Computers in Human Behavior 76 (2017) 641–655

Contents lists available at ScienceDirect

Computers in Human Behavior

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

Full length article

Using multi-channel data with multi-level modeling to assess in-game performance during gameplay with Crystal Island

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

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

ARTICLE INFO

Article history:
Received 1 July 2016
Received in revised form
12 January 2017
Accepted 22 January 2017
Available online 7 February 2017
An earlier version of this study was presented at the 13th International Conference on Intelligent Tutoring Systems (ITS 2016) in Zagreb, Croatia and published in A. Micarelli, J. Stamper, & K. Panourgia (Eds.), *Proceedings of the 13th International Conference on Intelligent Tutoring Systems—Lecture Notes in Computer Science 9684* (pp. 240–246). The Netherlands: Springer.

Keywords:
Cognitive strategies
Metacognitive monitoring
Game-based learning environments
Eye tracking
Log files
Self-regulated learning

ABSTRACT

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

1. Introduction

Game-based learning environments (GBLEs) have been touted as a solution for failing educational outcomes across several domains. Learning with GBLEs can be particularly effective for learning because they are designed to foster engagement during learning (e.g., Sabourin, 2013; Sabourin & Lester, 2014). Additionally, many games require self-regulated learning, in addition to other learning processes, such as scientific reasoning (Millis et al., 2011). Scientific reasoning involves generating and testing hypotheses, and therefore students use self-regulatory processes to assist in generating and testing these hypotheses. For example, during gameplay with Crystal Island, students are required to gather clues, and create and test hypotheses, to solve a mystery. It can thus be beneficial to situate theories of SRL with scientific reasoning to investigate learning with GBLEs.

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

* Corresponding author. North Carolina State University, Department of Psychology, 2310 Stinson Drive, Room 640, Raleigh, NC 27695-7650, United States.
E-mail addresses: mtaub@ncsu.edu (M. Taub), nvmudric@ncsu.edu (N.V. Mudrick), razeved@ncsu.edu (R. Azevedo), gcmillar@ncsu.edu (G.C. Millar), jprowe@ncsu.edu (J. Rowe), lester@ncsu.edu (J. Lester).

http://dx.doi.org/10.1016/j.chb.2017.01.038
0747-5632/

cognitive, affective, metacognitive, and motivational processes simultaneously during gameplay to understand their roles and impact other than the typical approach of using pre-to post-test measures and self-reports of motivation and engagement (Mayer, 2014). In this study, we address some of these major issues by using eye movements and log files to examine the cognitive and metacognitive self-regulatory processes deployed by college students while playing Crystal Island, a GBLE that incorporates microbiology content and scientific reasoning to solve the mystery of what disease has spread through a fictional remote island.

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

1.1. Theoretical framework

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

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

Our use of Winne and Hadwin's (1998, 2008) model is advantageous because even though it has yet to be empirically tested, it is the only model that assesses SRL as an event that temporally unfolds over time (Winne & Azevedo, 2014). The temporality of SRL is especially important during learning with GBLEs because students are presented with complex material, and their use of SRL strategies can change depending on the context. For example, during learning of complex text within a hypermedia-learning environment, we operationalize judgments of learning (JOLs) as assessing one's understanding of the text by having them make that judgment, followed by a content quiz (Greene & Azevedo, 2009). When making these judgments, there is a valence associated with it, such that a high rating of understanding would be a JOL+, and a JOL- would be indicative of low understanding. Although this might seem specific to hypermedia-learning environments, this can be applied to learning with GBLEs as well. During gameplay with Crystal Island, one activity students can engage in is reading books, which are associated with embedded assessments called concept matrices (Rowe, Shores, Mott, & Lester, 2011). In each concept matrix, there are questions that test students on their understanding of the material in the book. Therefore, this can be seen as a JOL because students can self-evaluate their understanding of the text, and go on to complete the concept matrix to test if they did understand the text. Furthermore, we can investigate the valence of the JOL based on the correctness of the responses. Thus, as opposed to the valence being associated with the student's judgment of their understanding, the valence can be associated with how well the student performed on the assessment, such that low performance has a negative valence, and high performance has a positive valence. Therefore, we can apply IPT, specifically metacognitive monitoring and control to gameplay and scientific reasoning with GBLEs.

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

2.1. Game-based learning environments

The effectiveness of GBLEs across domains (e.g., math, computer science, biology, psychology, etc.) has come into question as several meta-analyses (Clark, Tanner-Smith, & Killingsworth, 2016; Connolly, Boyle, MacArthur, Hainey, & Boyle, 2012; Girard, Escalle, & Magnan, 2012; Mayer, 2014; Wouters, van Nimwegen, van Oostendorp, & van der Spek, 2013) have revealed different results. More specifically, they have revealed that learning with GBLEs results in small to medium effect sizes for knowledge acquisition ($d = 0.29$, $p < 0.01$; Wouters et al., 2013), yet moderate to large effect sizes for knowledge acquisition ($g = 0.33$, 95% CI, [0.19, 0.48], $k = 57$, $n = 209$; Clark et al., 2016) and retention ($d = 0.36$, $p < 0.01$; Wouters et al., 2013), compared to conventional instructional methods (e.g., PowerPoint, classroom based learning). However, findings did not show significant effects for learning with GBLEs on motivation ($d = 0.11$, $p > 0.05$; Wouters et al., 2013). In addition, Mayer (2014) determined that learning is most effective with games when the topic is science or second language learning, however games for teaching math and language arts is no more effective than using traditional classroom approaches (however a notable limitation is that few studies were investigated). Moreover, adventure games were found to be the most effective ($d = 0.72$),

followed by simulations ($d = 0.62$), and puzzle games ($d = 0.45$). Additionally, games were found to be the most effective for adults or college students ($d = 0.74$), then secondary students ($d = 0.58$), and elementary school students ($d = 0.34$). As such, Mayer (2014) posits that games for learning are effective, however the strength of the effect depends on the context, game type, and population. Several contributing factors explain these effects including a lack of operational definitions, age of learners, learning phenomena and instructional approach (e.g., learning, problem solving, engagement, motivation), number and type of assessment methods (e.g., learning outcomes, transfer measures, self-reports), etc., which is due to a lack of theoretical grounding (Plass, Homer, & Kinzer, 2015; Qian & Clark, 2016; Tsai, Huang, Hou, Hsu, & Chiou, 2016). Filsecker and Kerres (2014) argue that games research needs to move beyond motivational and cognitive processes. Other processes, such as those underlying metacognitive monitoring, affect, and SRL have been largely unexamined (e.g., Sabourin & Lester, 2014).

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

2.2. Assessment of learning

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

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

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

Despite the potential of eye tracking for making inferences about the cognitive and metacognitive processes underlying successful learning (Mayer, 2010; van Gog, Jarodzka, Scheiter, Gerjets, & Paas, 2009), limited research has examined eye movements during learning with GBLEs. Specifically, analyzing eye movements can reveal what students are attending to and for how long during learning (Mayer, 2010; Scheiter & van Gog, 2009; van Gog & Jarodzka, 2013). This methodology has recently been used to study multimedia and hypermedia-based intelligent tutoring systems (ITSs, e.g., Bondareva et al., 2013; Jaques, Conati, Harley, & Azevedo, 2014; Taub & Azevedo, 2016) to examine learners' knowledge acquisition by assessing the total number of fixations on, and transitions between, areas of interest (AOIs) to assess content understanding (D'Mello, 2016; Hyönä & Nurminen, 2006). Lastly, the limited research published on eye movements and GBLEs suffers from serious conceptual, theoretical, methodological, and analytical issues (e.g., Plass et al., 2015).

2.4. Crystal Island: a synthesis of previous work

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

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

of in-game performance, as well as gender. These prior studies are distinguished from the current work by focusing primarily on datasets comprised of game trace logs and student self-reports, rather than multimodal data streams.

More recently, Min et al. (2016) investigated the application of multimodal learning analytics to devise models of companion agent behavior during game-based learning. By using two sequence-labeling techniques, long short-term memory networks (LSTMs) and conditional random fields, with multimodal data, including game trace logs, electrodermal activity, and facial action units (FAUs), they modeled companion agent behavior recorded during a Wizard-of-Oz study with middle school students. They found that utilizing FAUs and game trace logs in concert with LSTM models yielded the best performing model of companion-agent behavior, yielding a 43.9% marginal improvement in predicting the wizard's dialogue-act behaviors compared to a baseline approach. However, Min et al. did not collect eye gaze data, nor did they focus on student strategy use in their work. In sum, the studies conducted by Lester and colleagues with Crystal Island are theoretically-based and converge learning outcomes, self-report measures, and trace data to examine certain SRL processes with adolescents.

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

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

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

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

3.1. Research questions and hypotheses

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

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

matrix submission attempts, with higher levels of all variables resulting in fewer attempts, and thus greater performance.

4. Methods

4.1. Participants

Fifty ($N = 50$[1]) non-biology majors (56% female) from a large public university located in the southeast region of the US participated in a 1-session laboratory study. The participants' mean age was 19.9 ($SD = 1.69$). Participants were 62% Caucasian ($n = 31$), with the remaining racial and ethnic percentage including Native American, Asian/Asian American, African American, and Hispanic. 20% of participants reported level of proficiency and background with playing videogames to be skilled (from a range of not at all skilled to very skilled). The mean pre-test score was 55.81% ($SD = 2.99$), which indicates they had little prior knowledge in microbiology related topics covered in Crystal Island, the game-based learning environment used in this study. Participants were monetarily compensated $10/hour and received up to $25 for completing the study.

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

4.2. Materials

Participants completed several self-report measures, including a demographics questionnaire, and others related to emotions and motivation, including the Emotion-Values Questionnaire (Azevedo, Harley, Trevors, Feyzi-Behnagh, Duffy, Bouchet, & Landis, 2013; Pekrun, Elliot, & Maier, 2006), the Achievement Goal Questionnaire (Elliot & Murayama, 2008), the Intrinsic Motivation Inventory (Ryan, 1982), the Perceived Interest Scale (Schraw, 1997), and Presence Questionnaire (Witmer & Singer, 1998; Witmer, Jerome, & Singer, 2005). Participants' prior knowledge about microbiology was assessed with a researcher-developed four-choice, 21 multiple-choice question pre-test and a similar post-test (adopted from Nietfeld et al., 2014) immediately following game play. The questions contained 12 factual (e.g., declarative knowledge) and 9 procedural (e.g., *You observe a biological agent and notice that it does not have a nucleus. What type of agent might you be looking at?*) questions. These materials were chosen because we wanted to establish participants' levels of prior content knowledge and prior gaming experience, as well as establish a baseline measure of levels of emotions and motivation. However, the questionnaires and content tests were not analyzed for this study.

4.3. Apparatuses

4.3.1. Workstation

The equipment used in the study included a Dell Precision T7910 Workstation with 32 GB of memory and a 2.40 GHz Intel® Xeon® CPU E5-2630 v3. This computer was connected to an external monitor with a resolution of 1680x1050 with the SMI EYERED 250 eyetracker attached below. Video recording of the session was captured using a Logitech 920 webcam positioned over the monitor and processed using the FACET module for iMotions Attention Tool (2016). The participant was seated in front of the external monitor. The workstation also captured log-file data (i.e., information recorded about learner or system interactions, such as mouse clicks). This includes information regarding location, item, activity, and characters involved in any activity. For this study, we only analyzed eye-tracking and log-file data.

4.3.2. Eye tracking

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

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

4.4. Crystal Island

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

Crystal Island is a narrative-centered learning environment that combines both inquiry-based learning as well as direct instruction

[1] Out of a total of 77 participants; however 27 were not used for this analysis because they were in a control condition for which trace data were not collected. In this condition, participants viewed a video of someone playing and narrating while playing Crystal Island, and therefore participants did not engage in the in-game activities, thus not yielding any trace data.

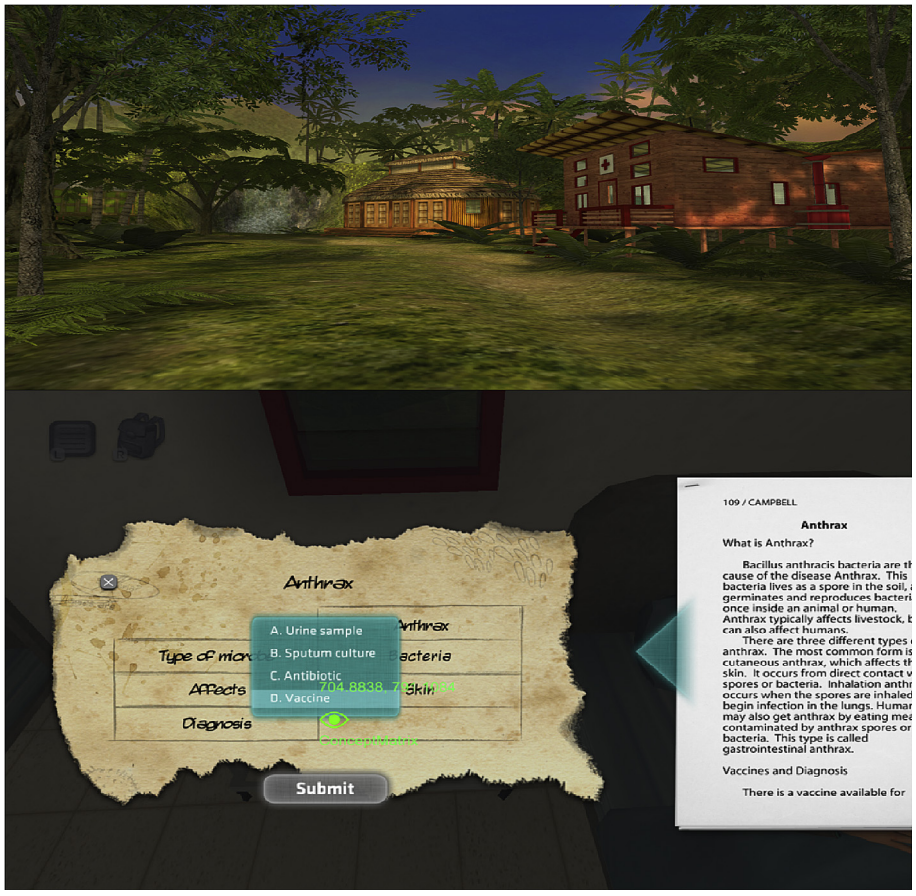

Fig. 1. Overview of Crystal Island (top) and book with concept matrix with eye tracking overlay (bottom).

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

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

4.4.1. Buildings

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

4.4.2. In-game activities

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

4.4.2.1. Conversations with non-player characters. The conversations throughout the game are multimodal representations of communication—they employ spoken language, facial expressions and text, as well as presenting the player with text dialogue options to choose from when communicating with the characters. The dialogue options are fixed dichotomies supplied by voice actors (Rowe et al., 2011). There are multiple non-player characters (NPCs)

that participants can converse with during gameplay. Kim is the camp nurse and teaches the player how to successfully use the diagnosis worksheet and advises that speaking with the sick patients is a good start to establishing a hypothesis for a diagnosis. There are three sick patients: Teresa, Greg, and Sam, who report on their symptoms and previously eaten food. The player may also converse with Bryce, Ford and Robert, who are experts in topics related to microbiology (such as viruses and bacteria), to gain a more sophisticated understanding of microbiology. The player can also talk to Quentin, the camp cook, who informs them of foods that inhabitants have been eating recently, in pursuance of a solidified and coherent hypothesis regarding the source of the illness. Finally, Elise works in the laboratory, and can assist with testing food items.

4.4.2.2. Testing food items. When participants collect food items, they can bring them to the laboratory to engage in hypothesis testing and test them for being potential transmission sources of the illness. The food items can be tested for contamination as viruses, bacteria, mutagens, or carcinogens. In addition, the scanner requires participants to input the reason for testing the item, for example, sick patient reported eating it. Reasons for testing food items include: "Sick members ate/drank it"; "It wasn't stored properly"; "It often carries disease"; "It looked dirty"; and "Sick members touched it". The scanner will then identify if the test is positive or negative, and if the test is positive, the participant can confirm hypotheses that the illness is of a particular type (e.g., virus), and that the transmission source is that positively tested item (e.g., milk). When these data are collected, the participant can gather these clues by entering them in their diagnosis worksheet.

4.4.2.3. Diagnosis worksheet. During investigation, the participant uses a diagnosis worksheet to track and organize pertinent information (findings, hypotheses, and final diagnosis). The diagnosis worksheet serves as a scaffolding element within the game, such that it is designed to support problem-solving processes as well as a space for the participant to record and offload notes and possible diagnoses (Lester, Spires, Nietfeld, Minogue, Mott, & Lobene, 2014). In this worksheet, participants can review their clues and reason about the diagnosis, transmission source, and treatment. Thus, the worksheet is a valuable monitoring tool participants can use to self-regulate while engaging in scientific reasoning processes to solve the science mystery. Once the player has narrowed down the diagnosis, transmission source, and treatment, they must then travel back to the infirmary in order to hand in the completed diagnosis worksheet to Kim, the camp nurse. If the diagnosis proves to be incorrect, Kim will point out the error and recommend the player to reconsider either the transmission source of the illness or the treatment plan. This feedback is valuable considering the player can use it to assess how close they are to solving the mystery and how to correct the original diagnosis. Once the player has listed the correct disease, source and treatment plan, and submitted the diagnosis to Kim, the science mystery is solved.

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

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

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

4.5. Experimental procedure

The study took place during a single session. Depending on condition assigned, the session lasted anywhere from one to two and a half hours ($M = 90.39$ min, $SD = 20.98$). At the start of the session, the participants were greeted in the laboratory, and asked to have a seat in front of the workstation. Once seated, they were presented and asked to sign the informed consent form. After the consent form was signed, participants were instrumented, and then presented with an overview of the study. Following the overview, participants began answering the pre-session questionnaires including the demographics questionnaire, Emotion-Values Questionnaire, Achievement Goal Questionnaire, and the pre-test about microbiology.

After completion of the questionnaires, the researcher began calibration of the eye-tracking equipment. A 9-point calibration was accomplished by asking the participant to keep their head still while focusing on a white dot that moved around the screen. Calibration was repeated until the participant reached a satisfactory level (offset of eyemovements that are less than 0.05 mm) or made 5 attempts. Once the participant completed eye-tracking calibration in Crystal Island, eye tracking was calibrated within Attention Tool. A baseline was then established for the facial recognition of emotion software as well as for the EDA bracelet in Attention Tool.

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

Following completion of the game, the participants were asked to complete a series of questionnaires including the Emotion-

Values Questionnaire, Intrinsic Motivation Inventory, Perceived Interest Scale, Presence Questionnaire, followed by the microbiology content post-test. Upon completion of the questionnaires, participants were debriefed, thanked, and paid for their participation.

4.5.1. Experimental conditions

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

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

4.6. Coding and scoring

For this study, we analyzed data from the log files and eye tracking, both of which were automatically generated from our trace data pipeline, which was developed to calculate a preset list of variables related to gameplay, learning, problem solving, scientific reasoning, and self-regulated learning (e.g., session duration, fixation duration, number of books, book duration, etc.). These data were calculated at each instance level, for each activity. An activity is one of the in-game activities described above (e.g., book and concept matrix or testing lab items), and an instance is one occurrence of that activity (e.g., opening book number 1 and opening book number 12 are separate instances within the book activity). More specifically, we have instance-level data for books and concept matrices, conversations with non-player characters, diagnosis worksheet opens and edits, testing lab items, etc. Therefore, for each activity, we have a data set for each participant, at every instance of engaging in that activity. For example, if a student read 20 books, we would have a list of variables for each of those 20 instances, yielding 20 rows of data for that participant. As such, since each participant read a number of books, this yielded 1198 total rows of data. For this analysis, we analyzed log files and eye-tracking data from the data pipeline, at the book instance level.

4.6.1. Log files

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

4.6.2. Eye tracking

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

$$\frac{\textit{Book Content Fixation Duration per Book Instance}}{\textit{Total Book Content Fixation Duration}} \quad (1)$$

and

$$\frac{\textit{Book Concept Matrix Fixation Duration per Book Instance}}{\textit{Total Book Concept Matrix Fixation Duration}}. \quad (2)$$

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

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

[2] In the *Partial Agency* condition, participants were required to read each book at least once, however this was not the focus of this study.

one another, as the data stems from the same participant. By using MLM, we can investigate multiple instances of events within students without having to satisfy this assumption, making MLM an ideal statistical technique when investigating learning and gameplay with GBLEs, as the typical behavior is to engage in cognitive and metacognitive processes multiple times. In addition, MLM allows for us to test for both between- and within-subject variance in the variable being investigated. For example, if we investigate the number of times a student reads a particular book, we can determine if this student reads books multiple times at the beginning of the learning session, compared to the end of the session, where they may read each book only once. This informs us of differences in reading behavior at different time points during the session, revealing within-subject variance in frequency of each book read. In addition, to compare between-subjects, we can investigate how one student's book reading behavior differs from another student. More specifically, we can determine if student *A* reads each book more frequently compared to student *B*, who only read each book once. As such, we can investigate both of these types of subject variance simultaneously when using MLM analyses. Therefore, for this analysis, we used MLM to investigate how students read multiple books, and how students completed multiple associated concept matrices during gameplay with Crystal Island.

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

5. Results

We used SAS software 9.4 (SAS Institute Inc., 2012) to run our analyses, with a restricted maximum likelihood (REML) estimation method, and a variance components covariance structure. The first step in using MLM requires a fully unconditional model, which allows us to determine if there is sufficient between- and within-subjects variance in our dependent variable, and the intra-class correlation coefficient (i.e., the percentage of variance explained at the between- and within-subject levels), prior to running models with predictor variables (i.e., a null model). For this study, our fully unconditional model was run with number of concept matrix attempts as our dependent variable, using the following equation with the individual and book instance levels, where β_{0ib} = the slope for the dependent variable, γ_{00} is the estimate of the grand mean of the number of concept matrix submission attempts, and r_{ib} and u_{0i} are the error terms at levels 1 and 2, respectively:

$$\textit{Level 1: Matrix Attempts}_{ib} = \beta_{0ib} + r_{ib}$$
$$\textit{Level 2: } \beta_{0i} = \gamma_{00} + u_{0i}$$

Results indicated that the grand mean (i.e., fixed effect) of the number of concept matrix submission attempts was significantly different from zero; $\gamma_{00} = 1.01$, $t = 19.75$, $p < 0.0001$. In addition, random effects results indicated that there was significant between- ($\tau_{00} = 0.16$, $z = 4.06$, $p < 0.0001$) and within-subjects (0.56, $z = 23.96$, $p < 0.0001$) variance in the number of concept matrix attempts, informing us that participants had different numbers of concept matrix attempts compared to each other, and had varying numbers of attempts at different points during gameplay. In addition, this model revealed that 17% of the variance in the number of concept matrix submission attempts was between-subjects, and 83% of the variance in the number of concept matrix submission attempts was within-subjects. Therefore, based on these results, there was sufficient variance at both levels of the dependent variable to proceed to running models with predictor variables.

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

We addressed this research question by running a random-intercept model with a fixed slope, with predictor variables at both levels 1 (frequency by title) and 2 (number of books). This model had the following equations, where β_{1ib} is the slope for the frequency by title predictor, γ_{01} is the association between the number of books and concept matrix attempt submissions, γ_{10} is the association between the frequency of book by title, and γ_{11} is the cross-level interaction between the number of books and the frequency of book by title on concept matrix submission attempts:

$$\textit{Level 1: Matrix Attempts}_{ib} = \beta_{0ib} + \beta_{1ib}(\textit{Frequency by Title}) + r_{ib}$$
$$\textit{Level 2:}$$
$$\beta_{0i} = \gamma_{00} + \gamma_{01}(\textit{Number of Books}) + u_{0i}$$
$$\beta_{1i} = \gamma_{10} + \gamma_{11}(\textit{Number of Books})$$

Results from this model (see Table 1 for an overview) revealed a significant association between number of books and concept matrix attempts; $\gamma_{01} = -0.014$, $t = -2.20$, $p = 0.0329$, such that an increase in books is associated with a decrease in concept matrix attempts, demonstrating that opening books leads to better performance on the concept matrices. Results also indicated a significant association between the frequency of books by title; $\gamma_{10} = -0.44$, $t = -4.94$, $p < 0.0001$, with an increase in books associated with a decrease in concept matrix attempts, revealing that if students read a book multiple times, they will perform better on the concept matrices. However, a significant cross-level interaction (see Fig. 2); $\gamma_{11} = 0.005$, $t = 1.98$, $p = 0.047$, revealed that the association between number of books, frequency of books by title, and concept matrix attempts is especially strong for fewer books and higher frequencies of books by title, such that the fewest attempts, and thus the best performance, was for participants who read fewer books, but read each book more frequently. In contrast, participants with the most concept matrix submission attempts, and thus performed the worst, read fewer books overall, and read each book less frequently. This model accounted for 46.14% of the between-subjects variance, and 6.14% of the within-subjects variance in concept matrix submission attempts.

Table 1
Unstandardized coefficients for number of books and frequency by title by concept matrix attempts.

Fixed effects	Estimate (std. error)	*t*
Matrix attempts, β_0		
Intercept, Υ_{00}	1.88 (0.18)	10.56^{***}
No. books, Υ_{01}	−0.014 (0.007)	-2.20^{*}
Freq. by title slope, β_1		
Intercept, Υ_{10}	−0.44 (0.089)	-4.94^{***}
Books*freq. by title, Υ_{11}	0.005 (0.003)	1.98^{*}
Random effects	**Estimate (std. error)**	***z***
Matrix attempts (τ_{00})	0.062 (0.019)	3.26^{**}
Within-person fluctuation (σ^2)	0.52 (0.022)	23.85^{***}

$^{*}p < 0.05$, $^{**}p < 0.001$, $^{***}p < 0.0001$.

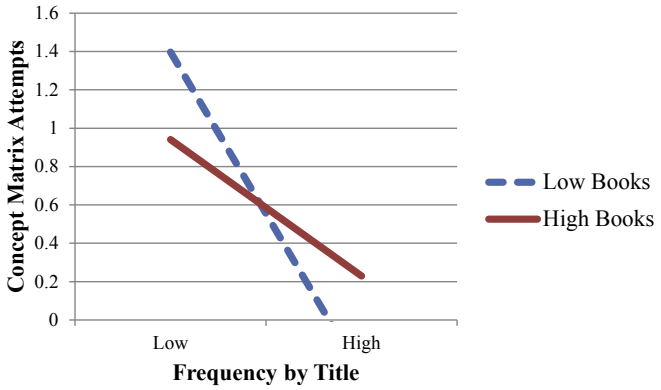

Fig. 2. Cross-level interaction between books and frequency of book by title on concept matrix submission attempts.

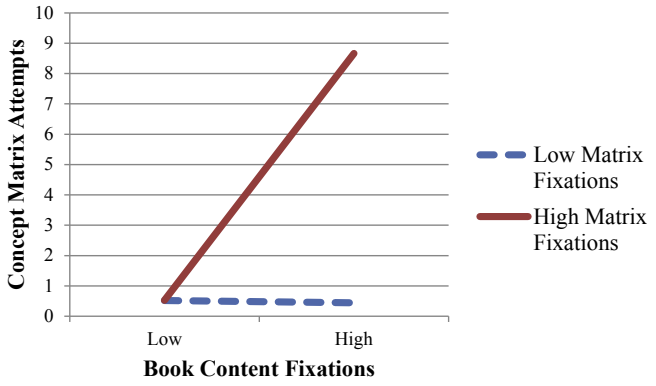

Fig. 3. Interaction between the proportions of fixations on book content and book concept matrices on concept matrix submission attempts.

5.2. Research question 2: Is there an association between the proportion of fixations on book content and on book concept matrices with the number of concept matrix submission attempts?

To address this research question, we ran a level 1 moderation model with constrained slopes, with the following equations, where β_{1ib} is the slope and Υ_{10} the effect of proportion of fixations on book content, β_{2ib} is the slope and Υ_{20} is the effect of proportion of fixations on book concept matrices, and β_{3ib} is the slope and Υ_{30} is the effect of the interaction:

$$\textit{Level 1: Matrix Attempts}_{ib} = \beta_{0ib} + \beta_{1ib}(\textit{fixC}) + \beta_{2ib}(\textit{fixM}) + \beta_{3ib}(\textit{fixC}^{*}\textit{fixM}) + r_{ib}$$

Level 2:

$$\beta_{0i} = \Upsilon_{00} + u_{0i}$$
$$\beta_{1i} = \Upsilon_{10}$$
$$\beta_{2i} = \Upsilon_{20}$$
$$\beta_{3i} = \Upsilon_{30}$$

Results revealed (see Table 2) no significant association between the proportion of fixations on book content (γ_{10} = -0.080, t = -0.56, $p = 0.57$) and no significant association between the proportion of fixations on book concept matrices ($\gamma_{20} = 0.013$, $t = 0.06$, $p = 0.95$). However, there was a significant interaction ($\gamma_{30} = 10.17$, $t = 14.31$, $p < 0.0001$), revealing that the effect between proportions of fixations on book content and book concept matrices is especially detrimental for poorer performance, such that students with high proportions of fixations on both the book content and book concept matrices had the highest number of concept matrix submission attempts (see Fig. 3). Overall, this model accounted for 1.39% of the between-subjects variance and 23.1% of the within-subjects variance in concept matrix submission attempts.

Table 2
Unstandardized coefficients for proportions of fixations on book content and book concept matrices by concept matrix attempts.

Fixed effects	Estimate (std. error)	t
Matrix attempts, β_0		
Intercept, Υ_{00}	0.52 (0.082)	6.26^{***}
Content fixations slope, β_1		
Intercept, Υ_{10}	−0.080 (0.14)	−0.56
Matrix fixations slope, β_2		
Intercept, Υ_{20}	0.013 (0.21)	0.06
Interaction slope, β_3		
Intercept, Υ_{30}	10.17 (0.71)	14.31^{***}
Random effects	**Estimate (std. error)**	z
Matrix attempts (τ_{00})	0.11 (0.028)	4.09^{***}
Within-person fluctuation (σ^2)	0.43 (0.018)	23.90^{***}

$^{*}p < 0.05$, $^{**}p < 0.001$, $^{***}p < 0.0001$.

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

To test for the cross-level interaction, we first examined the unique associations between each predictor variable and the number of concept matrix submission attempts, and then tested for the interaction between all four predictor variables and the number of concept matrix submission attempts. We used the following equations, with the cross-level interaction slope as β_{4ib}, and the interaction effect as Υ_{41}:

$$\textit{Level 1: Matrix Attempts}_{ib} = \beta_{0ib} + \beta_{2ib}(\textit{FreqbyTitle}) + \beta_{2ib}(\textit{fixC}) + \beta_{3ib}(\textit{fixM}) + \beta_{4ib}(\textit{FreqbyTitle}^{*}\textit{fixC}^{*}\textit{fixM}) + r_{ib}$$

Level 2:

$$\beta_{0i} = \Upsilon_{00} + \Upsilon_{01}(\textit{Number of Books}) + u_{0i}$$
$$\beta_{1i} = \Upsilon_{10} + \Upsilon_{11}(\textit{Number of Books})$$
$$\beta_{2i} = \Upsilon_{20} + \Upsilon_{21}(\textit{Number of Books})$$
$$\beta_{3i} = \Upsilon_{30} + \Upsilon_{31}(\textit{Number of Books})$$
$$\beta_{4i} = \Upsilon_{40} + \Upsilon_{41}(\textit{Number of Books})$$

For this model, we only investigated the unique associations between each predictor and the DV and the cross-level interaction, and not all other combinations of interactions (e.g., frequency by title*content fixations*number of books, etc.), thereby not obtaining all possible gamma values (i.e., Υ_{11}, Υ_{21}, Υ_{31}, and Υ_{40} were not included).[3] Results revealed a significant association between number of books (γ_{01} = -0.011, $t = -2.17$, $p = 0.035$), frequency of books by title (γ_{10} = -0.35, $t = -12.01$, $p < 0.0001$), proportion of fixations on book content ($\gamma_{20} = 0.50$, $t = 4.06$, $p < 0.0001$), and proportion of fixations on book concept matrices ($\gamma_{30} = 1.04$, $t = 5.50$, $p < 0.0001$) and concept matrix submission attempts (see Table 3), revealing all significant unique associations between each predictor variable and the dependent variable. Additionally, we found a significant interaction; $\gamma_{41} = 0.077$, $t = 10.41$, $p < 0.0001$, demonstrating the importance of including all of these variables in

[3] This was done to simplify our analyses, as if we had examined all combinations (e.g., interactions between a combination of some level 1 predictors and number of books), this would have yielded far too many results.

Table 3
Unstandardized coefficients for number of books, frequency by title, and proportions of fixations on book content and book concept matrices by concept matrix attempts.

Fixed effects	Estimate (std. error)	*t*
Matrix attempts, v		
Intercept, Υ_{00}	1.32 (0.142)	9.32***
No. books, Υ_{01}	−0.011 (0.005)	−2.17*
Freq. by title slope, β_1		
Intercept, Υ_{10}	−0.35 (0.029)	−12.01***
Content fixations slope, β_2		
Intercept, Υ_{20}	0.50 (0.12)	4.06***
Matrix fixations slope, β_3		
Intercept, Υ_{30}	1.04 (0.19)	5.50***
Interaction slope, β_4		
Interaction, Υ_{41}	0.077 (0.007)	10.41***
Random effects	**Estimate (std. error)**	***z***
Matrix attempts (τ_{00})	0.070 (0.020)	3.56**
Within-person fluctuation (σ^2)	0.44 (0.018)	23.85***

*$p < 0.05$, **$p < 0.001$, ***$p < 0.0001$.

predicting in-game assessment performance during gameplay with Crystal Island. Finally, this model accounted for 38.99% of the between-subjects variance and 21.82% of the within-subjects variance in concept matrix submission attempts.

6. Discussion

Overall, results from the log-file data indicate that the number of books read and the frequency of reading each book were both negatively related to concept matrix submission attempts, when assessing their unique associations with the number of concept matrix submission attempts. In addition to main effects of each variable, there was a significant interaction between both variables, such that reading fewer books, but reading each book more frequently was associated with fewer attempts, and thus better performance. With regards to eye tracking, neither variable yielded a significant main effect, however the interaction between the two was significant, such that participants with high proportions of fixations on both the book content and book concept matrices had the most matrix attempts, implying worse performance on the matrices. However, when we combined the log files and eye tracking, we found a significant main effect for each of the four predictor variables, and found an interaction effect, revealing that our most significant results were those that included online trace data from both log files and eye tracking. These findings demonstrate the importance of using different types of data to investigate how participants engage in scientific reasoning and self-regulated learning during gameplay with GBLEs.

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

For this research question that used log-file data, we hypothesized that there would be a significant negative association between the number of books read and the frequency of books read by title. Results revealed that there was a significant negative association (i.e., main effect) between the number of books and the number of concept matrix attempts, and a significant negative association between the frequency of books by title and the number of concept matrix attempts, which supports our hypothesis because we predicted that reading more would result in fewer attempts (i.e., negative associations). However, we also found a significant interaction, which partially supports our hypothesis because the fewest number of concept matrix submission attempts occurred when participants read fewer books (positive association) and higher frequencies of books by title (negative association). These results emphasize the importance of quality vs. quantity, such that reading more books did not lead to better performance, but reading each book several times did lead to better performance. As such, better performance was not predicted by reading a large amount of different material; better performance occurs when participants were reading fewer, but specific materials more often, and potentially more in-depth.

As such, our study extends on investigating the quality of performance, such that we analyzed how participants performed on each specific concept matrix, as opposed to investigating the quantity of concept matrices completed, or investigating the quality of completing all concept matrices in one composite score. Studies presented in this paper have investigated the quantity of performance, such as Sabourin (2013) who examined the frequency of use of cognitive and metacognitive SRL processes during gameplay with Crystal Island. However, other studies have investigated the quality of learning behaviors, such as studies conducted by Snow et al. (2015), who found that more controlled behaviors were associated with better learning outcomes; and Nietfeld et al. (2014), who scored participants' responses in the diagnosis worksheet in Crystal Island based on thoroughness and accuracy, and found that higher scores were associated with better overall learning. Although these two studies did investigate the quality of performance, they did not assess the quality for each specific activity.

The results from this study are supported by Winne and Hadwin's (1998, 2008) model because the model focuses on how SRL temporally unfolds during learning, and so the use of a strategy can increasingly improve the more it is used, as opposed to using many different strategies only one time, as this would not allow for participants to improve how they use these strategies. For example, if a student only reads each book one time and thus completes one concept matrix, their performance cannot improve on the matrix because they do not allow for more opportunities to complete the matrix. However, frequency data would show us that this student is reading a lot of material, even though performance is low. In contrast, if the student selects an appropriate number of books, and reads those books multiple times, they can improve on how they complete the matrices, resulting in better performance, even though the total frequency of individual books is not higher than the other student (i.e., better quality, not higher quantity). Therefore, the distinction between quality and quantity can lead to more in-depth analyses and results, revealing more specifically how participants are strategically learning during gameplay with GBLEs.

6.2. Research question 2: Is there an association between the proportion of fixations on book content and on book concept matrices with the number of concept matrix submission attempts?

We investigated eye-tracking data for this research question, and hypothesized that more fixations on the book content and the book concept matrices would lead to fewer concept matrix submission attempts, and thus better performance (i.e., a negative association). However, our results were only marginally partially supported, such that there was no unique association between the proportions of fixations on book content with concept matrix submission attempts, and no unique association between the proportions of fixations on book concept matrices with concept matrix submission attempts, thus not supporting our hypothesis. However, we did find a significant interaction effect, where the fewest concept matrix submission attempts (i.e., best performance) were associated with low proportions of fixations on book content and low proportions of fixations on book concept matrices. This does not support the direction of our hypothesis since we found a

positive association between the variables, and we predicted a negative association. The lack of significant main effects, but a significant interaction effect demonstrates the importance of investigating cognitive and metacognitive processes together, and how they jointly predict performance, as opposed to singling them out. In addition, the low proportions of fixations predicting better performance may result from participants strategizing by examining the questions in the concept matrix and going back to the text to search for the correct responses instead of reading through the entire text and then answering the concept matrices, however we cannot confirm this because we did not sequentially analyze behaviors within each book instance, which we will aim to do in future studies.

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

In addition, this strategizing behavior seen in our study could relate to how participants were processing the information in the books and concept matrices. Specifically, instead of reading a large range of material, participants spent more of their time processing smaller amounts of information more thoroughly. Thus, participants seemed to be strategizing and selecting the most relevant material for completing the concept matrices, and then spending valuable time reading and processing that specific content. This relates back to our findings from the first research question regarding quality vs. quantity, such that participants who performed better on the concept matrices were not reading larger quantities of information. Rather, the quality of what they were reading showed more thorough investigation of the content. This once again supports Winne and Hadwin's IPT model because participants seem to be taking their time and potentially using operations, referred to as SMART (searching, monitoring, assembling, rehearsing, and translating) during those shorter durations. Therefore, participants are selecting specific elements within the text, which are relevant to completing the concept matrices correctly, to apply these operations to. Searching for relevant material is known as the metacognitive strategy of content evaluation (CE; Azevedo, 2009; Greene & Azevedo, 2009); therefore, these participants could be using CEs as well, which is resulting in better performance. Overall, it seems that they are monitoring the material and controlling what they choose to read depending on what is relevant for them to perform well on the assessments. As such, this behavior demonstrates that participants are engaging in monitoring and control strategies, which is also supported by the IPT model and other studies assessing the role of cognitive and metacognitive processes with other advanced learning technologies (e.g., Azevedo et al., 2013, 2015).

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

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

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

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

6.4. Limitations

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

Additionally, we conceptualized JOLs with Crystal Island using a valence that is based on the correctness of performance on the concept matrices, as opposed to asking them to judge their understanding of the material for that book. As such, we are inferring correctness based on performance, as well as inferring that participants are ready to complete the assessment (and have thus judged that they have read enough material to complete the assessment) when they switch from the book to the concept matrix. This is in contrast to assessing JOLs with hypermedia-learning environments where participants either select or are prompted to make a JOL (e.g., Azevedo, 2014; Greene & Azevedo, 2009). Thus, these participants do not choose to make the assessment, rather they choose to make the judgment, which is followed by the assessment. Our measure of JOLs within Crystal Island did not actually allow for participants to overtly make a judgment that we could measure without inferring it based on their behavior.

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

6.5. Implications and future directions

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

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

Finally, these results can inform the design of GBLEs to adaptively support individual learner's cognitive and metacognitive processes during scientific reasoning. For example, if the participant is not reading a lot, and is not completing the matrices, the system can prompt them to open relevant books and complete the associated assessment. In addition, we can use multi-channel data to drive real-time scaffolding for participants. For example, if the eye-tracking data indicates that a participant is spending a significant amount of time reading text in a non-strategic fashion (e.g., no evidence of metacognitive judgments and use of cognitive strategies), the system can model a more efficient way to read books and complete concept matrices, which can result in better performance on the matrices. Specifically, we can model efficient eye movements (e.g., D'Mello, 2016; Jarodzka, van Gog, Door, Scheiter, & Gerjets, 2013) that will scaffold participants to use more effective self-regulated learning strategies. The ultimate goal for designing these GBLEs is to foster the most effective learning for students, which can be demonstrated through high overall performance and accurate metacognitive monitoring and effective cognitive strategy use (Azevedo et al., in press). As such, if adding an adaptive component to these environments can improve in-game and overall performance, which we can detect by examining participants' performance on in-game assessments and post-test scores at the end of gameplay, this confirms that these environments should continue to be used for all different types of learners, in different educational settings to foster and scaffold effective learning, gameplay, SRL, and scientific reasoning.

Acknowledgment

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

References

Adams, D. M., Mayer, R. E., McNamara, A., Koenig, A., & Wainess, R. (2012). Narrative games for learning: Testing the discovery and narrative hypotheses. *Journal of Educational Psychology, 104*, 235–249.

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

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

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

Azevedo, R., Harley, J., Trevors, G., Duffy, M., Feyzi-Behnagh, R., Bouchet, F., et al. (2013). Using trace data to examine the complex roles of cognitive, metacognitive, and emotional self-regulatory processes during learning with multi-agent systems. In R. Azevedo, & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 427–449). Amsterdam, The Netherlands: Springer.

Azevedo, R., Johnson, A., Chauncey, A., & Graesser, A. (2011). Use of hypermedia to convey and assess self-regulated learning. In B. Zimmerman, & D. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 102–121). New

York, NY: Routledge.
Azevedo, R., Taub, M., & Mudrick, N. (2015). Technologies supporting self-regulated learning. In M. Spector, C. Kim, T. Johnson, W. Savenye, D. Ifenthaler, & G. Del Rio (Eds.), *The SAGE Encyclopedia of educational technology* (pp. 731–734). Thousand Oaks, CA: SAGE.
Azevedo, R., Taub, M., & Mudrick, N. V. Using multi-channel trace data to infer and foster self-regulated learning between humans and advanced learning technologies. In. D. Schunk & Greene, J.A (Eds.), Handbook of self-regulation of learning and performance (second edition). New York, NY: Routledge, in press.
Bondareva, D., Conati, C., Feyzi-Behnagh, R., Harley, J., Azevedo, R., & Bouchet, F. (2013). Inferring learning from gaze data during interaction with an environment to support self-regulated learning. In C. H. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Proceedings of the 16th international conference on artificial intelligence in education—lecture notes in artificial intelligence 7926* (pp. 229–238). Berlin, Heidelberg: Springer-Verlag.
Cagiltay, N. E., Ozcelik, E., & Ozcelik, N. S. (2015). The effect of competition on learning in games. *Computers & Education, 87*, 35–41.
Calderón, A., & Ruiz, M. (2015). A systematic literature review on serious games evaluation: An application to software project management. *Computers & Education, 87*, 396–422.
Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2016). Digital games, design, and learning: A systematic review and meta-analysis. *Review of Educational Research, 86*, 79–122.
Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., & Boyle, J. M. (2012). A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education, 59*, 661–686.
Cromley, J. C., & Azevedo, R. (2009). Location information within extended hypermedia. *Educational Technology Research and Development, 57*, 287–313.
D'Mello, S. K. (2016). Giving eyesight to the blind: Towards attention-aware AIED. *International Journal of Artificial Intelligence in Education, 26*, 645–659.
D'Mello, S. K., & Graesser, A. C. (2012). Dynamics of affective states during complex learning. *Learning and Instruction, 22*, 145–157.
Elliot, A. J., & Murayama, K. (2008). On the measurement of achievement goals: Critique, illustration, and application. *Journal of Educational Psychology, 100*, 613–628.
Filsecker, M., & Kerres, M. (2014). Engagement as a volitional construct: A framework for evidence-based research on educational games. *Simulation & Gaming, 45*, 450–470.
Girard, C., Escalle, J., & Magnan, A. (2012). Serious games as new educational tools: How effective are they? A meta-analysis of recent studies. *Journal of Computer Assisted Learning, 29*, 207–219.
Graesser, A. C. (2015). Deeper learning with advances in discourse science and technology. *Policy Insights from Behavioral and Brain Sciences, 2*, 42–50.
Greene, J. A., & Azevedo, R. (2009). A macro-level analysis of SRL processes and their relations to the acquisition of sophisticated mental models. *Contemporary Educational Psychology, 34*, 18–29.
Ha, E. Y., Rowe, J. P., Mott, B. W., & Lester, J. C. (2011). Goal recognition with Markov logic networks for player-adaptive games. In V. Bulitko, & M. Riedl (Eds.), *Proceedings of the Seventh AAAI conference on artificial intelligence and interactive digital entertainment* (pp. 32–39). Menlo Park, CA: AAAI Press.
Hyönä, J., & Nurminen, A. M. (2006). Do adult readers know how they read? Evidence from eye movement patterns and verbal reports. *British Journal of Psychology, 97*, 31–50.
iMotions. (2016). *Attention tool (version 6.0)* [Computer software]. Boston, MA: iMotions Inc.
Jaques, N., Conati, C., Harley, J., & Azevedo, R. (2014). Predicting affect from gaze data during interaction with an intelligent tutoring system. In S. Trausan-Matu, K. E. Boyer, M. Crosby, & K. Panourgia (Eds.), *Proceedings of the 12th international conference on intelligent tutoring systems—lecture notes in computer science 8474* (pp. 29–38). Amsterdam, The Netherlands: Springer.
Jarodzka, H., van Gog, T., Door, M., Scheiter, K., & Gerjets, P. (2013). Learning to see: Guiding students' attention via a model's eye movements fosters learning. *Learning and Instruction, 25*, 62–70.
Lee, S., Mott, B., & Lester, J. (2011). Director agent intervention strategies for interactive narrative environments. In M. Si, D. Thue, E. André, J. Lester, & J. Tanenbaum (Eds.), *Proceedings of the 4th international conference on interactive digital storytelling* (pp. 140–151). Vancouver, Canada: Springer-Verlag.
Lee, S., Rowe, J., Mott, B., & Lester, J. (2014). A supervised learning framework for modeling director agent strategies in educational interactive narrative. *IEEE Transactions on Computational Intelligence and AI in Games, 6*, 203–215.
Lester, J., Ha, E. Y., Lee, S., Mott, B., Rowe, J., & Sabourin, J. (2013). Serious games get smart: Intelligent game-based learning environments. *AI Magazine, 34*, 31–45.
Lester, J. C., Mott, B. W., Robinson, J. L., & Rowe, J. P. (2013). Supporting self-regulated science learning in narrative-centered learning environments. In R. Azevedo, & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 471–483). Amsterdam, The Netherlands: Springer.
Lester, J., Spires, H. A., Nietfeld, J., Minogue, J., Mott, B., & Lobene, E. (2014). Designing game-based learning environments for elementary science education: A narrative-centered learning perspective. *Information Sciences, 264*, 4–18.
Mayer, R. E. (2010). Unique contributions of eye-tracking research to the study of learning with graphics. *Learning and Instruction, 20*, 167–171.
Mayer, R. E. (2014). *Computer games for learning: An evidence-based approach*. Cambridge, MA: MIT Press.
Mayer, R. E. (2015). On the need for research evidence to guide the design of computer games for learning. *Educational Psychologist, 50*, 349–353.
McNamara, D. S., & Shapiro, A. M. (2005). Multimedia and hypermedia solutions for promoting metacognitive engagement, coherence, and learning. *Journal of Educational Computing Research, 33*, 1–29.
McQuiggan, S., Rowe, J., & Lester, J. (2008). The Effects of empathetic virtual characters on presence in narrative-centered learning environments. In M. Czerwinski, A. Lund, & D. Tan (Eds.), *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1511–1520). New York: ACM Press.
Millis, K., Forsyth, C., Butler, H., Wallace, P., Graesser, A., & Halpern, D. (2011). Operation ARIES: A serious game for teaching scientific inquiry. In M. Ma, A. Oikonomou, & L. C. Jain (Eds.), *Serious games and entertainment applications* (pp. 169–195). London, UK: Springer-Verlag.
Min, W., Mott, B., Rowe, J., Liu, B., & Lester, J. (2016). Player goal recognition in open-world digital games with long short-term memory networks. In G. Brewka, & S. Kambhampati (Eds.), *Proceedings of the 25th international joint conference on artificial intelligence*. Menlo Park, CA: AAAI Press.
Min, W., Rowe, J., Mott, B., & Lester, J. (2013). Personalizing embedded assessment sequences in narrative-centered learning environments: A collaborative filtering approach. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Proceedings of the 16th international conference on artificial intelligence in education—Lecture notes in artificial intelligence 7926* (pp. 369–378). Berlin, Germany: Springer Verlag.
Min, W., Wiggins, J., Pezzullo, L., Vail, A., Boyer, K. E., Mott, B., Frankosky, M., Wiebe, E., & Lester, J. (2016). Predicting dialogue acts for intelligent virtual agents with multimodal student interaction data. In T. Barnes, M. Chi, & M. Feng (Eds.), *Proceedings of the 9th international conference on educational data mining*. Raleigh, NC: International Educational Data Mining Society.
Mott, B., & Lester, J. (2006). Narrative-centered tutorial planning for inquiry-based learning environments. In M. Ikeda, K. Ashley, & T.-W. Chan (Eds.), *Proceedings of the 8th international conference on intelligent tutoring systems* (pp. 675–684). Berlin: Springer-Verlag.
Nietfeld, J. L., Shores, L. R., & Hoffmann, K. F. (2014). Learning environment self-regulation and gender within a game-based learning environment. *Journal of Educational Psychology, 106*, 961–973.
Pekrun, R., Elliot, A. J., & Maier, M. A. (2006). Achievement goals and discrete achievement emotions: A theoretical model and prospective test. *Journal of Educational Psychology, 98*, 583–597.
Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 451–502). San Diego, CA: Academic Press.
Plass, J. L., Homer, B. D., & Kinzer, C. K. (2015). Foundations of game-based learning. *Educational Psychologist, 50*, 258–284.
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.
Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
Rowe, J., Ha, E. Y., & Lester, J. (2008). Archetype-driven character dialogue generation for interactive narrative. In H. Prendinger, J. Lester, & M. Ishizuka (Eds.), *Proceedings of the 8th international conference on intelligent virtual agents* (pp. 45–58). Berlin, Heidelberg: Springer-Verlag.
Rowe, J., & Lester, J. (2010). Modeling user knowledge with dynamic bayesian networks in interactive narrative environments. In G. M. Youngblood, & B. Bulitko (Eds.), *Proceedings of the 6th artificial intelligence and interactive digital entertainment conference* (pp. 57–62). Menlo Park, CA: AAAI Press.
Rowe, J., & Lester, J. (2015). Improving student problem solving in narrative-centered learning environments: a modular reinforcement learning framework. In C. Conati, N. Heffernan, A. Mitrovic, & M. F. Verdejo (Eds.), *Proceedings of the 17th international conference on artificial intelligence in education—lecture notes in artificial intelligence 9112* (pp. 419–428). Basel, Switzerland: Springer International.
Rowe, J., Shores, L., Mott, B., & Lester, J. (2010). Individual differences in gameplay and learning: a narrative-centered learning perspective. In I. Horswill, & Y. Pisan (Eds.), *Proceedings of the 5th international conference on foundations of digital games* (pp. 171–178). Monterey, CA: ACM Press.
Rowe, J., Shores, L., Mott, B., & Lester, J. (2011). Integrating learning, problem solving, and engagement in narrative-centered learning environments. *International Journal of Artificial Intelligence in Education, 21*, 115–133.
Ryan, R. M. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology, 43*, 450–461.
Sabourin, J. L. (2013). *Stealth assessment of self-regulated learning in game-based learning environments* (Doctoral dissertation). Retrieved from Dissertation abstracts international. (2586200).
Sabourin, J. L., & Lester, J. C. (2014). Affect and engagement in game-based learning environments. *IEEE Transactions on Affective Computing, 5*, 45–56.
Sabourin, J., Mott, B., & Lester, J. (2011). Modeling learner affect with theoretically grounded dynamic Bayesian networks. In S. D'Mello, A. Graesser, B. Schuller, & J. C. Martin (Eds.), *Proceedings of the 4th international conference on affective computing and intelligent interaction* (pp. 286–295). Berlin, Heidelberg: Springer Verlag.
Sabourin, J., Mott, B., & Lester, J. (2013). Discovering behavior patterns of self-regulated learners in an inquiry-based learning environment. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Proceedings of the 16th international conference on artificial intelligence in education—lecture notes in artificial intelligence 7926* (pp. 209–218). Berlin, Heidelberg: Springer-Verlag.
Sabourin, J., Rowe, J., Mott, B., & Lester, J. (2012). Exploring inquiry-based problem-

solving strategies in game-based learning environments. In S. A. Cerri, W. J. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Proceedings of the 11th international conference on intelligent tutoring systems—lecture notes in computer science 7315* (pp. 59–64). Amsterdam, The Netherlands: Springer.

SAS Institute Inc. (2012). *SAS software (version 9.4)* [Software]. Cary, NC: SAS Institute Inc.

Scheiter, K., & van Gog, T. (2009). Using eye tracking in applied research to study and stimulate the processing of information from multi-representational sources. *Applied Cognitive Psychology, 23*, 1209–1214.

Schraw, G. (1997). Situational interest in literary text. *Contemporary Educational Psychology, 22*, 436–456.

Schunk D. H. & Greene J. A., (Eds.), *Handbook of self-regulation of learning and performance*, (2nd ed.). New York, NY: Routledge (in press).

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer Games and Instruction, 55*, 503–524.

Shute V. J,. & Moore G. R. Consistency and validated in game-based stealth assessment, In: H. Jiao R.W. Lissitz (Eds.), *Technology enhanced innovative assessment: Development, modeling, and scoring from an interdisciplinary perspective* Charlotte, NC: Information Age Publisher (in press).

Shute, V. J., & Ventura, M. (2013). *Measuring and supporting learning in games: Stealth assessment*. Cambridge, MA: The MIT Press.

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

Snow, E. L., Allen, L. K., Jacovina, M. W., & McNamara, D. S. (2015). Does agency matter? Exploring the impact of controlled behaviors within a game-based environment. *Computers & Education, 82*, 378–392.

Spires, H. A., Rowe, J. P., Mott, B. W., & Lester, J. C. (2011). Problem solving and game-based learning: Effects of middle grade students' hypothesis testing strategies on learning outcomes. *Journal of Educational Computing Research, 44*, 453–472.

Taub, M., & Azevedo, R. (2016). Using eye-tracking to determine the impact of prior knowledge on self-regulated learning with an adaptive hypermedia- learning environment? In A. Micarelli, J. Stamper, & K. Panourgia (Eds.), *Proceedings of the 13th international conference on intelligent tutoring systems—lecture notes in computer science 9684* (pp. 34–47). The Netherlands: Springer.

Tsai, M. J., Huang, L. J., Hou, H. T., Hsu, C. Y., & Chiou, G. L. (2016). Visual behavior, blow and achievement in game-based learning. *Computers & Education, 98*, 115–129.

Unity Game Engine (Version 5) [Software]. (2015). San Francisco, CA, USA: Unity Technologies.

VanLehn, K. (2016). Regulative loops, step loops and task loops. *International Journal of Artificial Intelligence in Education, 26*, 107–112.

van Gog, T., & Jarodzka, H. (2013). Eye tracking as a tool to study and enhance cognitive and metacognitive processes in computer-based learning environments. In R. Azevedo, & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 143–156). Amsterdam, The Netherlands: Springer.

van Gog, T., Jarodzka, H., Scheiter, K., Gerjets, P., & Paas, F. (2009). Attention guidance during example study via the model's eye movements. *Computers in Human Behavior, 25*, 785–791.

Winne, P., & Azevedo, R. (2014). Metacognition. In K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (2nd ed., pp. 63–87). Cambridge, MA: Cambridge University Press.

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

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

Witmer, B. G., Jerome, C. J., & Singer, M. J. (2005). The factor structure of the presence questionnaire. *Presence, 14*, 298–312.

Witmer, B. G., & Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence, 7*, 225–240.

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*, 249–265.

Zimmerman, B., & Schunk, D. (Eds.). (2011). *Handbook of self-regulation of learning and performance*. New York, NY: Routledge.