defined based on the depth of the written response, with emphasis placed on whether the participant was observing, contemplating, and inspecting their thoughts, ideas, beliefs, perspectives as they related to achieving a learning goal and whether they generated a clear hypothesis (i.e., beliefs about a particular direction for their problem solution or learning) to inform their next set of actions. Three prompt types were built into Crystal Island requiring learners to reflect on different aspects of their learning and problem solving during game-based learning: (1) the important information they had learned, (2) their current approach to finding a solution, and (3) a different solution approach. These prompts were triggered when participants completed actions deemed critical for achieving the learning goals (see Table 1 for details). Results showed the average reflection quality across the three reflection types was positively related to learning outcomes (Carpenter et al., 2020).

This study highlights that examining the quality at which adolescents' reflect provides insight into the role of reflection on learning with GBLEs; but the authors failed to consider the other learning goal, i.e., solving the mystery to study how reflection quality impacted successful problem solving. For this very reason, we argue that a model of reflection that accounts for the goals or objectives built into the learning environment needs to be adopted in order gain insight into whether learners were observing, contemplating, and inspecting their beliefs, actions, thoughts, etc. in relation to achieving their learning goals. As highlighted in the discussion above, gaps remain in literature about whether different reflection prompt types (e.g., open, directed, etc.) impact adolescents' ability to practice reflection and how the quality of their reflections relate to learning and problem solving as they align with particular learning goals and objectives (Carpenter et al., 2020; Geden et al., 2020; Johnson & Mayer, 2010; Fiorella & Mayer, 2012). To address these challenges, we guided our research using McAlpine and others (1998) model of reflection relative to other models since it emphasizes that learning goals guide reflection which impacts reasoning, decision making, monitoring, and knowledge acquisition.

Model of Reflection

The model of reflection (McAlpine et al., 1999) describes six main components: goals, knowledge, action, monitoring, decision making, and a corridor of tolerance, where reflection functions as a continuous and dynamic interaction between knowledge and action (see Figure 10)

Specifically, the model explains reflection is driven by learning goals (e.g., solve the mysterious illness versus learn as much as possible about microbiology with Crystal Island). Once goals are identified and established, the learner constructs plans to achieve set goals based on their knowledge (i.e., cognitive structures that range from surface- to deep-level which are built from training and previous experiences), subsequently leading to actions which are continuously revised and evaluated via metacognitive monitoring that is based on the learner’s corridor of tolerance. The corridor of tolerance is a mechanism that determines whether monitoring will result in decision making that leads to change such as a change in strategy use (Shavelson & Stern, 1981).It is hypothesized that the corridor of tolerance determines whether change should be implemented based on the learner’s threshold for what is acceptable, such that the cues the learner is monitoring will fall within acceptable or unacceptable boundaries based on how the learner defines progress toward achieving their goals. Depending on where this evaluation falls, a decision will be made about adjusting or continuing actions based on observing, contemplating, and inspecting plans, ideas, hypotheses, etc. that the learner has developed (McAlpine et al., 1999) as they relate to learning goals. Further, during inquiry reflection serves to control inference making and reasoning while searching for information to inform a solution (Dewey, 1933; McAlpine et al., 1999).

It is critical to emphasize that in the model, learning goals drive inquiry, reflection, inference making and reasoning. We argue McAlpine and others’ (1999) model of reflection is appropriate for studying reflection during game-based learning, since GBLEs are designed with clearly structured learning goals that require learners to complete in order to be successful (Plass et al., 2020). Plass and others’ (2020) explains that “...games for learning may be defined as games with specific learning goals.” (p. 3); thus, the influence of goals on problem-solving actions and learning information is evident. Significant gaps exist in literature as this model has not been used to game-based learning or learning analytics research (Kovanović, Joksimović, Mirriahi, Blaine, Gašević, Siemens, & Dawson, 2018; Scheffel, Drachsler, Stoyanov, & Specht, 2014; Carpenter et al., 2020; Geden et al., 2021; Perez-Colado, Perez-Colado, Freire-Moran, Martínez-Ortiz, & Fernández-Manjón, 2017; Taub et al., 2020). We addressed this gap by examining the role of the quantity and quality of reflection as they related to learning goals structured within Crystal Island to advance our understanding of reflection during game-based learning. Our study is outlined below.

Current Study

In this study, we examined the role of reflection prompt types on adolescents' problem-solving and learning outcomes with Crystal Island. Our research was guided by the model of reflection by McAlpine and others (1999), as challenges are prevalent in literature examining reflection prompts with GBLEs (Carpenter et al., 2020; Moreno & Mayer, 2005; Fiorella & Mayer, 2012; Johnson & Mayer, 2010; ter Vgute et al., 2015; Plass et al., 2015; Taub et al., 2020). Since studies show reflection develops in adolescence (Moshman, 2011), it is critical to study how to effectively build reflection prompts within GBLEs to enhance learning and problem solving across learners. As such, the objectives of the current study were to analyze the effectiveness of three reflection prompt types scaffolding: (1) the important information previously learned, (2) the current approach to finding a solution, and (3) a different solution approach on two different learning goals, i.e., problem solving and knowledge acquisition with Crystal Island. To align our research questions with the model of reflection (McAlpine et al., 1999), we hypothesized reflection prompt types would impact learning and problem solving differently because they varied by goals (see Coding and Scoring subsection for details).

What is the likelihood that a learner who solves the problem is related to both his/her quantity of reflections with Crystal Island?

While reflection positively impacted performance with GBLEs (Johnson & Mayer, 2010; Moreno & Mayer, 2005; Fiorella & Mayer, 2012), samples consist largely of young adults who may be more developmentally capable of reflecting with GBLEs. A study by ter Vgute and others (2015) found that reflection prompts did not have an effect on adolescents' performance with GBLEs. Yet, a study by Wu & Looi (2012) found a positive effect of adolescents' quantity of reflection on performance with a learning-by-teaching environment. Since the effect of reflection prompt types built into Crystal Island on adolescents' problem solving has yet to be investigated, we adopted a non-directional hypothesis and expected the quantity of the three reflection prompt types to impact adolescents' problem-solving performance.

To what extent does the quantity of reflections predict post-test scores while controlling for pre-test scores with Crystal Island?

We expect the quantity of the three reflection prompt types to impact adolescents' learning outcomes via pre/post-test assessments based on current literature (Johnson & Mayer, 2010; Moreno & Mayer, 2005; Fiorella & Mayer, 2012; Taub et al., 2020); however, we adopted a non-directional hypothesis due to mixed findings on adolescents' reflection on learning, and since the effect of the three reflection prompt types built into Crystal Island on adolescents' learning outcomes has yet to be investigated.

What is the likelihood that a learner who solves the problem is related to both his/her average quality of reflections with Crystal Island?

Studies have found a positive effect of prompts on the depth at which adolescents' reflect and knowledge acquisition (Carpenter et al., 2020; Wu & Looi, 2012). Yet, studies have yet to investigate the effect of reflection prompt types that target different learning goals on adolescents' problem solving. As such, we adopted a non-directional hypothesis because the empirical evidence is unclear about adolescents' capacity to reflect with GBLEs and gaps exist on the effectiveness of reflection prompt types built into Crystal Island on adolescents' problem-solving performance. Specifically, we expected the average quality across reflection prompt types to impact adolescents' problem-solving performance.

To what extent does the average quality of reflections predict post-test scores while controlling for pre-test scores with Crystal Island?

Studies have found a positive effect of reflection prompts on relations between adolescents' depth of reflection and learning (Carpenter et al., 2020; Wu & Looi, 2012). Yet, studies have yet to investigate the effect of reflection prompt types designed to target different learning goals on adolescents' learning. Because of this, we adopted a non-directional hypothesis and expected the average quality across the three reflection prompt types to impact adolescents' learning outcomes.

Methods

Participants and Materials

A sample of 120 secondary-school learners (51% female; Age: *M* = 13.57, *SD* = 0.54) were recruited from several middle-school classrooms to solve a mystery during their science class using Crystal Island, a narrative-based GBLE designed to foster (1) higher-order thinking skills and (2) microbiology knowledge (Rowe et al., 2011; Taub et al., 2020). Most participants identified as 'White/Caucasian' (54%), while the remaining participants identified as 'Black/African American' (27%), 'Hispanic or Latino' (19%), or 'Other' (16%). Most of the sample reported occasionally playing video games (33%), while others reported frequently

(28%) and very frequently (20%) playing video games. The majority also reported possessing average video game skills (39%), while the remaining reported being skilled (29%) or of limited skills (14%). The majority of the sample also reported playing 0-2 hours of video games per week (38%), while others reported playing 5-10 (23%) and 3-5 hours (20%) of video games per week. None of the participants reported they had learned with Crystal Island before the study.

A 21-item, 4-option multiple choice, pre/post-test assessment was administered before and after game-based learning with Crystal Island regardless of whether the mystery was solved to capture microbiology knowledge. The pre/post assessments were created by the research team and middle-school teachers, which contained 12 factual (e.g., “What is the smallest type of living organism?”) and 9 procedural questions (e.g., “What is the difference between bacterial and viral reproduction?”). Several self-report questionnaires were also administered to participants before and after game-based learning to gauge emotions, motivation, and cognitive load. We obtained ethics committee approval before recruiting and collecting data. There was no experimental manipulation in this study as our aim was to explore how to best leverage Crystal Island to help learners engage in effective problem solving, knowledge construction, and higher-order thinking using reflection prompts and system features.

Crystal Island Learning Environment

The Crystal Island system is single-player game-based learning environment built to create a rich story-centered, problem-solving experience using game features (e.g., tools) and embedded reflection prompts (Carpenter et al., 2020; Carpenter, Cloude, Rowe, Azevedo, & Lester, 2021; Geden et al., 2021). The science content provided in the game was aligned with the Standard Course of Study Essential Standards for Eighth-Grade Microbiology (McQuiggan, Goth, Ha, Rowe, & Lester, 2008). Upon starting the game, participants were required to adopt the role of a Center for Disease Control and Prevention agent who has been sent to a tropical island with a recent illness outbreak. To successfully complete their mission, participants needed to identify (1) the pathogen, (2) the source of the pathogen (e.g., salmonellosis transmitted through eggs), and (3) a correct treatment plan. Tools and resources were designed into the system to provide clues and information about the illness such as how the pathogen was contracted using non-player characters (NPCs) like the camp nurse explaining patients' symptomatology, the ability to scan food items for pathogens, opportunity to read books, research articles, and posters that describe information about various pathogens and diseases (e.g., bacteria versus virus characteristics; Figure 11), and document hypotheses and findings using a diagnosis worksheet (Figure 12). The diagnosis worksheet was also a tool that participants used to submit their final diagnosis, source of contamination, and treatment solution.

The prompts embedded into the system were designed to foster reflection and triggered using event-based, production rules (Table 7). Further, the prompts asked participants to reflect on their problem solving once they completed actions deemed critical for solving the mystery via open-ended, short responses (e.g., *"Please describe the most important things that you've learned so far, and what is your plan moving forward?"*) and rating scale items (i.e., *"On a scale of 1-10, how well is your investigation going?"*; Figure 13). Three types of reflection prompts were built into the system which correspond to different learning goals: (1) progress plan, (2) solution approach, and (3) different solution approach, where each was designed to raise awareness to different aspects of problem solving. Specifically, the progress-plan prompts were triggered once a participant completed an action deemed critical for solving the illness, such as talking to the camp nurse to gather information on sick patients' symptoms. In response to actions like this, the learner was immediately prompted to describe the most important thing they had learned thus far in their problem solving toward identifying the pathogen. As such, the progress-plan prompt is distinctive from other prompts in that it was triggered before participants indicated they had reached a final solution during game-based learning to foster their reflection on progress toward solving the mystery. Conversely, both solution-approach and different-approach prompts were triggered when the learner submitted a correct diagnosis and treatment plan for the illness (i.e., solved the mystery), or if they ran out of class time. Thus, learners only completed one solution- and different-solution approach prompts. Learners were first prompted to describe what problem-solving actions and information contributed to their success or failure (i.e., solution-approach). Afterwards, they were prompted to describe what problem-solving actions and strategies they would do differently (or the same) if given the chance to solve the mystery again (i.e., different approach). We would like to emphasize that the three reflection prompts were triggered at different points during game-based learning (i.e., before finding the solution and after) to scaffold reflection and awareness to their problem-solving approach and knowledge of microbiology.

Procedure

Before data collection, parents and participants provided written consent and assent to allow their questionnaire, interaction log, and performance data to be used for research purposes. Consent/assent were obtained using forms distributed in person or electronically via email by researchers and middle-school teachers. At least one day prior to playing Crystal Island, participants who provided both consent/assent completed a series of online questionnaires gauging experience with video games, emotions, metacognition, motivation, and microbiology knowledge via a 21-item, multiple-choice pre-test assessment using school computers during regular school hours. On the following day, participants were introduced to the game in their classroom by a researcher using a short video that explained the story line of Crystal Island. Participants were told that the objective of the game was to solve a mysterious illness plaguing a remote, tropical island and learn about microbiology.

Afterwards, participants interacted with the game on individual laptops provided by the research team for approximately 100 minutes during their normal science class time over the course of two class periods. Participants completed the study once the 2 hours had passed on the second day, whether or not they solved the mystery illness. They did not receive the correct solution if they did not correctly solve the mystery to ensure that it did not confound their performance on the post-test knowledge assessment. Next, they completed a series of questionnaires similar to the pre-test session, including a 21-item, multiple-choice post-test to capture microbiology knowledge.

Coding and Scoring

Outcome Variables

McAlpine and others' (1999) model of reflection guided our variable operational definitions. Since learning goals guide reflection according to this model, we created two outcome variables which differed by learning goals: (1) microbiology knowledge acquisition and (2) successful problem solving. For instance, the goal of Crystal Island was to solve the mystery illness (i.e., problem solving); yet, performing well on the post-test required learning as much as possible about microbiology which may include information irrelevant to solving the mystery. Because of this, we defined outcome variables separately to study their relation to adolescents' reflection. To measure microbiology knowledge acquisition, i.e., learning, we used pre/post-test scores that were defined based on the ratio of correct answers over total items on the assessment. Successful problem solving--i.e., performance was defined based on whether the learner solved the mystery illness using a binary format (1 = solved, 0 = unsolved), which required the learner to not only identify the correct pathogen (e.g., salmonellosis), but also the source of contamination (e.g., transmitted via eggs), and provide a correct treatment solution (e.g., rest) using the diagnosis worksheet.

Reflection

To capture reflections, we extracted data from on-line behavioral traces, and reflection was defined in two ways: (1) quantity and (2) the average quality (or depth) of reflection responses. Overall, participants spent an average of 10 minutes (*SD* = 3.08) reflecting and on average spent 86.46 minutes (*SD* = 10.60) learning with Crystal Island. To define the quality (or depth), a rubric developed by Carpenter and others (2020) was used to score the written reflection responses on a scale from one to five (1 = *"no depth"*, 5 = *"a reflective response with a high-quality sequence of abstract plans for problem solving"*; see Table 8 for the rubric). While it is common for quantitative models of reflection to capture both depth (e.g., non-reflective, shallowly reflective, or highly reflective) and breadth (e.g., attending to feelings, validation, or justification) aspects of reflection (Kovanović et al., 2018; Ullmann, 2018), this rubric focused exclusively on depth. This was done because the reflections collected during game-based learning in Crystal Island were brief (*M* = 20.22 words, *SD* = 15.56) and therefore inherently limited in reflective breadth. The rubric of reflective depth was developed by two researchers using a grounded theory approach.

To determine what constituted a reflection rating of one, the two researchers searched through the reflections until several that seemed particularly weak were identified. Based on these reflections, the researchers determined that the weakest reflections were those that either (1) lacked both a plan of action and commentary on relevant knowledge, (2) too abstract to be very useful, (3) largely unactionable, or (4) entirely unrelated to the learning experience (e.g., *"Yeah cool game I learned science"*; see Table 2 for more examples). Next, the researchers selected several reflections that seemed exceptionally strong and used them as the basis for a reflection rating of five (e.g., *"I will continue to test the foods the sick people touched or previously ate to see if it is contaminated"*). The researchers determined that reflections which received a score of 5 either presented a clear hypothesis that was supported by strong evidence and reasoning, or provided a high-quality abstraction of the problem that demonstrated an understanding of the most important information the student had obtained (see Table 2). A similar process was used to complete the rest of the rubric. In total, 20 reflections (four per reflection depth rating) were used to develop the rubric (Carpenter et al., 2020). Once the rubric was developed, another 20 reflections were randomly sampled from the dataset and separately coded by each researcher. The ratings for these reflections were then discussed and any differences were reconciled, thus ensuring that both researchers had a shared understanding of the rubric. Finally, both researchers separately coded the remaining 708 reflections and an intra-class correlation of 0.669 was achieved, indicating moderate inter-rater reliability (Cohen, 1960). The final reflection depth ratings used in this work were calculated by averaging the scores assigned by the two researchers (*M* = 2.41, *SD* = 0.86).

Statistical Analyses

The data were processed using a pipeline created in Python (van Rossum & Drake, 2010) and then analyzed in R Version 3.6.2 (R Core Team, 2019). Packages 'read xl' (Wickham & Bryan, 2017). 'dplyr' (Wickham, Francois, Henry, & Muller, 2018), and 'reshape2' (Wickham, 2007) were used for data manipulating and wrangling. The 'fitdistrplus' and 'logspline' packages were used to determine the distribution of the data before model building. To build different types of models, we used the ‘glm’ and ‘lm’ functions from the ‘stats’ package (R Core Team, 2019). Next, the ‘stepAIC’ function from the ‘MASS’ package was used to conduct variable selection via stepwise Akaike Information Criterion (AIC; Venables, & Ripley, 2002). ‘Base’ and ‘lmtest’ packages were used to access model indices, while the ‘ggplot2’ package was used to visualize relationships among variables (Wickham, 2016).

For the analysis, binomial logistic regression models were calculated to assess relationships between the quantity and quality of reflection types and likelihood a learner solved the mystery. Multiple linear regression equations were used to assess relationships between the quantity and quality of reflections types and pre/post-test proportional scores. Stepwise AIC was used to select the final models which demonstrated the lowest AIC for each research question. Since stepwise AIC extends to generalized linear models, stepwise AIC was selected relative to other techniques (Yamashita, Yamashita, & Kamimura, 2007). Prior to initial analyses, we removed 30 participants from the dataset because they did not complete the post-test assessment. We also investigated whether the quantity and quality of reflections types contained significant outliers using Grubbs (1969) approach. Three participants were removed due to outlying data points which were further emphasized using boxplots.

Results

Table 9 provides descriptive statistics. Alpha coefficients for pre/post-test assessments met satisfactory reliability (α= 0.69; Cronbach, 1951). Prior to model building, we conducted preliminary analyses to examine whether there were relationships between solving the mystery in Crystal Island and scores on pre/post-test assessments. Two separate t-tests were calculated using a Bonferroni correction ($p < 0.05/2 = 0.025$) and revealed significant differences in pre, $t(115) =$

-2.33, *p* = 0.021, and post-test scores, *t*(112) = -2.65, *p* = 0.009, between learners who solved (pre: *M* = 0.35; post: *M* = 0.38) or did not solve the mystery (pre: *M* = 0.29; post: *M* = 0.30; see Tables 9 and 10). This suggests that an understanding of microbiology played a role in whether or not the learner was able to solve the mystery during learning with Crystal Island.

Research Question 1: What is the likelihood that a learner who solves the problem is related to both his/her quantity of reflections during learning with Crystal Island?

AIC metrics indicated that the best model was present using one predictor: progress-plan reflection instances. A simple binomial logistic model was fit to the data (Table 11) to test the research hypothesis regarding the relationship between the likelihood that a learner solves the mystery illness and the quantity of solution-approach and progress-plan reflections during learning with Crystal Island. We calculated the final model which showed that predicted logit of (Solved) = -3.14 + (-0.06)*Progress plans. According to the model, the likelihood of a learner solving the mystery illness was related to the quantity of completing progress-plan reflections ($p < 0.025$; Figure 14). In other words, the more progress-plan reflections completed by the learner during game-based learning, the more likely the learner solved the mystery. Specifically, the odds of a learner solving the mystery illness were 0.94 times greater if they completed more progress-plan reflections compared to a learner who completed less progress-plan reflections. This finding was partially consistent with our hypothesis and the model of reflection (McAlpine et al., 1999), where we expected the quantity of all reflection types to predict the likelihood that a learner solved the mystery. Please note that these findings are exploratory and have correlational implications. However, learners were not required to initiate specific actions during game-based learning that triggered reflection prompts and so the more often that learners initiated actions that were relevant to achieving the learning goals, the more often they were prompted to reflect.

Research Question 2: To what extent does the quantity of reflections predict post-test scores while controlling for pre-test scores after learning with Crystal Island?

Upon applying the AIC method, the best model indices were present using one predictor. A simple linear equation was fit to the data (Table 12) to test the research hypothesis regarding the relationships between pre- and post-test scores after learning with Crystal Island. The result showed that the predicted average of post-test scores = 0.139 + $(0.633)^{*}$Pre-test scores. According to the model, there were significant relationships between post-test scores and pre-test scores ($\beta = 0.63$, $p < 0.025$), but the model did not fit the data when including the quantity of reflection type variables, suggesting no relationship existed with post-test scores. Specifically, the fitted model estimated that the average post-test score increased by 0.633 for each point increase on the pre-test, where pre-test scores explained 27% of the variance in post-test scores (Figure 16). This finding was inconsistent with our hypothesis and previous research, where we expected the quantity of reflection types to predict post-test scores.

Research Question 3: What is the likelihood that a learner who solves the problem is related to both his/her average quality of reflections during learning with Crystal Island?

Upon applying the AIC method, the best model indices were present using one predictor. A one-predictor binomial logistic model was fit to the data (Table 13) to test the research hypothesis regarding the relationship between the likelihood that a learner solves the mystery illness and the average quality of solution-approach reflections during learning with Crystal Island. We calculated the final model which showed that the predicted logit of (Solved) = -2.20 + $(0.90)^{*}$Solution approach. According to the model, the likelihood of a learner solving the mystery illness was related to the average quality of their solution-approach reflections during game-based learning ($p < 0.025$; Figure 17). In other words, the higher average quality of solution-approach reflections completed during game-based learning, the more likely a learner solved the mystery. Specifically, the odds of a learner solving the mystery illness were 2.46 times greater if they completed, on average, higher quality solution-approach reflections compared to a learner who completed, on average, lower quality solution-approach reflections. This finding was partially consistent with our hypothesis and the model of reflection (McAlpine et al., 1999), where we expected the average quality of all reflection types to predict the likelihood that a learner solved the mystery.

Research Question 4: To what extent does the average quality of reflections predict post-test scores while controlling for pre-test scores after learning with Crystal Island?

Upon applying the AIC method, the best model indices were present using two predictors. A multiple linear equation was fit to the data (Table 13) to test the research hypothesis regarding the relationships between the average quality of solution-approach and progress-plan reflections and post-test scores while controlling for pre-test scores after learning with Crystal Island. The results showed that the predicted average of post-test scores = 0.012 + $(0.533)^{*}$Pre-test scores + $(0.054)^{*}$Progress plan. According to the model, pre-test scores (β = 0.53, $p < 0.025$) and the average quality of progress-plan reflections (β = 0.025, $p < 0.025$) predicted post-test scores, but there were no relationships between the average quality of solution-approach reflections and post-test scores ($p > 0.025$). Specifically, the fitted model estimated that the average post-test score increased by 0.53 for each point increase on the pre-test and 0.05 for each point increase in the average quality of progress-plan reflections during game-based learning, where pre-test scores and average quality of progress-plan reflections explained 33% of the variance in post-test scores (Figure 18) This finding was partially consistent with our hypothesis and the model of reflection (McAlpine et al., 1999), where we expected the average quality of all reflection types to predict post-test scores.

Discussion

In this study, we examined the role of different reflection types on adolescents' problem-solving performance and knowledge acquisition with Crystal Island. Since challenges are prevalent in literature examining reflection prompts with GBLEs, where a theoretical lens and clear operational definition on reflection are often lacking our research was guided by the model of reflection by McAlpine and others (1999) (Carpenter et al., 2020; Moreno & Mayer, 2005; Fiorella & Mayer, 2012; Johnson & Mayer, 2010; ter Vgute et al., 2015; Plass et al., 2015; Taub et al., 2020). The objective of the current study was to analyze data from adolescents to assess the effectiveness of three reflection prompt types on different learning goals: (1) problem solving and (2) knowledge acquisition with Crystal Island. To align our research questions with the model of reflection (McAlpine et al., 1999), we hypothesized reflection prompt types impacted learning and problem solving differently because they varied by learning goals. It is important to note research questions 1 & 3 and 2 & 4 were combined in this section to highlight findings by the learning goal, i.e., solving the mystery versus learning about microbiology.

What is the likelihood a learner who solves the mystery illness is related to both his/her quantity and quality of reflecting with Crystal Island?

The first research question examined whether the quantity of reflection prompt types were related to the likelihood a learner solved the mystery illness with Crystal Island. Our models found the more often learners were prompted to reflect on their progress plans, the more likely they were to solve the problem in Crystal Island. These findings are partially consistent with our hypothesis where we expected all reflection prompt types to impact performance, as empirical evidence and the model of reflection (McAlpine et al., 1999) suggests reflection enhances problem solving (Fiorella & Mayer, 2012; Johnson & Mayer, 2010; Moreno & Mayer, 2010; Wu & Looi, 2012). We would also like to note that these findings are exploratory and there was no control group to compare learners who were prompted versus those who were not prompted at all. However, learners were not required to initiate specific actions during game-based learning which triggered the reflection prompts, and so the more often that learners initiated actions that were relevant to achieving the learning goals, the more often they were prompted to reflect.

While we found that learners who reflected on their progress plans more often were more likely to solve the mystery, we did not find an effect of prompting learners to reflect on their solution approach or a different solution approach. A possible explanation for this could result from (1) prompting adolescents to reflect on their problem-solving approach after finishing the game which might suggest that learners had active learning goals to solve the mystery, thus according to McAlpine and others (1999), there would be no drive to engage in reflection and (2) prompting adolescents only one time, which captures no variability and thus leaves little room for predictive models. As such, future studies should investigate whether designing GBLEs that prompt adolescents' to contemplate their problem-solving approach at multiple instances during game-based learning has an effect on reflection and problem-solving outcomes. If learners are actively pursuing a goal to solve the mystery, the prompts could be more impactful in scaffolding reflection and enhancing performance. Future studies should also consider the role that reflection plays in self-regulated learning by examining different theoretical perspectives such as that proposed by Zimmerman (2013) and capturing self-regulated learning processes and strategies over the course of game-based learning (Carpenter et al., 2021; Siadaty, Gasevic, & Hatala, 2016; Viberg, Khalil, & Baars, 2020) and capturing self-regulated learning processes and strategies over the course of game-based learning.

The third research question examined whether the average quality of reflection prompt types were related to the likelihood a learner solved the illness with Crystal Island. Our models suggested the higher average quality (i.e., depth) of reflection responses on their solution approach, the more likely they were to solve the problem with Crystal Island. These findings highlight the importance of scaffolding adolescents' reflection on their approach to problem solving with GBLEs when learners engage in problem-solving tasks. Further, the average quality of progress-plan and different solution-approach prompts did not impact problem solving. These findings are partially consistent with our hypothesis where we expected all reflection prompt types to impact problem solving (Fiorella & Mayer, 2012; Johnson & Mayer, 2010; McAlpine et al., 1999; Moreno & Mayer, 2005; Wu & Looi, 2012). However, ter Vgute and others (2015) also found that some reflection prompts did not impact performance and suggested the prompts were too demanding for adolescents. As such, a possible explanation for this finding could be that different-solution approach prompts were too taxing for learners since it requires knowledge of how to problem solve. Thus, future studies should aim to assess adolescents' problem-solving ability which could explain why contemplating on different problem solving approaches.

In terms of progress plans, a possible explanation could be related to the design of the prompts such that they were built to scaffold the learner to contemplate the most important information they had learned about microbiology rather than their problem solving. As such, could this prompt be related to the learning goal of learning about microbiology and thus, according to the model of reflection (McAlpine et al., 1999), would not have driven the learner to reflect on their progress toward successful problem solving, potentially explaining the lack of effect of reflection quality on performance? A potential reason for finding an effect for the quantity of progress-plan prompts could be related to the number of times the GBLE was designed to prompt which was considerably higher than the other reflection prompt types (Table 9). Further, could the quantity of progress-plan prompts simply have had an effect because they outnumbered the other prompts? Based on these findings, future studies should consider initiating solution-approach prompts at critical points during problem solving instead of progress-plan prompts to assess if these prompts further enhance adolescents' reflection and problem solving.

Future research should also aim to capture multimodal data during game-based learning, as the information may provide more insight into the impact of prompts on adolescents' capacity to reflect. For example, what might facial expressions of emotions or eye-tracking data provide in explaining if, when, and how reflection prompts are effective in scaffolding reflection and fostering successful problem solving? These findings have implications for the classroom, where the models could serve to inform instruction and guide personalized interventions based on the quality and quantity of adolescents' reflection during game-based learning. However, considerable research is still required. Implementing game-learning analytics in the classroom for instructional decision making will also require additional support and technological resources, such as a dashboard to illustrate data visualizations for sense making that are currently unavailable to most teachers (Perez-Colado et al., Cloude, Dever, Wiedbusch, & Azevedo, 2020; Roll & Winne, 2015; Wiedbusch, Kite, Yang, Park, Chi, Taub, & Azevedo, 2021).

To what extent does the quantity and quality of reflections predict post-test scores while controlling for pre-test scores with Crystal Island?

For our second research question, we examined whether the quantity of reflection prompt types were related to post-test scores while controlling for pre-test scores. We found the frequency of reflection prompts did not have an impact on pre/post-test scores. Instead, prior knowledge was the best predictor of post-test scores. This finding was partially inconsistent with our hypothesis as well as previous research (Fiorella et al., 2012; Johnson & Mayer, 2010; Moreno & Mayer, 2005; Wu & Looi, 2012), where we expected the quantity of reflection prompts to impact learning. A possible explanation could be that assessing the quantity of reflection has no bearing on whether learners engage in reflection with GBLEs. Future studies should aim to assess the depth of reflection and its relation to learning goals to study relations between reflection and outcomes with GBLEs. However, this finding highlights the importance of prior knowledge about the topic as it relates to post-test performance. For example, an instructor may want to pay close attention to learners who have demonstrated less understanding of the learning materials prior to engaging with a GBLE to inform their instructional decisions in the classroom, particularly in regard to scaffolding and intervening in a classroom with a lot of students, which is more often the rule instead of the exception. For instance, if an instructor knows that prior knowledge is related to learning outcomes, then they can monitor the subset of students who demonstrate less understanding relative to those who demonstrate more understanding.

For the final research question, we examined whether the average quality of reflection prompt types were related to post-test scores while controlling for pre-test scores. We found the average quality of progress-plan prompts and pre-test scores positively predicted post-test scores.

This finding was partially consistent with our hypothesis as well as previous research (Fiorella et al., 2012; Johnson & Mayer, 2010; Moreno & Mayer, 2005; Wu & Looi, 2012), where we expected the quality of all reflection prompt types to positively impact learning. A possible explanation about the lack of effect for solution-approach and different-solution approach prompts could be the pre/post-test assessments were built to capture knowledge about microbiology; yet, the objective of Crystal Island was to solve a problem in which solution-approach and different-solution approach reflection prompt types were created to scaffold (e.g., *"Well done, Agent! You've saved everyone on the island. Now that you are finished, we would like to ask a couple of final questions. Please explain how you approached solving the mystery."*). As such, the prompt types had no relevance to learning information about microbiology, but rather scaffolded learners to reflect on problem solving. This explanation is aligned with the model of reflection (McAlpine et al., 1999), suggesting learning goals play a role in driving reflection with GBLEs. Further, the progress-plan reflection prompts required the learner to think about the most important information acquired (e.g., *"Agent, it looks like you are making progress on diagnosing the illness, but you're not quite there yet. In your own words, please describe the most important things that you've learned so far, and what is your plan moving forward?"*), potentially targeting information covered on the pre/post-test assessments. These findings provide directions for examining the role of reflection on performance and learning. When is the ideal time to prompt learners to reflect, and does that time vary based on the learning goal and environment? Researchers should also consider the role of adolescents' motivation, cognitive load, or level of knowledge of problem solving, as game-learning analytics data could provide insight to pinpoint when learners are not engaging in reflection (e.g., the quality of solution-based reflection and motivation is low) to inform instructional decision making (Winne, Teng, Chang, Lin, Marzouk, Nesbit,... & Vytasek, 2019; Winne, 2017; Cloude et al., 2020; Wiedbusch et al., 2021).

Limitations

For this study, we did not assess individual characteristics outside of prior knowledge, such as motivation or interest, which may have played a role in adolescents' quantity and quality of reflection during game-based learning. Additionally, we did not capture the temporal components related to learners' reflection, such as the amount of time engaging in reflection across the different prompt types to assess its relation to performance and learning. Another confounding factor could have been adolescents' experience and familiarity playing video games.

Final concluding statements

In this study, we examined the role of different reflection types on adolescents' problem-solving performance and knowledge acquisition with Crystal Island. Findings suggested the quantity and quality of adolescents' reflection was related to problem solving and learning with a GBLE, but the effectiveness of different reflection prompt types varied based on the learning goal they targeted. We emphasize the importance of adopting a theoretical perspective when studying reflection and highlight how learning goals built into GBLEs drive adolescents' reflection. Future studies should aim to investigate the role of time and individual factors on reflection during game-based learning. Implications for building reflection prompts within GBLEs to enhance adolescents' reflection, problem solving, and learning and using game-learning analytics to inform instruction in the classroom are provided.

Declaration of Conflicting Interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The publication of this article received financial support from the National Science Foundation (DRL-1661202). Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of NSF.

CHAPTER FIVE: INTEGRATED DISCUSSION

As we identify significant gaps in learners' capacity to acquire and apply KSAs deemed critical for the 21st century, we are drawn to investigate the structure of the U.S. educational system and its ability to prepare the next generation of the workforce across *all* populations of learners. In this dissertation, one book chapter and two journal articles were presented which attempted to leverage MLA grounded in SRL theory to understand how human learning unfolds across contexts, tasks, domains, and populations and their relation to developing KSAs with emerging technologies.

Specifically, the first book chapter highlights the importance of utilizing MLA grounded in metacognition and self-regulation to transform medical education. There is still a heavy reliance on static measurements of medical students’ clinical reasoning captured before and after learning activities that is often isolated from the application of clinical reasoning in the real world and misses information on the role that metacognition and self-regulation could play in augmenting clinical reasoning across medical scenarios and contexts. We propose that to train medical students to clinically reason effectively about medical cases, curricula need to equip students with adequate knowledge and skills in metacognition and self-regulation, such that they can monitor their reasoning approach and adapt their strategies when they detect biases, errors in thinking, anchoring etc., using MLA situated in SRL theory with emerging technologies. We demonstrate the possibility of doing this by referring to the socio-cognitive cyclic model of self-regulation (SCT; Winne, 2019). We reviewed studies and found that while some studies leverage MLA to study medical education performance, few studies failed to use a theoretical lens to guide their MLA procedures and techniques (e.g., operationally defining constructs, guiding modeling techniques based on theoretical assumptions, etc.). For example, most research has studied the performance phase of the SCT model using MLA data. We argue that to train medical students to use metacognition and SRL during their clinical reasoning practice, capturing MLA over the course of learning activities, such as during simulation-based training, could detect metacognitive, SRL, and clinical-reasoning abilities over time, contexts, populations, tasks, etc., providing teachers and instructors with the ability to model, trace, and foster metacognition and SRL.

The third and fourth journal articles presented empirical evidence of using MLA to study SRL, 21st century KSAs and their relation to learning and performance for two populations: undergraduates and middle-school students. Within these two chapters, emerging technologies called Crystal Island (Taub et al., 2021) and Crystal Island: REFELCT were to define whether learners were (1) thinking scientifically about information or (2) reflecting on the information within the environment based on their interaction with game elements. The value in using emerging technologies, specifically game-based learning environments, is that they provide a platform to capture how learners are interacting with the interface which can be a direct measure of their learning process. For instance, using both logfiles and eye movements to define scientific thinking was a novel MLA approach because it indicated when learners were not only interacting with the game elements deemed relevant to scientific thinking such as opening a research article via logfiles, but also ensured that the learner was fixating on the game elements at the time of their interaction. Merging these two data channels and guiding the operational definitions of MLA data to define scientific-thinking constructs using the SDDS theory (Dunbar & Klahr, 2012), such that game elements were broken down into three different categories related to scientific-thinking constructs: (1) gathering information, (2) generating hypotheses, and (3) experimentally testing those hypotheses (see Tables 1-2) is a prime example of theoretically grounded MLA with emerging technologies.

Similarly, chapter four emphasizes ways to define and measure how adolescents reflected on information presented within the GBLE and its relation to clearly structured goals presented within the environment that are guided by the model of reflection (see Figure 10 where goals drive reflection). Two goals were built into Crystal Island: REFLECT. The first being that learners were instructed to learn about microbiology and pathology concepts. This goal was further emphasized by requiring learners to complete both pre/post-test knowledge assessments to gauge the amount of information they learned about microbiology and pathology. The second goal involved learners being instructed to solve a mysterious illness plaguing the island. The novel approach used in this theoretically based MLA technique was that, theoretically reflection is driven by learning goals, and since Crystal Island: REFLECT presents two learning goals, while interrelated to each other, the modeling approach was based on assessing relationships between both learning goals: (1) learning outcomes about microbiology (i.e., core 21st century knowledge) and (2) problem solving (i.e., core 21st century skill).

These two chapters are similar in their approach of using learning theory, whether rooted in reflection or scientific thinking, to guide the MLA approach including the operationalization of learning constructs as well as the modeling technique of multimodal data. Further, I argue that there are significant similarities between chapters 2-4 since they all involve some degree of self-regulation. It is rather obvious the presence of SRL in chapter 2 since a specific theory on SRL was used to guide the literature review, discussion, and implications for medical education, while chapters 3 and 4 may be less obvious since a model of reflection and SDDS theory were used. It is essential to highlight the presence of SRL within these chapters to further emphasize the importance of studying MLA grounded in SRL theory with emerging technologies.

Specifically, scientific thinking according to SDDS involves gathering information, generating hypotheses, and experimentally testing those hypotheses, and at the very heart of SRL resides the monitoring and controlling of cognitive, metacognitive, motivational, social, and affective processes as it relates to a learning goal. One could argue that to effectively gather information and generate hypotheses, a learner needs to monitor their level of understanding about the topic—e.g., feeling of knowing, such as monitoring their level of understanding about a virus versus bacteria so that they may be able to hypotheses what the pathogen source could be based on clues gathered from patients who demonstrate symptoms synonymous with a particular disease. This was emphasized further based on the findings where prior knowledge played an integral role in how learners interacted with game elements related to gathering information (see Figure 10), where prior knowledge moderated relationships between time engaging with elements captured via logfiles and time fixating on elements captured via eye movements. Additionally, when testing hypotheses, learners would then need to control or regulate their approach if their food items were incorrect or did not possess the disease, they hypothesized was the pathogen infecting the patients on the island. At its core, *effective* and *accurate* scientific thinking requires some element of SRL.

Similarly, reflection is deeply rooted in self-regulation. In fact, chapter four takes special care to define reflection as distinctive from metacognition since reflection plays an integral role in one’s ability to use metacognitive processes. There is much agreement within the scientific community that metacognition is the beating heart of self-regulation given its role in monitoring and controlling cognitive processes (e.g., metamemory framework by Nelson & Narens, 1994).

Reflection is the foundation of metacognition because it trains learners to contemplate their approach toward solving a problem or engaging in activities. In order to reflect, a learner needs to be conscious of what they are doing in relation to a learning goal and then gauge how well their approach is being leveraged toward meeting that goal, giving rise to the foundation for building high order thinking skills like metacognitive monitoring mechanisms; Tarricone, 2011). Further, learners needed to leverage reflective thinking as it relates to successfully solving the mystery which encompassed scientific reasoning and thinking (e.g., gathering information, hypothesis generation, and experimental testing).

Overall, the objective of this dissertation is to argue that we can mitigate the educational challenges plaguing the United States by moving away from a standardized approach to assess 21st century KSAs by integrating emerging technologies into curricula to capture MLA grounded in SRL and define developing 21st century KSA competencies as they are being applied across tasks, contexts, populations, time, etc. If learning was captured using multiple streams of data to inform when deficiencies in learning occur, this methodological approach could provide a means to measure, model, and foster learning processes with emerging technologies to acquire the KSAs necessary for meeting the demands of the 21st century. Specifically, theoretically driven MLA techniques and integrating emerging technologies into the classroom to facilitate personalized instruction and learning, broadly impacting society by equalizing the educational space and developing methods that quantify non-standardized models of learning that support learner-centered active roles in educational settings by including complex problems to be solved as opposed to requiring in routine performance on exams and assignments that fail the learner, and ultimately do not define and measure the complexity of 21st century KSAs that learners might possess beyond the ‘ideal’ ratio of correct items on an assessment (Blikstein & Worsley, 2016).

Limitations and Future Directions

It is important to acknowledge the significant gaps that exist as the field in moving toward developing innovative tools to deal with the new challenges associated with collecting, processing, cleaning, analyzing, and making sense of multimodal time series data. Common tools used to analyze multimodal data are traditional linear statistics (e.g., regression, correlation, analysis of variance, etc.), which test relations between outcome and predictor variables using a linear function and often fail to account for individual differences, context, and time. Additionally, these tools require adhering to three primary assumptions to reveal meaningful results: (1) observation independence, (2) a normal frequency distribution, and (3) an equal amount of variation. The very nature of the multimodal data captured across multiple sensors, often, fails to meet these assumptions for running a linear model. It also assumes that learning processes unfold in a linear function *always* over time, such as the more time spent interacting with this game elements equates to more knowledge gained about the information presented within the game element. Linear techniques are also based on the notion that averaging across data points reveals the truth about how learning processes unfold in the real world (Laplace, 1812). I am drawn to question this assumption since studies also reveal that the fluctuating and adaptive nature of self-regulated learning processes impact learning and performance (Bakhtiar, Webster, & Hadwin, 2018; Cloude et al., 2020; Goetz, Sticca, Pekrun, Murayama, & Elliot, 2016; Moeller, Ivcevic, Brackett, & White, 2018). Further, the assumptions needed for linear models do not align with the assumptions of SRL, which assumes that learning processes unfold dynamical and nonlinearly across time, tasks, individuals, situations, and contexts.

Since contemporary theoretical models of SRL emphasize the dynamic, interdependent, and adaptive nature of learning (Winne, 2018), it is imperative that we leverage modeling techniques built to handle the complex nature of SRL especially as it relates to grounding MLA in SRL theory to partition the potential measurement errors from multimodal dataset from intentional regulation that may exist across learning activities. If contemporary frameworks of SRL explain learning as dynamical and fluctuating, then it is appropriate to leverage modeling techniques and tools that represent those characteristics. A potential solution for incorporating these distinctive features might be those outlined in complexity science, or non-linear dynamical systems (NDS; Favela, 2020; Amon, Vrzakova, & D'Mello, 2019). Complexity science can be explained as an interdisciplinary approach for studying complex systems, where complexity is defined as having many interacting parts (Allen, 2001; Guastello, Koopmans, & Pincus, 2009).

A component of complex systems theory is non-linear dynamical systems theory (NDS; Guastello, Koopmans, & Pincus, 2009; Schuster, 1984) is defined as systems change over time and provides frameworks/constructs that could significantly enrich our understanding of learning constructs that illustrate different types of change such as those that exist in SRL. In particular, the four properties that define a complex systems theory involve (1) emergence (i.e., global patterns of behavior or data points not reducible to individual components), (2) interaction dominance (i.e., occurs when there is casual dependence between a whole and its parts, or order that results from non-order without a central controlling agent), (3) self-organization (i.e., a system that is far from equilibrium, yet the type of structure allows it to achieve high energy efficiency), and (4) universality (i.e., they occur in different contexts and substrates; Favela, 2020). Complexity sciences has deep roots in the laws of nature and has historical influences in physics and physical sciences, and more recently, researchers have applied to the study of many disciplines including cognitive sciences (Schiavo, Prinari, Saito, Shoji, & Benight, 2019), team science (Gorman, Dunbar, Grimm, & Gipson, 2017), among many others.

One could argue that SRL is highly relevant to non-linear dynamical systems theory. At its core, NDS defines dynamics as a set of interconnected elements that undergo change (Schuster, 1984), while nonlinearity refers to the proportion of input and outputs of the system, such that one change to an element does not produce the same changes to other elements within the system. This definition is relatively synonymous with the monitoring and regulation of cognitive, metacognitive, affective, social, and motivational processes in pursuit of learning goals (Panadero, 2017). Further, NDS emphasize the importance of the intrinsic fluctuations of the systems while also account for the external components such as context and time. Vallacher and Nowak (1999) recognized this similarity between SRL and NDS theories and went as far as to propose that learning goals within SRL could serve as attractors in nonlinear dynamical systems theory. The four properties of attractors include (1) a fixed point, (2) limit cycle, (3) toroidal attractors, and (4) chaotic attractors (Guastello & Leibovitch, 2009). These ideas are essential for advancing theory of SRL especially as we are shifting into the direction of metamotivation, a research topic that recognize that goals are not static but rather dynamical (Miele & Scholer, 2018), and recent studies have found that changes in learning goals impact use of SRL processes and ultimately impact performance outcomes (Cloude, Wortha, Wiedbusch, & Azevedo, 2021; Cloude et al., 2018).

Future studies need synthesize literature and reconsider how learning is being operationalized using multiple data channels and how each channel relates to different learning theories (e.g., self-regulated learning versus scientific discovery as dual search), and then map these data channels to various learning constructs. Once similarities and contradictions are made between different data channels and learning constructs across studies and theoretical frameworks, the synthesis could provide a specific direction for where future directions need to focus (e.g., capturing affect using physiology during collaborative problem solving) based on either a lack of or very few studies using a particular data channel and/or modalities (e.g., heart rate) capture learning and how it fluctuates over time. This is where interdisciplinary methods could be used to address challenges in educational research and learning sciences.

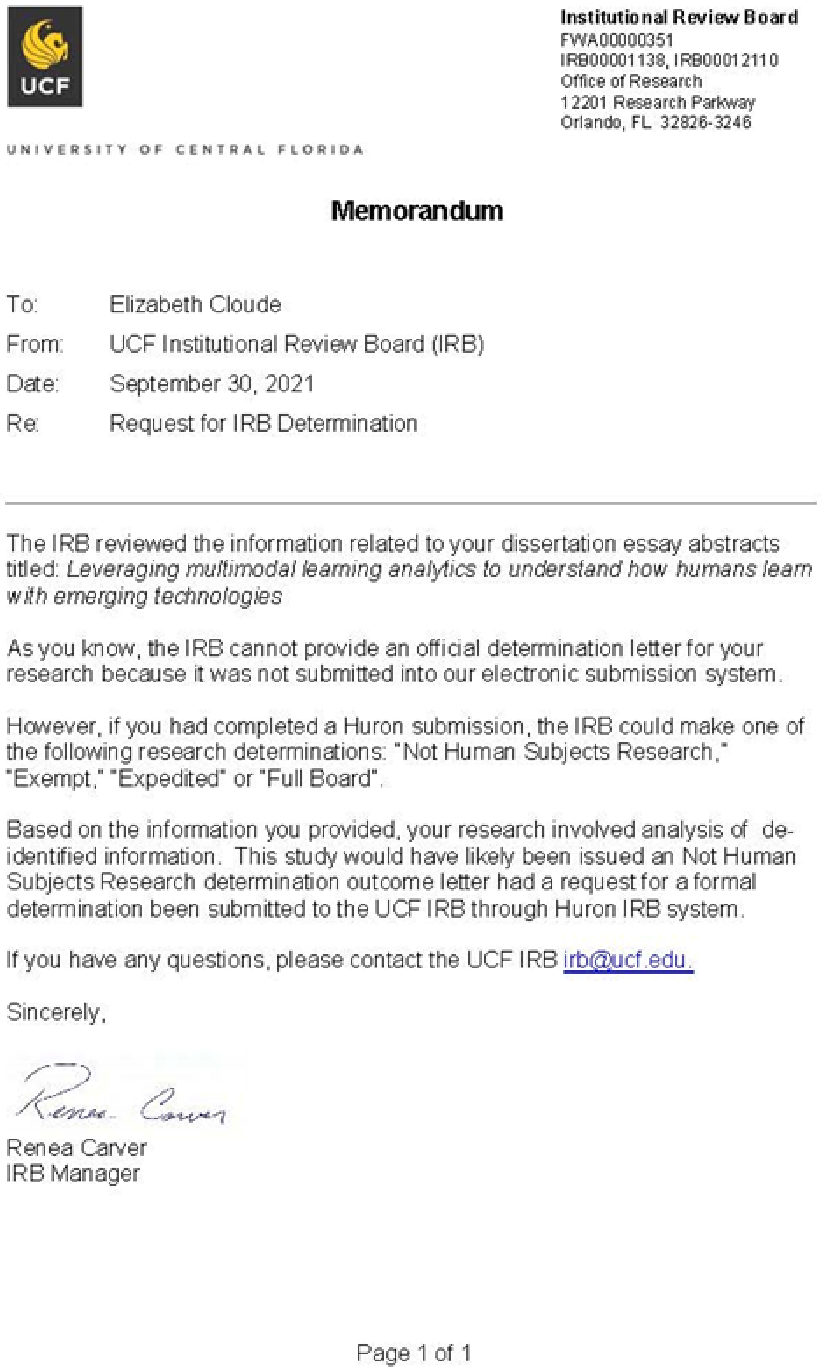
Intellectual Merit

This program of research will advance our understanding of the role of SRL on developing 21st century KSAs with emerging technologies through adopting a dynamical systems perspective and leveraging MLA. Studying SRL and its relation to 21st century KSA as a dynamical system offers insight into how learning processes change over time without neglecting the multiple levels of the system (i.e., physiological responses, facial expressions, individual, dyad of individuals, problem solving strategies, situation constraints, contextual factors, time on task, etc.). NDS theory offers insight beyond static and linear assumptions made in traditional cognitive and learning sciences research paradigms by focusing on if, when, and how learning changes over time and modeling as it exists in the real-world. Modeling how learning processes change using MLA grounded in SRL theory and NDS theory could advance our knowledge of how SRL impacts 21st century KSAs across contexts and time, allowing us to pinpoint if, when, and how learning processes emerge and impact developing KSAs with goals to enhance their outcomes using emerging technologies.

Broader Implications

Leveraging theoretically based MLA to understand how humans learn with emerging technologies could have broader implications in our society, including redesigning K-12 curricula for classrooms and high education and incorporating innovative use of emerging technologies and media in the classroom such as building intelligent dashboards that repurpose data back to instructors, teachers, learners, and potentially researchers. Further, this program of research could significantly benefit society by providing solutions to solve societal grand challenges by preparing the next generation of *all* learners to acquire and apply 21st century KSAs. This includes opening opportunities for the full participation of women, persons with disabilities, and underrepresenting minority groups including BIPOC and LBGTQIA+ groups to engage in STEM and law careers, improve education and educator preparation across all levels, improved well-being for all members of society, and increase the economic competitiveness of the United States by developing a diverse and globally competitive workforce that represents *all* members of society. A diverse and globally competitive workforce could help solve problems such as climate change, systemic racism, spread of misinformation, among many others that go beyond the classroom. When we invest in and equip the next several generations of learners with the capacity to effectively communicate and collaborate on solving a variety of novel and challenging problems, critically think about issues, and practice resilience and flexibility, we improve our world and those who exist in it.

APPENDIX: UCF IRB LETTER



UCF

UNIVERSITY OF CENTRAL FLORIDA

Institutional Review Board
FWA00000351
IRB00001138, IRB00012110
Office of Research
12201 Research Parkway
Orlando, FL 32826-3246

Memorandum

To: Elizabeth Cloude

From: UCF Institutional Review Board (IRB)

Date: September 30, 2021

Re: Request for IRB Determination

The IRB reviewed the information related to your dissertation essay abstracts titled: *Leveraging multimodal learning analytics to understand how humans learn with emerging technologies*

As you know, the IRB cannot provide an official determination letter for your research because it was not submitted into our electronic submission system.

However, if you had completed a Huron submission, the IRB could make one of the following research determinations: "Not Human Subjects Research," "Exempt," "Expedited" or "Full Board".

Based on the information you provided, your research involved analysis of de-identified information. This study would have likely been issued an Not Human Subjects Research determination outcome letter had a request for a formal determination been submitted to the UCF IRB through Huron IRB system.

If you have any questions, please contact the UCF IRB irb@ucf.edu.

Sincerely,

Renea Carver

Renea Carver
IRB Manager

Page 1 of 1

APPENDIX: FIGURES AND TABLES

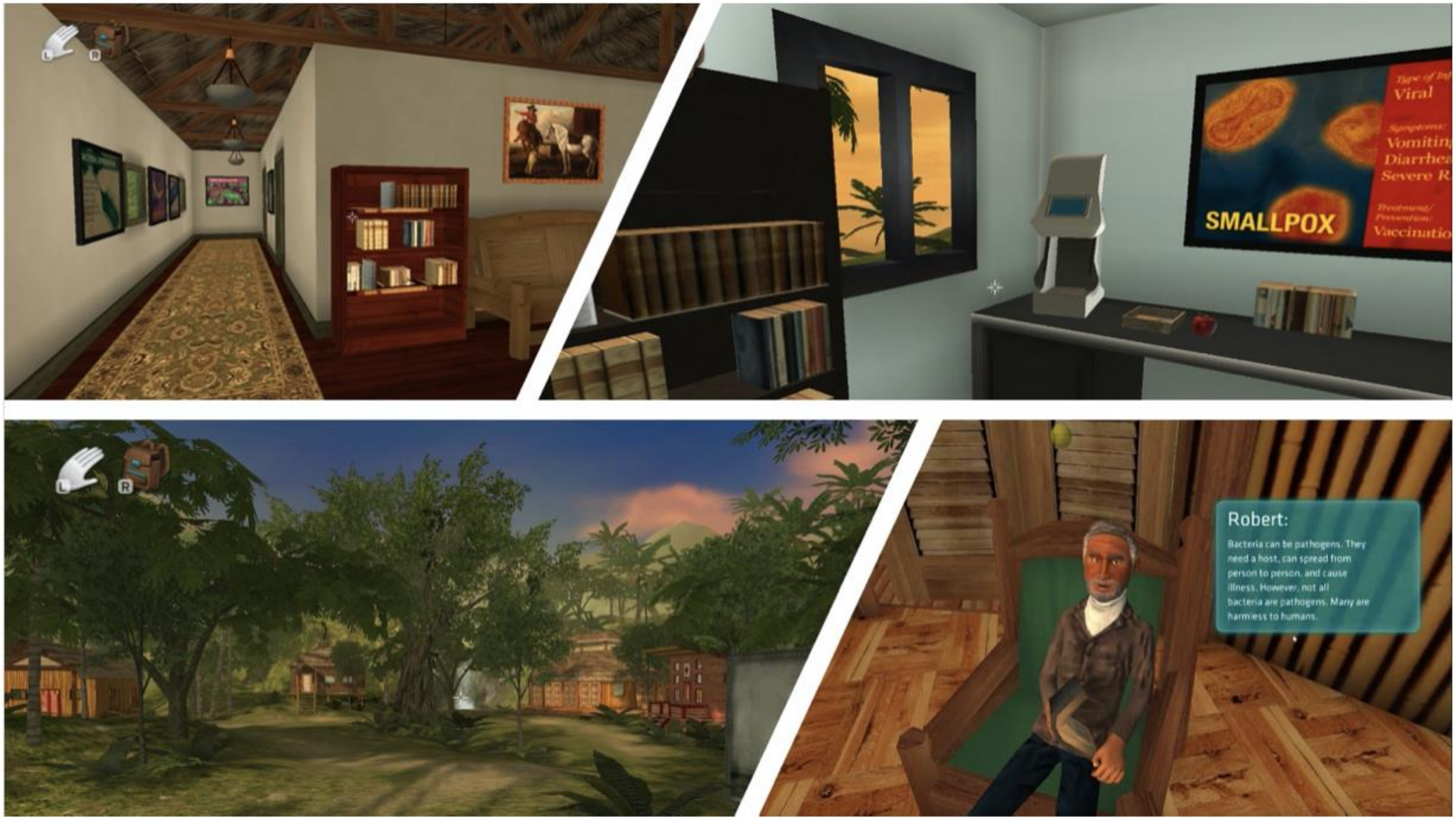
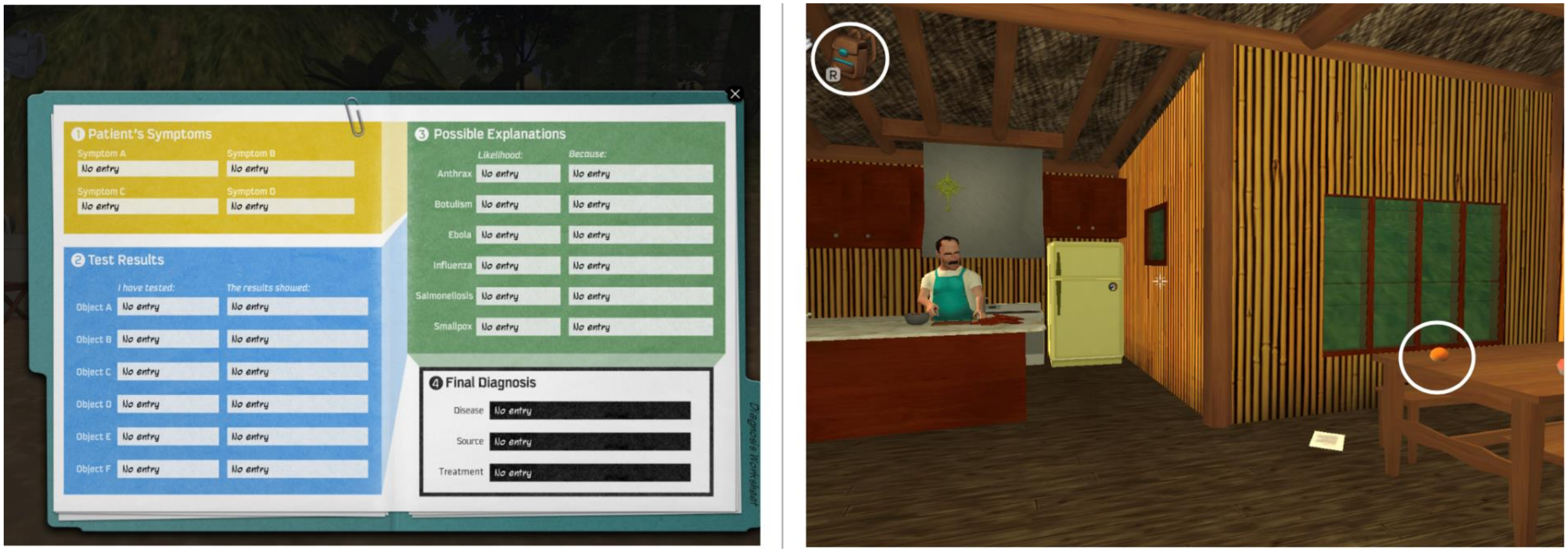

L
R
Type of Inf
Viral
Symptoms:
Vomiting
Diarrhea
Severe R
Treatment/
Prevention:
Vaccinatio
SMALLPOX
L
R
Robert:
Bacteria can be pathogens. They need a host, can spread from person to person, and cause illness. However, not all bacteria are pathogens. Many are harmless to humans.

Figure 1 Crystal Island environment; top left = living quarters; top right = laboratory; bottom left =island; bottom right = dormitory.



1 Patient's Symptoms

Symptom A	No entry
Symptom B	No entry
Symptom C	No entry
Symptom D	No entry

2 Test Results

	I have tested:	The results showed:
Object A	No entry	No entry
Object B	No entry	No entry
Object C	No entry	No entry
Object D	No entry	No entry
Object E	No entry	No entry
Object F	No entry	No entry

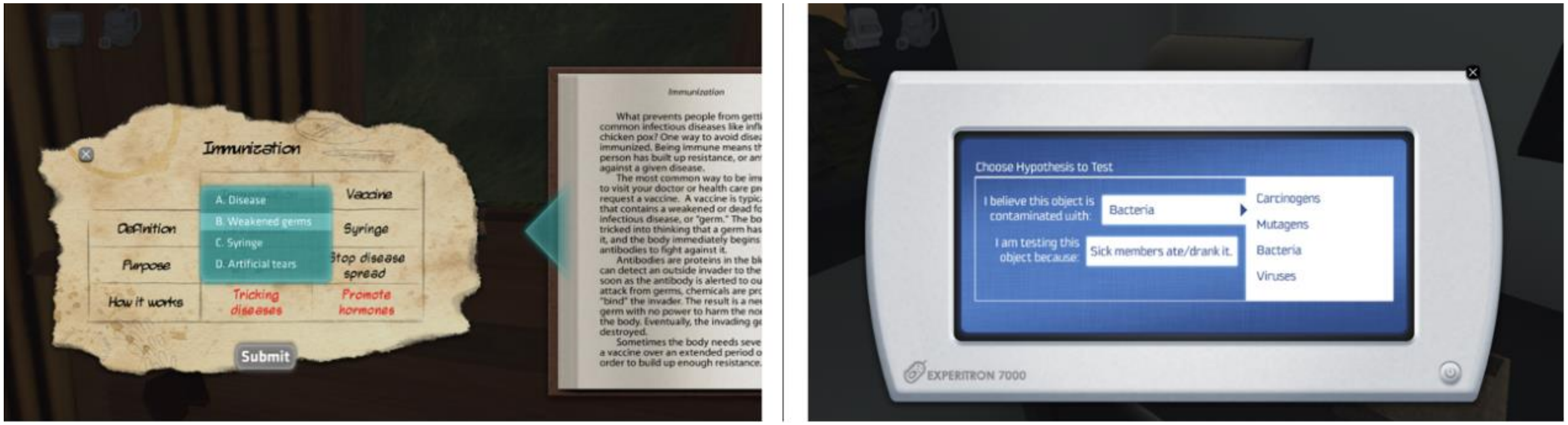
3 Possible Explanations

	Likelihood:	Because:
Anthrax	No entry	No entry
Botulism	No entry	No entry
Ebola	No entry	No entry
Influenza	No entry	No entry
Salmonellosis	No entry	No entry
Smallpox	No entry	No entry

4 Final Diagnosis

Disease	No entry
Source	No entry
Treatment	No entry

Figure 2 Generating hypotheses related game elements; left = diagnostic worksheet; right = dining hall where participants can select food items; circled game elements = Areas of Interest for a food item and backpack.



Immunization

		Vaccine
Definition		Syringe
Purpose		Stop disease spread
How it works	Tricking diseases	Promote hormones

Immunization

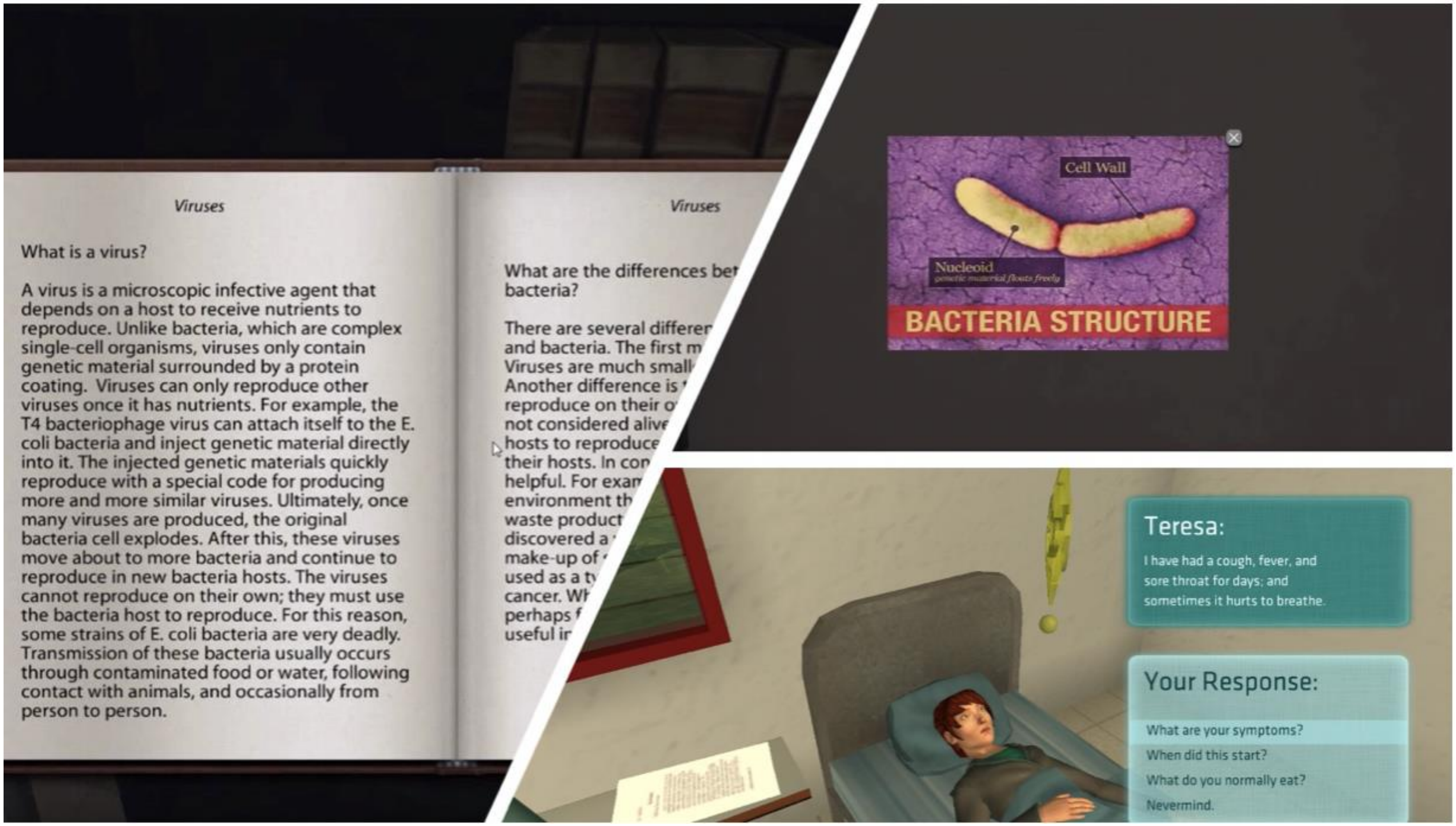
What prevents people from getti
common infectious diseases like infl
chicken pox? One way to avoid dise
immunized. Being immune means th
person has built up resistance, or an
against a given disease.
The most common way to be im
to visit your doctor or health care pr
request a vaccine. A vaccine is typic
that contains a weakened or dead fo
infectious disease, or "germ." The bo
tricked into thinking that a germ has
it, and the body immediately begins
antibodies to fight against it.
Antibodies are proteins in the bl
can detect an outside invader to the
soon as the antibody is alerted to ou
attack from germs, chemicals are pro
"bind" the invader. The result is a ne
germ with no power to harm the no
the body. Eventually, the invading ge
destroyed.
Sometimes the body needs seve
a vaccine over an extended period o
order to build up enough resistance.

Choose Hypothesis to Test

I believe this object is contaminated with: Bacteria

I am testing this object because: Sick members ate/drank it.

Figure 3 Experimental testing related game elements; left = concept matrix content and feedback (displayed in red when participant has incorrect answers); right = scanner for testing food items in laboratory.



Viruses

What is a virus?

A virus is a microscopic infective agent that depends on a host to receive nutrients to reproduce. Unlike bacteria, which are complex single-cell organisms, viruses only contain genetic material surrounded by a protein coating. Viruses can only reproduce other viruses once it has nutrients. For example, the T4 bacteriophage virus can attach itself to the E. coli bacteria and inject genetic material directly into it. The injected genetic materials quickly reproduce with a special code for producing more and more similar viruses. Ultimately, once many viruses are produced, the original bacteria cell explodes. After this, these viruses move about to more bacteria and continue to reproduce in new bacteria hosts. The viruses cannot reproduce on their own; they must use the bacteria host to reproduce. For this reason, some strains of E. coli bacteria are very deadly. Transmission of these bacteria usually occurs through contaminated food or water, following contact with animals, and occasionally from person to person.

Viruses

What are the differences bet
bacteria?

There are several differer
and bacteria. The first m
Viruses are much small
Another difference is
reproduce on their o
not considered alive
hosts to reproduce
their hosts. In con
helpful. For exam
environment th
waste product
discovered a
make-up of
used as a ty
cancer. Wh
perhaps f
useful ir

Teresa:

I have had a cough, fever, and sore throat for days; and sometimes it hurts to breathe.

Figure 4 Gathering information related game elements; left = book; top right = poster; bottom right =sick patient (non-player character) in infirmary.

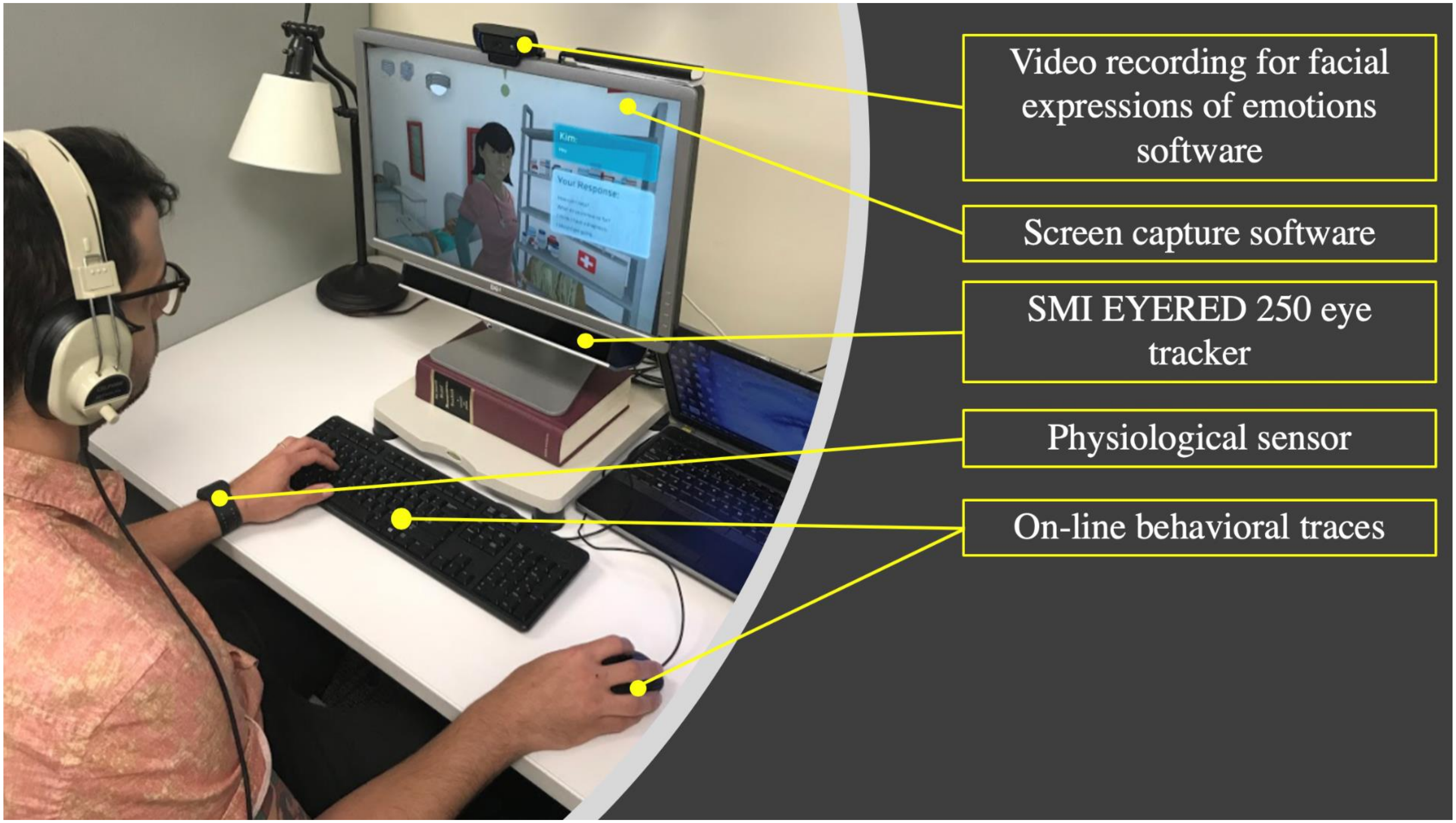
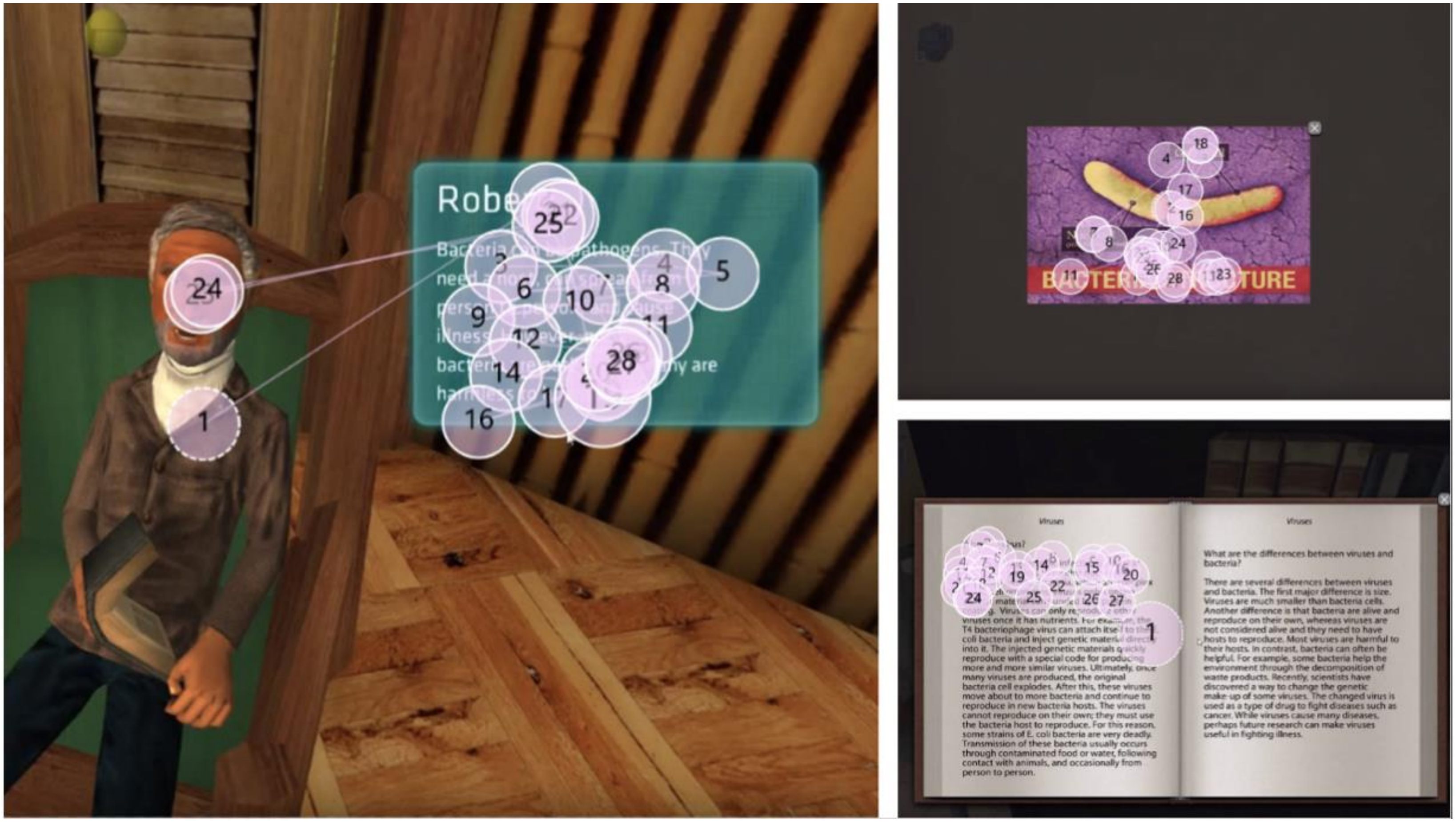

Video recording for facial expressions of emotions software
Screen capture software
SMI EYERED 250 eye tracker
Physiological sensor
On-line behavioral traces

Figure 5 Experimental setup and data channels captured during game-based learning.



Viruses

[illegible] virus?

[illegible] coating. Viruses can only reproduce other viruses once it has nutrients. For example, the T4 bacteriophage virus can attach itself to the E. coli bacteria and inject genetic material directly into it. The injected genetic materials quickly reproduce with a special code for producing more and more similar viruses. Ultimately, once many viruses are produced, the original bacteria cell explodes. After this, these viruses move about to more bacteria and continue to reproduce in new bacteria hosts. The viruses cannot reproduce on their own; they must use the bacteria host to reproduce. For this reason, some strains of E. coli bacteria are very deadly. Transmission of these bacteria usually occurs through contaminated food or water, following contact with animals, and occasionally from person to person.

Viruses

What are the differences between viruses and bacteria?

There are several differences between viruses and bacteria. The first major difference is size. Viruses are much smaller than bacteria cells. Another difference is that bacteria are alive and reproduce on their own, whereas viruses are not considered alive and they need to have hosts to reproduce. Most viruses are harmful to their hosts. In contrast, bacteria can often be helpful. For example, some bacteria help the environment through the decomposition of waste products. Recently, scientists have discovered a way to change the genetic make-up of some viruses. The changed virus is used as a type of drug to fight diseases such as cancer. While viruses cause many diseases, perhaps future research can make viruses useful in fighting illness.

Figure 6 Circles indicate eye fixations while numbers indicate sequence of fixations.

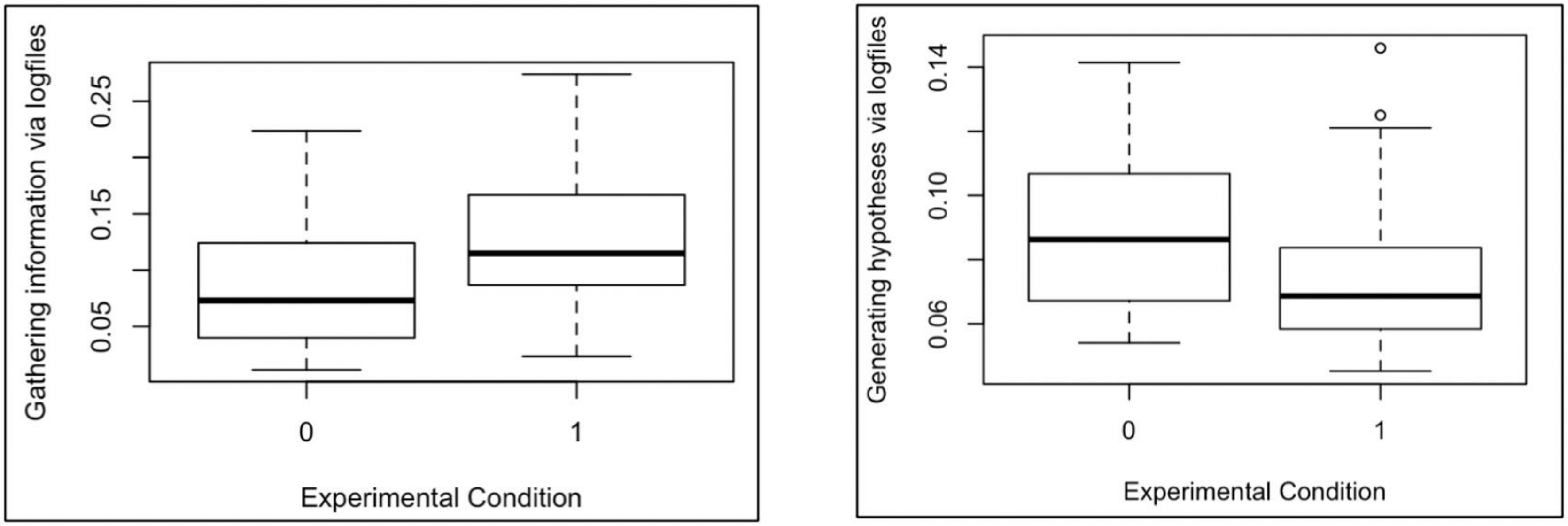


Figure 7 Significant differences in proportion of time spent interacting with elements related to gathering information and generating hypotheses between experimental conditions; 0 = full agency, 1 = partial agency.

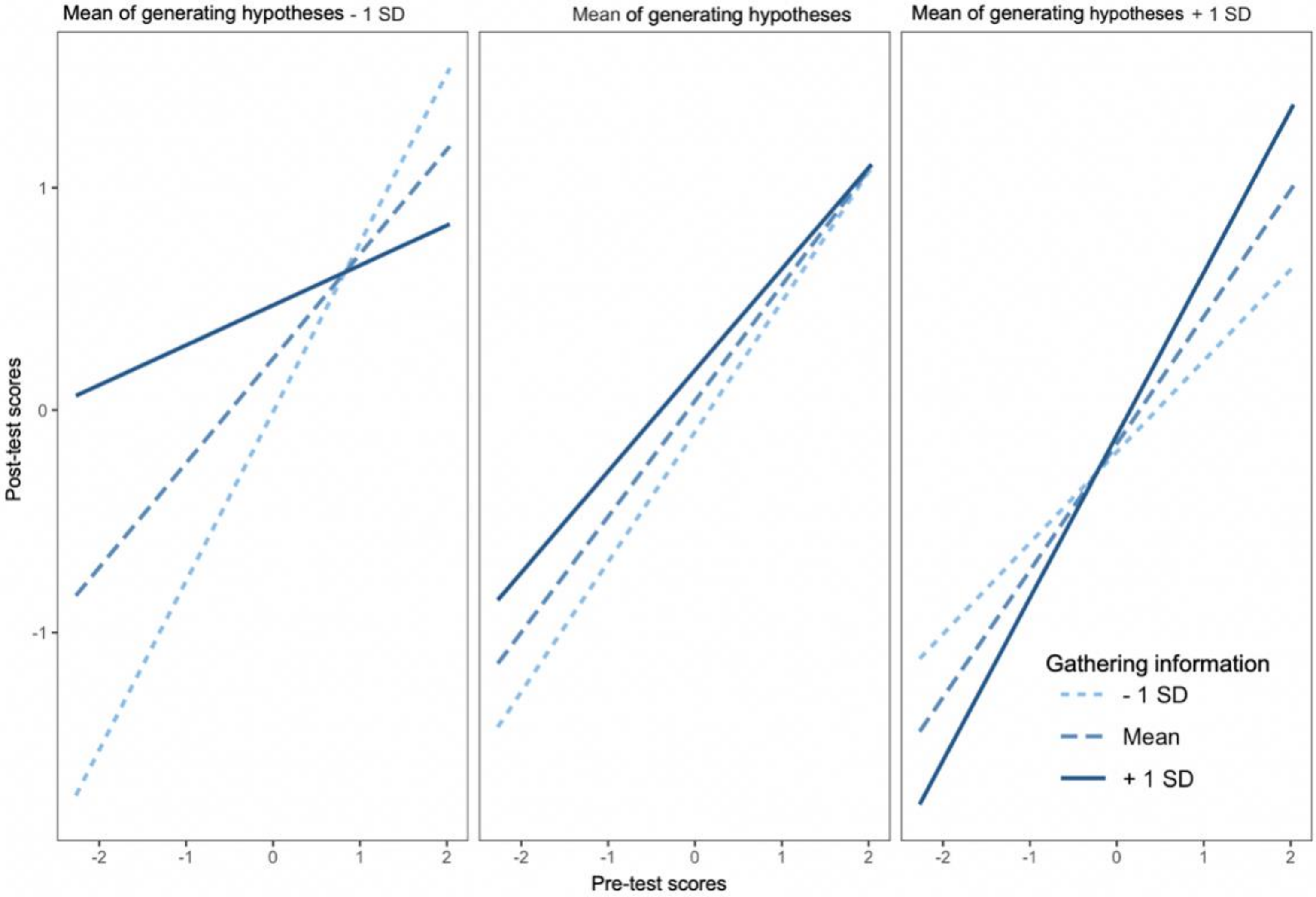


Figure 8 Three-way interaction between pre/post-test scores, proportion of time interacting with game elements related to gathering information and generating hypotheses.

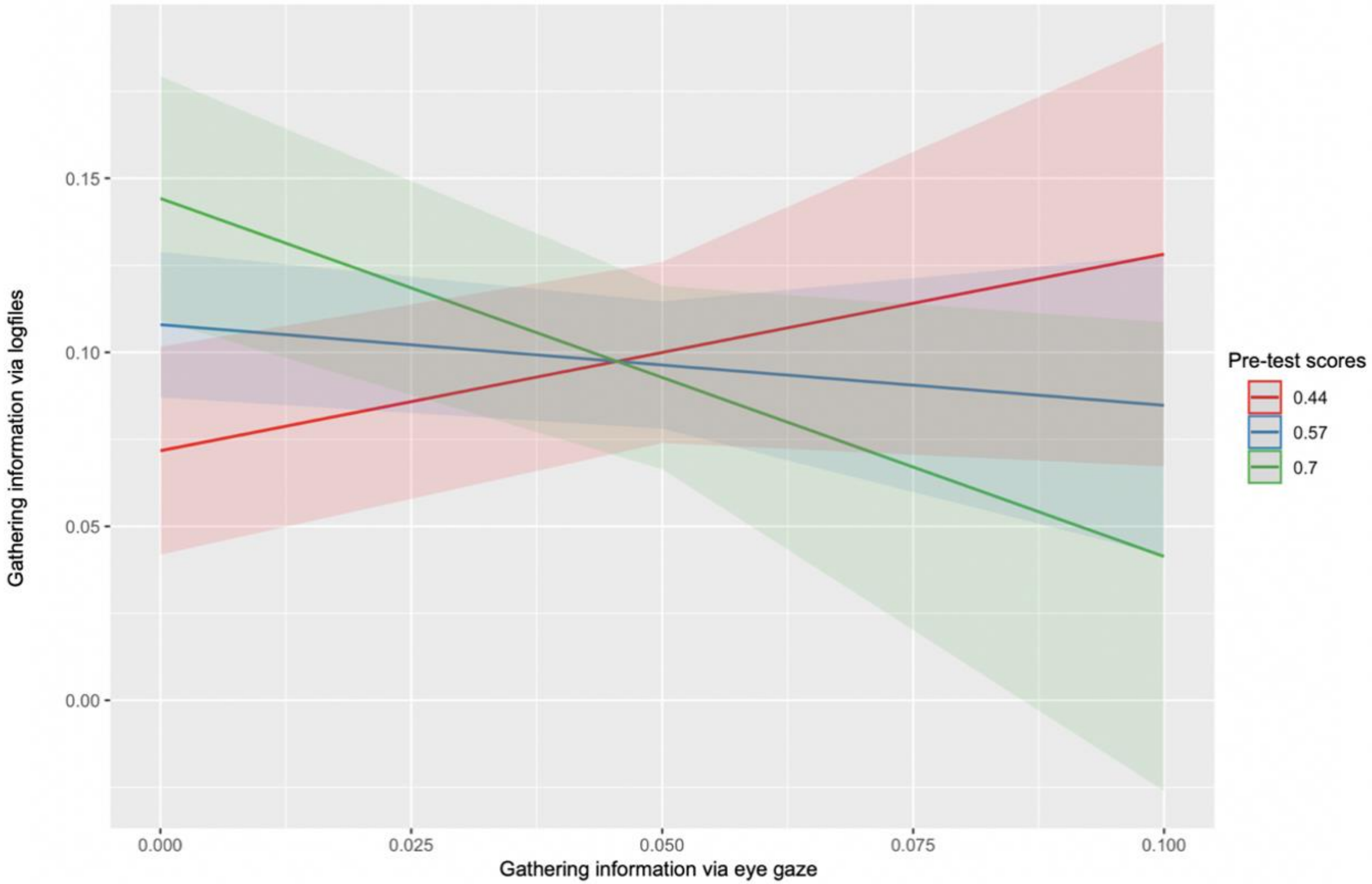


Figure 9 Two-way interaction where pre-test scores moderate relationships between proportion of time interacting with and fixating on elements related to gathering information.

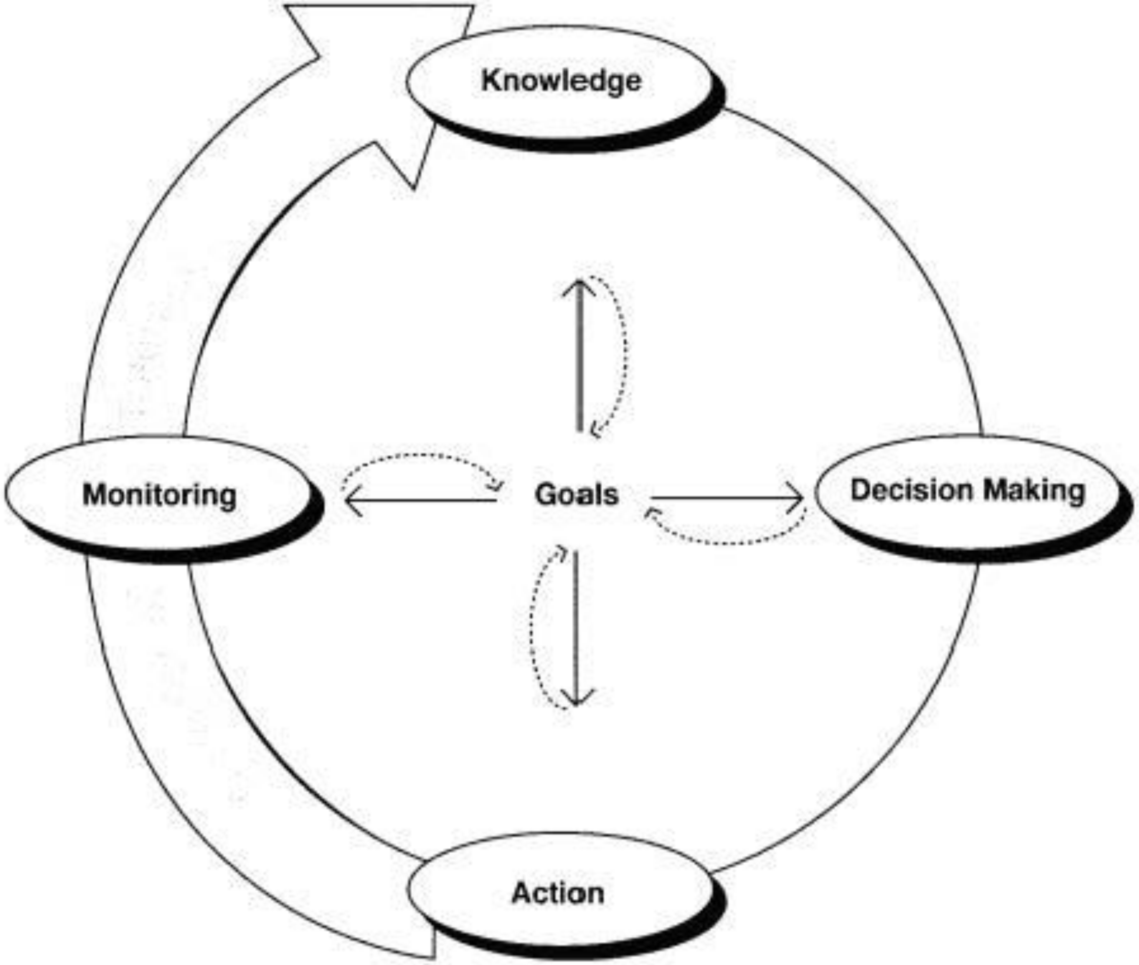


Figure 10 Model of reflection by McAlpine et al. (1999).

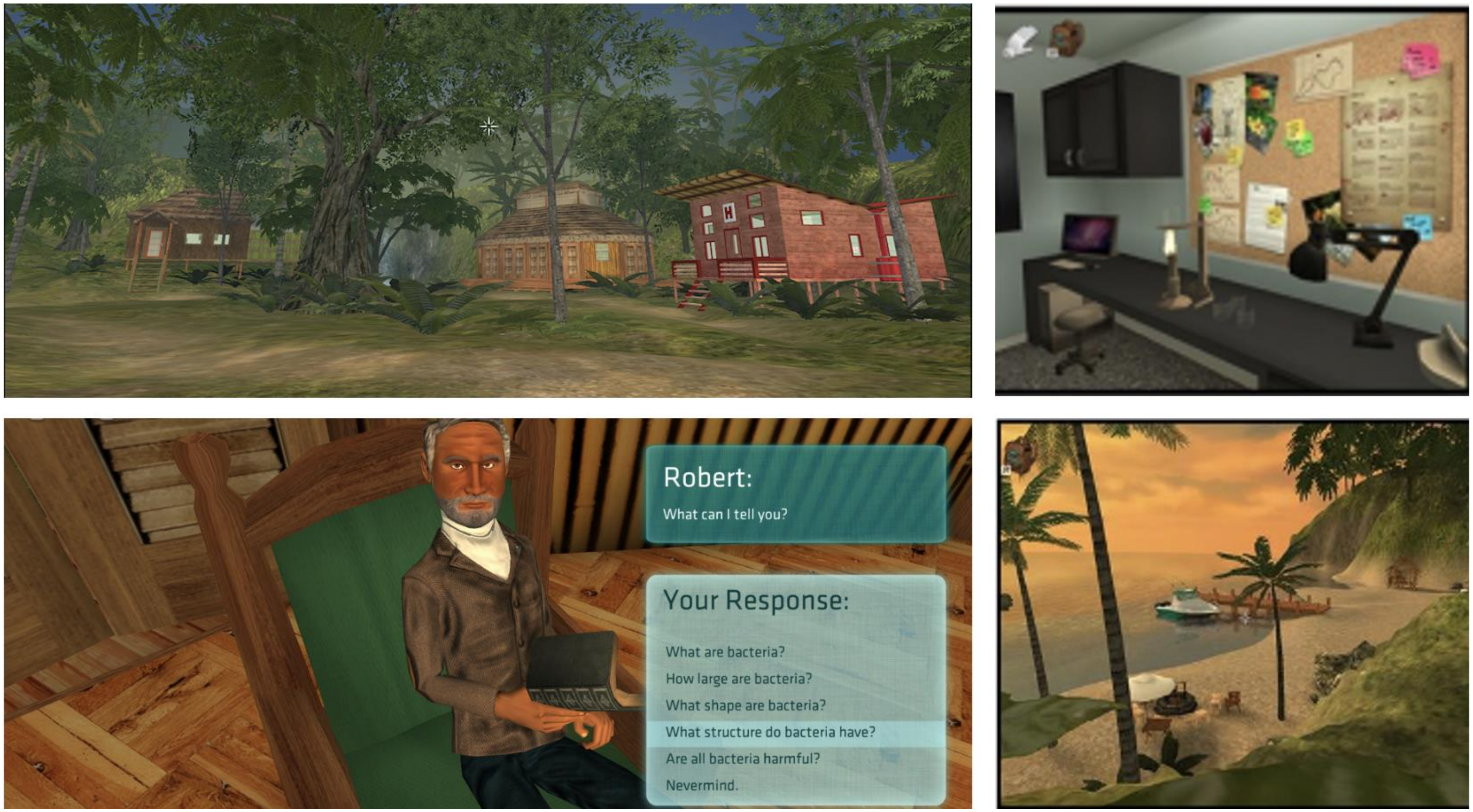
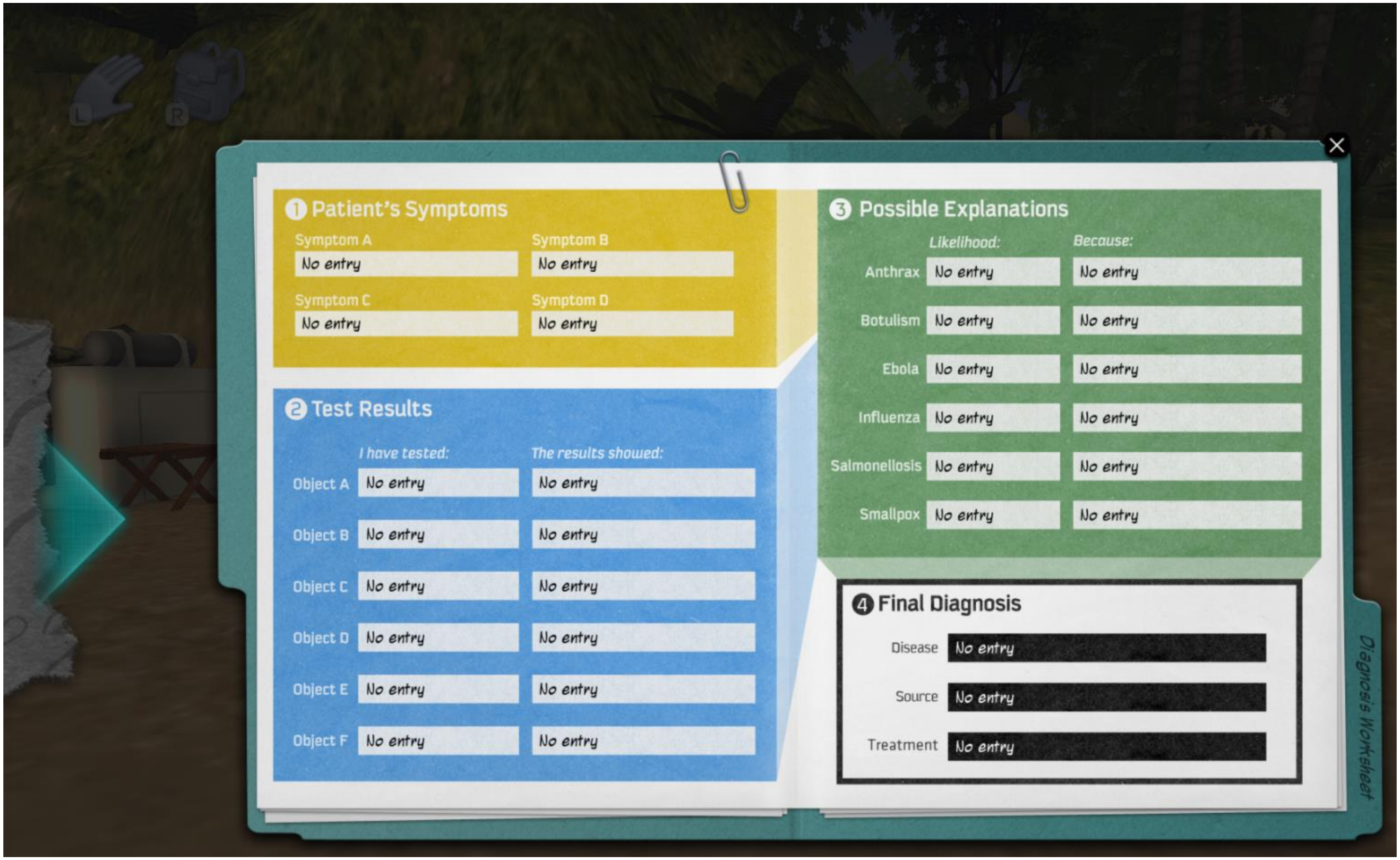


Figure 11 Screenshots of game-based learning environment, Crystal Island.



1 Patient's Symptoms

Symptom A	Symptom B
No entry	No entry
Symptom C	Symptom D
No entry	No entry

2 Test Results

	I have tested:	The results showed:
Object A	No entry	No entry
Object B	No entry	No entry
Object C	No entry	No entry
Object D	No entry	No entry
Object E	No entry	No entry
Object F	No entry	No entry

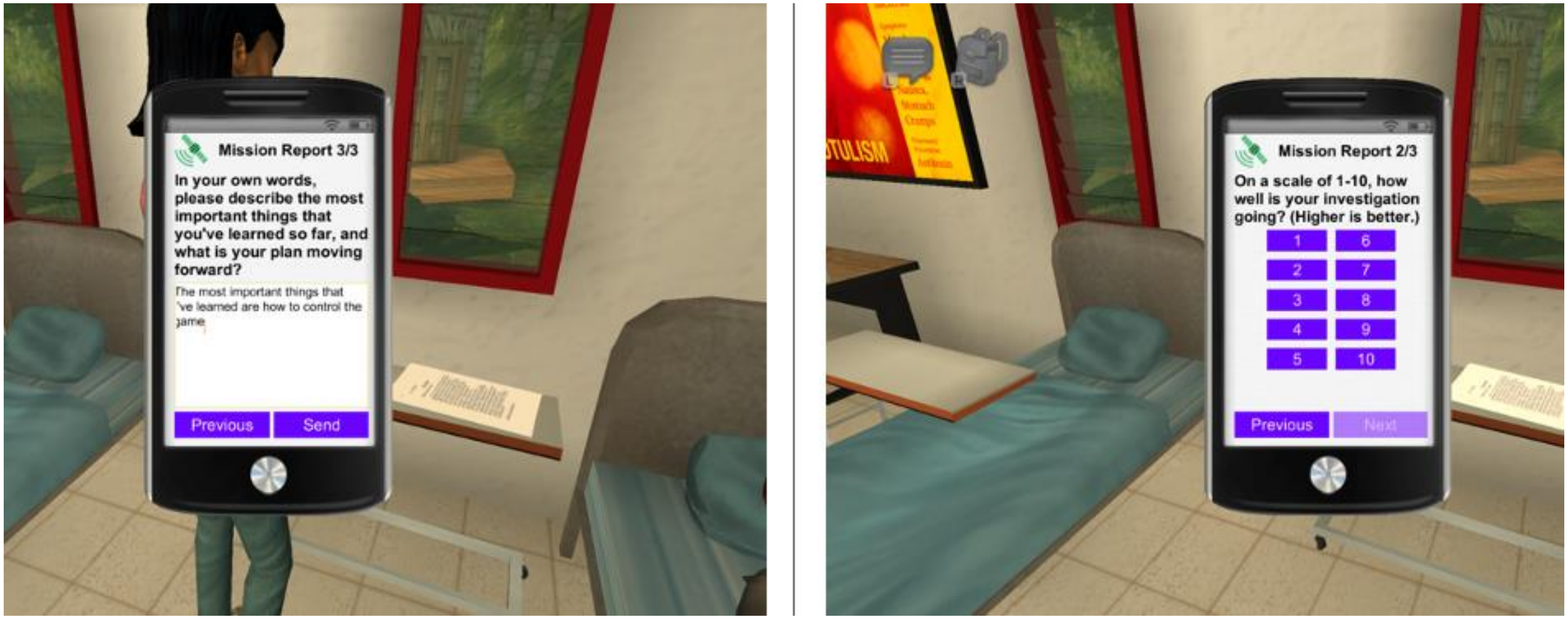
3 Possible Explanations

	Likelihood:	Because:
Anthrax	No entry	No entry
Botulism	No entry	No entry
Ebola	No entry	No entry
Influenza	No entry	No entry
Salmonellosis	No entry	No entry
Smallpox	No entry	No entry

4 Final Diagnosis

Disease	No entry
Source	No entry
Treatment	No entry

Figure 12 Diagnosis worksheet.



Mission Report 3/3

In your own words, please describe the most important things that you've learned so far, and what is your plan moving forward?

The most important things that 've learned are how to control the game

Mission Report 2/3

On a scale of 1-10, how well is your investigation going? (Higher is better.)

Figure 13 Reflection prompt with written component on the left and scale on the right.

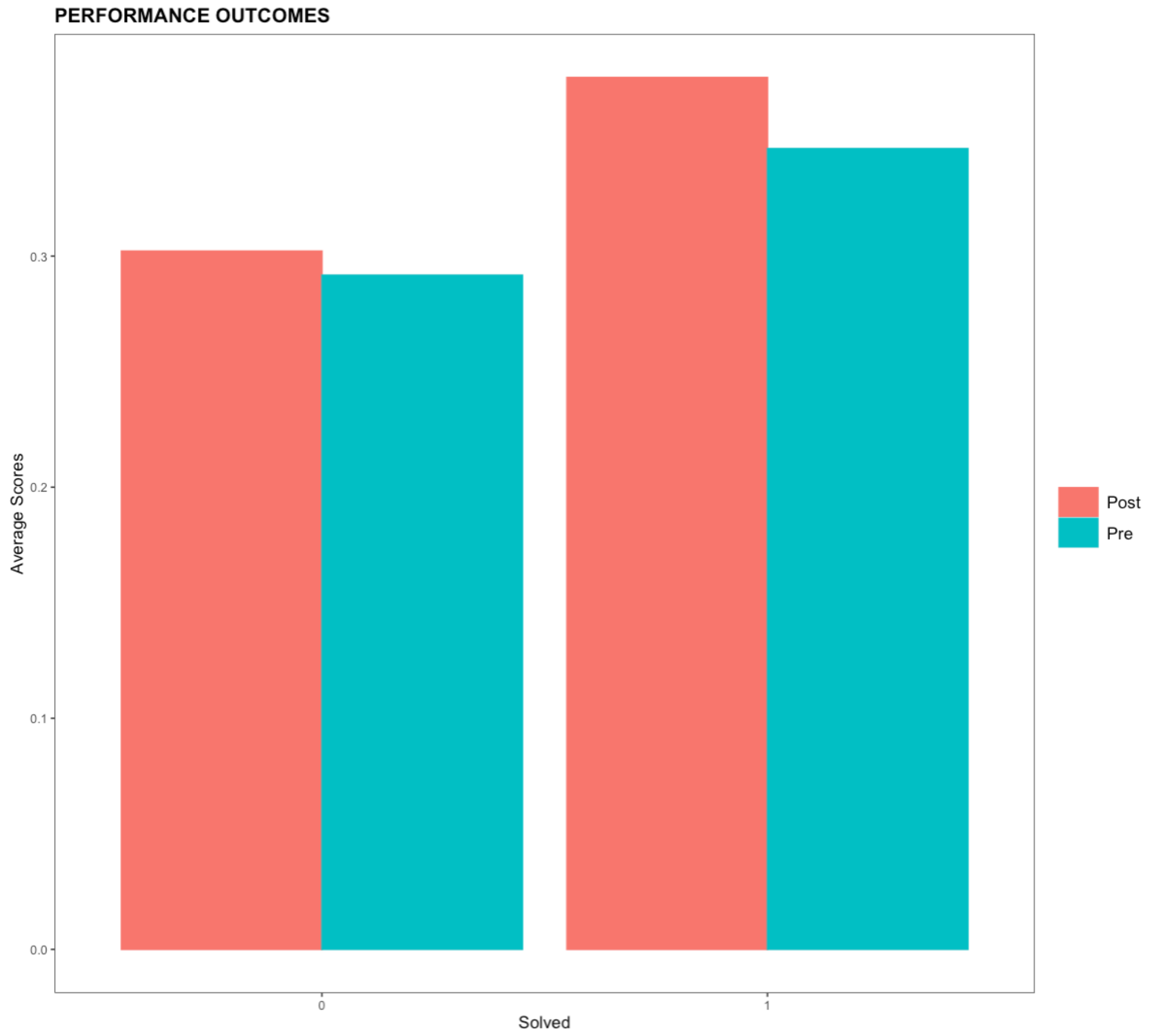


Figure 14 Differences in pre/post-test scores between learners who solved and did not solve the mystery; 1 = solved, 0 = unsolved.

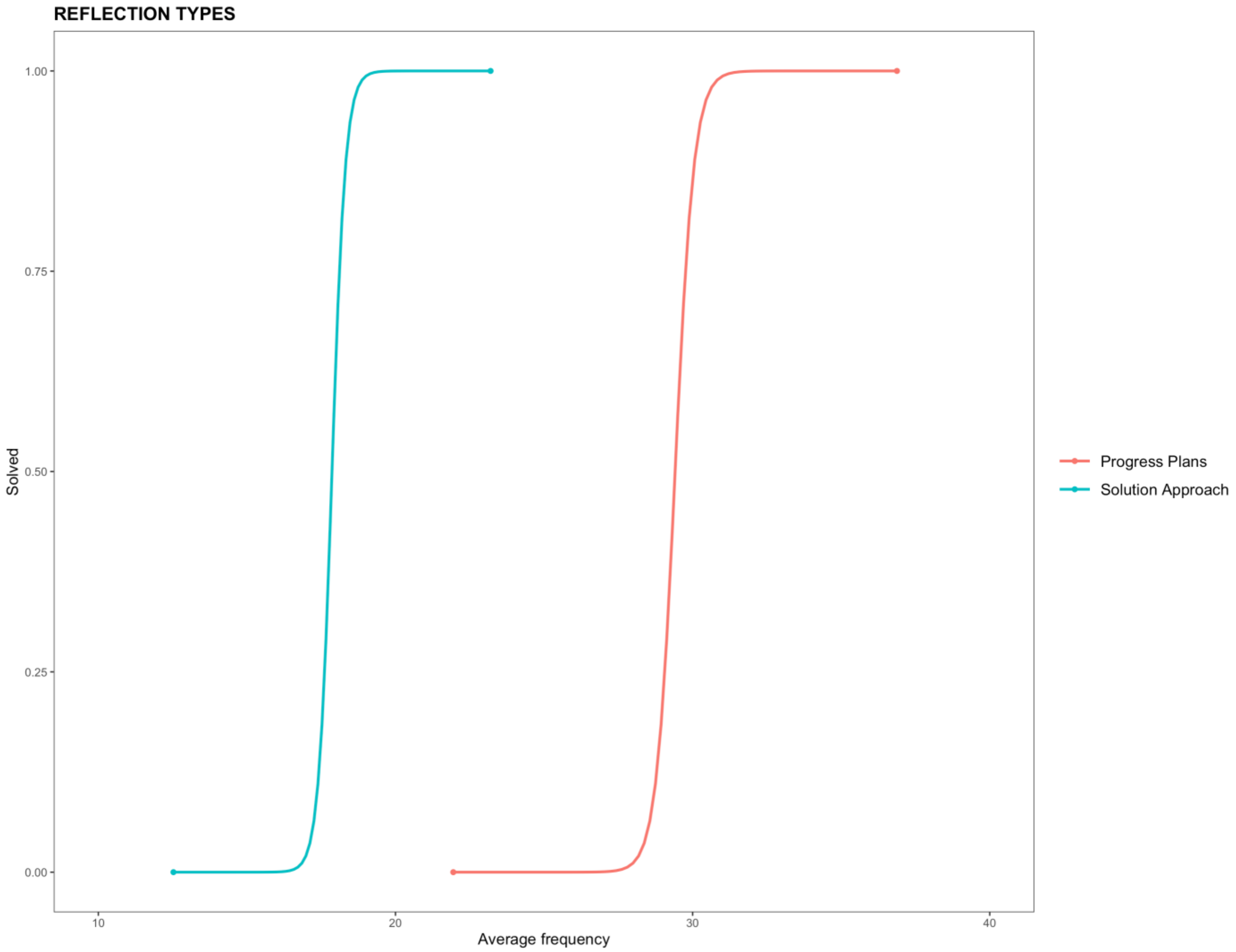


Figure 15 Line graph illustrating relationships between the average frequency of reflection types between learners who did and did not solve the mystery (solved = 1; unsolved = 0).

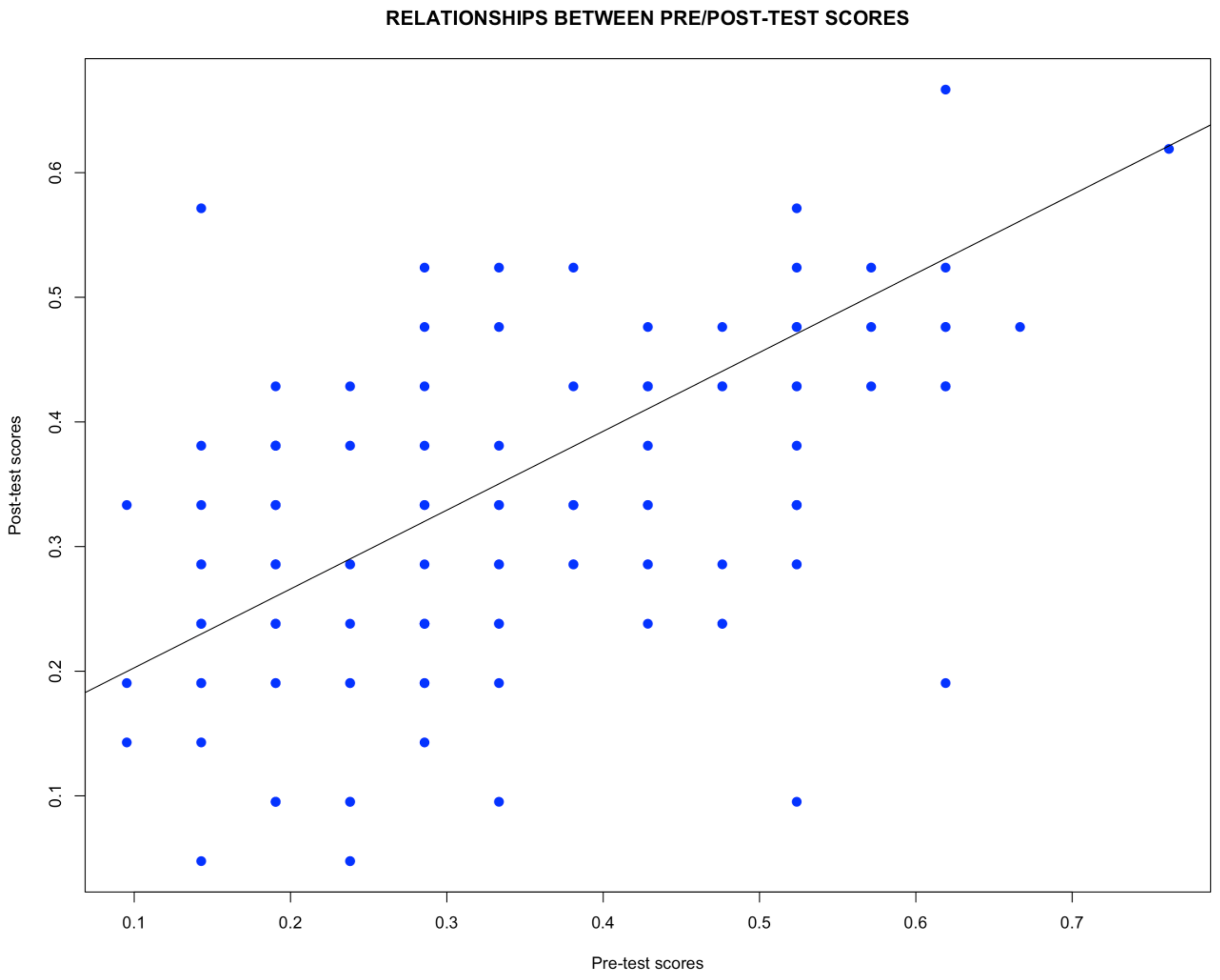


Figure 16 Scatterplot and regression line illustrating relations between pre and post-test scores.

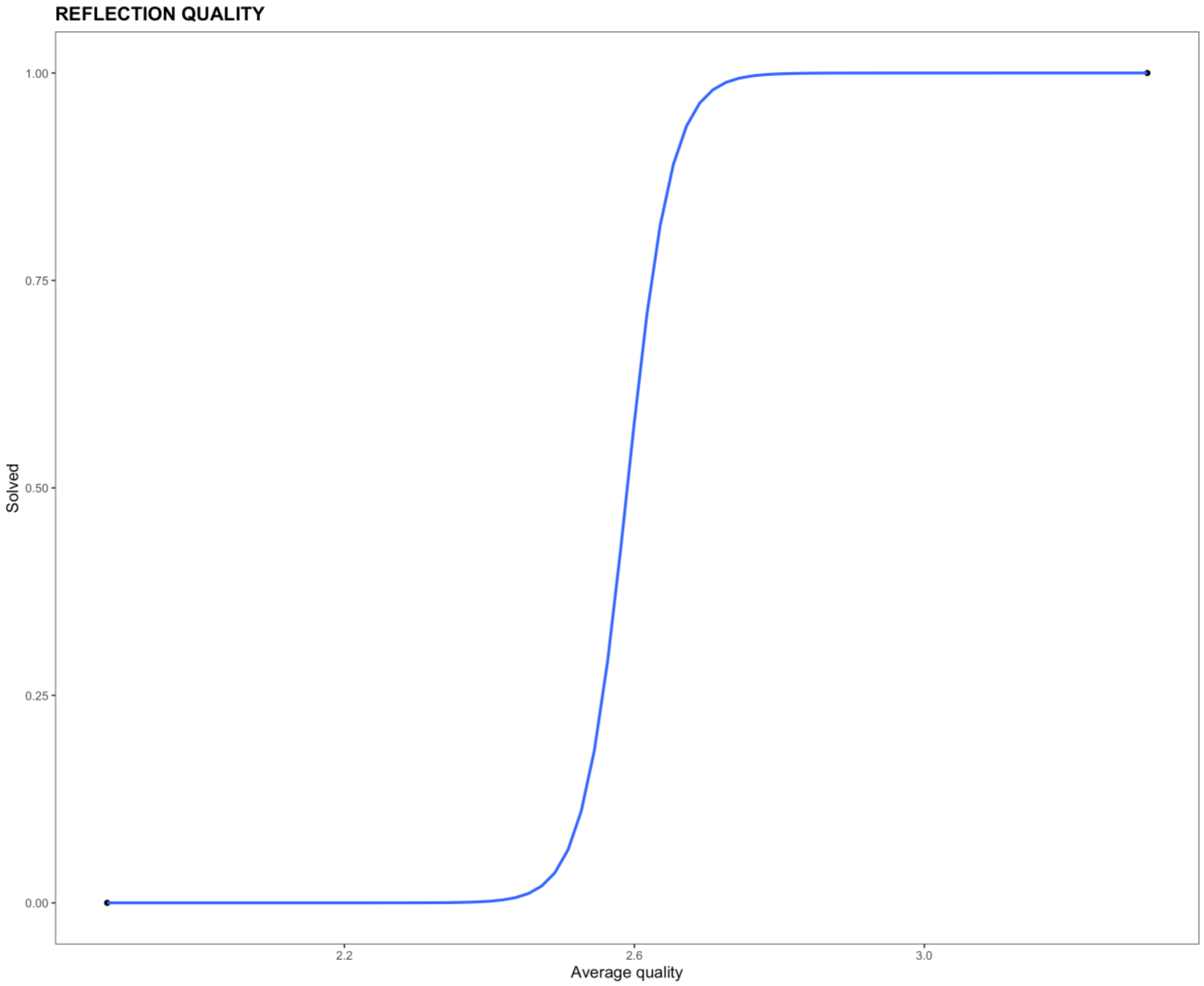


Figure 17 Line graph illustrating relationships between the average quality of solution-approach reflection between learners who did and did not solve the mystery (solved = 1; unsolved = 0).

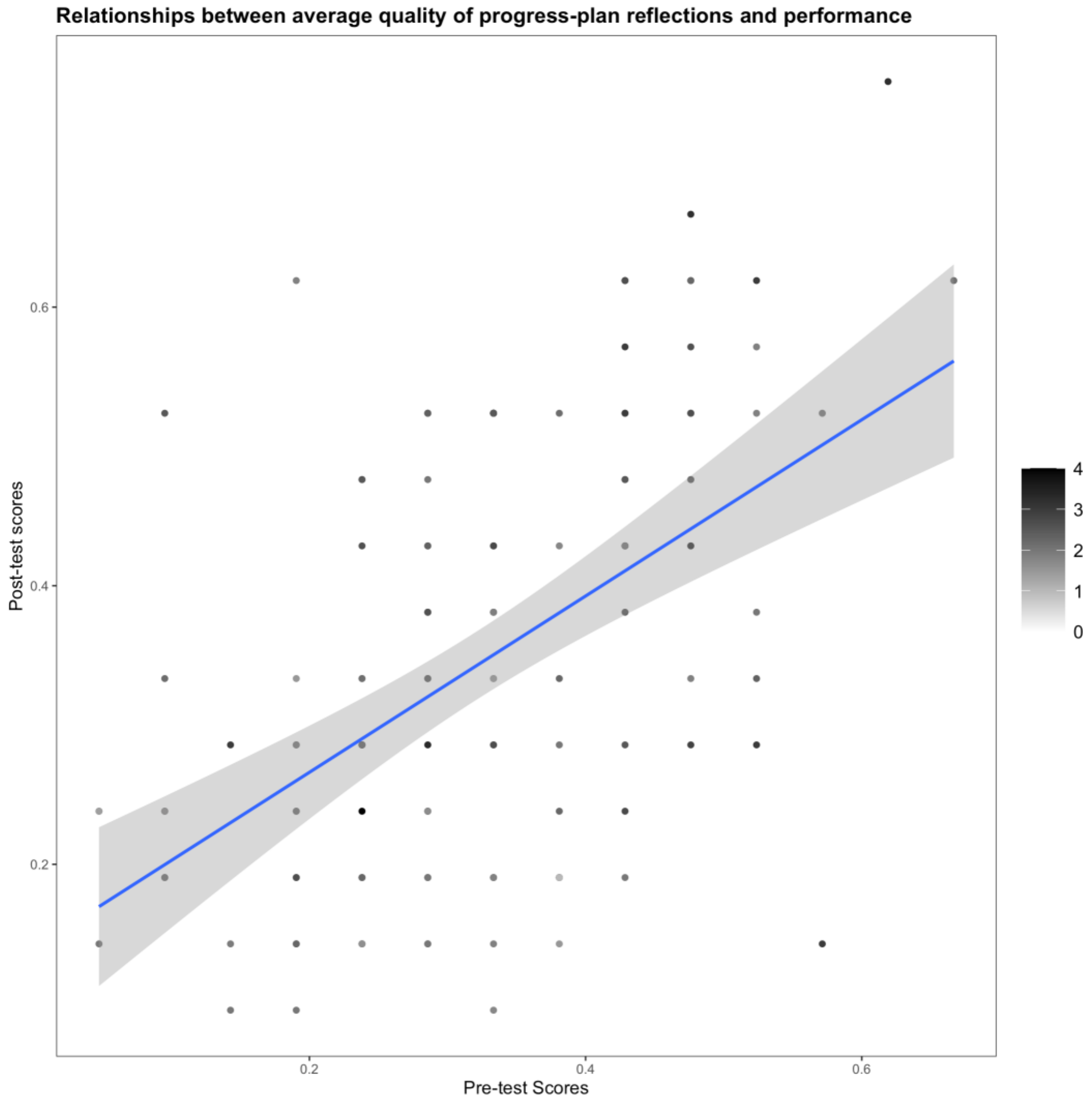


Figure 18 Scatterplot and regression line illustrating relations between pre and post-test scores.

Table 1 Fixations on game elements related to scientific and non-scientific reasoning

Variables	Areas of Interest on Game Elements	Operational Definitions
Information gathering	All books, research articles, posters, and non-player character	Time fixating on game elements for more than 250 milliseconds without selecting, opening, editing, or viewing content of game elements/total time during game-based learning
Hypothesis generation	Food items, diagnosis worksheet, and backpack	
Experimental testing	Concept matrices and scanner	
Non-scientific reasoning	Doors, buildings, trees, plants, pots, settings, rocks, keys, menu icons, lights (e.g., bulbs), turbine, firepit, treads, trophies, water, kitchen, appliances (e.g., refrigerator), lab equipment (e.g., flasks), windows, and furniture	

Table 2 Interactions with game elements related to scientific and non-scientific reasoning

Variables	**Areas of Interest on Game Elements**	**Operational Definitions**
Information gathering	All book content, research article content, poster content, and non-player character dialogue	Time fixating on game elements for more than 250 milliseconds while selecting, opening, editing, or viewing content of game elements/total time during game-based learning
Hypothesis generation	Food items scanned and diagnosis worksheet	
Experimental testing	Concept matrix content, concept matrix feedback, and scanner	
Non-scientific reasoning	Movement to different locations during game-based learning	

Table 3 Descriptive statistics

Variables	**Condition**	**Mean (SD)**	**Median**	**1st Quartile**	**3rd Quartile**
Information gather AOI	1	0.03 (0.03)	0.02	0.014	0.042
	0	0.02 (0.02)	0.02	0.007	0.038
Information gather LOG	1	0.12 (0.06)	0.1	0.070	0.152
	0	0.08 (0.05)	0.08	0.041	0.113
Hypothesis generation AOI	1	0.04 (0.03)	0.02	0.012	0.044
	0	0.03 (0.03)	0.03	0.017	0.041
Hypothesis generation LOG	1	0.09 (0.03)	0.07	0.061	0.106
	0	0.09 (0.03)	0.09	0.068	0.111
Experimental test AOI	1	0.00 (0.00)	0.00	0.00	0.003
	0	0.00 (0.00)	0.00	0.00	0.001
Experimental test LOG	1	0.21 (0.09)	0.18	0.142	0.279
	0	0.22 (0.1)	0.19	0.145	0.261
Pre-test scores	1	0.56 (0.14)	0.62	0.476	0.667
	0	0.55 (0.13)	0.52	0.48	0.62

Post-test scores	1	0.73 (0.12)	0.76	0.667	0.810
	0	0.70 (0.13)	0.71	0.62	0.76
Total game play	1	5586.15 (967.59)	6639	4616	6247
	0	4556.51 (895.45)	4345	3974	5098

Table 4 Modeling relationships with post-test scores; LOG = fixation and interactions with game elements; AOI = fixations on game elements without interaction; $^{*}p= 0.05$; $^{***}p<0.05$

Predictor Variables	*β*	Standard Error	*t*
Information gathering AOI	---	---	---
Hypothesis generation AOI	---	---	---
Experimental testing AOI	---	---	---
Non-scientific reasoning AOI	---	---	---
Condition	---	---	---
Pre-test scores	0.43^{***}	0.10	4.48
Pre-test*Condition	---	---	---
Adjusted R^2		0.222	
F		20.08^{***}	

Table 5 Johnson-Neyman simple slope analysis modeling relationships with post-test scores

Pre-test Scores

Moderating Variables	β	Standard Error	t
Hypothesis generation -1 SD			
Information gathering -1 SD	0.76^{***}	0.25	2.99
Information gathering mean	0.47^{***}	0.16	2.91
Information gathering +1 SD	0.18	0.26	0.70
Hypothesis generation mean			
Information gathering -1 SD	0.58^{***}	0.16	3.77
Information gathering mean	0.52^{***}	0.11	4.75
Information gathering +1 SD	0.46^{***}	0.17	2.67
Hypothesis generation +1 SD			
Information gathering -1 SD	0.41^{***}	0.13	3.04
Information gathering mean	0.57^{***}	0.15	3.90
Information gathering +1 SD	0.73^{***}	0.23	3.17

Table 6 Modeling in-game actions related to scientific reasoning; $^{*}p= 0.05$; $^{***}p<0.05$

Predictor Variables	β	Standard Error	t
Information gathering AOI	---	---	---
Condition	-0.03***	0.12	-0.22
Pre-test scores	---	---	---
Pre-test*Condition	---	---	---
Pre-test*Information AOI	-0.26***	0.12	-2.20
Adjusted R^2		0.189	
F		3.235***	
Hypothesis generation AOI	0.96	0.53	1.807
Condition	-0.02***	0.04	-0.425
Pre-test scores	---	---	---
Pre-test*Condition	---	---	---
Pre-test*Hypothesis AOI	---	---	---
Adjusted R^2		0.113	
F		2.23***	

Table 7 Reflection triggers, prompts, and types during game-play

Triggers	**Prompts**	**Types**
After talking to the camp nurse for the first time	*"Agent, it looks like you've spoken with the camp nurse. Before you continue, we'd like a report on your progress. In your own words, please describe the most important things that you've learned so far, and what is your plan moving forward?"*	Progress plan
After viewing 6 virtual texts in the game	*"Agent, it looks like you've found several materials that may be useful. Before you continue, we'd like a report on your progress. In your own words, please describe the most important things that you've learned so far, and what is your plan moving forward?"*	Progress plan
After obtaining a positive test result in the virtual laboratory	*"Agent, it looks like you found an object that tested positive for pathogenic contaminants. Before you continue, we'd like a report on your progress. In your own words, please describe the most important things that you've learned so far, and what is your plan moving forward?"*	Progress plan

After submitting diagnosis worksheet to camp nurse and getting it wrong	*“Agent, it looks like you are making progress on diagnosing the illness, but you’re not quite there yet. In your own words, please describe the most important things that you’ve learned so far, and what is your plan moving forward?”*	Progress plan
After solving the mystery	*“Well done, Agent! You’ve saved everyone on the island. Now that you are finished, we would like to ask a couple of final questions. Please explain how you approached solving the mystery.”*	Solution approach
After solving the mystery	*“If you were asked to solve a similar problem in the future, what would you do the same and/or differently?”*	Different problem approach
After time expires, but the learner has not solved the mystery	*“Thank you for playing Crystal Island. Now that you are finished, we would like to ask a couple of final questions. Please explain how you approached solving the mystery.”*	Solution approach
After time expires, but the learner has not solved the mystery	*“If you were asked to solve a similar problem in the future, what would you do the same and/or differently?”*	Different problem approach

Table 8 Rubric used to annotate the depth of reflection prompt responses (Carpenter, Geden, et al., 2020)

Ratings	**Characteristics**	**Examples**
1	Lacks both a plan and knowledge; abstract and largely meaningless; unactions.	*"Each clue will help with solving the problem"* or *"Yeah cool game I learned science"*
2	Presents a vague hypothesis or plan with no clear reasoning; simply restates information that was directly learned in the game without high-order thinking (e.g., inference making).	*"That the illness causing the people being sick might be a pathogen"* or *"I found out that the egg has bacteria"* or *"I think I am going to talk to other people"*
3	Presents a clear hypothesis or a plan, but does not provide reasoning behind it; demonstrates awareness of gaps in under-standing and presents a plan to address those gaps; organizes the importance of their knowledge or problem-solving strategies.	*"Getting more information about the food I think it has something to do with the food"* or *"The most important thing is how the illness is spreading"*
4	Presents a clear hypothesis or plan with reasoning; provide reasoning of the situation with	*"I plan on questioning the cook as they know more about the food and how it could be contaminated with viruses or*

	a plan; acknowledge what they have learned, why it is important, and what they plan to do with this information moving forward.	*bacteria"* or *"I need to learn more about what the sick people do on a day-to-day schedule"*
5	Present both a clear hypothesis and plan with reasoning; presents a high-quality sequence of abstract plans for problem solving.	*"I think that it might have to do with salmonella because when I tested the milk it was positive with pathogenic bacteria. I think that I will test things that can be contaminated"* or *"I will continue to test the foods the sick people touched or previously ate to see if it is contaminated"*

Table 9 Descriptive statistics.

Variables	***M***	***SD***	**Range**
Different Problem Approach Quantity	1*	0	1
Different Problem Approach Quality	1.97	1.04	4.5
Solution Approach Quantity	1*	0	1
Solution Approach Quality	2.65	1.42	5
Progress Plan Quantity	24*	23.04	36
Progress Plan Quality	2.21	0.61	4
Pre-test scores	0.32	0.13	0.62
Post-test scores	0.34	0.16	0.67

Note. * = median.

Table 10 Frequencies on mystery solving.

	Frequency	
	Yes	No
Solved the Mystery	65 (54.62%)	55 (45.38%)

Table 11 e^{β} = exponentiated beta or odds ratio; *SE*= standard error; $^{*}p < 0.025$.

Predictor	***β***	***SE***	***p***	**e^{β}**
Constant	-3.14	0.81	0.000	0.04
Progress plans	-0.06^{*}	0.03	0.000	0.94
Test		χ^2	***df***	***p***
Overall model evaluation				
Likelihood ratio test		43.145	2	< 0.001
Goodness-of-fit test				
Hosmer & Lemeshow		8.5243	8	0.384
McFadden R^2	0.261			

Note. e^{β} = exponentiated beta or odds ratio; *SE*= standard error; $^{*}p < 0.025$.

Table 12 Multiple linear regression analysis of post-test scores by frequency of reflection types and pre-test scores.

Predictor	***β***	***SE***
Constant	1.39*	0.033
Pre-test scores	0.633*	0.095
Test		
F		44.44*
df		1,118
Adjusted R^2		0.267

Note. *SE* = standard error; $^*p < 0.025$.

Table 13 Logistic regression analysis of solving mystery by quality of reflection types.

Predictor	$\boldsymbol{\beta}$	***SE***	$\mathbf{e}^{\beta}$
Constant	-2.20*	0.53	0.11
Solution approach	0.902*	0.19	2.46
Test		χ^2	*p*
Likelihood ratio test		43.145	< 0.001
Goodness-of-fit test			
Hosmer & Lemeshow		3.8462	0.8707
McFadden R^2	0.205		

Note. e^{β} = exponentiated beta or odds ratio; *SE*= standard error; $^{*}p < 0.025$.

Table 14 Multiple linear regression analysis of post-test scores by average quality of reflection types and pre-test scores.

Predictor	***β***	***SE***
Constant	0.010	0.050
Pre-test scores	0.533*	0.095
Progress plans	0.054*	0.022
Solution approach	0.014	0.009
Test		
F		20.75*
df		3,116
Adjusted R^2		0.3324

LIST OF REFERENCES

Accreditation Council of Graduate Medical Education (ACGME). (2021). *Common program requirements (residency)*. https://www.acgme.org/What-We-Do/Accreditation/Common-Program-Requirements/. Last accessed on 01 July 2021.

Ahmidi, N., Ishii, M., Fichtinger, G., Gallia, G. L., & Hager, G. D. (2012). An objective and automated method for assessing surgical skill in endoscopic sinus surgery using eye-tracking and tool-motion data. *International Forum of Allergy & Rhinology*, 2(6), 507-515).

Ainley, M., & Ainley, J. (2019). Motivation and learning: Measures and methods. In K. A. Renninger & S. E. Hidi (Eds.), *The Cambridge handbook of motivation and learning* (pp. 665-688). Cambridge University Press.

Ak, O., & Kutlu, B. (2017). Comparing 2D and 3D game-based learning environments in terms of learning gains and student perceptions. *British Journal of Educational Technology*, *48*(1), 129-144.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723.

Alonso-Fernández, C., Cano, A. R., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2019). Lessons learned applying learning analytics to assess serious games. *Computers in Human Behavior*, *99*, 301-309.

Alzaid, M., & Hsiao, I. H. (2018). Effectiveness of reflection on programming problem solving self-assessments. In FIE'18: *Proceedings of IEEE Frontiers in Education Conference* (pp. 1-5). IEEE.

Amon, M. J., Vrzakova, H., & D'Mello, S. K. (2019). Beyond dyadic coordination: Multimodal behavioral irregularity in triads predicts facets of collaborative problem solving. *Cognitive Science*, *43*(10), e12787.

Anderson, J. R., Betts, S., Bothell, D., & Lebiere, C. (2021). Discovering skill. *Cognitive Psychology*, *129*, 101410.

Artino Jr, A. R., Cleary, T. J., Dong, T., Hemmer, P. A., & Durning, S. J. (2014). Exploring clinical reasoning in novices: A self-regulated learning microanalytic assessment approach. *Medical Education*, *48*(3), 280-291.

Ashraf, H., Sodergren, M. H., Merali, N., Mylonas, G., Singh, H., & Darzi, A. (2018). Eye-tracking technology in medical education: A systematic review. *Medical Teacher*, *40*(1), 62-69.

Azevedo, R. (2020). Reflections on the field of metacognition: Issues, challenges, and opportunities. *Metacognition and Learning*, *15*, 91-98.

Azevedo, R., Feyzi-Behnagh, R., Duffy, M., Harley, J., & Trevors, G. (2012). Metacognition and self-regulated learning in student-centered learning environments. In D. Jonassen & S. Land (Eds.), *Theoretical foundations of learning environments* (pp. 171-197). Routledge.

Azevedo, R., & Gašević, D. (2019). Analyzing multimodal multichannel data about self-regulated learning with advanced learning technologies: Issues and challenges. *Computers in Human Behavior*, *96*, 207-210.

Azevedo, R. & Lajoie, S. P. (1998). The cognitive basis for the design of a mammography interpretation tutor. *International Journal of Artificial Intelligence in Education*, *9*, 32-44.

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

Azevedo, R., Taub, M., & Mudrick, N. V. (2018). Understanding and reasoning about real-time cognitive, affective, and metacognitive processes to foster self-regulation with advanced learning technologies. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (pp. 254-270). Routledge.

Bakhtiar, A., Webster, E. A., & Hadwin, A. F. (2018). Regulation and socio-emotional interactions in a positive and a negative group climate. *Metacognition and Learning*, *13*(1), 57-90.

Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 423-443.

Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual Review of Psychology*, *52*(1), 1-26.

Bannert, M., & Reimann, P. (2012). Supporting self-regulated hypermedia learning through prompts. *Instructional Science*, *40*(1), 193-211.

Batalden, P., Leach, D., Swing, S., Dreyfus, H., & Dreyfus, S. (2002). General competencies and accreditation in graduate medical education. *Health Affairs*, *21*(5), 103-111.

Birt, J., Moore, E., & Cowling, M. (2017). Improving paramedic distance education through mobile mixed reality simulation. *Australasian Journal of Educational Technology*, *33*(6), 69-83.

Biswas, G., Segedy, J. R., & Bunchongchit, K. (2016). From design to implementation to practice a learning by teaching system: Betty’s Brain. *International Journal of Artificial Intelligence in Education*, *26*(1), 350-364.

Blikstein, P., & Worsley, M. (2016). Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. *Journal of Learning Analytics*, *3*(2), 220-238.

Blikstein, P., Worsley, M., Piech, C., Sahami, M., Cooper, S., & Koller, D. (2014). Programming pluralism: Using learning analytics to detect patterns in the learning of computer programming. *Journal of the Learning Sciences*, *23*(4), 561-599.

Bondareva, D., Conati, C., Feyzi-Behnagh, R., Harley, J. M., Azevedo, R., & Bouchet, F. (2013). Inferring learning from gaze data during interaction with an environment to support self-regulated learning. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), AIED’13: *Proceedings from the International Conference on Artificial Intelligence in Education* (pp. 229-238). Springer.

Boulet, J. R., & Durning, S. J. (2019). What we measure… and what we should measure in medical education. *Medical Education*, *53*(1), 86-94.

Bric, J. D., Lumbard, D. C., Frelich, M. J., & Gould, J. C. (2016). Current state of virtual reality simulation in robotic surgery training: A review. *Surgical Endoscopy*, *30*(6), 2169-2178.

Brown, A. L. (1982). *Learning, remembering, and understanding* (Technical Report No. 244). Wiley.

Byrnes, J. P., & Dunbar, K. N. (2014). The nature and development of critical-analytic thinking. *Educational Psychology Review*, *26*(4), 477-493. https://doi.org/10.1007/s10648-014-9284-0

Carpenter, D., Geden, M., Rowe, J., Azevedo, R., & Lester, J. (2020). Automated analysis of middle school students' written reflections during game-based learning. In I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), AIED'20: *Proceedings of the International Conference on Artificial Intelligence in Education* (pp. 67-78). Springer.

Carpenter, D., Cloude, E., Rowe, J., Azevedo, R., & Lester, J. (2021). Investigating Student Reflection during Game-Based Learning in Middle Grades Science. In I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Mill´an (Eds.), LAK'21: *Proceedings of the 11th International Learning Analytics and Knowledge Conference* (pp. 280-291). Association of Computing Machinery.

Chan, T., Sebok-Syer, S., Thoma, B., Wise, A., Sherbino, J., & Pusic, M. (2018). Learning analytics in medical education assessment: The past, the present, and the future. *AEM Education and Training*, *2*(2), 178-187.

Chetwood, A. S., Kwok, K. W., Sun, L. W., Mylonas, G. P., Clark, J., Darzi, A., & Yang, G. Z. (2012). Collaborative eye tracking: A potential training tool in laparoscopic surgery. *Surgical Endoscopy*, *26*(7), 2003-2009.

Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2016). Digital games, design, and learning: A systematic review and meta-analysis. *Review of Educational Research*, *86*(1), 79-122. https://doi.org/10.3102/0034654315582065

Cleary, T. J., Durning, S. J., & Artino, A. R. (2016). Microanalytic assessment of self-regulated learning during clinical reasoning tasks: recent developments and next steps. *Academic Medicine*, *91*(11), 1516-1521.

Cleary, T. J., Konopasky, A., La Rochelle, J. S., Neubauer, B. E., Durning, S. J., & Artino, A. R. (2019). First-year medical students' calibration bias and accuracy across clinical reasoning activities. *Advances in Health Sciences Education*, *24*(4), 767-781.

Cloude, E. B., Ballelos, N. A. M., Azevedo, R., Castiglioni, A., LaRochelle, J., Andrews, A., & Hernandez, C. (2021). Designing intelligent systems to support medical diagnostic reasoning using process data. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, & V. Dimitrova (Eds.), AIED'21: *Proceedings of the International Conference on Artificial Intelligence in Education* (vol. 12749, pp. 109-113). Springer.

Cloude, E. B., Carpenter, D., Dever, D. A., Lester, J., & Azevedo, R. (2021). Game-based learning analytics for supporting adolescents' reflection. *Journal of Learning Analytics*, *8*(2), 51-72.

Cloude, E. B., Dever, D. A., Wiedbusch, M. D., & Azevedo, R. (2020). Quantifying scientific thinking using multichannel data with crystal island: Implications for individualized game-learning analytics. *Frontiers in Education*, *5*, 1-21.

Cloude, E., Taub, M., & Azevedo, R. (2018). Investigating the role of goal orientation: Metacognitive and cognitive strategy use and learning with intelligent tutoring systems. In R. Nkombu, R. Azevedo, & J. Vassileva (Eds.), ITS'18: *Proceedings of the International Conference on Intelligent Tutoring Systems* (pp. 44-53). Springer.

Cloude, E. B., Taub, M., Lester, J., & Azevedo, R. (2019). The role of achievement goal orientation on metacognitive process use in game-based learning. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), AIED'19: *Proceedings of the International Conference on Artificial Intelligence in Education* (pp. 36-40). Springer.

Cloude, E. B., Wortha, F., Wiedbusch, M. D., & Azevedo, R. (2021). Goals matter: Changes in metacognitive judgments and their relation to motivation and learning with an intelligent tutoring system. In P. Zaphiris & A. Ioannou (Eds.), HCII’21: *Proceedings of the International Conference on Human-Computer Interaction* (pp. 224-238). Springer.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46.

Connor, D. M., Durning, S. J., & Rencic, J. J. (2020). Clinical reasoning as a core competency. *Academic Medicine*, *95*(8), 1166-1171.

Cook, D. A., Andriole, D. A., Durning, S. J., Roberts, N. K., & Triola, M. M. (2010). Longitudinal research databases in medical education: Facilitating the study of educational outcomes over time and across institutions. *Academic Medicine*, *85*(8), 1340-1346.

Cook, D. A., Hatala, R., Brydges, R., Zendejas, B., Szostek, J. H., Wang, A. T., ... & Hamstra, S. J. (2011). Technology-enhanced simulation for health professions education: A systematic review and meta-analysis. *Jama*, *306*(9), 978-988.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334.

Custers, E. J. (2015). Thirty years of illness scripts: Theoretical origins and practical applications. *Medical Teacher*, *37*(5), 457-462.

Dawson, S., Joksimovic, S., Poquet, O., & Siemens, G. (2019). Increasing the impact of learning analytics. In D. Azcona & R. Chung (Eds.), LAK’19: *Proceedings of the 9th international conference on learning analytics & knowledge* (pp. 446-455). Association of Computing Machinery.

Deater-Deckard, K., Chang, M., & Evans, M. E. (2013). Engagement states and learning from educational games. *New Directions for Child and Adolescent Development*, *2013*(139), 21-30. https://doi.org/10.1002/cad.20028

DeBoer, J. (2012). Twentieth-century American education reform in the global context. *Peabody Journal of Education*, *87*(4), 416-435. https://doi.org/10.1080/0161956X.2012.705136

Deng, M., & Gu, X. (2021). Information acquisition, emotion experience and behaviour intention during online shopping: An eye-tracking study. *Behaviour & Information Technology*, *40*(7), 635-645.

Dewey, J. (1923). *Democracy and education: An introduction to the philosophy of education*. Macmillan.

Dewey, J. (1933). *How we think: A restatement of the relation of reflective thinking to the educative process*. D. C. Heath and Co.

D'Mello, S. (2017). Emotional learning analytics. In C. Lang, G. Siemens, A. Wise, & D. Gasevic, (Eds.), *Handbook of learning analytics* (pp. 115-127). SOLAR: Society for Learning Analytics Research.

Duffy, M. C., Azevedo, R., Sun, N. Z., Griscom, S. E., Stead, V., Crelinsten, L., ... & Lachapelle, K. (2015). Team regulation in a simulated medical emergency: An in-depth analysis of cognitive, metacognitive, and affective processes. *Instructional Science*, *43*(3), 401-426.

Dunbar, K. N. & Klahr, D. (2012). Scientific thinking and reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 701-718). Oxford University Press.

Dunlosky, J., & Tauber, S. U. K. (Eds.). (2016). *The Oxford handbook of metamemory*. Oxford University Press.

Dyre, L., & Tolsgaard, M. G. (2018). The gap in transfer research. *Medical Education*, *52*(6), 580–582. https://doi.org/10.1111/medu.13591

Ericsson, K. A. (2015). Acquisition and maintenance of medical expertise: A perspective from the expert-performance approach with deliberate practice. *Academic Medicine*, *90*(11), 1471-1486.

Ericsson, K. A. (2007). An expert-performance perspective of research on medical expertise: The study of clinical performance. *Medical Education*, *41*(12), 1124-1130.

Ericsson, K. A., Hoffman, R. R., Kozbelt, A., & Williams, A. M. (Eds.). (2018). *The Cambridge handbook of expertise and expert performance*. Cambridge University Press.

Erridge, S., Ashraf, H., Purkayastha, S., Darzi, A., & Sodergren, M. H. (2018). Comparison of gaze behaviour of trainee and experienced surgeons during laparoscopic gastric bypass. *Journal of British Surgery*, *105*(3), 287-294.

Fard, M. J., Ameri, S., Darin Ellis, R., Chinnam, R. B., Pandya, A. K., & Klein, M. D. (2018). Automated robot-assisted surgical skill evaluation: Predictive analytics approach. *The International Journal of Medical Robotics and Computer Assisted Surgery*, *14*(1), e1850.

Favela, L. H. (2020). Cognitive science as complexity science. Wiley Interdisciplinary Reviews: *Cognitive Science*, *11*(4), e1525.

Feltovich, P. J., Prietula, M. J., & Ericsson, K. A. (2018). Studies of expertise from psychological perspectives: Historical foundations and recurrent themes. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 59–83). Cambridge University Press.

Feyzi-Behnagh, R., Azevedo, R., Legowski, E., Reitmeyer, K., Tseytlin, E., & Crowley, R. S. (2014). Metacognitive scaffolds improve self-judgments of accuracy in a medical intelligent tutoring system. *Instructional Science*, *42*(2), 159-181.

Fiorella, L., & Mayer, R. E. (2012). based aids for learning with a computer-based game. *Journal of Educational Psychology*, *104*(4), 1074-1082.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*(10), 906–911.

Forsberg, E., Ziegert, K., Hult, H., & Fors, U. (2014). Clinical reasoning in nursing, a think-aloud study using virtual patients–A base for an innovative assessment. *Nurse Education Today*, *34*(4), 538-542.

Fox, S.-E. & Faulkner-Jones, B. E. (2017). Eye-tracking in the study of visual expertise: Methodology and approaches in medicine. *Frontline Learning Research*, *5*(3), 29-40.

Freire, M., Serrano-Laguna, Á., Manero, B., Martínez-Ortiz, I., Moreno-Ger, P., & Fernández-Manjón, B. (2016). Game learning analytics: learning analytics for serious games. In M. J. Spector, B. B. Lockee, & M. D. Childress (Eds.), *Learning, Design, and Technology* (pp. 1-29). Springer.

Geden, M., Emerson, A., Carpenter, D., Rowe, J., Azevedo, R., & Lester, J. (2021). Predictive student modeling in game-based learning environments with word embedding representations of reflection. *International Journal of Artificial Intelligence in Education*, *31*(1), 1-23.

Giannakos, M. N., Sharma, K., Pappas, I. O., Kostakos, V., & Velloso, E. (2019). Multimodal data as a means to understand the learning experience. *International Journal of Information Management*, *48*, 108-119.

Goetz, T., Sticca, F., Pekrun, R., Murayama, K., & Elliot, A. J. (2016). Intraindividual relations between achievement goals and discrete achievement emotions: An experience sampling approach. *Learning and Instruction*, *41*, 115-125.

Gorman, J. C., Dunbar, T. A., Grimm, D., & Gipson, C. L. (2017). Understanding and modeling teams as dynamical systems. *Frontiers in Psychology*, *8*, 1053.

Greene, J. A., Bolick, C. M., & Robertson, J. (2010). Fostering historical knowledge and thinking skills using hypermedia learning environments: The role of self-regulated learning. *Computers & Education*, *54*(1), 230-243.

Greene, J. A., Deekens, V. M., Copeland, D. Z., & Yu, S. (2018). Capturing and modeling self-regulated learning using think-aloud protocols. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (pp. 323–337). Routledge/Taylor & Francis Group.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, *11*(1), 1-21.

Guastello, S. J., Koopmans, M., & Pincus, D. (Eds.). (2009). *Chaos and complexity in psychology: The theory of nonlinear dynamical systems*. Cambridge University Press.

Guastello, S. J. & Leibovitch, L. S. (2009). Introduction to nonlinear dynamics and complexity. In S. J. Guastello, M. Koopmans, & D. Pincus (Eds.), *Chaos and complexity in psychology: The theory of nonlinear dynamical systems* (pp. 1-40). Cambridge University Press.

Hall, J. C., Ellis, C., & Hamdorf, J. (2003). Surgeons and cognitive processes. *Journal of British Surgery*, *90*(1), 10-16.

Harrell, F. E. (2015). Multivariable modeling strategies. *Regression modeling strategies* (pp. 63-102). Springer.

Henneman, E. A., Cunningham, H., Fisher, D. L., Plotkin, K., Nathanson, B. H., Roche, J. P., ... & Henneman, P. L. (2014). Eye tracking as a debriefing mechanism in the simulated setting improves patient safety practices. *Dimensions of Critical Care Nursing*, *33*(3), 129-135.

Hermens, F., Flin, R., & Ahmed, I. (2013). Eye movements in surgery: A literature review. *Journal of Eye Movement Research*, *6*(4), 1-11.

Hillaire, S., Lécuyer, A., Breton, G., & Corte, T. R. (2009). Gaze behavior and visual attention model when turning in virtual environments. In S. N. Spencer (Ed.), VRST'09: *Proceedings of the 16th ACM Symposium on Virtual Reality Software and Technology* (pp. 43-50). Association for Computing Machinery.

Holstein, K., McLaren, B. M., & Aleven, V. (2019). Co-designing a real-time classroom orchestration tool to support teacher–AI complementarity. *Journal of Learning Analytics*, *6*(2), 27-52.

Huang, C. M., Andrist, S., Sauppé, A., & Mutlu, B. (2015). Using gaze patterns to predict task intent in collaboration. *Frontiers in Psychology*, *6*, 1-12.

iMotions. (2016). Attention tool (version 6.2) [computer software]. In Boston: iMotions Inc.

Izu, C., & Alexander, B. (2018). Using unstructured practice plus reflection to develop programming/problem-solving fluency. *Proceedings of the 20th Australasian Computing Education Conference*, 25–34. https://doi.org/10.1145/3160489.3160496

Järvelä, S. & Bannert, M. (2021). Temporal and adaptive processes of regulated learning- What can multimodal data tell? *Learning and Instruction*, *72*. 101268.

Johnson, C. I., & Mayer, R. E. (2010). Applying the self-explanation principle to multimedia learning in a computer-based game-like environment. *Computers in Human Behavior, 26*(6), 1246-1252.

Johnson, P. O., & Neyman, J. (1936). Tests of certain linear hypotheses and their application to some educational problems. *Statistical Research Memoirs*, *1*, 57-93.

Joksimović, S., Kovanović, V., & Dawson, S. (2019). The journey of learning analytics. *HERDSA Review of Higher Education*, *6*, 27-63.

Josephsen, J. M. (2017). A qualitative analysis of metacognition in simulation. *Journal of Nursing Education*, *56*(11), 675-678.

Keiser, N. L., & Arthur Jr, W. (2021). A meta-analysis of the effectiveness of the after-action review (or debrief) and factors that influence its effectiveness. *Journal of Applied Psychology*, *106*(7), 1007-1032.

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

Klahr, D., Zimmerman, C., & Matlen, B. J. (2019). Improving students' scientific thinking. In J. Dunlosky & K. A. Rawson (Eds.), *The Cambridge handbook of cognition and educatio*n (pp. 67–99). Cambridge University Press. https://doi.org/10.1017/9781108235631.005

Kok, E. M., & Jarodzka, H. (2017). Before your very eyes: The value and limitations of eye tracking in medical education. *Medical Education*, *51*(1), 114-122.

Koochaki, F., & Najafizadeh, L. (2021). A data-driven framework for intention prediction via eye movement with applications to assistive systems. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *29*, 974-984.

Kovanović, V., Joksimović, S., Mirriahi, N., Blaine, E., Gašević, D., Siemens, G., & Dawson, S. (2018). Understand students' self-reflections through learning analytics. In LAK'18: *Proceedings of the 8th international conference on learning analytics and knowledge* (pp. 389-398). Association for Computing Machinery.

Krupat, E. (2018). Critical thoughts about the core entrustable professional activities in undergraduate medical education. *Academic Medicine*, *93*(3), 371-376.

Krupinski, E. A., Tillack, A. A., Richter, L., Henderson, J. T., Bhattacharyya, A. K., Scott, K. M., ... & Weinstein, R. S. (2006). Eye-movement study and human performance using telepathology virtual slides. Implications for medical education and differences with experience. *Human Pathology*, *37*(12), 1543-1556.

Kuehn, B. M. (2018). Virtual and augmented reality put a twist on medical education. *Jama*, *319*(8), 756-758.

Kuhl, P. K., Lim, S., Guerriero, S., & Damme, D. (2019). *Developing minds in the digital age: Towards a science of learning for 21st century education*. Educational Research and Innovation. OECD Publishing. https://doi.org/10.1787/562a8659-en

Lajoie, S. P., Li, S., & Zheng, J. (2021). The functional roles of metacognitive judgement and emotion in predicting clinical reasoning performance with a computer simulated environment. *Interactive Learning Environments*, 1-12. https://doi.org/10.1080/10494820.2021.1931347

Lajoie, S. P., Pekrun, R., Azevedo, R., & Leighton, J. P. (2020). Understanding and measuring emotions in technology-rich learning environments. *Learning and Instruction*, *70*, 101272.

Lajoie, S. P., Poitras, E. G., Doleck, T., & Jarrell, A. (2015). Modeling metacognitive activities in medical problem-solving with BioWorld. In A. Peña-Ayala, (Ed.), *Metacognition: Fundaments, applications, and trends* (pp. 323--343). Springer.

Lang, C., Siemens, G., Wise, A., & Gasevic, D. (Eds.). (2017). *Handbook of learning analytics*. SOLAR, Society for Learning Analytics and Research.

Laplace, P. S. marquis de. (1812). *Théorie analytique des probabilités.* V. Courcier.

Lazonder, A. W., Hagemans, M. G., & De Jong, T. (2010). Offering and discovering domain information in simulation-based inquiry learning. *Learning and Instruction*, *20*(6), 511-520. https://doi.org/10.1016/j.learninstruc.2009.08.001

Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning: Effects of guidance. *Review of Educational Research*, *86*(3), 681-718.

Lewis, K. O., & Baker, R. C. (2007). The development of an electronic educational portfolio: An outline for medical education professionals. *Teaching and Learning in Medicine*, *19*(2), 139-147.

Loderer, K., Pekrun, R., & Lester, J. C. (2020). Beyond cold technology: A systematic review and meta-analysis on emotions in technology-based learning environments. *Learning and Instruction*, *70*, 101162. https://doi.org/10.1016/j.learninstruc.2018.08.002

Lodge, J. M., Panadero, E., Broadbent, J., & de Barba, P. G. (2018). Supporting self-regulated learning with learning analytics. In J. M. Lodge, J. C. Horvath, & L. Corrin (Eds.), *Learning analytics in the classroom* (pp. 45--55). Routledge.

Long, J. A. (2019). Interactions: comprehensive, user-friendly toolkit for probing interactions (version 1.1.0.) [computer software], In Vienna, Austria.

Lüdecke, D. (2018). sjmisc: Data and variable transformation functions. *Journal of Open Source Software*, *3*(26), 1-2.

Lüdecke, D. (2018). sjPlot: Data visualization for statistics in social science. (Version 2.1.0) [computer software], In Vienna, Austria.

Luo, T., & Baaki, J. (2019). Scaffolding problem-solving and instructional design processes: Engaging students in reflection-in-action and external representations in three online courses. In M. Boboc & S. Koc (Eds.), *Student-Centered Virtual Learning Environments in Higher Education* (pp. 40-69). IGI Global.

Mangaroska, K., Sharma, K., Gaševic, D., & Giannakos, M. (2020). Multimodal learning analytics to inform learning design: Lessons learned from computing education. *Journal of Learning Analytics*, *7*(3), 79-97.

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

McAlpine, L., Weston, C., Beauchamp, C., Wiseman, C., & Beauchamp, J. (1999). Building a metacognitive model of reflection. *Higher Education*, *37*(2), 105-131.

McQuiggan, S. W., Goth, J., Ha, E., Rowe, J. P., & Lester, J. C. (2008). Student note-taking in narrative-centered learning environments: Individual differences and learning effects. In B. P. Woolf, E. Aïmeur, R. Nkambou, & S. Lajoie (Eds.), ITS'08: *Proceedings of the International Conference on Intelligent Tutoring Systems* (pp. 510-519). Springer.

Menekse Dalveren, G. G., & Cagiltay, N. E. (2020). Distinguishing intermediate and novice surgeons by eye movements. *Frontiers in Psychology*, *11*, 2330.

Meyers, L., Mahoney, B., Schaffernocker, T., Way, D., Winfield, S., Uribe, A., ... & Lipps, J. (2020). The effect of supplemental high Fidelity simulation training in medical students. *BMC Medical Education*, *20*(1), 1-7.

Miele, D. B., & Scholer, A. A. (2018). The role of metamotivational monitoring in motivation regulation. *Educational Psychologist*, *53*(1), 1-21.

Miller Singley, A. T., & Bunge, S. A. (2018). Eye gaze patterns reveal how we reason about fractions. *Thinking & Reasoning*, *24*(4), 445-468.

Molenaar, I., & Knoop-van Campen, C. A. (2018). How teachers make dashboard information actionable. *IEEE Transactions on Learning Technologies*, *12*(3), 347-355.

Moos, D. C. (2014). Setting the stage for the metacognition during hypermedia learning: What motivation constructs matter?. *Computers & Education*, *70*, 128-137.

Moreno, R., & Mayer, R. E. (2005). Role of guidance, reflection, and interactivity in an agent-based multimedia game. *Journal of Educational Psychology*, *97*(1), 117-128.

Morris, B., Croker, S., Zimmerman, C., Gill, D., & Romig, C. (2013). Gaming science: The "Gamification" of scientific thinking. *Frontiers in Psychology*, *4*(607). 1-16. https://doi.org/10.3389/fpsyg.2013.00607

Moshman, D. (2013). Adolescent rationality. In R. M. Lerner & J. B. Benson (Eds.), *Embodiment and epigenesis: Theoretical and methodological issues in understanding the role of biology within the relational developmental system* (pp. 155-183). Elsevier.

Mourad, A., Jurjus, A., & Hussein, I. H. (2016). The what or the how: a review of teaching tools and methods in medical education. *Medical Science Educator*, *26*(4), 723-728.

Mudrick, N. V., Azevedo, R., & Taub, M. (2019). Integrating metacognitive judgments and eye movements using sequential pattern mining to understand processes underlying multimedia learning. *Computers in Human Behavior*, *96*, 223-234.

Mulder, Y. G., Lazonder, A. W., & de Jong, T. (2011). Comparing two types of model progression in an inquiry learning environment with modelling facilities. *Learning and Instruction*, *21*(5), 614-624. https://doi.org/10.1016/j.learninstruc.2011.01.003

Munoz, J., Yannakakis, G. N., Mulvey, F., Hansen, D. W., Gutierrez, G., & Sanchis, A. (2011). Towards gaze-controlled platform games. In CIG'11: *Proceedings from the IEEE Conference on Computational Intelligence and Games* (pp. 47-54). IEEE.

National Academies of Sciences, Engineering, and Medicine. (2018). *How people learn II: Learners, contexts, and cultures*. The National Academies Press. https://doi.org/10.17226/24783

National Research Council. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Committee on defining deeper learning and 21st century skills, James W. Pellegrino & Margaret L. Hilton (Eds.), Board on Testing and Assessment and Board on Science Education, Division of Behavioral Sciences and Education. The National Academies Press.

National Research Council. (2015). *Guide to implementing the next generation science standards*. National Academies Press.

National Research Council. (2011). *Learning science through computer games and simulations*. Committee on science learning: Computer games, simulations, and education, Margaret A. Honey & Margaret L. Hilton (Eds.), Board on Science Education, Division of Behavioral and Social Sciences and Education. The National Academies Press.

National Research Council. (2007). *Rising above the gathering storm: Energizing and employing America for a brighter economic future*. National Academies Press.

National Research Council. (2010*). Rising above the gathering storm, revisited: Rapidly approaching category 5*. National Academies Press.

Nelson, T. O. & Narens, L. (1994). Why investigate metacognition. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1--25). MIT Press.

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. The National Academies Press.

Norris, M., & Gimber, P. (2013). Developing nursing students' metacognitive skills using social technology. *Teaching and Learning in Nursing*, *8*(1), 17-21.

Norman, G. R., Grierson, L. E. M., Sherbino, J., Hamstra, S. J., Schmidt, H. G., & Mamede, S. (2018). Expertise in medicine and surgery. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 331–355). Cambridge University Press.

Noroozi, O., Alikhani, I., Järvelä, S., Kirschner, P. A., Juuso, I., & Seppänen, T. (2019). Multimodal data to design visual learning analytics for understanding regulation of learning. *Computers in Human Behavior*, *100*, 298-304.

Ochoa, X. (2017). Multimodal learning analytics. In C. Lang, G. Siemens, A. Wise, D. Gašević (Eds.), *The handbook of learning analytics* (pp. 129-141). SOLAR: Society for Learning Analytics Research.

Organisation for Economic Co-operation and Development. (2007). *Economic survey of the United States 2007*. OECD Publishing.

Organisation for Economic Co-operation and Development. (2016). *PISA 2015 results (volume I): Excellence and equity in education*. OECD Publishing.

Orlosky, J., Itoh, Y., Ranchet, M., Kiyokawa, K., Morgan, J., & Devos, H. (2017). Emulation of physician tasks in eye-tracked virtual reality for remote diagnosis of neurodegenerative disease. *IEEE Transactions on Visualization and Computer Graphics*, *23*(4), 1302-1311.

Paas, F., & van Merriënboer, J. J. (2020). Cognitive-load theory: Methods to manage working memory load in the learning of complex tasks. *Current Directions in Psychological Science*, *29*(4), 394-398.

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology, 8*, 422.

Park, H., Lee, S., Lee, M., Chang, M. S., & Kwak, H. W. (2016). Using eye movement data to infer human behavioral intentions. *Computers in Human Behavior*, *63,* 796-804.

Patel, V. L., Kaufman, D. R., & Kannampallil, T. G. (2018). Diagnostic reasoning and expertise in health care. In P. Ward, J. M. Schraagen, J. Gore, & E. M. Roth (Eds.), *The Oxford handbook of expertise* (1st edn, pp. 618-641). Oxford University Press.

Patel, N., Baker, S. G., & Scherer, L. D. (2019). Evaluating the cognitive reflection test as a measure of intuition/reflection, numeracy, and insight problem solving, and the implications for understanding real-world judgments and beliefs. *Journal of Experimental Psychology: General*, *148*(12), 2129-2153.

Perez-Colado, I. J., Perez-Colado, V. M., Freire-Moran, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2017). Integrating learning analytics into a game authoring tool. In H. Xie, E. Popescu, G. Hancke, & B. Fernández Manjón (Eds.), *Proceedings of the International Conference on Web-Based Learning* (pp. 51-61). Springer.

Peterson, R. A., & Cavanaugh, J. E. (2019). Ordered quantile normalization: A semiparametric transformation built for the cross-validation era. *Journal of Applied Statistics, 47*(13-15), 2312-2327.

Petrizzo, M. C., Barilla-LaBarca, M. L., Lim, Y. S., Jongco, A. M., Cassara, M., Anglim, J., & Stern, J. N. (2019). Utilization of high-fidelity simulation to address challenges with the basic science immunology education of preclinical medical students. *BMC medical Education*, *19*(1), 1-8.

Pintrich, P. R. (2002). The role of metacognitive knowledge in learning, teaching, and assessing. *Theory into Practice*, *41*(4), 219-225.

Plass, J. L., Homer, B. D., & Kinzer, C. K. (2015). Foundations of game-based learning. *Educational Psychologist, 50*(4), 258-283.

Plass, J. L., Mayer, R. E., & Homer, B. D. (Eds.). (2020). *Handbook of game-based learning*. MIT Press.

Plummer, P., DeWolf, M., Bassok, M., Gordon, P. C., & Holyoak, K. J. (2017). Reasoning strategies with rational numbers revealed by eye tracking. *Attention, Perception, & Psychophysics*, *79*(5), 1426-1437.

Qian, M., & Clark, K. R. (2016). Game-based learning and 21st century skills: A review of recent research. *Computers in Human Behavior*, *63*, 50-58.

Raij, A. B., & Lok, B. C. (2008). Ipsviz: An after-action review tool for human-virtual human experiences. In M. Lin, A. Steed, & C. Cruz-Neira (Eds.), *Proceedings of the IEEE Virtual Reality Conference* (pp. 91-98). IEEE.

R Core Team. (2013). R: A language and environment for statistical computing (Version 3.6.2) [Computer software]. https://www.R-project.org/

R Core Team. (2019). R: A language and environment for statistical computing (Version 3.6.2) [Computer software]. https://www.R-project.org/

Rajendran, R., Kumar, A., Carter, K. E., Levin, D. T., & Biswas, G. (2018). Predicting learning by analyzing eye-gaze data of reading behavior. In K. E. Boyer & M. Yudelson (Eds.), EDM'18: *Proceedings from the International Educational Data Mining Society* (pp. 455-461). ERIC.

Raven, J. C. (1941). Standardization of progressive matrices, 1938. *British Journal of Medical Psychology*, *19*(1), 137-150.

Raven, J. (2000). The Raven's progressive matrices: change and stability over culture and time. *Cognitive Psychology*, *41*(1), 1-48.

Rayner, K. (2009). The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, *62*(8), 1457-1506.

Rencic, J., Trowbridge, R. L., Fagan, M., Szauter, K., & Durning, S. (2017). Clinical reasoning education at US medical schools: Results from a national survey of internal medicine clerkship directors. *Journal of General Internal Medicine*, *32*(11), 1242-1246.

Richstone, L., Schwartz, M. J., Seideman, C., Cadeddu, J., Marshall, S., & Kavoussi, L. R. (2010). Eye metrics as an objective assessment of surgical skill. *Annals of Surgery*, *252*(1), 177-182.

Roll, I., & Winne, P. H. (2015). Understanding, evaluating, and supporting self-regulated learning using learning analytics. *Journal of Learning Analytics*, *2*(1), 7-12.

Rosenthal, D. (2000). Introspection and self-interpretation. *Philosophical Topics*, *28*(2), 201-233.

Rowe, J. P., Shores, L. R., Mott, B. W., & Lester, J. C. (2011). Integrating learning, problem solving, and engagement in narrative-centered learning environments. *International Journal of Artificial Intelligence in Education*, *21*(1-2), 115-133.

Salvucci, D., & Goldberg, J. (2000). Identifying fixations and saccades in eye-tracking protocols. *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*, 71–78. https://doi.org/10.1145/355017.355028

Scheffel, M., Drachsler, H., Stoyanov, S., & Specht, M. (2014). Quality indicators for learning analytics. *Journal of Educational Technology & Society*, *17*(4), 117-132.

Scheiter, K., & Eitel, A. (2017). The use of eye tracking as a research and instructional tool in multimedia learning. In F. Yang & C. Zimmerman (Eds.), ETRA'17: *Proceedings of the Eye-tracking technology applications in educational research* (pp. 143-164). IGI Global.

Schunk, D. H., & DiBenedetto, M. K. (In press). Motivation and social cognitive theory. *Contemporary Educational Psychology*, *60*, 101832.

Schunk, D. H. & Greene, J. A. (Eds.), *Handbook of self-regulation of learning and performance*. Routledge/Taylor & Francis Group.

Schuster, H. G. (1984). *Deterministic chaos*. Physik Verlag.

Sharma, K., & Giannakos, M. (2020). Multimodal data capabilities for learning: What can multimodal data tell us about learning?. *British Journal of Educational Technology*, *51*(5), 1450-1484.

Sharma, Papamitsiou, Z., Olsen, J., & Giannakos, M. (2020). Predicting learners' effortful behaviour in adaptive assessment using multimodal data. *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, 480–489. https://doi.org/10.1145/3375462.3375498

Shavelson, R. J., & Stern, P. (1981). Research on teachers' pedagogical thoughts, judgments, decisions, and behavior. *Review of Educational Research*, *51*(4), 455-498.

Shepard, L. A., Penuel, W. R., & Pelligrino, J. W. (2013). Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment. *Educational Measurement: Issues and Practice*, *37*(1), 21-34. https://doi.org/10.1111/emip.12189

Shinnick, M. A. (2016). Validating eye tracking as an objective assessment tool in simulation. *Clinical Simulation in Nursing*, *12*(10), 438-446.

Shute V., Rahimi S., & Lu X. (2019). Supporting learning in educational games: Promises and challenges. In P. Díaz, A. Ioannou, K. Bhagat, & J. Spector J. (Eds.), *Learning in a digital world. Smart computing and intelligence*. Springer. https://doi.org/10.1007/978-981-13-8265-9_4

Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior*, *63*, 106-117.

Siadaty, M., Gasevic, D., & Hatala, M. (2016). Trace-based micro-analytic measurement of self-regulated learning processes. *Journal of Learning Analytics*, *3*(1), 183-214.

Singh, R., Miller, T., Newn, J., Sonenberg, L., Velloso, E., & Vetere, F. (2018). Combining planning with gaze for online human intention recognition. In M. Dastani, G. Sukthankar, E. Andre, & S. Koenig (Eds.), *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems* (pp. 488-496). International Foundation for Autonomous Agents and Multiagent Systems.

SMI Experiment Center 3.4.165 [Apparatus and Software]. SensoMotoric Instruments, Boston, Massachusetts, USA (2014).

Smith, G., Shute, V., & Muenzenberger, A. (2019). Designing and validating a stealth assessment for calculus competencies. *Journal of Applied Testing Technology*, *20*(S1), 52-59.

Suter, L. E., & Camilli, G. (2019). International student achievement comparisons and US STEM workforce development. *Journal of Science, Education, and Technology*, *28*(1), 52-61. https://doi.org/10.1007/s10956-018-9746-0

Tarricone, P. (2011). *The taxonomy of metacognition*. Psychology Press.

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

Taub, M., Azevedo, R., Bradbury, A. E., & Mudrick, N. V. (2020). Self-regulation and reflection during game-based learning. In J. L. Plass, R. E. Mayer, & B. D. Homer (Eds.), *Handbook of game-based learning* (pp. 239-262). MIT Press.

Taub, M., Sawyer, R., Smith, A., Rowe, J., Azevedo, R., & Lester, J. (2020). The agency effect: The impact of student agency on learning, emotions, and problem-solving behaviors in a game-based learning environment. *Computers & Education*, *147*, 103781. https://doi.org/10.1016/j.compedu.2019.103781

ter Vrugte, J., de Jong, T., Wouters, P., Vandercruysse, S., Elen, J., & van Oostendorp, H. (2015). When a game supports prevocational math education but integrated reflection does not. *Journal of Computer Assisted Learning*, *31*(5), 462-480.

Tien, T., Pucher, P. H., Sodergren, M. H., Sriskandarajah, K., Yang, G. Z., & Darzi, A. (2014). Eye tracking for skills assessment and training: A systematic review. *Journal of Surgical Research*, *191*(1), 169-178.

Ullmann, T. D. (2019). Automated analysis of reflection in writing: Validating machine learning approaches. *International Journal of Artificial Intelligence in Education*, *29*(2), 217-257.

U.S. Department of Education, Office of Educational Technology. (2017). *Reimagining the role of technology in higher education: A supplement to the national education technology plan*.

Vallacher, R. R., & Nowak, A. (1999). The dynamics of self-regulation. In R. S. Wyer, C. S. Carver, & M. F. Scheier (Eds.), *Perspectives on behavioral self-regulation* (pp. 241-259). Lawrence Erlbaum Associates.

van Rossum, G., & Drake, F. L. (2010). *The python language reference*. Python Software Foundation.

van Zoest, W., Van der Stigchel, S., & Donk, M. (2017). Conditional control in visual selection. *Attention, Perception, & Psychophysics*, *79*(6), 1555-1572.

Venables, W. N., & Ripley, B. D. (2002). Random and mixed effects. *Modern applied statistics with S* (pp. 271-300). Springer.

Vendetti, M. S., Starr, A., Johnson, E. L., Modavi, K., & Bunge, S. A. (2017). Eye movements reveal optimal strategies for analogical reasoning. *Frontiers in Psychology*, *8*, 1-9.

Viberg, O., Khalil, M., & Baars, M. (2020). Self-regulated learning and learning analytics in online learning environments: a review of empirical research. *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, 524–533. https://doi.org/10.1145/3375462.3375483

Virginia Department of Education, Commonwealth of Virginia. (2021). *Standards of learning (SOL) & testing*. doe.virginia.gov. https://www.doe.virginia.gov/testing/index.shtml

Wesiak, G., Steiner, C. M., Moore, A., Dagger, D., Power, G., Berthold, M., ... & Conlan, O. (2014). Iterative augmentation of a medical training simulator: Effects of affective metacognitive scaffolding. *Computers & Education*, *76*, 13-29.

West, M. (2012). Global lessons for improving US education. *Issues in Science and Technology*, *28*(3), 37-44. https://www.jstor.org/stable/43315668

Wickham, H. (2016). *Ggplot2: Elegrant graphics for data analysis*. Springer.

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, *21*(12), 1-20.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., ... & Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software, 4*(43), 1-5.

Wickham, H., & Bryan, J. (2017). readxl: Read excel files. R package (Version 1.0.0). [Computer software]. https://CRAN.993R-project. org/package= readxl

Wickham, H., Francois, R., Henry, L., & Muller, K. (2018). dyplyr: A grammar of data manipulation (Version 0.7.6) [Computer software]. https://CRAN.R-project.org/package=dplyr

Wiedbusch, M. D. & Azevedo, R. (2020). Modeling metacomprehension monitoring accuracy with eye gaze on informational content in a multimedia learning environment. In Stephen N. Spencer (Ed.), ETRA'20: *Proceedings of the international symposium on eye tracking research and applications* (pp. 1-9). Association of Computing Machinery.

Wiedbusch, M. D., Kite, V., Yang, X., Park, S., Chi, M., Taub, M., & Azevedo, R. (2021). A theoretical and evidence-based conceptual design of MetaDash: An intelligent teacher dashboard to support teachers decision making and students self-regulated learning. *Frontiers in Education*, *6*, 1-14.

Winne, P. H. (2017). Learning analytics for self-regulated learning. In C. Lang, G. Siemens, A. Wise, & D. Gasevic (Eds.), *Handbook of learning analytics* (pp. 241-249). SOLAR, Society for Learning Analytics and Research.

Winne, P. H. (2019). Paradigmatic dimensions of instrumentation and analytic methods in research on self-regulated learning. *Computers in Human Behavior*, *96*, 285-289.

Winne, P. H. & Hadwin, A. F. (1998). Studying as self-regulated engagement in learning. In D. Hacker, J. Dunlosky, & A. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277-304). Erlbaum.

Winne, P. H., Teng, K., Chang, D., Lin, M. P. C., Marzouk, Z., Nesbit, J. C., ... & Vytasek, J. (2019). nStudy: Software for learning analytics about processes for self-regulated learning. *Journal of Learning Analytics*, *6*(2), 95-106.

Winne, P. H. (2018). Theorizing and researching levels of processing in self-regulated learning. *British Journal of Educational Psychology*, *88*(1), 9-20.

Winner, J., & Millwater, T. L. (2019). Evaluating human patient simulation fidelity and effectiveness for combat-medical training. In HFES'19: *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care* (pp. 176–180). SAGE. https://doi.org/10.1177/2327857919081043

Wood, T. J., & Pugh, D. (2020). Are rating scales really better than checklists for measuring increasing levels of expertise?. *Medical Teacher*, *42*(1), 46-51.

Worsley, M., Abrahamson, D., Blikstein, P., Grover, S., Schneider, B., & Tissenbaum, M. (2016). Situating multimodal learning analytics. In C. Looi, J. Polman, U. Cress, & P.

Reimann (Eds.), ICLS'16: *12th International Conference of the Learning Sciences* (pp. 1346-1349).

Worsley, M., & Blikstein, P. (2015). Leveraging multimodal learning analytics to differentiate student learning strategies. *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*, 360–367. https://doi.org/10.1145/2723576.2723624

Wouters, P., van Nimwegen, C., van Oostendorp, H., & van der Spek, E. D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology*, *105*(2), 249-265. https://doi.org/10.1037/a0031311

Wu, L., & Looi, C. K. (2012). Agent prompts: Scaffolding for productive reflection in an intelligent learning environment. *Journal of Educational Technology & Society*, *15*(1), 339-353.

Yamashita, T., Yamashita, K., & Kamimura, R. (2007). A stepwise AIC method for variable selection in linear regression. *Communications in Statistics—Theory and Methods*, *36*(13), 2395-2403.

Zheng, B., & Zhang, Y. (2020). Self-regulated learning: The effect on medical student learning outcomes in a flipped classroom environment. *BMC Medical Education*, *20*(1), 1-7.

Zimmerman, B. J. (2013). From cognitive modeling to self-regulation: A social cognitive career path. *Educational Psychologist*, *48*(3), 135-147.

Zimmerman, C., & Croker, S. (2014). A prospective cognition analysis of scientific thinking and the implications for teaching and learning science. *Journal of Cognitive Education and Psychology*, *13*(2), 245-257. https://doi.org/10.1891/1945-8959.13.2.245

Zimmerman, B. J. & Moylan, A. R. (2009). Self-regulation: Where metacognition and motivation intersect. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 311-328). Routledge.