Development of Self-Regulated Learning Skills Within Open-Ended Computer-Based Learning Environments for Science

Yang Jiang

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2018

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

ABSTRACT

Development of Self-Regulated Learning Skills Within Open-Ended Computer-Based Learning Environments for Science

Yang Jiang

Over the past decade, open-ended computer-based learning environments have been increasingly used to facilitate students' learning of complex scientific topics. The non-linearity and open-endedness of these environments create learning opportunities for students, but can also impose challenges in terms of extraneous cognitive load and greater requirements for self-regulated learning (SRL). SRL is crucial for academic success in various educational settings. This dissertation explores how self-regulatory skills develop and the role of gender in the development of SRL skills in Virtual Performance Assessments (VPA), an immersive, open-ended virtual environment designed to assess middle school students' science inquiry skills. Findings from three analyses combining educational data mining techniques with multilevel modeling indicated that students developed self-regulatory behaviors and strategies as they used VPA. For example, experience with VPA prepared students to adopt more efficient note-taking and note-reviewing strategies. Students who used VPA for the second time engaged in note-taking more frequently, noted a significantly higher quantity of unique information, used the control of variables strategy more frequently in note-taking, and reproduced more domain-specific declarative information in notes than students who used VPA for the first time, all of which have been found to be positively associated with science inquiry performance. Students also learned to exploit more available sources of information by applying learning strategies, in order to either solve inquiry problems, or to monitor and evaluate their solutions. Compared to

the second-time users who focused primarily on answering the core inquiry question and selectively collected data, the first-time users’ behaviors showed the repetition and combination of exploratory actions such as talking with non-player characters and collecting data. In addition, consistent gender differences in SRL were observed in this study. Female students were more likely to take notes than male students; they took notes and reviewed notes more frequently and recorded a higher quantity of information in notes, especially information from the research kiosk. Females were also more likely to review notes or read research pages to assist them with the problem-solving and decision-making process than their male counterparts. Possibly due to the higher quantity of information recorded by female note-takers and their tendency to review notes over males, female students’ performance on science inquiry tasks improved across the course of using the two scenarios of VPA, whereas the male students’ science inquiry skills did not show improvement. Results from this dissertation study provide insights into the instructional design of personalized open-ended learning environments to facilitate self-regulated learning for both male and female students.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Dr. Ryan Baker, for his unparalleled mentorship and unwavering support throughout my doctoral study. I am extremely grateful to you for guiding me to the field of learning analytics and educational data mining, which I am highly passionate about and am ready to embark on my career in. I am deeply honored to be one of your students; I cannot thank you enough for teaching me what research means, challenging me to think more and deeper, and inspiring me to create new ways to analyze data. This dissertation would not have been possible without your insights, help, and support. You have made my Ph.D. journey special and rewarding. Your guidance and advice, no matter in research, career, or life, have shaped me as a better researcher and person; they are lifelong treasures to me.

I would also like to extend my sincere appreciation to my dissertation committee members for their thoughtful questions and invaluable feedback throughout the different phases of the process. I would like to thank Dr. John Black, who was the first professor I met at Columbia University and who funded me during my first three years of doctoral study. Thank you for guiding me to Teachers College and for all your help and support. To Dr. Jody Clarke-Midura, with whom I had the pleasure to work since I was a graduate student at Harvard University, thank you for generously sharing the data you collected and keeping inspiring me and providing me with insightful advice. I am grateful that I have met you at Harvard and that our collaboration continued as I pursued Ph.D. at Columbia University. This work would not have been possible without you and your data. My sincere thanks also goes to Dr. Stephen Peverly, who shared his profound knowledge and valuable suggestions on note-taking that have

shaped this dissertation and future directions. To Dr. Bryan Keller, from whom I learned a lot in his three statistics courses: Statistical Inference, Computational Statistics, and Linear Models and Experimental Design – your constructive feedback on research methods helped shape the data analysis for this work and trained me as a more rigorous researcher. Your courses set the foundation of my knowledge on statistics, and are extremely helpful as I apply it to this research and future analyses.

To Dr. Luc Paquette – I am so grateful to have you mentor me when I started my doctoral program. You provided me with the most insightful suggestions on each piece of my work and unyielding support whenever I need help. You are such a wonderful mentor, collaborator, and friend. I would also like to thank Dr. Jaclyn Ocumpaugh for her help, both professionally and personally, especially when I woke up to find my apartment building surrounded by flood water last summer. To members and alumni of the Educational Data Mining (EDM) lab – Dr. Sweet San Pedro, Dr. Jeanine Defalco, Dr. Elle Yuan Wang, Mia Almeda, Shimin Kai, Aaron Hawn, Miggy Andres, Stefan Slater, Zhulin Yu, Shamya Karumbaiah, Alexis Andres, and Chad Coleman – thank you for your wisdom, support, and friendship; they meant a lot to me during the past few years. I also want to thank all my colleagues from Columbia University outside the EDM lab for their help, encouragement, and friendship.

A special thank you to Carolyn Hsu and the Kuo Ping Wen Scholarship, the Ben Woods Fellowship, the Susan A. and Robert S. Diamond Scholarship, and TC Doctoral Fellowships – thank you for supporting and funding me through my doctoral study.

Last but not least, I would like to express my sincerest gratitude to my parents and my family. To my mother, Huiyan Shen, and my father, Shaolin Jiang, thank you for your unconditional love, unyielding support, and constant faith in me. I am extremely fortunate to be your daughter. Your love and your encouragements sustained me during my moments of difficulty and stress and stimulated me to pursue my dream. Thank you for teaching me to work hard and be a kind person. I will try my best to become a person and scholar whom you are proud of. To my family, including my grandparents, Chunying Hu and Mengguang Shen, and my dear uncles and aunts – thank you for everything you have done for me. I feel so much love and warmth in this big family. Special thanks to my uncle Xiaofan Shen, who was my first math teacher and from whom I have learned a lot. Lastly, to Qing Xu, thank you for your love and support.

DEDICATION

For Mom and Dad.

CHAPTER I.

INTRODUCTION

Background

Self-regulated learning (SRL) is important for academic success in various educational settings (B. J. Zimmerman & Schunk, 2001). SRL is defined as “an active, constructive process whereby learners set goals for their learning and then attempt to monitor, regulate, and control their cognition, motivation and behavior, guided and constrained by their goals and the contextual features in the environment” (Pintrich, 2000, p. 453). Advantages for students who are able to regulate their own learning compared to those with insufficient self-regulatory skills have been well documented (B. J. Zimmerman, 1990). However, research has indicated that even undergraduate students usually lack sufficient SRL skills and ability and are often faced with difficulties in using SRL (Moos & Azevedo, 2008b). It has therefore developed as an important goal for many K-12 teachers to help their students develop into learners who can regulate their own learning with effective SRL strategies (Perry, Phillips, & Dowler, 2004).

Self-regulated learning is a complex and multifaceted construct that involves the regulation of various components (e.g., cognition, metacognition, affect, motivation, behaviors) and multiple cyclical phases (Panadero, 2017; Schraw, Crippen, & Hartley, 2006; Winne, 2011; Winne & Hadwin, 2009; B. J. Zimmerman, 1990). This makes measuring SRL challenging. Traditional measures of SRL include self-report questionnaires, structured interviews, teacher and parent ratings, think-aloud protocols, and observations (Winne & Perry, 2000). More recently, researchers have increasingly applied action trace analysis on the log data produced as students learn with computer-based learning environments to study self-regulated learning (Schraw, 2010). Action traces from log data (e.g., sequences of behaviors executed by students during their interaction with computers) are unobtrusive, fine-grained measures that are scalable and show great potential in revealing how SRL strategies and skills manifest (Winne & Baker, 2013).

Recent years have seen rapid advancement in the use of technology and computers for science learning and classrooms. One increasingly popular strategy for fostering SRL is to use computer-based environments such as open-ended learning environments (OELEs) (Azevedo, 2005). OELEs are learner-centered, technology-based learning environments that support problem-solving and inquiry by presenting learners with authentic contexts, complex and challenging learning tasks, and a set of tools and resources to explore and manipulate (Land, 2000; Segedy, Kinnebrew, & Biswas, 2015). Over the past decade, OELEs have transformed traditional K-12 science classrooms by fostering learning of complex scientific topics and assessing science inquiry skills (Clarke-Midura & Dede, 2010; Land, 2000). In OELEs, learners set their own learning goals; generate, test, and modify hypotheses; utilize and manipulate tools and resources; construct solutions to problems and reflect on solutions and inquiry process (Kinnebrew, Segedy, & Biswas, 2014; Land, 2000; Segedy et al., 2015).

However, there have been conflicting attitudes towards the effectiveness of OELEs. On the one hand, researchers argue that the authentic context, non-linearity, and open-endedness of these environments create learning opportunities for students, making OELEs effective in enhancing science inquiry skills, boosting self-regulated learning, and preparing students for future learning (Land, 2000). On the other hand, some researchers contend that OELEs impose challenges on learners in terms of extraneous cognitive load and greater requirements for self-regulated learning (Azevedo, 2005; Moos, 2009; Moos & Azevedo, 2008b), and learning could be limited in OELEs because of these challenges. In OELEs, learners have to deploy self-regulatory processes and strategies in order to complete tasks and learn complex topics (Azevedo, 2005; Segedy et al., 2015). The lack of structure and guidance in these open-ended learning environments may lead students with insufficient self-regulatory skills in planning, executing, and monitoring their learning activities to struggle and be less successful in these environments (Alfieri, Brooks, Aldrich, & Tenenbaum, 2011; Kinnebrew, Loretz, & Biswas, 2013; Kirschner, Sweller, & Clark, 2006). This is true especially for younger populations (e.g., middle school students), who usually lack sophisticated SRL skills (Greene, Moos, Azevedo, & Winters, 2008; Pintrich & Zusho, 2002), as they engage in learning and inquiry in OELEs independently without scaffolding. Therefore, meaningful questions that are worth investigating include whether OELEs without embedded scaffolding will promote the development of SRL skills for young students, and if so, how do self-regulatory processes and strategies manifest and develop in these environments. The wealth of log data in OELEs also affords the opportunity to automatically and unobtrusively measure and detect self-regulated learning in a fine-grained manner.

Furthermore, individual differences, such as gender differences, may play important roles in the development of self-regulatory skills within OELEs. Gender differences have been extensively studied and well documented in science education (D. Baker, 2002; Benbow & Arjmand, 1990; Britner, 2008; Erwin & Maurutto, 1998; Jovanovic, Solano-Flores, & Shavelson, 1994; Mullis, Martin, Fierros, Goldberg, & Stemler, 2000). Studies indicate that gender differences in science achievement and attitudes towards science are present as early as elementary school (Curran & Kellogg, 2016; Halpern, 2004; Mullis et al., 2000; Neuschmidt, Barth, & Hastedt, 2008). For example, examination of data from large-scale assessments such as the National Assessment of Educational Progress (NAEP), the Trends in International

Mathematics and Science Study (TIMSS), and the Early Childhood Longitudinal Study (ECLS-K) showed consistent advantages for males in science academic achievement (Cunningham, Hoyer, & Sparks, 2015; Curran & Kellogg, 2016; Else-Quest, Mineo, & Higgins, 2013; National Center for Education Statistics, 2016; Quinn & Cooc, 2015; Reilly, Neumann, & Andrews, 2015). Similar gender gaps were found on the motivation and attitudes towards science, and career and major selection (Benbow & Arjmand, 1990; X. Chen & Weko, 2009; Cunningham et al., 2015; Erwin & Maurutto, 1998; Griffith, 2010; Jovanovic et al., 1994). Beyond this, males were documented as more capable of using computers and hold more positive attitudes towards computer use than females (Kay, 1992, 2008; Kay & Lauricella, 2011; Whitley Jr., 1997). In contrast, research investigating gender-related differences in self-regulatory skills have shown mixed results (Basol & Balgalmis, 2016). Some studies indicated that females were more highly self-regulated and reported using self-regulatory strategies more often than males (Lee, 2002; Matthews, Ponitz, & Morrison, 2009; Pajares, 2002; Yukselturk & Top, 2013; B. J. Zimmerman & Martinez-Pons, 1990), while other studies did not find significant difference in SRL between males and females (Yukselturk & Bulut, 2009). However, research exploring the relationship between gender and development of self-regulatory skills is lacking in the context of open-ended computer-based learning environments for science learning.

Statement of the Problem

This dissertation aims to detect and trace how self-regulatory skills manifest and develop in an open-ended learning environment for middle school science named Virtual Performance Assessment, and the role of gender in the development of self-regulatory skills and strategies.

Despite its importance, the measurement and detection of SRL is challenging. Many of the extensive studies on SRL predate the wide use of computer-based learning environments.

Traditional measures of SRL such as self-report questionnaires test learners' perceived self-regulation by asking them to report their use of typical self-regulatory strategies and behaviors (e.g., how often they engage in SRL activities). These measures view SRL as an aptitude and do not reflect the individual SRL actions and their real-time development (Winne & Perry, 2000). In addition, they rely on self-reporters' honesty and may not be accurate and reliable (Schraw, 2010). As Winne and Jamieson-Noel (2002) found, students tended to overestimate their use of SRL behaviors in self-reports. On the other hand, although other measures such as think-aloud protocols could address this concern, they are intrusive and may change the behaviors that naturally occur in important ways and reduce their representativeness. In the meantime, some observation-based measures such as observations and teacher and parent ratings may be susceptible to observer bias. As such, applying novel methods to unobtrusively and accurately measure self-regulated learning is important but challenging. The rich action log data produced as students learn in computer-based learning environments such as open-ended learning environments provide potential to investigate self-regulatory behaviors and traces in real time and in an unobtrusive and fine-grained manner (Schraw, 2010).

Meanwhile, the open-ended nature of open-ended learning environments makes it more challenging to identify and measure SRL than in other computer-based learning environments. Unlike other environments with high restrictions on student actions, OELEs do not have constraints on behaviors and there are many possible paths and strategies that could lead to correct answers and success. Previous studies on SRL in open-ended learning environments usually examine how specific SRL behaviors and processes manifest in these environments, focusing on one or a few component(s) such as goal setting (Sabourin, Mott, & Lester, 2013), metacognitive monitoring and knowledge acquisition (Kinnebrew et al., 2013; Taub, Azevedo, Bradbury, Millar, & Lester, 2017), help-seeking behaviors (Aleven, Roll, McLaren, & Koedinger, 2010), and note-taking behaviors (Trevors, Duffy, & Azevedo, 2014). According to its definition, SRL is a multifaceted construct comprised of multiple components and cyclical phases (Pintrich & Zusho, 2002; Winne, 2011; Winne & Hadwin, 1998, 2009; B. J. Zimmerman & Martinez-Pons, 1990; B. J. Zimmerman & Schunk, 2001). However, not many studies have comprehensively examined the behaviors and strategies related to all the phases listed in established SRL frameworks in the context of OELEs. This dissertation aims to adopt the SRL framework proposed by Winne and Hadwin (2009) and identify the behaviors that represent each phase of SRL in the model (i.e., understanding learning goal, planning, use of learning and self-regulatory strategies, and monitoring and self-evaluation).

In addition, research on SRL in OELEs has mainly focused on how SRL manifests in OELEs by comparing the behavior patterns of students with high versus low SRL skills. However, little research has examined whether students' self-regulatory skills develop over the use of OELEs or not. Studies have shown that students often show ineffective self-regulatory paths and strategies, and these students with insufficient self-regulatory skills may struggle in OELEs (Azevedo, 2005; Moos, 2009; Moos & Azevedo, 2008b). Will this struggle interfere with learning in OELEs with limited guidance, or will students develop robust self-regulatory skills from this process? Consequently, it is meaningful to explore whether the use of OELEs would help the students low in SRL ability to acquire self-regulatory skills and strategies over time. Furthermore, studying the dynamic development of SRL strategies and skills over time within OELEs is especially needed for younger populations such as middle school students, who typically show less sufficient self-regulatory skills (Greene et al., 2008; Pintrich & Zusho, 2002).

Note-taking is an important SRL strategy that is frequently studied in SRL literature (Trevors et al., 2014). Research has shown that paper-based note-taking from lectures or texts is associated with positive learning outcomes (Armbruster, 2009). However, results from studies examining the role of note-taking as an SRL strategy in open-ended learning environments are mixed, sometimes agreeing with and sometimes contradicting the results found in the literature on traditional note taking (Bouchet, Harley, Trevors, & Azevedo, 2013; McQuiggan, Goth, Ha, Rowe, & Lester, 2008; Trafton & Trickett, 2001; Trevors et al., 2014). Many of these studies on note-taking in OELEs did not distinguish the encoding function of note-taking from the external storage function (i.e., note-reviewing) (Di Vesta & Gray, 1972), and did not examine the content and quality of notes taken in these computer-based environments in depth. Further, previous literature on note-taking has a strong emphasis on examining undergraduates' or adults' note-taking strategies. Middle school students typically exhibit less sophisticated note-taking skills than older populations. Limited research has studied note-taking as a self-regulatory strategy in OELEs among middle school students. Studies that investigate the *development* of note-taking strategies in OELEs are also needed. In order to better understand the development of SRL processes and strategies, this dissertation comprehensively investigates note-taking as an SRL strategy in the context of OELEs by exploring both the development of quantitative measures of note-taking/reviewing (e.g., frequency, duration) as well as the content of notes (e.g., the level of cognitive processing involved in notes) as students learned with an open-ended learning environment.

Last but not least, previous studies exploring gender differences in self-regulatory skills have mainly used self-report questionnaires or interviews in the context of traditional instructional settings or online courses (Lee, 2002; Yukselturk & Bulut, 2009; Yukselturk &

Top, 2013; B. J. Zimmerman & Martinez-Pons, 1990). These studies typically compare the mean scores on self-reported SRL measures instead of the development of SRL over time. To my knowledge, no study has systematically examined gender-related differences in the development of self-regulatory behaviors and strategies as students learn science inquiry in open-ended learning environments. On the other hand, despite the extensive literature on gender difference in note-taking, most of the previous research studied paper-based note-taking in classroom lecture-based contexts among adults. Limited studies have examined gender differences in computer-based note-taking for the younger population. In addition, although females were sometimes found to be more highly self-regulated (Lee, 2002; Matthews et al., 2009; Pajares, 2002; Yukselturk & Top, 2013; B. J. Zimmerman & Martinez-Pons, 1990), males generally surpassed females in ability, achievement, and motivation in science (X. Chen & Weko, 2009; Cunningham et al., 2015; Curran & Kellogg, 2016; Halpern, 2004; Mullis et al., 2000; National Center for Education Statistics, 2016; Neuschmidt et al., 2008; Quinn & Cooc, 2015; Reilly et al., 2015) and computer use (Kay, 1992, 2008; Kay & Lauricella, 2011; Whitley Jr., 1997). As such, it is unclear whether males and females regulate their learning differently and develop self-regulatory skills and strategies such as note-taking at different rates in open-ended computer-based learning environments for science. Accordingly, this dissertation aims to address these issues by investigating the development of self-regulatory skills and strategies in an open-ended learning environment and potential gender differences in SRL.

Research Questions

This dissertation research involves three analyses to answer the following research questions in the context of an open-ended computer-based learning environment named Virtual Performance Assessments:

Analysis 1:

1a). Does student performance on science inquiry tasks, which is closely related to self-regulated learning, improve across the course of using Virtual Performance Assessments (VPA)?

1b). Are there any gender-related differences in the development of science inquiry expertise in the open-ended learning environment?

Analysis 2:

2a). How do students' skills in using self-regulatory processes and strategies develop in VPA?

2b). Are there any gender-related differences in the development of SRL behaviors and strategies in VPA?

Analysis 3:

3a). How do students' note-taking and note-reviewing strategies, including the quantity of note-taking/reviewing behaviors and the content of notes, develop in VPA?

3b) Are there any gender-related differences in the development of note-taking and note-reviewing strategies in the open-ended learning environment?

The goal of this dissertation is to answer these research questions by combining an established SRL theoretical framework with educational data mining methods such as sequential pattern mining and feature engineering, and traditional statistical methods such as multilevel modeling.

CHAPTER II.

REVIEW OF LITERATURE

Self-Regulated Learning

Self-regulated learning (SRL) is important for academic success in various educational settings (B. J. Zimmerman & Schunk, 2001). There is extensive evidence that individuals who actively monitor and regulate their own learning are likely to be more successful in academic performance and learning tasks (Boekaerts, Pintrich, & Zeidner, 2000; B. J. Zimmerman, 1990).

While researchers have developed many theoretical models of SRL (see Pintrich, 2000; B. J. Zimmerman & Schunk, 2001), most models and definitions agree that the cognitive and metacognitive operations used in SRL require effort (Winne, 2011), and characterize learners as actively monitoring and controlling cognitive, motivational, metacognitive, and behavioral processes. In an attempt to integrate all the definitions, Pintrich (2000) organized published research around a set of phases of SRL. He described self-regulated learning as “an active, constructive process whereby learners set goals for their learning and then attempt to monitor, regulate, and control their cognition, motivation and behavior, guided and constrained by their goals and the contextual features in the environment” (p. 453).

The widely adopted SRL models all assume that SRL phases are cyclical. Zimmerman’s (2000) framework adopts social cognitive perspective and describes SRL as a cyclical process composed of three phases – forethought, performance, and self-reflection (see Figure 1). The *forethought* phase in his model involves task analysis, where individuals set their learning goals and develop plans to achieve their goals by identifying appropriate strategies relevant to the goals. Their self-motivation beliefs, such as goal orientation, self-efficacy, outcome expectations, and task interest/valuing, also play important roles in this phase. In the *performance* phase, learners execute the strategies they identified during the forethought phase that will assist learning (e.g., note-taking, cognitive mapping, help-seeking). They also metacognitively monitor their performance and strategy use and/or keep records of their progress (i.e., self-observation). The *self-reflection* phase entails self-judgment and self-reaction. In this phase, learners evaluate multiple dimensions of their learning (e.g., learning process and learning outcomes) to standards or goals, and attribute causal significance to them. The self-judgment will trigger further steps, such as self-modification. This phase may also lead students to change their goals and plans accordingly in the forethought phase, which makes the SRL process cyclical.

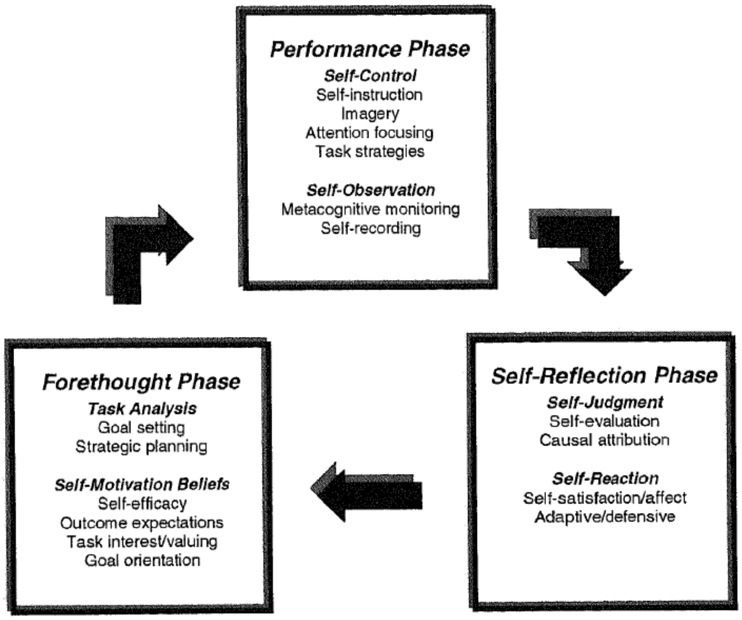


Figure 1. Three phases of self-regulated learning in Zimmerman's (2000) SRL model.

Winne and Hadwin's (1998, 2009) framework (see Figure 2) proposes four distinguishable but recursively linked stages that SRL encompasses: 1) task definitions; 2) goal setting and planning; 3) enacting study tactics and strategies; and 4) metacognitively adapting studying. In these phases, students develop an understanding of the learning task, set goals and construct plans to achieve their learning goals, execute various learning tactics and strategies, metacognitively monitor and reflect on their learning process, and adapt their plans, behaviors, and strategies accordingly. The SRL framework that Winne (2011) described in the 2011 Handbook of Self-Regulation maintained the same processes and components for SRL as in Figure 2. This framework offers a metacognitive view of SRL that integrates a more complex cognitive architecture (Greene & Azevedo, 2007; Panadero, 2017; Winne, 2011), and has been adopted to study SRL in other open-ended learning environments (Moos, 2009; Moos & Azevedo, 2008b). Given the interactive and open-ended nature of open-ended learning environments, this dissertation applies Winne & Hadwin's model of SRL to the context of an open-ended learning environment.

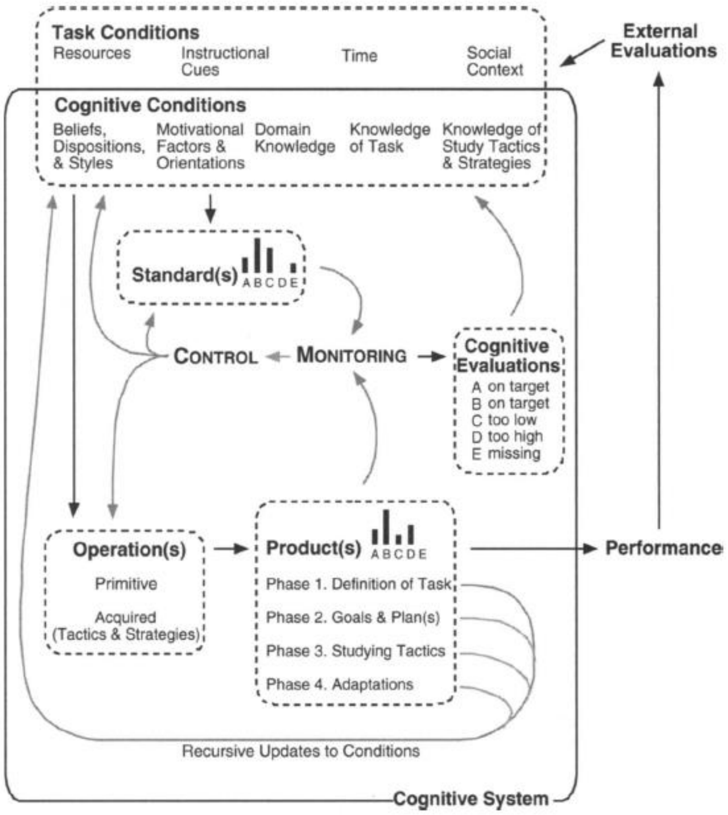


Figure 2. Winne and Hadwin's (1998, 2009) model of self-regulated learning.

Open-Ended Learning Environments

One of the important goals for K-12 science education is to help students develop the scientific knowledge and skills needed to actively and effectively engage in science inquiry (Kuhn & Pease, 2008). Over the past decade, open-ended learning environments (OELEs) have transformed traditional K-12 science classrooms by fostering learning of complex scientific topics and assessing science inquiry skills (Clarke-Midura & Dede, 2010; Land, 2000). OELEs are learner-centered, technology-based learning environments that support problem-solving and inquiry by presenting learners with authentic contexts, complex and challenging learning tasks, and a set of tools and resources to explore and manipulate (Land, 2000; Segedy et al., 2015). In OELEs, learners set their own learning goals; generate, test, and modify hypotheses; utilize and manipulate tools and resources; construct solutions to problems and reflect on solutions and inquiry process (Kinnebrew et al., 2014; Land, 2000; Segedy et al., 2015). The open-endedness of OELEs is represented by the limited external directions provided in the environment, and the control and responsibility learners assume in their own problem-solving process — they pursue unique learning goals, create unique plans, and execute unique inquiry paths and learning sequences to accomplish learning goals (Hannafin, 1995; Hannafin, Land, & Oliver, 1999). For example, in the open-ended learning environment used in this dissertation – Virtual Performance Assessment – learners have to take responsibility and make decisions on their inquiry processes as they gather data, make observations, test hypothesis, develop solutions to the problem and reflect on their solutions and learning process. To be successful in this environment, learners have to create plans, set goals, execute learning strategies, and adaptively monitor their solutions and inquiry processes. These skills comprise the foundation of self-regulated learning and scientific inquiry.

Researchers argue that the non-linearity and open-endedness of these environments create learning opportunities for students. Accumulated evidence shows that OELEs provide an authentic learning context and are effective in enhancing science inquiry skills, boosting self-regulated learning, and preparing students for future learning (Jiang, Paquette, Baker, & Clarke-Midura, 2015; Land, 2000). Popular OELEs that have been found to assist science learning include virtual environments (e.g., Clarke-Midura & Dede, 2010), science microworlds (e.g., Gobert, Sao Pedro, Baker, Toto, & Montalvo, 2012), teachable agents (e.g., Leelawong & Biswas, 2008), games (e.g., Shute, Ventura, & Kim, 2013), and hypermedia (e.g., Azevedo, 2005).

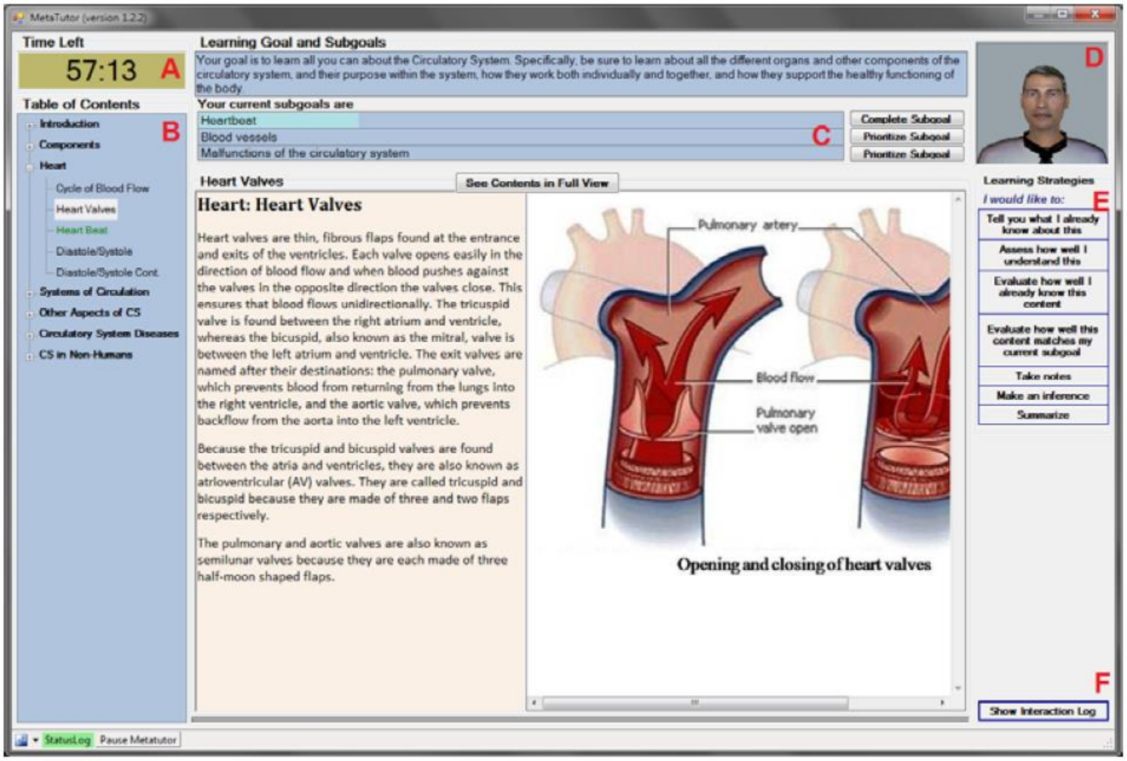
Despite its benefits, there has been debates on the effectiveness of these open-ended learning environments. One challenge of open-ended learning environments, even environments designed to personalize based on student knowledge, is that learners have to deploy self-regulatory processes and strategies in order to complete tasks and learn complex topics (Azevedo, 2005; Segedy et al., 2015). Not every learner is sufficiently competent in self-regulatory skills to plan, execute, and monitor their learning activities (Azevedo et al., 2015; Lester, Rowe, & Mott, 2013). Researchers have found that the lack of structure and guidance in open-ended learning environments may lead students with insufficient self-regulatory skills to be less successful, possibly leading to floundering, confusion, and frustration (Alfieri et al., 2011; Kinnebrew et al., 2013; Kirschner et al., 2006). This is true especially for younger populations (e.g., middle school students), who usually lack sufficient SRL skills (Greene et al., 2008; Pintrich & Zusho, 2002), as they engage in learning and inquiry in OELEs without scaffolding.

Therefore, it is crucial to make sure that students with low self-regulatory skills can also learn from OELEs, and more importantly, that they can develop self-regulatory skills during the inquiry process. Studying SRL in OELEs also provides a wealth of fine-grained action log data, which allow researchers to study and measure SRL unobtrusively in real time.

Self-Regulated Learning in OELEs

With the development of OELEs and the opportunities they afford to study SRL in an unobtrusive and fine-grained way, there is a surge of interest in examining how various SRL processes and strategies manifest in OELEs. Researchers have developed different ways to detect, track, measure, and model SRL in computer-based learning environments, including OELEs (see Winters, Greene, & Costich, 2008 for a review). Students' self-regulation has an impact on the observable behaviors that they exhibit, with students of different degrees of competence in SRL demonstrating different frequencies of behaviors during learning (Sabourin, Mott, et al., 2013). Therefore, researchers have studied students' observed actions and sequences of behaviors in computer-based learning environments to assess their SRL (B. J. Zimmerman, 2008).

One line of these efforts involves studying students' use of interface features that directly externalize and prompt their SRL processes and strategies in OELEs. For example, in Bouchet et al. (2013), undergraduate students learned a challenging science topic (i.e., human circulatory system) with an open-ended hypermedia learning environment named MetaTutor. The OELE provides students with an SRL palette (see Figure 3) where they can select and deploy SRL processes, including planning, monitoring, and self-regulatory strategies such as note-taking. For instance, students could use the palette to indicate that they want to read and select their learning goal and subgoals, judge their learning, assess their understanding, take notes, summarize information, etc. The pedagogical agents embedded in MetaTutor explicitly prompt students to set goals, enact learning strategies, and monitor their learning regularly to ensure the use of the palette. Thus, clicks of buttons in the palette directly link to the various SRL processes and strategies. This allowed the researchers to measure SRL in a straightforward approach by examining the frequency of using different options of the tool. In their study, the researchers applied clustering to classify participants into three clusters based on their performance and their use of the SRL tool. Results indicated that students with high SRL skills who frequently monitored their performance showed higher prior content knowledge, tended to set goals more frequently and distributed more time to goal setting, and accessed more pages compared to the other two groups of students who regulated their learning less effectively. Students who engaged in a higher amount of monitoring behaviors also took significantly fewer notes and spent relatively less time taking notes, while they spent more time checking their notes than the other students.



Your goal is to learn all you can about the Circulatory System. Specifically, be sure to learn about all the different organs and other components of the circulatory system, and their purpose within the system, how they work both individually and together, and how they support the healthy functioning of the body.

Heart: Heart Valves

Heart valves are thin, fibrous flaps found at the entrance and exits of the ventricles. Each valve opens easily in the direction of blood flow and when blood pushes against the valves in the opposite direction the valves close. This ensures that blood flows unidirectionally. The tricuspid valve is found between the right atrium and ventricle, whereas the bicuspid, also known as the mitral, valve is between the left atrium and ventricle. The exit valves are named after their destinations: the pulmonary valve, which prevents blood from returning from the lungs into the right ventricle, and the aortic valve, which prevents backflow from the aorta into the left ventricle.

Because the tricuspid and bicuspid valves are found between the atria and ventricles, they are also known as atrioventricular (AV) valves. They are called tricuspid and bicuspid because they are made of three and two flaps respectively.

The pulmonary and aortic valves are also known as semilunar valves because they are each made of three half-moon shaped flaps.

Figure 3. SRL palette in MetaTutor (Bouchet et al., 2013).

Other researchers study SRL in OELEs by building predictive models of the components and strategies related to SRL. These studies include predicting help-seeking behavior (Aleven et al., 2010), cognitive tool use (Shores, Rowe, & Lester, 2011), goal setting and monitoring (Sabourin, Shores, Mott, & Lester, 2012), and affective states and behavior (R. S. Baker, Clarke-Midura, & Ocumpaugh, 2016). For instance, Sabourin and colleagues (2012) had participants self-report on their moods and status regularly in an OELE for middle school science named Crystal Island. The researchers coded the self-reported statements about their mood and status based on how well students set goals and monitor their progress towards the goals. Machine-learned models were then developed to predict the high, medium and low SRL categories. They found that the highly self-regulated students were generally high-performing, and selectively chose the tests to run in the environment while low SRL students were more likely to game the system.

Segedy and colleagues (2015) tracked SRL by analyzing the coherence between students' observable actions in an open-ended learning environment for middle school students called Betty's Brain. Examples of coherent action sequences include marking a causal link between two concepts in a concept map correct after reading relevant information on the relationships between the two concepts, or removing incorrect links from the concept map if quiz results reveal that they are incorrect. They found that highly self-regulated students showed higher levels of coherence in their actions, and demonstrated higher learning gains and performance in Betty's Brain.

Using Sequential Pattern Mining to Study SRL in OELEs

Examining students' observable behavior patterns to infer self-regulatory processes and use of strategies is unobtrusive, fine-grained, and could be more accurate than the other measures

(Aleven et al., 2010; Taub et al., 2017; Winne & Baker, 2013; B. J. Zimmerman, 2008). Sequential Pattern Mining (Agrawal & Srikant, 1995), a methodology that has been extensively used in Educational Data Mining (R. S. Baker & Yacef, 2009), has shown potential for discovering complicated patterns of SRL behaviors within open-ended learning environments (Bouchet, Azevedo, Kinnebrew, & Biswas, 2012; Taub et al., 2017; Winne & Baker, 2013). Sequential pattern mining is a popular data mining technique that automatically identifies frequent temporal patterns of actions in the data (Agrawal & Srikant, 1995). For example, Kinnebrew and colleagues (2014) applied differential pattern mining techniques, a form of sequential pattern mining where patterns are compared between different groups of individuals, to log data produced by students engaging in activities within the OELE Betty's Brain. This enabled them to study the differences in students' SRL behaviors by identifying frequent sequential patterns indicative of SRL strategies and determining which sequential patterns were characteristic of high-performing students as compared to low-performing students. Results indicated that high-performing students showed better employment of self-regulatory strategies such as monitoring compared to low-performing students. For instance, high-performing students were more likely to correct their errors in a concept map after testing that map than low-performers, indicating that they were evaluating their own progress.

Differential pattern mining was also used by Sabourin and colleagues (2013) to analyze the differences in inquiry behaviors utilized by learners depending on their level of self-regulation within the OELE Crystal Island. As in Sabourin et al. (2012), students were classified into low, medium, and high SRL groups based on their ability in goal setting and monitoring suggested by their self-reports. Differential pattern mining was then implemented to identify behavioral patterns characteristic of these low, medium, and high self-regulated learners. Results suggested that highly self-regulated students made better use of the resources and tools in the environment (e.g., by keeping track of relevant information in worksheets), and showed different patterns in monitoring their learning progress and reflecting on their science inquiry processes than students with insufficient SRL skills. Low-SRL students used information and resources presented in the environment less effectively and did not record the information in worksheets. Students with good SRL strategies also generated inferences and processed information more deeply compared to students with poor SRL skills.

In a more recent study, Taub and colleagues (2017) applied sequential pattern mining and differential pattern mining to compare the behavioral patterns of undergraduate students with different levels of *efficiency* in problem-solving and scientific reasoning in Crystal Island. Results indicated that more efficient learners made use of their time more efficiently and were more strategic in hypothesis testing compared to less efficient participants. Specifically, students with higher efficiency ran significantly fewer tests on objects that were partially relevant or irrelevant to the solution than the less efficient participants. The researchers further concluded that the difference was probably because the less efficient participants focused solely on hypothesis testing and executing learning strategies during the inquiry, while they engaged less in setting goals, planning, and monitoring and adapting their hypothesis testing strategies than the students who were more efficient in problem-solving. On the other hand, the more efficient students focused more on metacognitive processes and knowledge acquisition. In a similar study, Taub et al. (2014) compared the sequence patterns of SRL behaviors between undergraduate students with high versus low prior knowledge who used MetaTutor.

Despite the surging interest in applying educational data mining methods such as sequential pattern mining to action log data to study how self-regulated learning manifests in open-ended learning environments, very few studies have investigated how self-regulatory skills dynamically develop over the use of OELEs. This dissertation applied sequential pattern mining to identify the action sequences that corresponded to the four SRL processes in Virtual Performance Assessments (VPA): task definition, goal setting and planning, enacting study tactics and strategies, and metacognitively adapting study strategies. I then conducted differential pattern mining to study the development of SRL skills in the open-ended learning environment for middle school students.

Gender Differences in SRL

There has been extensive research exploring individual differences in self-regulatory skills and strategies. One factor that is frequently examined is gender. Research investigating gender-related differences in self-regulatory skills have shown mixed results (Basol & Balgalmis, 2016). Results from some studies indicated that females showed significantly higher perceived self-regulation and reported themselves using self-regulatory strategies more often than males (Lee, 2002; Matthews et al., 2009; Pajares, 2002; Yukselturk & Top, 2013; B. J. Zimmerman & Martinez-Pons, 1990). For example, Matthews et al. (2009) found that girls demonstrate better self-regulatory behaviors as early as kindergarten, according to both teacher and parent reports and structured observational tasks. Yukselturk and Top (2013) had university students from an online course complete questionnaires and report their use of SRL. Female learners reported themselves as more highly self-regulated than males in the online learning environment. In another study, Zimmerman and Martinez-Pons (1990) interviewed 5th, 8th, and 11th grade students and asked them to report their use of SRL strategies. Gender-related differences were found with girls reporting more use of SRL strategies such as goal setting and planning, record keeping and monitoring, and environmental structuring. On the other hand, some studies did not find significant differences in SRL between males and females (Yukselturk & Bulut, 2009).

It is worth noting that in most of the abovementioned studies, self-regulatory skills were measured by self-report measures, where students answered questionnaire or interview questions and self-reported on their use of typical self-regulatory strategies. A few other studies used traditional measures such as observations and teacher and parent ratings. Many studies were conducted to detect and measure SRL in the subject domain of psychology. To my knowledge, no study that I am aware of has applied sequence mining and other educational data mining strategies to examine whether middle school students of different sex exhibit different behavior patterns related to the self-regulatory phases and strategies in OELEs for middle school students.

Science Inquiry Skills and SRL

One of the important goals for science education is to help students develop the scientific knowledge and practices needed to actively and effectively engage in science inquiry (van der Graaf, Segers, & Verhoeven, 2015; C. Zimmerman, 2000, 2007). As such, science inquiry skills have been a critical component of the K-12 science curriculum standards (National Research Council, 2011). Scientific inquiry skills are not innate and are not learned immediately (Kuhn, 2010). Rather, these skills develop over repeated practice and engagement in science inquiry activities. Therefore, it is particularly crucial to understand and assess the development of students' science inquiry skills in 21st-century classrooms (Clarke-Midura, Dede, & Norton, 2011; Kuhn, 2010; Kuhn & Dean, 2005; Kuhn & Pease, 2008).

Science inquiry skills are closely related to self-regulated learning. Self-regulatory skills are crucial in inquiry-based tasks and learning environments that are typically complex and open-ended. During the scientific investigation, students are expected to plan, execute, and monitor their inquiry processes and strategies such as making observations, collecting data, conducting experiments, seeking knowledge, interpreting results, and making informed decisions (Kuhn & Dean, 2005). Research suggested that highly self-regulated learners also showed higher science inquiry skills in open-ended learning environments (Sabourin, Mott, et al., 2013).

Science inquiry skills have been evaluated from multiple perspectives. A large body of science inquiry literature has examined science inquiry and scientific reasoning based on the use of the control of variables strategy (CVS) in multivariable experiments. CVS refers to the scientific strategy of holding constant all other variables than the one under investigation to eliminate their influence on the outcome variable (Z. Chen & Klahr, 1999; Kuhn, 2010; Kuhn & Dean, 2005; van der Graaf et al., 2015). CVS is challenging for young learners to learn (Z. Chen & Klahr, 1999); specifically, the more variables that can be manipulated, the more difficult it is for young learners to utilize CVS in experiments (van der Graaf et al., 2015). Researchers have tested and proved the effectiveness of computer-based interventions that teach CVS on learning gains (Z. Chen & Klahr, 1999; Klahr & Nigam, 2004; Sao Pedro, Gobert, & Raziuddin, 2010; Siler, Klahr, Magaro, Willows, & Mowery, 2010; van der Graaf et al., 2015).

In other studies, CVS is evaluated using “knowledge engineered rules” (Sao Pedro, Baker, Gobert, Montalvo, & Nakama, 2013) by counting the number of trials (either consecutive or inconsecutive) students run that are controlled (e.g., trial 1 and trial 2 where only one variable is changed while the other variables are held constant) (Gobert & Koedinger, 2011; Kuhn & Pease, 2008; McElhaney & Linn, 2010). More recently, researchers have built machine-learned models to estimate and predict the acquisition of the ability to design controlled experiments in computer-based learning environments, including cases that are not obviously demonstrations of the CVS strategy (Gobert et al., 2012; Sao Pedro et al., 2013). Building on this work, Sao Pedro and colleagues (2014) developed extensions of Bayesian Knowledge Tracing to assess the transfer of skills in designing controlled experiments across different learning domains and contexts.

Educational Data Mining (EDM) approaches have also been adopted to assess other science inquiry skills in more open-ended learning environments. For example, Baker and colleagues (2016) distilled a set of features related to inquiry behavior from log data in Virtual Performance Assessments (VPA), an open-ended immersive virtual environment that is also used in the current study, to develop predictive models of student success on two science inquiry tasks, including identifying a correct final claim and what they referred to as designing causal explanations. The features that were predictive of science inquiry in machine-learned models provided insights into science inquiry skills. In another study, Clarke-Midura and Yudelson (2013) applied machine learning to automatically model students' causal reasoning in VPA and compared the machine-learned algorithm with expert scored rubrics of causal reasoning.

Note-Taking as an SRL Strategy

Winne and Hadwin (2009) have identified the utilization of various learning strategies as a key component of their SRL framework. In open-ended learning environments, students are expected to determine which learning strategies would be effective in assisting the achievement of learning goals, correspondingly adopt these strategies, continuously evaluate and adaptively modify the use of these strategies in real time to facilitate their learning process. One frequently studied strategy in SRL literature is note-taking (Trevors et al., 2014). Note-taking is a nearly ubiquitous academic strategy that is commonly used by learners and highly encouraged by educators (Bonner & Holliday, 2006; Weiss, Banilower, McMahon, & Smith, 2001). Research has shown that paper-based note-taking from lectures or texts is associated with positive learning outcomes (Armbruster, 2009). However, studies examining the role of note-taking strategy in open-ended learning environments are still emerging with mixed results. This section will discuss previous literature on note-taking.

Past research has shown that learners with different self-regulatory skills may demonstrate different note-taking/reviewing behaviors and show different patterns in the content of notes recorded (Trevors et al., 2014). In open-ended virtual environments that pose high demands on self-regulatory skills, regulating one's use of note-taking strategy effectively is challenging for students, especially for students with insufficient SRL skills (Moos, 2009). Given the importance and effectiveness of note-taking and note-reviewing as SRL strategies and the difficulty of implementing these strategies, this dissertation studies whether learning science through open-ended virtual environments fosters students' use of note-taking and note-reviewing strategies. Specifically, since previous research have revealed that the quantity and the content of notes are important components of SRL that are related to academic performance (Bretzing & Kulhavy, 1979; Cohn, Cohn, & Bradley, 1995; Fisher & Harris, 1973; Peverly, Brobst, Graham, & Shaw, 2003), the present study develops measures of both the quantity of note-taking/reviewing behaviors and the content of notes recorded by students to trace the growth of ability in applying these learning strategies over time. These measures may act as indicators of SRL, as students who develop better self-regulatory skills could be expected to not only exhibit a higher frequency of note-taking and note-reviewing, but also – and more importantly – to take notes that are of higher quality.

Research on Paper-Based Note-Taking

As a popular, nearly ubiquitous academic strategy, note-taking has been thoroughly studied. In particular, there has been extensive research on traditional paper-based note-taking in the context of classroom lectures or learning from texts. Educational studies have long documented the crucial role of note-taking in facilitating academic success, especially for college students (Armbruster, 2009; Crawford, 1925; Fisher & Harris, 1973; Kiewra, 1989; Peverly et al., 2007). Researchers have identified two basic functions of note-taking that could explain its beneficial role in enhancing learning and performance – taking notes (referred to as the encoding function because the process of recording information in notes supports encoding information cognitively and facilitates learning) and reviewing notes (referred to as the external storage function because these notes serve as external memory storage that can be reviewed afterwards and are beneficial for learning) (Di Vesta & Gray, 1972; Williams & Eggert, 2002).

Encoding function

Several researchers have argued that the process of selecting and recording information in notes is by itself beneficial for learning and performance. They propose that taking notes promotes learning as it attracts learner attention to instructional content (Di Vesta & Gray, 1972; Einstein, Morris, & Smith, 1985; Kiewra, 1989), facilitates translation of instructional content into text and one's own understanding (Conway & Gathercole, 1990; Piolat, Olive, & Kellogg, 2005), enables better construction of deep-level mental representations of content (Bui, Myerson, & Hale, 2013; Slotte & Lonka, 1999), and empowers elaborative and generative processing by encouraging learners to connect new content with existing prior knowledge (Einstein et al., 1985; Peper & Mayer, 1978). Meanwhile, as the information that originally needs to be stored in working memory has been stored in external storage (e.g., notebooks), the process of taking notes also offloads extraneous cognitive load imposed on students during learning (Moos, 2009; Piolat et al., 2005).

However, results of empirical studies on the benefits of encoding have been mixed (see Kiewra, 1985a; Kobayashi, 2005 for reviews). On the one hand, considerable research has indicated that students who took lecture notes generally outperformed non-note-takers who merely listened during lectures on various tasks (e.g., comprehension, recall, retention) in the absence of reviewing notes (e.g., Bretzing, Kulhavy, & Caterino, 1987; Crawford, 1925; Di Vesta & Gray, 1972; Einstein et al., 1985), supporting the encoding function hypothesis with overall small to modest positive effects (Kobayashi, 2005). On the other hand, a number of other studies have shown no significant difference in performance between note-takers who did not review notes and non-note-takers (e.g., Howe, 1970; Kiewra et al., 1991), or have indicated that taking notes can even interfere with learning (e.g., Peck & Hannafin, 1983). Kobayashi (2005) suggested that the mixed results might be moderated by the depth of processing involved in the note-taking process (e.g., whether a generative note-taking strategy was adopted or not).

External storage function

Findings of empirical studies testing the external storage function show higher consensus in favor of this hypothesis than research on the encoding function (Kiewra, 1989). In this context, notes produced by learners serve as "external storage" for subsequent review and study. According to a meta-analysis, reviewing notes produces overall large positive effects on performance (Henk & Stahl, 1984). Substantial evidence has demonstrated that students who reviewed notes (including notes provided to them) showed superior performance on measures of learning than students who did not review notes (Carter & Van Matre, 1975; Di Vesta & Gray, 1972; Fisher & Harris, 1973; Howe, 1970; Kiewra, 1985b; Kiewra et al., 1991; O'Donnell & Dansereau, 1993; Rickards & Friedman, 1978). These researchers have argued that reviewing and studying notes consolidates one's understanding of instructional content and assists in retention and learning, as it provides a second chance to study and relearn the content (Carter & Van Matre, 1975; Di Vesta & Gray, 1972), and reduces the risks of natural forgetting (Kiewra, 1989). Meanwhile, reviewing notes may help students organize and reconstruct the material through linking concepts recorded in notes (i.e., internal connections) and connecting the noted information with prior knowledge (i.e., external connections), ultimately leading to enhanced performance and learning (Peper & Mayer, 1978; Rickards & Friedman, 1978).

Furthermore, for studies comparing the relative importance of note-taking versus note-reviewing, reviewing notes as a form of external storage has been shown to be more beneficial than the encoding function for performance (e.g., Carter & Van Matre, 1975; Fisher & Harris, 1973; Kiewra et al., 1991; Rickards & Friedman, 1978). However, these two fundamental functions of note-taking, encoding and external storage, are not incompatible. Both functions contribute to positive learning outcomes. A combination of taking and reviewing notes benefits learning the most and leads to optimal achievement (Fisher & Harris, 1973; Kiewra, 1985a; Kobayashi, 2006).

Assessing Quantity and Content of the Notes Associated with Successful and Unsuccessful Learning

More recently, research on note-taking has developed beyond experimental studies testing the relative importance of the encoding and external storage functions and has begun to delve into the quantitative and qualitative differences of notes taken by students that are associated with successful and unsuccessful learning.

Note quantity and academic performance

Multiple studies have examined note-taking quantitatively, demonstrating that increased note-taking (e.g., measured by indicators like word count or number of important ideas in notes) is significantly positively associated with learning and test performance, whether or not students review the notes (Cohn et al., 1995; Kiewra & Fletcher, 1984; Slotte & Lonka, 1999). These findings align with the two functions of note-taking (Armbruster, 2009). As learners record a greater quantity of information in notes, the encoding function is strengthened. More complete notes might also suggest that more elaborative content is generated and a deeper level of processing is involved, leading to better achievement. In the meantime, more notes indicate a larger repository of information for review, maximizing the external storage benefit.

In addition to the extensive research that examines the quantity of notes encoded and its importance for learning, some studies on the repetition effect have investigated whether increasing the quantity of reviewing episodes can boost performance or not (Annis & Annis, 1987; Bromage & Mayer, 1986; English, Welborn, & Killian, 1934). These studies suggest that reviewing instructional material multiple times improves performance over listening to or reading instructional material during one single period. However, it is worth pointing out that the review sessions of lectures or texts in these studies are somewhat different from reviewing notes. When reaccessing and reviewing this type of instructional content, students listen to the entire lecture or reread passages. During note-reviewing periods, students reaccess and restudy their notes, which typically have lower completeness and accuracy of information, but usually contain chunks of information that they regard as important and may include notes reflecting the students' own understanding. Meanwhile, these studies on the repetition effect mainly focus on review sessions after the study is over while reviewing notes could occur during learning to assist with real-time problem solving, especially in computer-based learning environments. Therefore, more research should be conducted on the quantity of note-reviewing, including the frequency of reviewing notes as external storage and the amount of time spent on reviewing notes, not only after the study but also during study sessions. Further, it could be useful to explore how note-takers should distribute their time between taking and reviewing notes.

Note content and academic performance

In addition to the quantity of notes, the content of notes is also important for academic achievement. Numerous studies have examined the content of notes that are associated with successful and less successful learners from the perspective of the level of cognitive processing (Craik & Lockhart, 1972). Cognitive processing involved in note-taking ranges from the superficial level of verbatim copying of information to a relatively deeper level of cognitive processing that entails elaboration of instructional content (e.g., through inducing inferences, summarizing, generating hypothesis, constructing connections, self-questioning, concept mapping, etc.). Generative and elaborative note-taking (referred to as constructive by Chi, 2009) that involves deep cognitive processing, such as inference generation, was found to be associated with better performance than note-taking that involves relatively shallower processing such as verbatim copying (Armbruster, 2009; Bretzing & Kulhavy, 1979; Igo, Bruning, & McCrudden, 2005; Mueller & Oppenheimer, 2014; Slotte & Lonka, 1999). However, elaborative note-taking can be difficult and, as Kiewra and Fletcher (1984) have found, even undergraduate students can have difficulties in taking content elaborative notes despite being instructed to do so.

These results on the advantage of elaborative note-taking are consistent with the well-documented literature on the generation effect (Foos, Mora, & Tkacz, 1994; Peper & Mayer, 1978, 1986; Richland, Bjork, Finley, & Linn, 2005; Wittrock, 1974), which indicates that having learners generate information and meaning during study leads to increased retention and learning, compared to merely passively processing the information without generation. For example, note-taking is a generative activity when note-takers relate the instructional material to their prior knowledge and generate new information by making inferences or constructing connections. Thus, note-taking that involves generative strategies is more effective and instrumental in learning than non-generative note-taking. This finding is also included in Chi's (2009) Interactive-Constructive-Active-Passive (ICAP) framework. In ICAP, Chi posits that constructive activities are superior to active activities, based on this earlier evidence, which in turn are seen as better for learning than passive activities. Accordingly, she points out that the active process of taking notes, which is at minimum an active activity, is better in terms of learning outcomes than being passive and not taking notes. Elaborating on presented information and generating information and ideas that go beyond the meaning of the original content in notes, which constitutes a constructive activity, is therefore preferable to reproducing instructional content while taking notes, which comprises an active activity.

Note-Taking in Non Lecture-Based Contexts

In addition to the considerable quantity of research on note-taking in lecture-based settings, another line of paper-based note-taking research studies the effects of note-taking and note-reviewing in non-educational contexts such as the note-taking/reviewing by jurors in courtrooms (see Kiewra, 2016; Peverly & Wolf, in press for reviews). Experimental studies indicated that note-taking improved jurors' recall and recognition memory for trial information and promoted the decision making of jurors (Forsterlee & Horowitz, 1997; Forsterlee, Kent, & Horowitz, 2005; Thorley, Baxter, & Lorek, 2016). The improved recall and memory potentially mediated the relationship between paper-based note-taking and decision making (Forsterlee & Horowitz, 1997). Researchers also concluded that the encoding function of note-taking is more important than the external storage function of note-taking for jurors in courtrooms, as the note-takers demonstrated better recall for trial information and more effective decision making than non note-takers, whereas no significant difference was obtained between the jurors with and without access to notes for review during the deliberations and decision making (Forsterlee & Horowitz, 1997). This contradicts findings from the literature on note-taking in lecture-based contexts, where the external storage function was usually found to be more important than the encoding function (e.g., Carter & Van Matre, 1975; Fisher & Harris, 1973; Kiewra et al., 1991; Rickards & Friedman, 1978).

Computer-Based Note-Taking

Compared with the substantial literature on traditional paper-based note-taking that mostly predates the introduction of computers to science classrooms, computer-based note-taking is an emerging area of study with a growing number of studies (Bauer, 2008; Bauer & Koedinger, 2006; Crooks, White, & Barnard, 2007; Igo et al., 2005; Igo & Kiewra, 2007; Mueller & Oppenheimer, 2014; Robinson et al., 2006). Computer-based note-taking is different from traditional paper-based note-taking partly because typing speed on computers is typically faster than handwriting speed (Brown, 1988), probably resulting in a greater amount of information being recorded on computers. Additionally, the content and quality of notes recorded might also be different depending on how the notes are taken (Armbruster, 2009; Mueller & Oppenheimer, 2014).

A few researchers investigated the effect of computers on student note-taking from lectures compared to paper-based note-taking (Bui et al., 2013; Mueller & Oppenheimer, 2014). Bui and colleagues (2013) indicated that college students who took notes of a short lecture on computers recorded significantly more idea units and showed better performance on immediate recall tests than students who took notes by hand. On the other hand, Mueller and Oppenheimer (2014) have indicated that students who took notes on laptops tended to process information relatively shallowly and take more verbatim notes, which impaired learning and led to lower performance than taking notes by hand. However, these studies mainly focus on examining the effect of computers on student note-taking from lectures. Note-taking/reviewing of lectures on computers is different from taking and reviewing notes from open-ended learning environments from multiple perspectives (discussed in details in the following section).

Taking and reviewing notes in open-ended learning environments for science inquiry

Note-taking in OELEs is different from note-taking during lectures on computers in the following fashions. First, oral content is delivered during lectures while the instructional information in OELEs are usually distributed over various representations (e.g., animations, text, graphics, audios, videos, etc.). In a meta-analysis, Kobayashi (2005) indicated that the encoding effect is greater when the material is presented as audio than when the presentation mode is text or audio-visual, where the note-taking process interferes with visual attention to instructional material. Second, the information students listen to and simultaneously take notes of during lectures is linear and transient. On the contrary, the multimedia information in OELEs is nonlinear and does not have the time restriction inherent to lectures. Students can select, process, and record the information at their own pace (Slotte & Lonka, 1999). Third, reviewing notes in OELEs is different from note-reviewing during lectures, where review of notes mainly takes place after class when all notes have been taken. In OELEs, note-reviewing happens concurrently with note-taking during science inquiry, as students have accesses to their notes in real-time to support their problem-solving. Fourth, the non-linearity and the open-endedness of OELEs result in more flexibility and time for students to connect and coordinate representations from multiple disparate sources and record them in notes. Last, learners explore open-ended learning environments actively and assume active control of their learning and exploration. Accordingly, OELEs pose high demands on self-regulated learning skills, which in turn imposes high cognitive load on students (Moos, 2009). The processing of a large volume of multimedia information from OELEs also has the potential to tax students' limited cognitive processing capacity. Both of these processes may overload students and make note-taking in OELEs challenging. In contrast, lectures entail more passive listening to the linear content and less control by learners (O'Donnell & Dansereau, 1993). Thus, note-taking in OELEs poses different challenges on learners from note-taking during lectures. A summary of these differences is shown in Table 1.

Results from studies on computer-based note-taking in OELEs are mixed, sometimes agreeing and sometimes contradicting the results found in the literature on traditional note-taking. For example, undergraduate learners in Trafton and Trickett's (2001) study who used a digital notepad to take notes while solving scientific problems in an OELE for science outperformed those who did not use the notepad. Students who had used the notepad performed better even later when it was no longer available, a result comparable to previous findings on the positive effects of note-taking in traditional settings. No relationship between the quantity/content of notes and performance was explored in this study. On the other hand, results contradicting traditional note-taking literature have been found in other OELEs. For instance, McQuiggan and colleagues (2008) had students take and review notes while engaging in science inquiry tasks and solving a science mystery in Crystal Island. They did not find significantly different performance and learning gains between note-takers and non-note-takers.

Table 1

Differences between computer-based note-taking in traditional instructional environments and note-taking in open-ended learning environments (OELEs)

Computer-Based Note-Taking in Traditional Instructional Environments	**Note-Taking in OELEs**
• Oral content is delivered during lectures and notes are taken on that content.	• The instructional information in OELEs is usually distributed over various representations (e.g., animations, text, graphics, audios, videos, etc.), where the visual attention to text or visual information may interfere with the note-taking process (Kobayashi, 2005).
• The information students listen to and simultaneously take notes of during lectures is linear and transient.	• The multimedia information in OELEs is nonlinear and does not have the time restriction inherent to lectures. Students can select, process, and record the information at their own pace (Slotte & Lonka, 1999).
• Reviewing notes in traditional instructional environments mainly takes place after class when all notes have been taken.	• Note-reviewing happens concurrently with note-taking during inquiry, and students review notes to support their real-time problem-solving.
• Learners have relatively less time to take generative notes that connect instructional information with prior knowledge or with information transmitted earlier, as they are taking notes simultaneously with receiving direct instruction.	• The non-linearity and the open-endedness of OELEs result in more flexibility and time for students to connect and coordinate representations from multiple disparate sources and record them in notes.
• Lectures entail more passive listening of the linear content and less control by learners (O'Donnell & Dansereau, 1993).	• Learners assume active control of their learning and exploration in OELEs. The high requirements on self-regulated learning and the large volume of multimedia information in OELEs impose a high cognitive load on students and may make it challenging to allocate cognitive resources to note-taking.

A more recent analysis on note-taking in OELEs by Trevors and colleagues (2014) did not find any positive associations between the quantity and quality of notes and learning outcomes. For example, they found that the frequency of note-taking actions was negatively associated with subsequent learning outcomes in a hypermedia learning environment, which contradicts the positive correlations between note quantity and performance found in previous research. However, in this study, note-reviewing actions were not distinguished from note-taking actions. In addition, they coded notes qualitatively into content reproduction (notes where learners reproduce the instructional content through memorization or rehearsal strategies), and content elaboration (notes where learners elaborate on the underlying meaning and patterns of content). The number of content reproductive notes, which comprised the majority of the notes taken by students, was negatively associated with learning outcome. Meanwhile, no advantage of constructive and generative note-taking was found, as the number of content elaborative notes that involved a deep level of cognitive processing was not significantly associated with learning performance. The researchers argue that taking notes in OELEs is detrimental to learning and impedes performance because the cognitive overload caused by note-taking limits students' exploration of the representations and the learning environment.

In another study on the same OELE, Bouchet and colleagues (2013) applied clustering analysis to classify undergraduate learners based on their use of self-regulatory processes and strategies. Results suggested that students with higher self-regulatory skills and higher prior knowledge tended to take fewer notes and spend less time taking notes than students in the other clusters. Despite taking fewer notes, these students checked their notes more often.

With the lack of consensus in these studies, it is desirable for researchers to systematically probe into whether note-taking/reviewing is beneficial or detrimental for science performance in the context of open-ended learning environments and whether findings from classical paper-based note-taking literature can transfer to OELEs. Specifically, this dissertation examined both the quantity of note-taking/reviewing behavior and content of notes, their relationship with science inquiry performance, and the development of these skills for both male and female learners in an OELE.

Gender Differences in Note-Taking

Considering the critical role of note-taking/reviewing in SRL and academic success, researchers have also explored the individual differences that influence note-taking (e.g., gender, academic level, prior knowledge, cognitive characteristics such as working memory and motivation, etc.). As mentioned above, research suggests that there are gender differences in traditional paper-based note-taking (Cohn et al., 1995; Kiewra, 1984; McQuiggan et al., 2008; Williams & Eggert, 2002).

Quantity of note-taking/reviewing and gender

Cumulative evidence from studies on the relationship between paper-based note-taking and learner's gender suggest that females take a significantly greater quantity of notes than males (Cohn et al., 1995; Kiewra, 1984; Maddox & Hoole, 1975; Nye, 1978; Reddington, Peverly, & Block, 2015; Slotte, Lonka, & Lindblom-Ylänne, 2001; Williams & Eggert, 2002). In a study conducted by Nye (1978), female university students noted significantly more words and more major and minor points than their male counterparts while listening to an introductory psychology lecture on child development. However, females did not have better end-of-year performance in this course than males. Kiewra (1984) also found that female undergraduates were more complete note-takers; their notes contained more words and more critical points than notes by males. Moreover, females outperformed males in the subsequent delayed learning tests after taking and reviewing notes over the lecture on educational psychology. Nevertheless, it should be noted that the sample in this study consisted of 22 females and 7 males, which limits the generalizability of the conclusions. In a larger study where a total of 211 undergraduate students listened to a videotaped lecture on economics (Cohn et al., 1995), women recorded 40 more words and 5.1 more important lecture points on average than men. Yet, gender was not a significant predictor of performance on post-test when holding other variables constant. More recently, Reddington and colleagues (2015) examined notes taken by college students as they listened to a videotaped psychology lecture in a laboratory setting and reported that females were better lecture note-takers than males. Overall, females showed higher handwriting speed and recorded more content topics in notes. The higher quality and completeness of notes led to females' superior performance on written recall of the lecture over males.

Despite the consistent findings that females take more notes than males, literature on gender differences has rarely examined the note-reviewing process, although reviewing notes as external storage has been shown to be crucial for learning. To my knowledge, no study has compared the note-reviewing behaviors/activities engaged in by male and female learners, such as the timing, quantity, and duration of note-reviewing episodes/sessions. As indicated by studies in the previous paragraph, results on the relationship between test performance following taking/reviewing notes of lectures or texts and gender have been mixed and unclear (Kiewra, 1984). Therefore, it is important to explore whether students of different sexes review notes for subsequent tests differently or not, and if so, how this difference is related to performance on later tests.

Note content and gender

In addition to the links between note quantity and student's gender, a few of the aforementioned studies explored potential gender-related differences in note content by comparing the level of cognitive processing involved in notes taken by males and females. However, results from the limited studies are mixed and equivocal. For example, Maddox and

Hoole (1975) noticed that females not only took more complete notes than males but also tended to verbatim copy information transmitted in a university lecture more often than males, while males tended to abbreviate the information to fewer words. In Slotte et al. (2001), female high-school graduates relied more on note-taking while learning from texts on philosophy and statistics than male students. Further examination revealed that females were more likely to summarize the content from the statistical texts in notes than males, which entails a relatively deeper level of cognitive processing. On the other hand, males and females did not differ significantly in their tendency to verbatim copy content from the text. Despite female students engaging in more note-taking within this study, males performed as well as females on deep-level comprehension of the texts. Nonetheless, it is worth noting that in this research the corpus of a student's notes was coded as either verbatim notes or summarized notes, in the corpus's entirety. It is highly possible that a student takes both types of notes and it is the distribution of the contents that matters. More comprehensive research is needed to study note content at a more fine-grained level (e.g., sentence segment level) and compare the distribution of different contents of note segments by gender.

Hypotheses for why gender differences exist in note-taking

Further, researchers have explored the underlying cognitive and non-cognitive factors that might explain the existence of gender-related differences in note-taking. It is not yet clear why women take more notes than men. Below I will summarize some hypotheses raised by researchers that shed light on the differences. Carrier and colleagues (1988) postulate that the gender differences in note-taking are related to learners' beliefs in note-taking. According to their survey results, female college students deemed note-taking as more valuable and were more self-confident about their note-taking strategies than male students. Other researchers argue that the relationship between note-taking and gender reflect general gender differences in verbal ability, which usually is higher in females than males in previous research (Hartley & Davies, 1978; Reddington et al., 2015). For example, females were found to possess higher verbal fluency than their male counterparts and usually tended to excel at verbal related tasks, such as reading and writing, while males typically showed better visuospatial skills (Halpern, 1997, 2004). Reddington and colleagues (2015) investigated the cognitive and motivational factors that might illuminate the relationship between gender and lecture note-taking. They reported a significant interaction between language comprehension and gender on notes, such that females with high language comprehension produced significantly more idea units in notes than males with high language comprehension, while notes taken by females and males with low language comprehension were not significantly different in terms of the quantity of ideas recorded in notes.

Limitations of previous research on gender-related differences in note-taking

Note that most of the aforementioned studies on gender-related differences in note-taking have been conducted in the context of paper-based note-taking during lectures or text learning in psychology-related domains, among college students. In addition, student performance was mainly evaluated through simple paper-based tests (e.g., recall tests). Does gender difference in note-taking emerge as early as middle school, or even elementary school? Would similar trends in gender differences develop when students take and review notes of content in other domains, such as science? Does the relationship between gender and performance on subsequent tests extend to more complex tasks such as science inquiry tasks? These are all directions for further research. I aim to explore these questions in this dissertation study.

Relationship between gender and note-taking in OELEs

Moreover, research is limited on the relationship between gender and note-taking in computer-based learning environments, such as open-ended learning environments (OELEs) for science. One of the limited studies that examined the role of gender in student computer-based note-taking behaviors was conducted by Kay and Lauricella (2011), who examined student laptop usage in higher education classrooms and found that female students reported significantly more note-taking activities on laptops than males in a survey. In another study, McQuiggan et al. (2008) noted that female middle school students took significantly more notes in a digital notepad than male students while they tried to solve a science mystery in Crystal Island. Additionally, females took more notes containing the facts from the instructional content (e.g., definitions of scientific terms), as well as more notes on the narrative storyline. In contrast, males recorded more notes that did not contain any meaningful information than females. There was no significant gender-related difference in deeper-level notes such as hypothesis or procedural notes. These findings suggest that females take more notes in OELEs just like they do on paper, despite the more positive attitudes of males towards computers and science. However, further research is necessary to examine whether these results can be replicated in other science OELEs with different design goals and structures, and to study gender differences in note-taking within OELEs more comprehensively by analyzing more features such as the quantity and timing of note-taking/reviewing behaviors and the content of notes.

Summary of the Literature Review

In summary, the development of open-ended learning environments (OELEs) and the abundant interaction data in OELEs not only pose challenges to measure and identify self-regulated learning, but also afford an opportunity to apply Educational Data Mining (EDM) methods for analyzing self-regulated learning in an unobtrusive and fine-grained manner. There is a surging interest in applying sequential pattern mining methods to study how self-regulated learning manifests in open-ended learning environments. This is typically achieved by using clustering methods or human coding to classify students with various levels of self-regulatory skills and then comparing the behavioral patterns of these groups. However, research investigating the *development* of SRL strategies and skills over time within OELEs is lacking, especially for younger learners who typically lack sophisticated self-regulated learning skills (Greene et al., 2008; Pintrich & Zusho, 2002). Specifically, do open-ended learning environments foster the development of self-regulatory skills for middle school students? If so, how do self-regulatory processes and strategies develop over the use of OELEs? These are research questions that are worth investigating.

Note-taking is an important SRL strategy that learners frequently use. Research has shown that taking notes (encoding function) and reviewing notes (external storage function), which are nearly universal academic strategies across various educational contexts, are positively associated with academic success. In addition, both the quantity and the content of notes are important for learning performance. However, most of these studies focus on conventional paper-based note-taking from lectures or texts in classroom or laboratory settings for adults. Results from the growing number of studies on note-taking in science OELEs are mixed. Further research is needed regarding the correspondence between the quantity of note-taking/reviewing behavior and the content of notes within OELEs and success on complex academic tasks such as scientific inquiry among younger learners. Most importantly, research investigating the development of note-taking as an SRL strategy in OELEs is needed and will provide insights into the development of SRL processes and strategies.

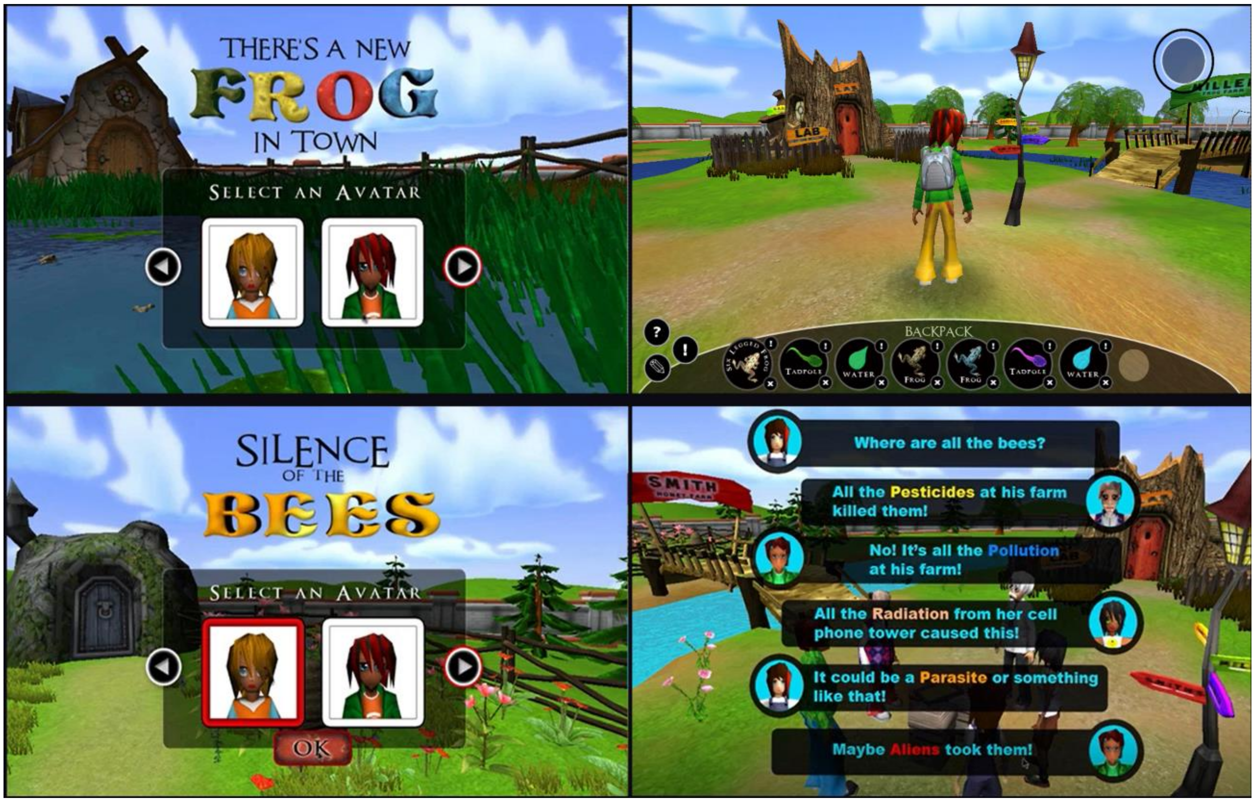
Gender has also been documented as being significantly associated with self-regulatory skills. Females reported themselves as using self-regulatory activities and strategies more often, and taking more notes and capturing more critical ideas in notes than males. However, few studies have investigated the role of gender in the development of self-regulatory skills within OELEs. Studies that systematically and comprehensively examine the relationship of a learner's gender with their development of key processes and strategies of self-regulated learning in science OELEs will be informative to us to understand whether males and females develop self-regulatory skills differently in OELEs, and to design personalized scaffolding accordingly.

CHAPTER III.

VIRTUAL PERFORMANCE ASSESSMENTS

The context for this dissertation research is the Virtual Performance Assessment Project (Clarke-Midura, McCall, & Dede, 2012). The open-ended learning environment, referred to as Virtual Performance Assessments (VPA) within this dissertation, is a 3-D immersive virtual environment that has the look and feel of a video game but is designed to assess middle school students' science inquiry skills *in situ* (McCall & Clarke-Midura, 2013; Scalise & Clarke-Midura, 2014). Within the environment, students engage in authentic scientific inquiry activities by navigating an avatar around the open-ended learning environment, making observations, gathering data, interacting with non-player characters (NPCs), reading kiosk informational pages for research, taking notes, and conducting virtual laboratory experiments. These actions are recorded automatically and unobtrusively on the back end in the form of process data (e.g., where they went and what they did in the open-ended learning environment) as well as product data (e.g., student notes and final claims).

The larger Virtual Performance Assessment Project provides students with multiple assessment scenarios. This research uses data from two VPA scenarios: the "frog scenario" and the "bee scenario" (see Figure 4). The two scenarios have similar structure and mechanics in order to allow researchers to assess performance of the same inquiry practices in different contexts. The difference between the two scenarios is the content of the problem that students are asked to solve and the surface features associated with the scenario. In both scenarios, students visit four virtual farms to determine the cause of distress to the creature in question (frogs or bees). In both, they are told that the possible causal factors are parasites, pesticides, pollution, radiation-induced genetic mutation, and space aliens. In each scenario, only one of these is correct.



Where are all the bees?

All the Pesticides at his farm killed them!

No! It's all the Pollution at his farm!

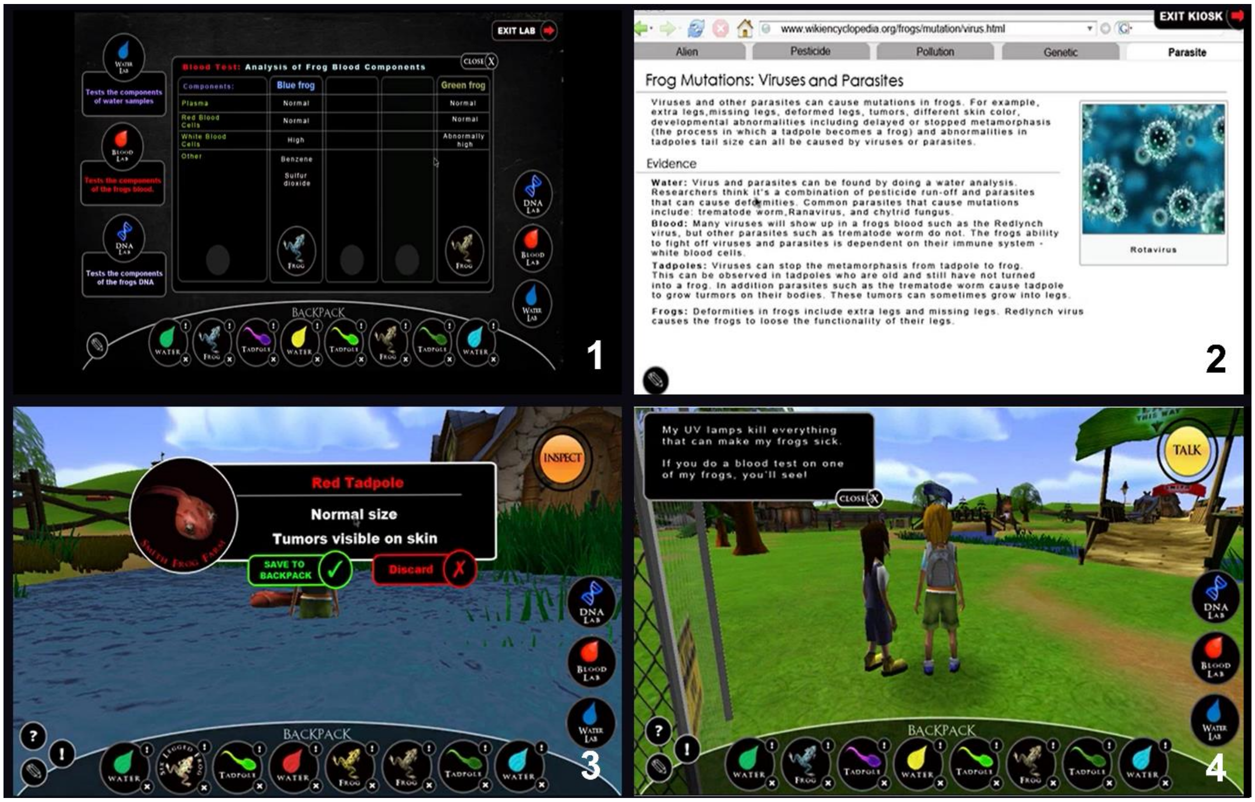
All the Radiation from her cell phone tower caused this!

It could be a Parasite or something like that!

Maybe Aliens took them!

Figure 4. Screenshots of the VPA frog scenario (top) and the VPA bee scenario (bottom).

The environment contains different types of data sources. Students can talk to NPCs from the four farms who provide conflicting opinions of what is causing the problem. They can also read informational pages about five possible causal factors from a research kiosk (see Figure 5). The information in the kiosk pages includes what types of tests and evidence can be found for each causal factor. For example, the page about parasites in the frog scenario contains information about water and blood tests and what type of results are evidence for parasites. Students can also conduct laboratory tests (see Figure 5) such as a water analysis that includes pH levels and contaminants and a blood test that reports on components such as plasma, red blood cells, and white blood cells of the samples they collect at the farms (e.g., frogs, tadpoles, water samples, bees, larvae, and nectar samples). These data provide evidence that parasites have caused the frog to grow six legs and radiation-induced genetic mutation is causing the bees to die.



Blood Test: Analysis of Frog Blood Components

Components:	Blue frog			Green frog
Plasma	Normal			Normal
Red Blood Cells	Normal			Normal
White Blood Cells	High			Abnormally high
Other	Benzene Sulfur dioxide			

www.wikiencyclopedia.org/frogs/mutation/virus.html

Frog Mutations: Viruses and Parasites

Viruses and other parasites can cause mutations in frogs. For example, extra legs,missing legs, deformed legs, tumors, different skin color, developmental abnormalities including delayed or stopped metamorphasis (the process in which a tadpole becomes a frog) and abnormalities in tadpoles tail size can all be caused by viruses or parasites.

Evidence

Water: Virus and parasites can be found by doing a water analysis. Researchers think it's a combination of pesticide run-off and parasites that can cause deformities. Common parasites that cause mutations include: trematode worm,Ranavirus, and chytrid fungus.

Blood: Many viruses will show up in a frogs blood such as the Redlynch virus, but other parasites such as trematode worm do not. The frogs ability to fight off viruses and parasites is dependent on their immune system - white blood cells.

Tadpoles: Viruses can stop the metamorphasis from tadpole to frog. This can be observed in tadpoles who are old and still have not turned into a frog. In addition parasites such as the trematode worm cause tadpole to grow turmors on their bodies. These tumors can sometimes grow into legs.

Frogs: Deformities in frogs include extra legs and missing legs. Redlynch virus causes the frogs to loose the functionality of their legs.

Red Tadpole

Normal size

Tumors visible on skin

My UV lamps kill everything that can make my frogs sick.

If you do a blood test on one of my frogs, you'll see!

Figure 5. Screenshots of the different data sources in the frog scenario: 1) laboratory test results, 2) research kiosk page, 3) field observation, and 4) conversation with NPCs.

One of the key tools that students have as they investigate these possible causes and as they keep track of their data is a digital notepad (Figure 6). Students can access the notepad any time they want to enter information or review their notes. When taking notes in their digital notepad, students are not able to simply copy and paste information from the environment (e.g., kiosk research pages, dialogue with NPCs, laboratory test results, observation, etc.). Instead, they must hold the information they obtain in working memory and type in text in the notepad. The notepad can only contain text; there is no way for students to enter pictures.

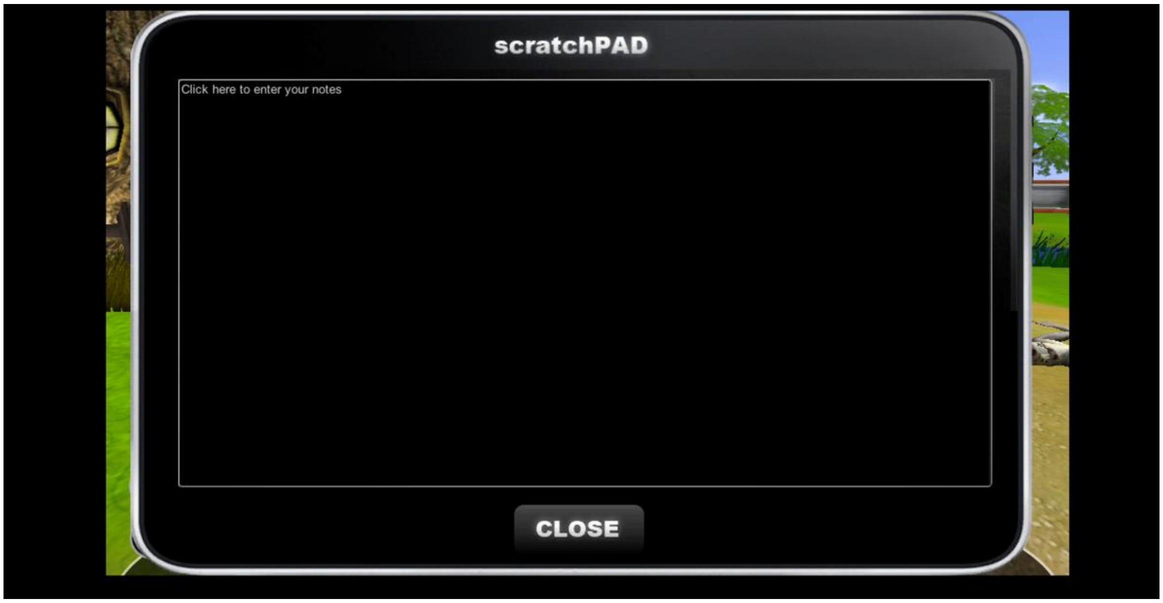


Figure 6. Screenshot of the digital notepad within VPA.

Once students think that they have collected sufficient data, they submit a final claim on the causal factor resulting in the frog mutation or bee deaths from the list of possible hypotheses and justify their conclusion with supporting evidence. These two submissions form the primary basis of VPA's assessment of science inquiry skills for each student.

Importance of Open-Ended Learning Environments for Learning and SRL

Research suggests that many middle school students do not use effective SRL processes and strategies as they learn in open-ended learning environments (Moos & Azevedo, 2008b).

The non-linearity and open-endedness of open-ended learning environments (OELEs) such as VPA create learning opportunities for students but can also impose challenges in terms of extraneous cognitive load and greater requirements for self-regulation (Azevedo, 2005; Moos, 2009; Moos & Azevedo, 2008b). In order to be successful in VPA, students need to understand their inquiry tasks, set goals and make corresponding plans. Such plans include deciding where to go in the open-ended learning environment, which information (e.g., research information from kiosk pages, information from conversations with NPCs, etc.) to collect and access, which activities to engage in, which resources to utilize, and in what sequence. At the same time, they must apply learning strategies such as recording information in the online notepad and reviewing their notes, reflect on their learning, and monitor their inquiry processes. Students with different levels of self-regulatory skills employ different SRL strategies and processes and exhibit unique behavior sequences. These behaviors and processes correspond to the recursive stages in Winne and Hadwin's (2009) SRL framework: understanding task definition, goal setting and planning, enacting study tactics and strategies, and metacognitively adapting studying. This study explores how self-regulatory skills manifest and develop in the open-ended learning environment, whether there are any gender-related individual differences in self-regulated learning, and how should educators design future environments to promote self-regulated learning for both male and female learners.

CHAPTER IV.

METHODOLOGY

This dissertation research investigates how the learning processes and activities in the open-ended learning environment for middle school science map to various self-regulated learning (SRL) processes, and studies how student skills in regulating their own learning develop in the open-ended learning environment. Furthermore, gender differences in the manifestation and development of SRL skills in the open-ended learning environment are explored. To investigate these issues, this research examined students' behaviors from more than 2,000 middle school as they used Virtual Performance Assessments (VPA) in their science classes. A combination of educational data mining techniques (e.g., sequential pattern mining, feature engineering) and multilevel analysis was applied on students' action log data to track how student behaviors demonstrate SRL, how self-regulatory skills develop in VPA, and whether the development of SRL processes and strategies is different between male and female learners.

Participants

This dissertation analyzes interaction log files produced by a total of 2,429 seventh and eighth-grade students (12-14 years old) who used VPA within their science classes at the end of the 2011-2012 school year. Two other students were excluded from analysis due to lack of data on their demographics (e.g., gender). These students were drawn from 130 classrooms that were taught by 39 teachers from a diverse selection of school districts in the Northeastern and Midwestern United States and Western Canada. Forty-seven percent of the students were males (*n* = 1,140), and 53% of them were females (*n* = 1,289).

Procedure

Students were randomly assigned to begin with either the frog scenario (*n* = 1,232) or the bee scenario (*n* = 1,197). Each student was assigned the other scenario two weeks later (bee: *n* = 824; frog: *n* = 753), subject to some attrition[1]. Prior to each assessment, students were shown a short introductory video that provided instructions on how to use the VPA. Following the video, students worked within each scenario until they had completed the analysis and produced a final answer for its underlying problem (e.g., why does this frog have extra legs or why are these bees dying). In sum, a total of 1,985 students completed the frog scenario and 2,021 students completed the bee scenario, with 1,577 students completing both scenarios.

Students spent approximately half an hour in each scenario (frog: *M* = 30 min., 47 sec., *SD* = 14 min., 6 sec.; bee: *M* = 26 min., 6 sec., *SD* = 12 min., 26 sec.). On average, each student completed 192 actions within the frog scenario, resulting in a total of 381,331 actions. In the bee scenario, students completed an average of 196 actions, producing 396,760 actions in total. During this time, student actions, notes, and performance in the virtual assessments were automatically logged and were used for analyses.

[1] Due to factors such as time arrangement, some students completed one scenario whereas others completed both scenarios in this study. Pre-intervention measures (e.g., gender, age, grade) were compared between students who completed both scenarios and students who only completed one scenario in order to test the potential effect of attrition on results. No significant difference was found for the comparison of the pre-intervention measures. To further ensure that the comparison of the first-time users and the second-time users is valid, I conducted the same analyses by excluding the students who only completed one scenario. Similar results were obtained when I only considered students who completed both scenarios. This indicated that the attrition occurred at random and that the first-time user group and the second-time user group might be equivalent.

Data Analysis

As mentioned in the previous section, students were randomly assigned to begin with either the frog or the bee scenario and were assigned to complete the other scenario two weeks later. Therefore, within each scenario, participants could be put into two groups – users who were using VPA for the first time (*first-time user* group) and users who had previously experienced the other VPA scenario and who were using VPA for the second time (*second-time user* group). Accordingly, among the 1,985 students who completed the frog scenario, 1,232 were first-time users (frog-first) and 753 were second-time users (frog-second). Among the students who completed the bee scenario, 1,198 were first-time users (bee-first) and 825 were second-time users (bee-second). This dissertation study explores students' development of self-regulatory skills while playing with VPA by comparing the first-time users and the second-time users.

In addition, the role of gender in this development process was examined by exploring whether the development of self-regulatory skills differs by gender or not. Specifically, I investigated the interaction of experience with VPA (whether a student is a first-time user or a second-time user) and gender (male vs. female) on each SRL-relevant measure, including science inquiry performance, behavioral patterns that are reflective of SRL, and note-taking/reviewing strategies. Main effects comparisons of differences between conditions were reported if no significant interaction was observed; otherwise simple effects for each subgroup were examined and reported if the interaction was statistically significant.

As mentioned in the previous section, three analyses were conducted to answer the research questions. In analysis 1, I ran a multilevel analysis to study the relationship between experience with VPA and gender towards student performance on science inquiry tasks. The comparisons of student performance would help us study the development of SRL skills across scenarios. In analysis 2, sequential pattern mining was applied on middle school students' action log data to track how their behavior patterns demonstrated SRL, and whether using VPA promoted students' use of self-regulatory processes and strategies. The frequency of behavior patterns that are representative of various SRL phases and strategies was compared between 1) the first-time users and the second-time users, and 2) male and female learners in each scenario. Analysis 3 first employs feature engineering and correlation mining to explore the relationship between note-taking and note-reviewing strategies in VPA, which are effective learning strategies and important components of SRL (Moos, 2009; Trevors et al., 2014), with science inquiry performance. Furthermore, the development of these strategies in VPA were studied by examining the interaction of experience and gender on the quantity of note-taking/reviewing behaviors and content of notes.

Accordingly, three different types of measures related to SRL processes within VPA were collected and developed for analyses: 1) Student science inquiry performance in VPA, including the correctness of the final claim (CFC) on the cause of the six-legged frog or the death of the bees, the student's success in identifying supporting evidence (ISE) to justify why that claim is correct, and the student's demonstration of the control of variables strategy (CVS) in their science inquiry behavior; 2) Frequency metrics of action sequences related to different SRL processes and strategies, identified through sequential pattern mining; 3) Variables related to note-taking and note-reviewing, including purely quantitative measures based on actions involving VPA's digital notepad (e.g., frequency of note-taking or note-reviewing) as well as measures developed through qualitative coding of the notes. These measures are discussed in detail in the following chapters.

Multilevel modeling was adopted in this dissertation to investigate the potential differences between the first-time users and the second-time users on the SRL-relevant measures, and the role of gender in this process. Multilevel models are linear statistical models that are applied to nested data (e.g., data where individuals are nested within classes, classes nested within teachers, teachers nested within schools, etc.) by allowing coefficients to vary randomly and vary at more than one level (Snijders & Bosker, 1999). Accounting for the associations among observations within levels, separate equations are specified and fit at each level in multilevel modeling to contain both fixed and random effects. Multilevel modeling is often used in educational research because it takes into account the effects of common contexts shared by individuals, such as students grouped within the same class. The multilevel approach is adopted in the dissertation study due to the hierarchical structure of the data, where the population consists of students nested within classes, and multiple classes that shared the same teacher.

Specifically, three-level logistic regression models were fitted with students in each scenario nested within classes, and classes nested within teachers for comparison of binary measures in each scenario. Similarly, three-level regression models with students in each scenario nested within classes, and classes nested within teachers were fitted to explore the relationship between the predictor variables and continuous measures. More details about the multilevel models can be found in the following sections. These three-level analyses, taking the hierarchy of data into consideration, enable the author to examine the relationship between variables (e.g., the relationship between student's experience with VPA and gender towards their use of SRL processes and strategies) after controlling for class- and teacher-level variability.

In this study, multilevel analyses were conducted for each SRL-relevant measure in each scenario and were implemented using the "lme4" package (Bates, Maechler, Bolker, & Walker,

2015) and the “lmerTest” package (Kuznetsova, Brockhoff, & Christensen, 2016) in the statistical software program R. Given the substantial number of statistical tests, the false discovery rate is controlled by applying Benjamini and Hochberg’s (1995) post-hoc correction method. This post-hoc control method avoids the substantial overconservation found in methods such as the Bonferroni correction (cf. Perneger, 1998). Benjamini and Hochberg’s method was used in all analyses throughout this dissertation to control for multiple statistical significance analyses.

In the following chapters, I will discuss the measures and methods used in each analysis and report the corresponding results.

CHAPTER V.

ANALYSIS 1: SCIENCE INQUIRY PERFORMANCE

Analysis one compares student performance on science inquiry tasks between 1) the first-time users and the second-time users, and 2) males and females within each VPA scenario. Three types of measures of student performance in VPA were collected and compared: 1) The correctness of the student's final claim (CFC); 2) Success in identifying supporting evidence (ISE); and 3) Ability in using the control of variables strategy (CVS) in science inquiry. SRL has been shown to be closely related to academic performance (B. J. Zimmerman, 1990), and the regulation of science inquiry strategy is a crucial part of SRL in this domain (Pintrich & Zusho, 2002).

Identifying Correct Final Claim (CFC)

In each VPA scenario, students submitted a final claim by choosing from five possible causal factors as the cause of the underlying problem. A student's final claim was considered correct and scored as 1 if the student concluded that parasites caused the mutation of the six-legged frog, or that the bee deaths were caused by radiation. All other claims were considered incorrect and scored as 0. Overall, 30% of students correctly concluded that parasites led the frog to have six legs, and 28% of students made a correct claim on what was killing the bee population.

A 3-level logistic regression was conducted to compare student CFC performance across the four groups of students in each scenario. In these models, student's CFC score serves as the dependent variable, while experience with VPA (whether students were using VPA for the first time (coded as 0) or had previous experience in the other VPA scenario (coded as 1)), gender (females coded as 0 or males coded as 1), and the interaction of experience and gender serve as the level-one factors. These hierarchical models are helpful for determining the potential interaction between experience and gender on student success on identifying a correct final claim after controlling for class- and teacher-level variability.

Results

Examination of the main effects in the bee scenario indicated that there was a statistically significant main effect for experience with VPA on the correctness of students' final claim (CFC) after controlling for class-level and teacher-level variability ($z = 4.94$, $p < .001$), with 35% of the second-time users who had previously used the frog scenario identifying correctly that radiation was killing the bees, while only 24% of the first-time users *without* prior experience in the frog scenario submitted the correct final conclusion. These results suggested that the students transferred what they learned about how to make a correct final claim from the previous frog scenario to the bee scenario. In addition, a significant gender effect was observed such that male students were more likely to identify the correct final claim than female students in the bee scenario (32% vs. 25%), $z = 4.29$, $p < .001$. There was no significant interaction between experience with VPA and gender on CFC score, $z = -1.54$, $p = .124$.

Table 2

Descriptive statistics (means with standard deviations in parentheses) of the measures on science inquiry performance for female first-time users (F-1), female second-time users (F-2), male first-time users (M-1), and male second-time users (M-2) by scenario

Scenario	**Frog**				**Bee**			
Measure	**F-1**	**F-2**	**M-1**	**M-2**	**F-1**	**F-2**	**M-1**	**M-2**
CFC	28%	38%	26%	28%	19%	33%	29%	38%
ISE	51 (23)	53 (24)	48 (23)	47 (24)	44 (18)	51 (23)	44 (21)	47 (24)
CVS-data	9.96 (6.47)	10.04 (6.76)	11.18 (6.75)	11.18 (7.00)	9.34 (6.65)	10.16 (6.99)	11.16 (7.36)	10.04 (7.23)
CVS CFC-data	2.09 (1.34)	2.12 (1.39)	2.33 (1.35)	2.29 (1.41)	1.97 (1.32)	2.15 (1.38)	2.28 (1.44)	2.08 (1.45)

On the contrary, there was a significant interaction between experience and gender on CFC in the frog scenario, $z = -2.14$, $p = .033$ (see Figure 7). Follow-up simple effect analyses revealed that among female students who completed the frog scenario, a statistically significantly higher percentage of the second-time users (38%) made a correct final claim than the first-time users (28%), $z = 3.67$, $p < .001$. There was no statistically significant difference in the likelihood of identifying a correct final claim between the male first-time users and the male second-time users (26% vs. 28%), $z = .51$, $p = .610$. Different from the bee scenario, female second-time users performed better in CFC than male second-time users (38% vs. 28%), $z = -2.21$, $p = .027$.

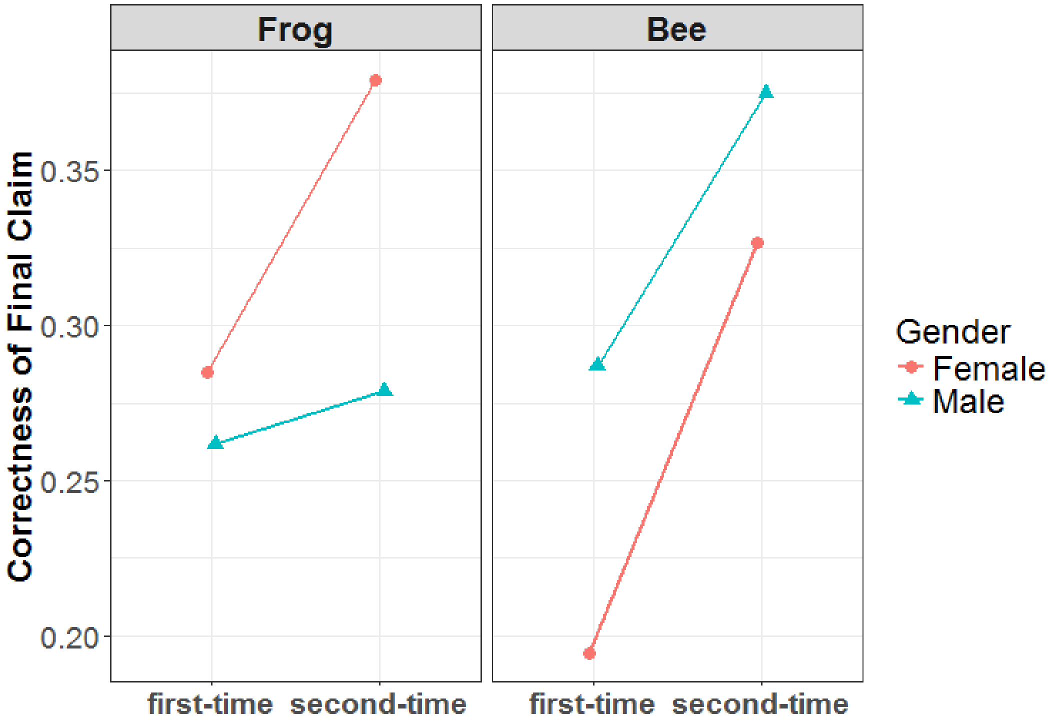


Figure 7. Marginal means plots of CFC score for the frog scenario (left) and the bee scenario (right).

Identifying Supporting Evidence (ISE)

The second measure of science inquiry skill, Identifying Supporting Evidence (ISE), evaluates student ability in supporting final conclusions with evidence even when they have not identified the correct final claim. Most of the evidence in the frog scenario was consistent with parasites being the cause of the 6-legged frog and in the bee scenario with radiation being the cause of the death of bee population. Three of the four other incorrect claims in each scenario had at least some evidence consistent with the claim, but the evidence against them was stronger. There was no evidence in support of the aliens claim in either scenario. While the non-causal data was strong enough to show that these claims were unlikely to be the cause, students were given partial credit if they provided supporting evidence for these claims.

The measure of students' ability in identifying supporting evidence was operationalized through assigning points based on whether the evidence they provided supported the claim they made. At the end of the assessment in each scenario, students were first asked to identify data that was supporting evidence of their final claims based on what they had collected in their backpack and the results of laboratory tests they had conducted. They were then allowed to choose from all possible data in the environment, to give students who may not have collected all the necessary data a chance to support their claim with evidence. Students indicated for each piece of data whether or not it was evidence for their claim, as well as identifying which farm was causing the problem. A rubric was developed by a content expert and researcher on the Virtual Performance Assessment project based on how the data supported the claim. Most evidence and the final claim were scored on a scale of 0–3 points. Student success in selecting supporting evidence was aggregated into a single composite outcome measure that ranges from

0–100, by averaging across the use of each piece of evidence. The data logging system kept track of all data students submitted and a back-end scoring engine automated a final score.

Therefore, even if students were unsuccessful in identifying the correct final claim, partial credit would be awarded to them for the quality and quantity of the causal evidence they identified in support of their claim from the non-causal data and results. Using this metric enabled researchers to better distinguish students who understood the principles of scientific inquiry (but were led astray by distractor information) from those who were completely unsuccessful at demonstrating science inquiry skills. The mean ISE score for the frog scenario was statistically higher than that for bee scenario (50, *SD* = 23 vs. 46, *SD* = 21), *t*(2390) = 5.76, *p* < .001[2].

A three-level regression was conducted to compare student ability in identifying supporting evidence between the first-time users and the second-time users in each scenario. The three-level regression models with students in each scenario nested within classes, and classes nested within teachers were fit to explore whether systematic differences exist between the groups on the continuous measure of student science inquiry performance — ISE score. In these models, the dependent variable is ISE score, while experience with VPA (first-time user vs. second-time user), gender (male vs. female), and interaction between experience and gender serve as the student-level predictor variables in each model.

[2] The statistical test comparing participants' performance between the two scenarios was based on their performance on the first assessment they used. Among the 2,429 participants, 1,232 students were randomly assigned to complete the frog scenario first, and 1,197 students were assigned to the bee scenario as their first assessment. A three-level regression comparing the performance of these two groups in their first assessment indicated that students showed a significantly higher supporting evidence score on average in the frog scenario than the performance of students in the bee scenario, *t*(2390) = 5.76, *p* < .001.

Results

Three-level regression results suggested that the interaction between experience and gender on ISE score was statistically significant in both the bee scenario (t(1977[3]) = −2.16, p = .031) and the frog scenario (t(1935) = −2.22, p = .026). Marginal means plots are presented in Figure 8. Among the females, students who had previously completed the other scenario achieved a significantly higher average ISE score than the first-time users in both scenarios (bee: *Ms* = 51 and 44, t(1054) = 4.44, p < .001; frog: *Ms* = 53 and 51, t(1049) = 2.21, p = .028). However, student performance in identifying supporting evidence for the male first-time users was not statistically significantly different from the male second-time users (bee: *Ms* = 44 and 47, t(937) = 1.31, p = .190; frog: *Ms* = 48 and 47, t(910) = −0.67, p = .503).

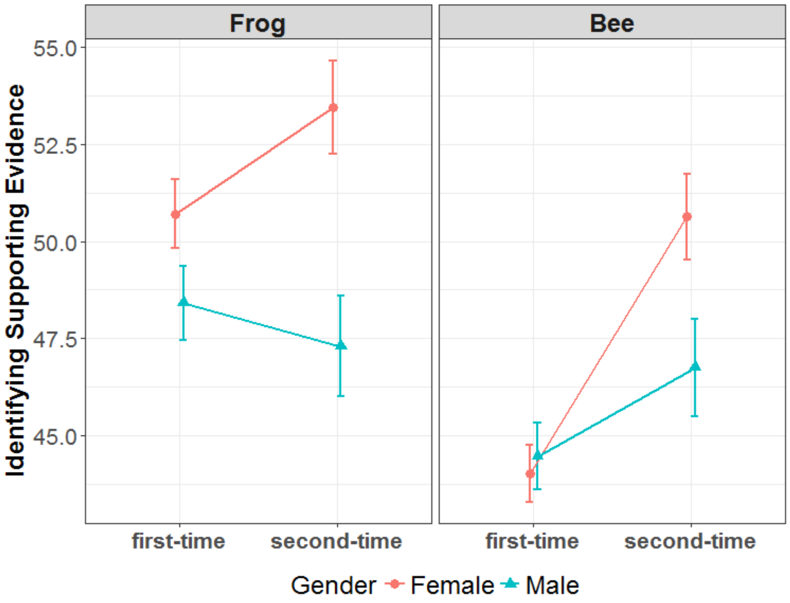


Figure 8. Plotted means for ISE score, on which a significant interaction between experience with VPA and gender was obtained for the frog scenario (left) and the bee scenario (right). Error bars represent standard errors.

[3] Satterthwaite approximations are applied to degrees of freedom for t-tests.

Control of Variables Strategy (CVS)

Two other measures of science inquiry performance evaluate students’ use of the control of variables strategy (CVS), which is a crucial component of science inquiry and scientific reasoning (Z. Chen & Klahr, 1999). In order to succeed in VPA, students are supposed to apply the control of variables strategy and conduct unconfounded experiments and observations to test each potential hypothesis during the scientific investigation. For example, in order to test the hypothesis of parasites and subsequently identify it as the causal factor leading to the frog mutation, students need to construct controlled comparisons – such as the comparison of the blood test results on the six-legged frog versus the red frog, comparison by inspecting the six-legged frog versus the red frog, and comparison of results from water tests on the control water versus the red water. Evidence obtained from these controlled comparisons, together with information students read from the research kiosk on parasites, would enable them to conclude that parasites is the causal factor resulting in the frog mutation. Similarly, evidence gathered from controlled comparisons between the six-legged frog versus the yellow frog and the control water versus the yellow water, together with the research information on pesticides, are necessary for students to test and exclude the hypothesis of pesticides. The hypothesis of pesticides could be excluded after the CVS process since the evidence/symptoms were inconsistent between the six-legged frog and the yellow farm, or with the research information on pesticides.

Information obtained from a total of 22 pairs of controlled comparisons or kiosk reading actions was identified as evidence necessary for students to use the control of variables strategy to test all five potential hypotheses (called CVS evidence, see APPENDIX for a complete list) in each scenario. Four pieces of the CVS evidence in each scenario, referred to as CVS CFC evidence, were evidence necessary for students to apply CVS to test the correct causal factor in each scenario (i.e., parasites in the frog scenario and radiation in the bee scenario). As such, the CVS CFC evidence is a subset of the CVS evidence in each scenario. In the frog scenario, in addition to the CVS CFC evidence, CVS evidence also includes evidence needed for CVS to test the other incorrect hypotheses (i.e., pesticides, pollution, radiation, and alien). Similarly, the CVS evidence in the bee scenario includes evidence needed for CVS use to test the hypotheses of parasites, pesticides, pollution and alien in addition to the CVS CFC evidence.

Two measures were developed to evaluate students' use of CVS during the science inquiry process based on the availability of the evidence. First, *CVS-data score* is the number of pieces of CVS evidence that were collected by the student. Additionally, *CVS CFC-data score* is the number of pieces of CVS CFC evidence that were collected by students during the science inquiry.

It is worth pointing out that these measures do not necessarily mean that the students actually followed CVS to test the hypotheses, which is impossible to be obtained as no direct information on students' mental process (e.g., think-aloud data) was collected in this study. However, the CVS measures suggest that the students executed behaviors that enabled them to collect the evidence necessary for CVS use, which is an indicator of the students' understanding of CVS use in science inquiry and problem-solving process. In previous studies on CVS, students were taught to run controlled trials where only one variable of interest is changed while all other extraneous variables are kept constant, in order to test the effects of the independent variable on a dependent variable (Z. Chen & Klahr, 1999; Klahr & Nigam, 2004). CVS is therefore typically evaluated by counting the number of trials (either consecutive or inconsecutive) students run that are controlled (Gobert & Koedinger, 2011; Kuhn & Pease, 2008; McElhaney & Linn, 2010). In more recent work, researchers built machine-learned models to estimate and predict the acquisition of CVS in computer-based learning environments (Gobert et al., 2012; Sao Pedro et al., 2013; Sao Pedro et al., 2014). Unlike the tasks and environments specifically designed to teach and assess CVS, VPA is more open-ended and its tasks are more complicated. Students need to collect data of various formats (e.g., field observation, laboratory test results, research kiosk information), conduct multiple controlled experiments to test each of the five potential causal factors, and synthesize the information to make inferences and final decisions. Due to the open-endedness of VPA, the same behaviors might represent different types of learners. For example, a student might bring the six-legged frog to the lab and run a blood test on it. After some interval (during which the student explored other farms), the student visited the red farm, brought the red frog to the lab, ran a blood test on it, and compared the blood test results on the red frog and the six-legged frog, with a plan to conduct an unconfounded comparison. However, another student might execute the same actions without comparing and constructing connections between the results of the two objects even though they were available to them. Other students might simply collect all possible information and run all possible experiments without making controlled comparisons. This is a limitation of considering inquiry within VPA solely in terms of CVS.

Results

In the bee scenario, there was a significant interaction between experience with VPA and gender on CVS-data score (*t*(1968) = −2.79, *p* = .005) and on CVS CFC-data score (*t*(1970) = −2.73, *p* = .006) (see Figure 9). For males, first-time users collected significantly more evidences such as results from controlled experiments that are necessary for CVS use to test the correct final claim of radiation (*Ms* = 2.28 and 2.08, *t*(923) = −2.14, *p* = .033) or to test all potential hypotheses (*Ms* = 11.16 and 10.04, *t*(921) = −2.40, *p* = .016) than second-time users. Among females, first-time users and second-time users did not collect a significantly different number of CVS CFC evidences (*Ms* = 1.97 and 2.15, *t*(1058) = 1.33, *p* = .183) or CVS evidences (*Ms* = 9.34 and 10.16, *t*(1055) = 1.22, *p* = .224) on average than the second-time users. As a result, the original advantage for males in adopting CVS to test hypotheses over females disappeared as students used VPA for the second time.

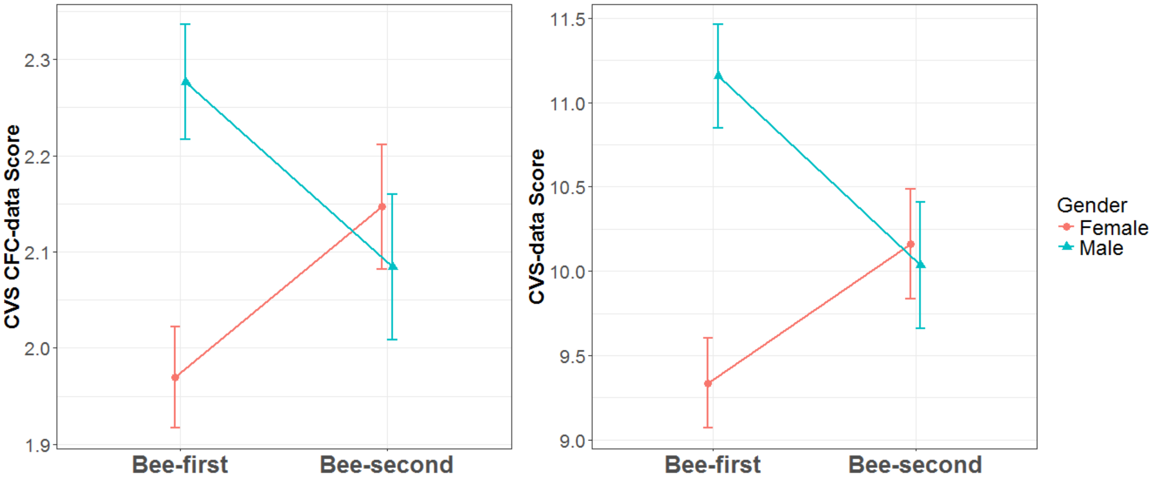


Figure 9. Marginal means plots for CVS CFC-data score (left) and CVS-data score (right) in the bee scenario.

In the frog scenario, first-time users and second-time users did not show significantly different CVS-data scores (*Ms* = 10.52 and 10.57, 50% and 50%, *t*(1958) = .17, *p* = .864) or CVS CFC-data scores (*Ms* = 2.20 and 2.20, 55% and 55%, *t*(1962) = .43, *p* = .667). Males generally conducted significantly more controlled comparisons to test the correct causal hypothesis (*Ms* = 2.31 and 2.11, 58% and 53%, *t*(1968) = 4.72, *p* < .001) and all potential hypotheses

(*Ms* = 11.29 and 9.83, 53% and 48%, *t*(1964) = 5.10, *p* < .001). No significant interactions were obtained for the CVS measures (see Figure 10).

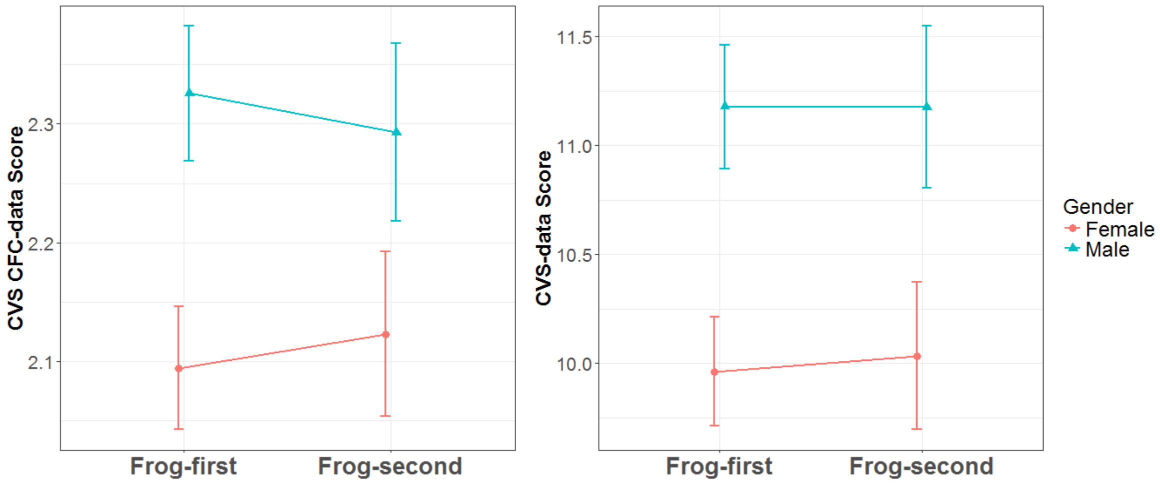


Figure 10. Marginal means plots for CVS CFC-data score (left) and CVS-data score (right) in the frog scenario.

Discussion

Analysis 1 revealed mixed results on the development of science inquiry skills for male and female students in the two scenarios. Among females, students with previous experience with the other VPA scenario outperformed the students who used VPA for the first time on science inquiry tasks such as identifying a correct final claim and supporting the final claim with evidence in both the frog scenario and the bee scenario. Considering that self-regulated learning has long been found to be positively associated with learning and performance, and that self-regulating one's performance is a critical component of SRL, it is reasonable to hypothesize that female students' SRL skills also developed as they transitioned from one scenario to another in VPA.

However, marked gender difference was observed in the development of science inquiry skills. Among male learners, the second-time users only showed better performance on identifying a correct final claim than the first-time users in the bee scenario. No significant difference between male first-time users and male second-time users was obtained for CFC performance in the frog scenario or ISE performance in either scenario. In addition, male second-time users executed fewer behaviors that enabled them to collect CVS evidence and CVS CFC evidence than male first-time users in the bee scenario.

In contrast, females' performance improved over time. As a result, despite the fact that there was no gender difference in CFC performance in the frog scenario and ISE performance in both scenarios for first-time users, a gender difference favoring females emerged as students used VPA for the second time. In addition, male first-time users showed advantages in CFC performance (in the bee scenario) and CVS-data and CVS CFC-data scores (in both scenarios) compared to female first-time users, which is consistent with previous literature showing the male advantages in science achievement and attitudes towards science (Curran & Kellogg, 2016; Halpern, 2004; Mullis et al., 2000; Neuschmidt et al., 2008). However, this difference in the bee scenario narrowed down or disappeared as students used VPA for the second time.

It is still unclear why female students' inquiry skills appeared to improve over time within VPA while male students did not seem to improve their science inquiry performance except for the CFC score in bee scenario. One possible explanation was that the males might be more vulnerable to novelty effect than female students. Clark (1983) argued that a novelty effect occurs when new computer programs are introduced. In those cases, the novel computer programs initially attract student attention, leading to increased efforts invested, persistence, motivation, and achievement gains. Previous studies (e.g., Cuban, 1986; Keller, 1999; Schofield,

1995) indicated that students showed greater initial enthusiasm and motivation in classrooms when novel educational technologies were introduced. This enthusiasm gradually diminished as students became more familiar with the technologies and the initial novelty effect wore off (Cuban, 1986; Keller, 1999). Therefore, as these students were first introduced to the novel 3D virtual environment, the initial attraction and attention led to a higher level of interest and effort invested in the tasks, which tended to decline when students became relatively experienced and familiar with the environment. This might explain the lack of improvement in performance on science inquiry tasks for males across scenarios, although it is still not clear why female students did not show the same pattern.

Another possibility is that gender difference in conscientiousness leads to the gender difference in development of science inquiry skills. Individuals who are conscientious tend to be careful, disciplined, thorough, persevering, and motivated to achieve goals (Costa & McCrae, 1992). Conscientiousness is positively related to academic achievement (Kappe & van der Flier, 2010; Poropat, 2009). Previous research has revealed gender differences in conscientiousness in favor of females over males (Feingold, 1994; Reddington et al., 2015). Therefore, it is likely that female students who were usually more conscientious tended to persevere despite their enthusiasm towards using VPA declined and thereby developed science inquiry skills over the use of VPA. On the other hand, it is also possible that male students were overconfident about their performance as they used the system for the second time and did not invest as much effort as they used the system for the second time.

Considering the difference in the development of science inquiry skills for males and females, it is worth further exploring whether students' self-regulatory behaviors and strategies evolved as they used VPA or not, and how students of different gender develop SRL skills differently. I attempt to further explore these issues in analysis 2 and analysis 3, studying whether better understanding student behaviors also increases understanding of the gender difference in science inquiry performance.

CHAPTER VI.

ANALYSIS 2: BEHAVIOR PATTERN ANALYSIS

Analysis 1 examined the development of science inquiry expertise, which is positively related to SRL performance, within each VPA scenario. In analysis two, I aim to go beyond simply looking at whether previous experience in VPA and gender are related to student inquiry performance, and instead delve into whether the second-time users used VPA differently than the first-time users and whether male students showed different behaviors than female students. Exploration of behaviors will enable better understanding of how students' self-regulatory behaviors and strategies develop over time.

Sequential Pattern Mining

This analysis investigates patterns in behavior by applying sequential pattern mining to identify and compare the frequent sequential patterns of student actions between the two groups (either first-time versus second-time user, or female versus male). Sequential pattern mining is a popular data mining technique that automatically identifies frequent temporal patterns of actions in data (Agrawal & Srikant, 1995). It can be used to detect differentially frequent behavioral patterns of different groups of students (Kinnebrew et al., 2013). An example sequential pattern in VPA is that students who talked to the NPC in a farm tended to pick up and inspect objects in the farm as a next step (i.e., *talk* → *inspect*). In sequential pattern mining, the most frequent sequential patterns are typically selected within the data set on the basis of two frequency metrics – support and confidence (Agrawal & Srikant, 1995). The support of a sequential pattern $A \rightarrow B$ corresponds to the percentage of transactions that contain the sequence $A \rightarrow B$. The confidence of the pattern $A \rightarrow B$ can be viewed as the conditional probability and is defined as the percentage of transactions that meet the pattern $A \rightarrow B$, divided by the percentage of transactions that contain *A* as the first element in the sequence. Short sequences with high confidence and support are combined into longer sequences, which are in turn checked for acceptably high confidence and support. Additional “interestingness” measures are further calculated to discover novel, interesting, and sometimes unexpected sequences of behaviors (Bazaldua, Baker, & San Pedro, 2014; Merceron & Yacef, 2008).

Data Pre-Processing

Prior to performing sequential pattern mining, detailed raw interaction log data were transformed into more abstract sequences. This involved three steps. First, a set of actions related to science inquiry were identified from the log files, including picking up and inspecting objects (e.g., frogs, tadpoles, bees, larvae, water sample, nectar sample) within VPA (*inspect*), saving objects to the backpack (*save*), discarding objects (*discard*), talking with NPCs (*talk*), opening and reading informational pages at the research kiosks (*read*), running laboratory tests (e.g., blood/protein test, water/nectar sample test, genetic test) (*test*), reviewing and looking at test results (*look*), accessing the notepad to take or review notes (*note*), opening the help page to review tasks (*help*), starting to answer final questions (*start final questions*), and submitting a final claim (*final claim*). Some actions that were irrelevant to the inquiry process, such as selecting an avatar and entering/exiting a specific area, were filtered out from the raw interaction data. Second, as in Kinnebrew et al. (2013), repeated actions that occurred more than once in succession were distinguished from a single action and were labeled as the “action” followed by the “-MULT” suffix, in order to distinguish brief behaviors from more intensive patterns of behavior. Last, the actions were represented as sequences of actions for each student in each group.

Two-Action Sequential Patterns

Simple two-action sequential patterns were identified using the arules package (Hahsler, Gruen, & Hornik, 2005) within the statistical software program R. Two-action sequential patterns are behavioral sequences that are comprised of two actions, such as viewing experiment results followed by reading research page at the kiosk (i.e., *look* → *read*). Arules was used to determine the most frequent two-action sequences by selecting the temporal associations of one specific action and a subsequent action with the highest support and confidence. In this analysis, sequential patterns of consecutive actions were selected with the cut-off thresholds of support = 0.0005 and confidence = 0.05. A total of 111 short sequential patterns (length = 2) were identified that met the minimum support and confidence constraints in the frog scenario and a total of 113 patterns were identified in the bee scenario. These patterns were similar across the four conditions, and most had support and confidence considerably higher than the threshold. They were then ordered according to their *Jaccard* similarity coefficient to find interesting sequential patterns. *Jaccard* was chosen as a measure of the pattern's interestingness (Merceron & Yacef, 2008) because this metric was found to be the most highly correlated with human judgments of whether a finding is interesting (Bazaldua et al., 2014). According to Bazaldua et al. (2014), lower *Jaccard* measures indicated higher interestingness for human raters, among rules already identified to have acceptably high support and confidence. Among the action sequences with high interestingness (i.e., low *Jaccard*), I then identified a subset of sequential patterns that I believe corresponded to self-regulatory processes and strategies, and compared their frequency between the two groups.

Differential Pattern Mining

To facilitate the comparison of the frequency metrics between the first-time users and the second-time users, the support and confidence for each pattern were calculated separately for each student. Three-level regression tests were then conducted, controlling for multiple comparisons with Benjamini & Hochberg (1995) corrections, to compare the metric values between the groups in each scenario. Similar to analysis 1, the dependent variable is each frequency measure, while experience with VPA, gender, and the interaction between experience and gender serve as the student-level predictor variables in each three-level model. Table 3 presents the comparison of the descriptive statistics of the support and confidence metrics of frequent sequential patterns identified as reflective of self-regulatory processes and strategies. Multilevel regression results are reported in Table 4.

Table 3

Comparisons of descriptive statistics of the support and confidence of frequent sequential patterns related to self-regulatory processes and strategies across conditions in each scenario

Pattern	**Scenario**	**Metric**	**F-1**	**F-2**	**M-1**	**M-2**
help → note	Frog	Supp	.0008 (.0036)	.0002 (.0020)	.0007 (.0034)	.0003 (.0027)
		Conf	.15 (.34)	.12 (.32)	.12 (.31)	.15 (.34)
	Bee	Supp	.0005 (.0028)	.0003 (.0024)	.0009 (.0038)	.0004 (.0029)
		Conf	.09 (.27)	.16 (.37)	.16 (.35)	.18 (.36)
read → note-MULT	Frog	Supp	.0086 (.0173)	.0135 (.0229)	.0036 (.0102)	.0056 (.0143)
		Conf	.31 (.37)	.40 (.40)	.19 (.32)	.24 (.37)
	Bee	Supp	.0084 (.0183)	.0130 (.0223)	.0042 (.0119)	.0046 (.0129)
		Conf	.33 (.39)	.38 (.4)	.20 (.33)	.19 (.31)
read-MULT → note-MULT	Frog	Supp	.0034 (.0088)	.0051 (.0115)	.0022 (.0063)	.0028 (.0083)
		Conf	.12 (.26)	.18 (.33)	.09 (.23)	.11 (.25)
	Bee	Supp	.0034 (.0079)	.0044 (.0101)	.0020 (.0064)	.0028 (.0092)
		Conf	.14 (.27)	.16 (.29)	.08 (.21)	.09 (.23)
test → note-MULT	Frog	Supp	.0038 (.0096)	.0038 (.0097)	.0024 (.0071)	.0022 (.0085)
		Conf	.15 (.27)	.20 (.34)	.09 (.21)	.08 (.22)
	Bee	Supp	.0031 (.0092)	.0039 (.0106)	.0017 (.0059)	.0020 (.0079)
		Conf	.11 (.23)	.19 (.32)	.07 (.19)	.11 (.27)
look → note	Frog	Supp	.0021 (.0072)	.0023 (.0076)	.0017 (.0054)	.0021 (.0076)

		Conf	.09 (.23)	.15 (.30)	.07 (.20)	.09 (.23)
	Bee	Supp	.0013 (.0053)	.0022 (.0089)	.0011 (.0044)	.0015 (.0063)
		Conf	.06 (.18)	.11 (.23)	.06 (.18)	.07 (.19)
look → read-MULT	Frog	Supp	.0011 (.0045)	.0015 (.0063)	.0013 (.0055)	.0019 (.0068)
		Conf	.04 (.13)	.06 (.18)	.04 (.15)	.09 (.22)
	Bee	Supp	.0006 (.0036)	.0011 (.0049)	.0013 (.0047)	.0016 (.0070)
		Conf	.03 (.13)	.08 (.22)	.06 (.18)	.08 (.22)
look-MULT → read	Frog	Supp	.0009 (.0040)	.0012 (.0050)	.0007 (.0038)	.0011 (.0048)
		Conf	.05 (.16)	.06 (.16)	.03 (.15)	.05 (.16)
	Bee	Supp	.0006 (.0034)	.0011 (.0048)	.0008 (.0034)	.0008 (.0048)
		Conf	.03 (.11)	.07 (.19)	.03 (.13)	.04 (.15)
note → final claim	Frog	Supp	.0028 (.0079)	.0040 (.0098)	.0008 (.0038)	.0020 (.0077)
		Conf	.06 (.19)	.09 (.21)	.03 (.12)	.05 (.17)
	Bee	Supp	.0017 (.0058)	.0027 (.0072)	.0013 (.0050)	.0022 (.0070)
		Conf	.04 (.15)	.07 (.17)	.04 (.15)	.08 (.22)
note-MULT → final claim	Frog	Supp	.0023 (.0070)	.0040 (.0096)	.0011 (.0046)	.0014 (.0059)
		Conf	.07 (.17)	.09 (.16)	.06 (.17)	.05 (.14)
	Bee	Supp	.0020 (.0064)	.0029 (.0084)	.0009 (.0040)	.0010 (.0049)
		Conf	.06 (.15)	.07 (.14)	.04 (.14)	.05 (.12)
read-MULT → final claim	Frog	Supp	.0044 (.0137)	.0046 (.0119)	.0025 (.0087)	.0025 (.0082)
		Conf	.12 (.25)	.12 (.26)	.08 (.21)	.06 (.16)
	Bee	Supp	.0042 (.0107)	.0048 (.014)	.0022 (.0081)	.0018 (.0071)
		Conf	.14 (.28)	.13 (.27)	.06 (.17)	.05 (.18)
final claim → note	Frog	Supp	.0017 (.0058)	.0021 (.0069)	.0006 (.0030)	.0017 (.0067)
		Conf	.25 (.43)	.24 (.43)	.16 (.36)	.32 (.47)
	Bee	Supp	.0008 (.0040)	.0018 (.0055)	.0006 (.0035)	.0011 (.0047)
		Conf	.08 (.21)	.13 (.23)	.08 (.18)	.14 (.26)
final claim → note-MULT	Frog	Supp	.0009 (.0038)	.0021 (.0061)	.0005 (.0028)	.0007 (.0036)
		Conf	.16 (.37)	.28 (.45)	.16 (.36)	.16 (.37)
	Bee	Supp	.0011 (.0041)	.0014 (.0051)	.0003 (.0022)	.0004 (.0027)
		Conf	.10 (.22)	.11 (.22)	.05 (.15)	.05 (.17)

Note. Average support/confidence values with standard deviation in parentheses are reported for female first-time users (F-1), female second-time users (F-2), male first-time users (M-1), and male second-time users (M-2) by scenario.

Results

Understanding Task Definition

One behavior pattern with high interestingness, confidence, and support was *help → note*, which I postulate to be related to understanding task definition in the SRL cycle. Note that students have access to the help button throughout their exploration process in VPA. A window would pop up to remind students of the ultimate goals they need to achieve when the button is clicked on. This pattern showed marginally significantly higher support for the first-time users than the second-time users in the frog scenario (frog: *Ms* = .0007 and .0003, *t*(1986) = −2.60, *p* = .009, adjusted *α* = .007; bee: *Ms* = .0007 and .0003, *t*(2017) = −1.42, *p* = .156, adjusted *α* = .019). Despite similar means between conditions, and similar standard deviations as well, the two scenarios achieved different degrees of statistical significance, possibly due to different patterns in the variables being controlled for (gender, interaction, class, and teacher). I speculate that the difference in the frog scenario was due to the novelty effect (Jiang et al., 2015). That is, due to the increased attention and enthusiasm as the novel VPA environment was first introduced to classrooms, students tended to take notes of the information they just read about what they were supposed to do in VPA more frequently than the second-time users, whereas previous experience in the other VPA scenario had familiarized second-time users with their tasks and they did not access the help page and take notes of it as often since the task information could be kept in mind.

Table 4

Relationship between experience with VPA and gender towards the support and confidence of frequent sequential patterns related to SRL in each scenario

			Experience		**Gender**		**Experience × Gender**	
Scenario	**Pattern**	**Metric**	**B (SE B)**	**t**	**B (SE B)**	**t**	**B (SE B)**	**t**
Frog	help → note	Supp	−.0005 (.0002)	−2.60	−.0001 (.0002)	−.45	.0002 (.0003)	.66
		Conf	−.03 (.05)	−.56	−.04 (.04)	−1.01	.06 (.08)	.76
	read → note-MULT	Supp	.0052 (.0010)	5.12 *	−.0045 (.0009)	−4.85 *	−.0033 (.0015)	−2.19
		Conf	.1 (.03)	3.34 *	−.11 (.03)	−3.77 *	−.05 (.05)	−1.01
	read-MULT → note-MULT	Supp	.0016 (.0005)	2.90 *	−.0012 (.0005)	−2.48	−.0011 (.0008)	−1.41
		Conf	.06 (.02)	3.25 *	−.04 (.02)	−2.14	−.04 (.03)	−1.49
	test → note-MULT	Supp	.00004 (.0006)	.07	−.0013 (.0005)	−2.61	−.0004 (.0008)	−.53
		Conf	.05 (.02)	2.26	−.06 (.02)	−3.15 *	−.06 (.03)	−1.74
	look → note	Supp	.0002 (.0004)	.54	−.0004 (.0004)	−.95	.0001 (.0006)	.19
		Conf	.05 (.02)	2.48	−.02 (.02)	−1.11	−.03 (.03)	−1.00

	look → read-MULT	Supp	.0004 (.0004)	1.01	.0003 (.0003)	.88	.0002 (.0005)	.41
		Conf	.02 (.01)	1.22	.002 (.01)	.13	.02 (.02)	1.07
	look-MULT → read	Supp	.0003 (.0003)	1.00	−.0001 (.0002)	−.40	.0001 (.0004)	.18
		Conf	.01 (.01)	.55	−.01 (.01)	−1.08	.01 (.02)	.43
	note → final claim	Supp	.0011 (.0005)	2.43	−.0019 (.0004)	−4.58 *	−.00002 (.0007)	−.03
		Conf	.03 (.01)	2.05	−.04 (.01)	−3.26 *	.0002 (.02)	.01
	note-MULT → final claim	Supp	.0017 (.0004)	3.82 *	−.0012 (.0004)	−3.11 *	−.0014 (.0006)	−2.24
		Conf	.02 (.01)	1.17	−.01 (.01)	−.81	−.02 (.02)	−.87
	read-MULT → final claim	Supp	.0003 (.0007)	.39	−.0016 (.0006)	−2.54	−.0003 (.0010)	−.34
		Conf	.003 (.02)	.16	−.05 (.02)	−2.98 *	−.02 (.03)	−.76
	final claim → note	Supp	.0004 (.0004)	1.00	−.0012 (.0003)	−3.61 *	.0008 (.0005)	1.52
		Conf	−.01 (.04)	−.33	−.09 (.05)	−1.99	.17 (.07)	2.36
	final claim → note-MULT	Supp	.0012 (.0003)	4.47 *	−.0004 (.0002)	−1.65	−.0011 (.0004)	−2.79 *
		Conf	.12 (.04)	3.07 *	−.01 (.04)	−.13	−.12 (.07)	−1.76
Bee	help → note	Supp	−.0003 (.0002)	−1.42	.0004 (.0002)	2.03	−.0002 (.0003)	−.57
		Conf	.07 (.05)	1.24	.07 (.03)	1.97	−.05 (.08)	−.58
	read → note-MULT	Supp	.0040 (.0010)	3.87 *	−.0039 (.001)	−4.04 *	−.0040 (.0015)	−2.69
		Conf	.04 (.03)	1.45	−.13 (.03)	−4.47 *	−.06 (.05)	−1.18
	read-MULT → note-MULT	Supp	.0008 (.0005)	1.65	−.0014 (.0005)	−2.84 *	−.00004 (.0008)	−.06
		Conf	.02 (.02)	.93	−.06 (.02)	−3.44 *	−.004 (.03)	−.14
	test → note-MULT	Supp	.0007 (.0005)	1.40	−.0014 (.0005)	−2.81 *	−.0005 (.0008)	−.59
		Conf	.07 (.02)	3.42 *	−.04 (.02)	−2.32	−.03 (.03)	−1.03
	look → note	Supp	.0008 (.0004)	2.13	−.0002 (.0004)	−.44	−.0005 (.0006)	−.81
		Conf	.04 (.02)	2.36	−.01 (.02)	−.33	−.03 (.03)	−1.05
	look → read-MULT	Supp	.0005 (.0003)	1.73	.0007 (.0003)	2.53	−.0002 (.0005)	−.52
		Conf	.05 (.02)	2.85 *	.03 (.01)	2.33	−.03 (.02)	−1.36
	look-MULT → read	Supp	.0004 (.0002)	1.70	.0002 (.0002)	.94	−.0004 (.0004)	−1.16
		Conf	.04 (.01)	3.17 *	.01 (.01)	.79	−.03 (.02)	−1.88
	note → final claim	Supp	.0010 (.0004)	2.50	−.0004 (.0004)	−1.16	−.0001 (.0006)	−.19
		Conf	.02 (.01)	1.85	−.0005 (.01)	−.04	.02 (.02)	.90
	note-MULT → final claim	Supp	.0009 (.0004)	2.33	−.0011 (.0004)	−3.23 *	−.0007 (.0006)	−1.29
		Conf	.01 (.01)	.53	−.02 (.01)	−1.59	.0007 (.02)	.04
	read-MULT → final claim	Supp	.0006 (.0006)	.93	−.0018 (.0006)	−2.93 *	−.0010 (.0009)	−1.11
		Conf	−.01 (.02)	−.51	−.08 (.02)	−4.64 *	−.0002 (.03)	−.01
	final claim → note	Supp	.001 (.0003)	3.51 *	−.0003 (.0003)	−1.00	−.0005 (.0004)	−1.19
		Conf	.06 (.02)	2.59	−.01 (.02)	−.28	.005 (.04)	.13
	final claim → note-MULT	Supp	.0004 (.0002)	1.60	−.0007 (.0002)	−3.43 *	−.0003 (.0003)	−.83
		Conf	.004 (.02)	.20	−.06 (.02)	−2.63	.005 (.03)	.14

Note. Coefficient of the predictor (B), standard error associated with the coefficient (SE B), and t-statistics (*t*) are reported for each term (experience, gender, and experience × gender). Statistically significant results after Benjamini and Hochberg's control are marked with *.

Enacting Study Tactics and Strategies

Interesting sequential patterns were also found for the application of note-taking and note-reviewing strategies after reading research pages or viewing laboratory test results. In the frog scenario, the pattern *read* → *note-MULT* had significantly higher support and confidence for second-time users than first-time users (support: *Ms* = .0098 and .0063, *t*(1964) = 5.12, *p* < .001, adjusted *α* < .001; confidence: *Ms* = .33 and .26, *t*(1040) = 3.34, *p* < .001, adjusted *α* = .003). In the bee scenario, the interaction between experience and gender on the support of this pattern was marginally significant, *t*(1995) = −2.69, *p* = .007, adjusted *α* = .006. Results of simple effects analysis indicated that the support was significantly higher for female second-time users than female first-time users (*Ms* = .0130 and .0084, *t*(1071) = 3.31, *p* < .001), whereas it was not statistically different between male first-time users and male second-time users (*Ms* = .0042 and .0046, *t*(932) = −0.03, *p* = .976). The support and confidence of this pattern was significantly higher for female students than male students in both scenarios (frog: support: *Ms* = .0105 and .0044, *t*(1974) = −4.85, *p* < .001, adjusted *α* < .001; confidence: *Ms* = .35 and .21, *t*(1031) = −3.77, *p* < .001, adjusted *α* = .002; bee: support: *Ms* = .0104 and .0043, *t*(1998) = −4.04, *p* < .001, adjusted *α* = .001; confidence: *Ms* = .35 and .20, *t*(1031) = −4.47, *p* < .001, adjusted *α* < .001).

A similar pattern *read-MULT* → *note-MULT* also had significantly higher support and confidence for the second-time users than the first-time users in the frog scenario (support: *Ms* = .0040 and .0029, *t*(1977) = 2.90, *p* = .004, adjusted *α* = .005; confidence: *Ms* = .15 and .11, *t*(1384) = 3.25, *p* = .001, adjusted *α* = .004). In the bee scenario, the second-time users and the first-time users did not differ significantly in the frequency metrics of this pattern (supp: *Ms* = .0037 and .0027, *t*(2016) = 1.65, *p* = .099, adjusted *α* = .016; conf: *Ms* = .13 and .11, *t*(1345) = .93, *p* = .354, adjusted *α* = .029), again despite comparable means and standard deviations between the two conditions. Consistent with previous results, this pattern showed marginally significantly higher support for female students than male students in the frog scenario (support: *Ms* = .0040 and .0024, *t*(1981) = −2.48, *p* = .013, adjusted *α* = .008; confidence: *Ms* = .15 and .10, *t*(1383) = −2.14, *p* = .032, adjusted *α* = .010), and significantly higher support and confidence for females than males in the bee scenario (support: *Ms* = .0038 and .0023, *t*(1999) = −2.84, *p* = .005, adjusted *α* = .006; confidence: *Ms* = .15 and .08, *t*(1300) = −3.44, *p* < .001, adjusted *α* = .003).

These results suggested that the second-time users and female students were more likely to open the notepad repeatedly to take or review notes after reading a research page (once or repeatedly). In other words, students who used VPA for the second time tended to show more utilization of the note-taking and note-reviewing strategies, suggesting their growing competence in enacting self-regulatory strategies. While taking notes of research information from the kiosk pages, students transferred the information presented in the kiosk to the digital notepad, which may have involved a generative process, strengthening student understanding of the domain-specific declarative information. Additionally, reviewing notes after reading kiosk pages may have helped students build connections between the notes previously recorded and the concepts they just read about. Furthermore, repeated access of the notepad most likely indicates more complete notes being encoded by users, further fostering student learning (Armbruster, 2009).

Similarly, second-time users were more likely to open the notepad to take or review notes multiple times after conducting laboratory experiments (*test* → *note-MULT)* or viewing test results (*look* → *note*). For the sequence *experiment* → *note-MULT*, the confidence for the second-time users was significantly higher than that for the first-time users in the bee scenario (frog: *Ms* = .14 and .12, *t*(1135) = 2.26, *p* = .024, adjusted *α* = .009; bee: *Ms* = .16 and .09, *t*(1074) = 3.42, *p* < .001, adjusted *α* = .003). The confidence for this pattern was also significantly higher for female students than male students in the frog scenario (frog: *Ms* = .17 and .09, *t*(1142) = −3.15, *p* = .002, adjusted *α* = .004; bee: *Ms* = .14 and .08, *t*(1063) = −2.32, *p* = .020, adjusted *α* = .009). The confidence value of the pattern *look* → *note* for the second-time users was marginally significantly higher than that for the first-time users in the frog scenario (frog: *Ms* = .12 and .08, *t*(1033) = 2.48, *p* = .013, adjusted *α* = .008; bee: *Ms* = .09 and .06, *t*(954) = 2.36, *p* = .019, adjusted *α* = .008). That is, second-time users and female students were more likely to access the notepad immediately after running laboratory tests or viewing the results of lab tests. These patterns appear to have represented effective learning strategies; opening the notebook in these contexts likely produced a second opportunity for students to understand the laboratory test results, elaborate on the results and make inferences, and connect them with other test results or the research information recorded in notepad. The information the students recorded or reviewed in the notepad on the laboratory tests also had the potential to help students with problem-solving and hypothesis generation in VPA.

Two other interesting sequential patterns corresponded to viewing laboratory test results (once or repeatedly), followed by reading informational pages (once or repeatedly) (i.e., *look* → *read-MULT, look-MULT* → *read*). The confidence for these patterns was significantly higher for the second-time users than the first-time users in the bee scenario (*look* → *read-MULT*: *Ms* = .08 and .05, *t*(953) = 2.85, *p* = .004, adjusted *α* = .006; *look-MULT* → *read: Ms* = .06 and .03, *t*(1027) = 3.17, *p* = .002, adjusted *α* = .004). Students who used VPA for the second time were more likely to read one or multiple research information page(s) on possible causal factors immediately after viewing the results of lab tests in the bee scenario. The higher relative frequency of reading research information might help second-time users interpret laboratory test results and facilitate the acquisition of domain-specific knowledge (Z. Chen & Klahr, 1999). This is consistent with results from previous studies on the development of expertise, where experts were found to be more opportunistic in using resources and exploit more available sources of information than novices (Gilhooly et al., 1997).

Learning strategies were not only applied during the inquiry process to assist with the understanding of instructional information or experiment results, but were also executed as part of the process of decision-making. Two relevant patterns, *note* → *final claim* and *note-MULT* → *final claim* indicated that students opened the notepad either once or repeatedly before submitting a final claim, possibly to review notes that had been taken and utilize the information for decision-making. The support of the pattern *note* → *final claim* was marginally significantly higher for the second-time users than the first-time users in both scenarios (frog: $Ms = .0030$ and .0019, $t(1981) = 2.43$, $p = .015$, adjusted $\alpha = .008$; bee: $Ms = .0025$ and .0015, $t(2018) = 2.50$, $p = .013$, adjusted $\alpha = .008$). The support of the pattern *note-MULT* → *final claim* was also significantly higher for the second-time users than the first-time users in the frog scenario ($Ms = .0028$ and .0018, $t(1980) = 3.82$, $p < .001$, adjusted $\alpha = .002$), but not in the bee scenario ($Ms = .0021$ and .0015, $t(2018) = 2.33$, $p = .020$, adjusted $\alpha = .009$). In other words, students who used VPA for the second time tended to make use of the notes that they took in the digital notepad, where the information from multiple sources had been recorded, to assist them with decision-making and the selection of the final claim.

Similarly, female students were more likely to access the notepad before submitting a final claim than male students. The support and confidence for the pattern *note* → *final claim* was significantly higher for the female students than their male counterparts in the frog scenario (frog: supp: *Ms* = .0032 and .0013, *t*(1903) = −4.58, *p* < .001, adjusted *α* < .001; conf: *Ms* = .07 and .04, *t*(1370) = −3.26, *p* = .001, adjusted *α* = .004; bee: supp: *Ms* = .0021 and .0016, *t*(1985) = −1.16, *p* = .247, adjusted *α* = .023; conf: *Ms* = .05 and .05, *t*(1301) = −.04, *p* = .969, adjusted *α* = .048). The support for the pattern *note-MULT* → *final claim* was also significantly higher for the female students than the male students in both scenarios (frog: *Ms* = .0030 and .0012, *t*(1931) = −3.11, *p* = .002, adjusted *α* = .004; bee: *Ms* = .0024 and .0009, *t*(1978) = −3.23, *p* = .001, adjusted *α* = .004). In addition to reviewing notes, female students were also more likely to read multiple kiosk pages before submitting the final claim (*read-MULT* → *final claim*), possibly to use the research information they read about to help them with the claim submission. This pattern showed a marginally significantly higher support and a significantly higher confidence for the females than the males in the frog scenario (supp: *Ms* = .0045 and .0025, *t*(1963) = −2.54, *p* = .011, adjusted *α* = .007; conf: *Ms* = .12 and .07, *t*(1364) = −2.98, *p* = .003, adjusted *α* = .005), and significantly higher support and confidence for the females than the males in the bee scenario (supp: *Ms* = .0044 and .0020, *t*(1977) = −2.93, *p* = .003, adjusted *α* = .005; conf: *Ms* = .14 and .06, *t*(1326) = −4.64, *p* < .001, adjusted *α* < .001).

Monitoring

The sequential patterns reflective of the monitoring process in SRL cycle involved making final claims (*final claim*) and accessing the digital notepad (*note*), such as *final claim* → *note* and *final claim* → *note-MULT*. These patterns indicated that students tended to open the notepad after submitting a final claim, perhaps to review the notes they had taken so far in order to self-evaluate and assess their final claim just submitted. Evaluating and reflecting on one's learning outcomes is an important part of the monitoring process in SRL frameworks (Winne & Hadwin, 2009; B. J. Zimmerman, 2000). These patterns appeared to have higher support for second-time users than first-time users. The pattern *final claim* → *note* showed significantly higher support and marginally significantly higher confidence for second-time users than first-time users in the bee scenario (support: *Ms* = .0015 and .0007, *t*(2016) = 3.51, *p* < .001, adjusted *α* = .003; confidence: *Ms* = .13 and .08, *t*(638) = 2.59, *p* = .010, adjusted *α* = .007). A similar pattern *final claim* → *note-MULT* also showed significantly higher confidence for second-time users in the frog scenario (confidence: *Ms* = .24 and .16, *t*(636) = 3.07, *p* = .002, adjusted *α* = .005). However, the interaction was statistically significant for the support of this pattern in the frog scenario, *t*(1979) = −2.79, *p* = .005, adjusted *α* = .006. Further simple effects tests indicated that the support was significantly higher for the second-time users than the first-time users among the females (*Ms* = .0021 and .0009, *t*(1063) = 3.85, *p* < .001), while it was not significantly different between male first-time users and male second-time users in the frog scenario (*Ms* = .0005 and .0007, *t*(911) = 0.55, *p* = .585). In addition, female students showed significantly higher support for the pattern *final claim* → *note* in the frog scenario (*Ms* = .0019 and .0010, *t*(1929) = −3.61, *p* < .001, adjusted *α* = .002) and significantly higher support for the pattern *final claim* → *note-MULT* in the bee scenario(*Ms* = .0012 and .0004, *t*(1951) = −3.43, *p* < .001, adjusted *α* = .003). This finding indicated that students who used VPA for the second time were more likely than students who used VPA for the first time to review their notes (both once or repeatedly), where the information they considered as important for decision-making was recorded, possibly to monitor their answers and reflect on previous steps (cf. Kuhn & Pease, 2008) after submitting a final claim. The notepad serves as a resource of combined information from various sources that students considered as important for problem-solving, and reviewing notes after submitting final claims could potentially help students check the claims and causal evidence they had just submitted.

Long Sequential Patterns

In addition to two-action patterns, a differential sequence mining technique developed by Kinnebrew and colleagues (2013) was utilized for identifying longer sequential patterns (length > 2) that occurred with significantly different frequencies between the groups. This methodology used sequence support (*s-support*) and instance support (*i-support*) as frequency measures. S-support is defined as the percentage of sequences in which the pattern occurs (Kinnebrew et al., 2013). It is different from the standard metric *support* in that *s-support* measures the percentage of students whose action sequence contained the specific pattern, regardless of the frequency of occurrence within each sequence for each student. The *i-support* corresponds to the number of times a given pattern occurs, without overlap, within a student's sequence of actions. A set of most frequent sequential patterns that met the s-support threshold was identified within each group by employing Kinnebrew et al.'s (2013) sequential pattern mining algorithm. The i-support value of each pre-identified pattern was then calculated for each sequence in each group, after which t-tests comparing the mean i-support between the groups were conducted and Benjamini and Hochberg's posthoc control method was applied to select significantly differentially frequent patterns. The mining of longer sequential patterns was conducted by using LASAT, a tool developed by Kinnebrew et al. (2013). One limitation of this analysis is that LASAT only allowed users to mine differentially frequent patterns between two groups. Therefore, I ran the differential pattern mining first between the first-time users and the second-time users, and then between male students and female students. Specifically, a cutoff s-support of 50% and a cutoff p-value of 0.05 were employed for selection and comparison of pattern usage between the first-time and second-time users, and a cutoff s-support of 10% and a cutoff p-value of 0.05 were employed for selection and comparison of pattern usage between males and females. Different cut-off values of s-support were applied in order to include an appropriate number of differentially frequent sequential patterns for each group. A cutoff s-support of 50% was initially tested to select and compare patterns between male and female students. However, only eight differentially frequent sequences were selected in the frog scenario and seven sequences were selected in the bee scenario. Therefore, the cutoff s-support value was adjusted to 10% accordingly.

Results

Twenty-five differentially frequent long patterns reached the minimum s-support and p-value in the frog scenario and 31 differentially frequent long patterns reached the minimum s-support and p-value in the bee scenario. Fourteen out of the 25 long patterns in the frog scenario and 16 out of the 31 long patterns in the bee scenario were common (i.e., met the 50% s-support threshold) for both groups, with relatively higher usage in the first-time user group. Eleven long patterns in the frog scenario and 15 in the bee scenario were frequently used only by the first-time users. All differentially frequent long patterns had a higher s-support and a significantly higher average i-support for the first-time users than the second-time users.

Table 5 presents the top five differentially frequent long patterns that were common to both groups and the top five that were frequently used only by the first-time users within each scenario. Most of these long sequential patterns entailed the repetition and combination of actions including inspecting objects, saving objects to backpack, discarding objects, and talking with NPCs.

Table 5

Top differentially frequent patterns between the first-time users (first) and the second-time users (second) in each scenario

Scenario	**Pattern**	**s-support**		**i-support**			**Frequent**
		first	**second**	**first**	**second**	**p**	
Frog	inspect → save → discard-MULT	0.62	0.37	1.01	0.54	<.001	first
	talk-MULT → inspect → save → inspect → save	0.58	0.36	0.78	0.45	<.001	first
	talk-MULT → inspect → save → inspect	0.59	0.37	0.79	0.46	<.001	first
	save → discard → inspect → save	0.53	0.36	0.74	0.48	<.001	first
	inspect → save → discard → inspect	0.53	0.36	0.75	0.49	<.001	first
	talk-MULT → inspect → save	0.78	0.53	1.25	0.70	<.001	both
	inspect → save → talk	0.78	0.60	1.50	0.99	<.001	both
	discard → inspect → save	0.82	0.62	1.97	1.31	<.001	both
	inspect → save → discard	0.78	0.60	1.74	1.19	<.001	both
	talk → inspect → save	0.78	0.63	1.56	1.10	<.001	both
Bee	talk-MULT → inspect → save → inspect → save → inspect	0.59	0.27	0.72	0.32	<.001	first
	talk-MULT → inspect → save → inspect → save → inspect → save	0.59	0.27	0.71	0.32	<.001	first
	talk-MULT → inspect → save → inspect → save	0.74	0.45	0.99	0.57	<.001	first
	talk-MULT → inspect → save → inspect	0.74	0.45	0.99	0.58	<.001	first
	inspect → save → discard-MULT	0.53	0.41	0.86	0.56	<.001	first
	talk-MULT → inspect → save	0.85	0.62	1.30	0.88	<.001	both
	inspect → save → inspect → save → inspect	0.82	0.60	1.83	1.18	<.001	both
	save → inspect → save → inspect → save	0.82	0.60	1.82	1.18	<.001	both
	inspect → save → inspect → save → inspect → save	0.82	0.59	1.81	1.17	<.001	both
	save → inspect → save → inspect	0.83	0.60	1.99	1.31	<.001	both

In terms of gender, 42 differentially frequent long patterns reached the minimum s-support and p-value for both genders in the frog scenario and 44 differentially frequent long patterns reached the minimum s-support and p-value for both genders in the bee scenario. Fifteen patterns were frequently used only by male students in the frog scenario, and 14 were frequently used only by female students in the frog scenario. Eight patterns were frequently used only by males in the bee scenario, and 20 were frequently used only by females in the bee scenario. Forty-four patterns in the frog scenario and 42 patterns in the bee scenario had a higher s-support and a significantly higher average i-support for male students than female students.

Table 6

Top differentially frequent patterns between the female students (F) and the male students (M) in each scenario

Scenario	**Pattern**	**s-support**		**i-support**			**Frequent**
		F	**M**	**F**	**M**	**p**	
Frog	read → note-MULT → read	0.19	0.07	0.27	0.08	<.001	female
	note-MULT → read → note-MULT	0.17	0.06	0.25	0.08	<.001	female
	read → note-MULT → read → note-MULT	0.14	0.04	0.19	0.06	<.001	female
	note-MULT → read → note-MULT → read	0.12	0.03	0.14	0.04	<.001	female
	note-MULT → read → note	0.12	0.05	0.13	0.06	<.001	female
	save → test-MULT → discard-MULT	0.09	0.15	0.12	0.19	<.001	male
	inspect → save → test-MULT → discard-MULT	0.09	0.15	0.12	0.19	.001	male
	inspect → save → inspect → save → test-MULT → discard-MULT	0.06	0.10	0.08	0.12	.005	male
	discard → inspect → save → test-MULT	0.10	0.14	0.11	0.16	.007	male
	save → inspect → save → test-MULT → discard-MULT	0.06	0.10	0.08	0.13	.008	male
	note → inspect → save	0.22	0.16	0.39	0.24	<.001	both
	talk-MULT → start final questions → final claim-MULT	0.17	0.24	0.17	0.24	<.001	both
	save → inspect → save → inspect → save → inspect	0.28	0.35	0.38	0.49	.001	both
	inspect → save → inspect → save → inspect → save → inspect → save	0.28	0.35	0.36	0.47	.001	both
	inspect → save → inspect → save → inspect → save → inspect	0.28	0.35	0.36	0.47	.001	both
Bee	note-MULT → read → note-MULT	0.18	0.07	0.27	0.10	<.001	female
	read → note-MULT → read → note-MULT	0.14	0.05	0.20	0.07	<.001	female
	read → note-MULT → read	0.18	0.08	0.27	0.11	<.001	female
	note-MULT → read → note-MULT → read	0.13	0.05	0.16	0.06	<.001	female
	talk → start final questions → note-MULT	0.11	0.04	0.11	0.04	<.001	female
	inspect → save → test-MULT → look-MULT	0.08	0.12	0.09	0.15	.001	male
	save → test-MULT → look-MULT	0.08	0.12	0.09	0.15	.001	male
	inspect → save → inspect → save → inspect → save → inspect → save → talk	0.08	0.12	0.08	0.13	.002	male
	save → inspect → save → inspect → save → inspect → save → talk	0.08	0.12	0.08	0.13	.002	male
	test-MULT → discard-MULT → inspect → save → inspect	0.07	0.10	0.07	0.11	.004	male
	inspect → save → test-MULT	0.38	0.47	0.62	0.84	<.001	both
	inspect → save → inspect → save → inspect → save → inspect → save	0.28	0.36	0.36	0.50	<.001	both
	inspect → save → inspect → save → inspect → save → inspect	0.28	0.36	0.36	0.50	<.001	both
	save → inspect → save → inspect → save → inspect → save	0.29	0.36	0.36	0.50	<.001	both
	note → inspect → save	0.20	0.14	0.36	0.21	<.001	both

Note. The top five differentially frequent long patterns that were common (i.e., met the 10% s-support threshold) to both female and male students, the top five that were frequently used by male students, and the top five that were frequently used by female students were listed for each scenario.

Table 6 presents the top five differentially frequent long patterns that were common (i.e., met the 10% s-support threshold) to both groups, the top five that were frequently used by male students, and the top five that were frequently used by female students within each scenario. Consistent with previous results on two-action sequential patterns, the long sequential patterns frequently used by female students involved opening the notepad repeatedly for note-taking, especially after reading kiosk information, probably to take notes of the research information they just read about. On the other hand, the male students showed a higher frequency of long sequential patterns that entailed actions such as running laboratory experiments, inspecting objects, saving objects to backpack, discarding objects, and talking with NPCs. In other words, male students were more likely to explore the environment, collect data, and run experiments on the items they collected.

Discussion

In summary, the analysis of student behavior patterns within VPA suggested that experience with learning in VPA stimulated students to make better use of learning strategies, and better self-monitor and self-evaluate their learning and performance during the exploration and assessment process, both of which are important components of self-regulated learning. New to the environment and probably not entirely understanding what they were supposed to do, students who were introduced to VPA for the first time tended to access the help page that reminded them of their tasks and took notes on it more often than the second-time users in the frog scenario. This might suggest that students were more familiar with their tasks and did not need to record this information the second time they used VPA. On the other hand, second-time users generally showed better strategy usage than their counterparts who used VPA for the first time during science inquiry. They tended to open the notepad more frequently after reading research information or running and viewing experiment results, probably to record information that they thought was important into the notepad. The activity of encoding the information might facilitate the acquisition and understanding of domain-specific knowledge presented in the research kiosk and the interpretation of the laboratory test results. The notes recorded in the notepad were also reviewed more frequently by second-time users in order to identify the final claim on frog mutation or death of bee population. Second-time users were also more likely to read kiosk research information after viewing laboratory test results, possibly to interpret the results. Furthermore, second-time users were more likely to access the notepad, where information from various sources could be recorded and synthesized, potentially enabling them to review their notes to monitor and reflect on their final claims (cf. Kuhn & Pease, 2008) immediately after submitting a final claim. Exploration of longer sequential patterns indicated that students who had not used VPA before executed more sequences comprised of exploratory behaviors such as talking with NPCs and collecting data, while the second-time users focused primarily on what was necessary to answer the core inquiry question and selectively collected data. These results suggested the development of self-regulatory skills across the course of using the two scenarios of VPA.

Significant gender differences were also found in this analysis. Female students demonstrated more expert-like SRL behaviors and strategies compared to males — they tended to make use of the notepad more frequently and exploited more available sources of information (e.g., laboratory test results, research information) to help them solve inquiry problems than their male counterpart. Females were more likely to review notes or read multiple research pages before submitting the final claim, which might assist them with the decision-making process. Female students also engaged in more self-monitoring and self-assessment than male students, as they made use of their notes taken during learning to monitor and reflect on their learning and solutions. In conclusion, previous findings that female students showed advantages in paper-based note-taking in traditional classroom lecture settings (Cohn et al., 1995; Kiewra, 1984; Maddox & Hoole, 1975; Nye, 1978; Reddington et al., 2015; Slotte et al., 2001; Williams & Eggert, 2002) transferred to the computer-based note-taking in the open-ended learning environment. These results were also consistent with previous literature that female students reported themselves as using self-regulatory strategies more often than males (Lee, 2002; Matthews et al., 2009; Pajares, 2002; Yukselturk & Top, 2013; B. J. Zimmerman & Martinez-Pons, 1990). Gender difference in self-regulatory skills favoring females existed for middle school students in open-ended learning environments for science, even though males typically showed higher science achievement and higher motivation towards science and computers according to previous literature (Kay & Lauricella, 2011; Reilly et al., 2015). On the other hand, male students tended to collect more data and run more experiments on the items they collected. The gender difference in behavioral patterns of students might also explain the gender difference in the development of science inquiry skills, which suggested that the science inquiry skills of female students whose behavioral patterns showed a higher level of self-regulatory skills developed over the use of VPA whereas male students' science inquiry performance did not improve over time.

CHAPTER VII.

ANALYSIS 3: NOTE-TAKING/REVIEWING STRATEGIES

The results from analysis 2 indicated that the second-time users and female students were more likely to utilize the digital notepad for note-taking or note-reviewing purposes after reading research pages, after obtaining test results within the virtual environment, and before and after submitting their final claims than the first-time users and male students. Taking and reviewing notes are popular learning strategies that have been deemed as beneficial for academic success (Armbruster, 2009) and are also critical elements of self-regulated learning (Azevedo, 2005; Moos, 2009). However, limited studies have been conducted on the effects of note-taking/reviewing as an SRL strategy in computer-based learning environments such as OELEs on student academic success, and the development of note-taking/reviewing strategies in these environments (Armbruster, 2009). Analysis 3 aims to investigate: 1) The relationship between the quantity of note-taking/reviewing behaviors and the content of notes by middle school students in a digital notepad with their performance on complex science inquiry tasks within Virtual Performance Assessments; 2) How VPA fosters the development of note-taking/reviewing strategies by comparing both the quantity of note-taking/reviewing behaviors and the content of notes taken by a) first-time and second-time users, and b) male and female students.

Quantitative Measures of Note-Taking/Reviewing Behavior

Within VPA, students could click on the digital notepad to take or review notes. Quantitative measures representing the quantity of note-taking/reviewing behaviors were computed for each note-taker who made use of the digital notepad and were used in later analysis. A description of the full set of measures on notepad use can be found in Table 7. They include features that represent general notepad access, such as the number of times students opened the notepad window (i.e., notepad access frequency), the total amount of time in minutes that notepad was open (i.e., notepad time), and the percentage of total time in VPA that the student was using the notepad (i.e., percent of time on notepad). In addition, some measures were calculated by distinguishing between note-taking (where students recorded information or changed previous content) and note-reviewing (where students opened the notepad without adding or changing content, indicating that the student likely reviewed the notes that had been taken without storing new information or editing previous notes) (Di Vesta & Gray, 1972; Kiewra, 1989).

Table 7

List of features related to note-taking/reviewing quantity that were distilled from log data

Feature	**Description**
Notepad access frequency	Frequency of opening the notepad window
Notepad time	Total amount of time in minutes that notepad was open
Percent of time on notepad	Total amount of time on notepad divided by total time in VPA
Word count in note	Number of words in note-taker's note
Segment count in note	Number of sentence segments in note-taker's note
Note-taking frequency	Frequency of note-taking actions
Note-reviewing frequency	Frequency of note-reviewing actions
Percent note-taking actions	Frequency of note-taking divided by frequency of notepad access
Percent note-reviewing actions	Frequency of note-reviewing divided by frequency of notepad access
Note-taking duration	Total amount of time (in minutes) spent on taking notes
Note-reviewing duration	Total amount of time (in minutes) spent on reviewing notes
Avg note-taking duration	Average duration (in minutes) of a note-taking action
Avg note-reviewing duration	Average duration (in minutes) of a note-reviewing action
Note-taking to notepad time	Ratio of time spent on note-taking actions and total time on notepad
Note-reviewing to notepad time	Ratio of time spent on note-reviewing actions and total time on notepad

Measures of Note Content

Beyond simply studying the quantity of note-taking/reviewing and time spent on this activity, the content of students' notes was also studied, following the procedures recommended by Chi (1997) and Trevors et al. (2014). Each student's notes were automatically parsed into sentential segments (i.e., sentence-based units) (Chi, 1997; Trevors et al., 2014), using the Stanford CoreNLP tool (Manning et al., 2014). These segments were then checked manually, and inappropriate segmentation was adjusted. For example, if a student placed a period or a line break in the middle of a sentence, the sentence was manually recombined in the second-round adjustment. Similarly, comma-splices (the use of a comma to connect two independent clauses) were manually split into multiple segments. This process resulted in the identification of 9,983 segments in the frog scenario and 9,738 segments in the bee scenario.

All segments were then coded by two raters using three coding schemes: (1) The *type of note* coding scheme, which is partially adapted from the coding scheme developed by Trevors et al. (2014), differentiates between *content reproduction* (verbatim or paraphrased content; Trevors et al., 2014), *content elaboration* (the introduction of new semantic information or ideas; Trevors et al., 2014), *metacognition* (reflection on learning process, experience, or knowledge), and *other*. (2) The *source of note content* coding scheme labels note segments according to their origin within the system, including research *kiosks*, lab *tests*, field *observations*, and *dialogues* with NPCs. Segments which reflect a mixture of these sources were given both the label *combination* and secondary codes reflecting which sources are combined. Segments whose source could not be determined were labeled as *unknown*. (3) Finally, the *hypothesis* or *conclusion* coding scheme differentiates between segments that make a *hypothesis* about the possible causal factors for the final assessment (e.g., hypothesizing that pollution was causing the frog mutation) and segments that draw a lower-level *conclusion* from data collected (e.g., linking a farm with a bad-smelling water sample with possible pollution). Segments that do not belong to either of these categories are coded as *other*. Examples of the coding schemes are shown in Table 8.

Table 8

Coding schemes for note content. Description of each category of the measures and relevant examples are provided.

Scheme	Category	Description	Example
Type of Note	Content Reproduction	Note segment is a verbatim copy or close paraphrase of the content presented in the environment that does not introduce new semantic information or ideas.	Ethonal [sic] is a natural chemical produced by plants
	Content Elaboration	Note segment introduces new semantic information/ideas/meaning to content immediately available in the environment (e.g., making an inference, connecting information with prior knowledge, identifying underlying patterns of data, constructing internal connections, etc.).	The tadpole from Jones pond had a short tail and missing an eye, a reaction to the pesticides in the water .
	Metacognitive	Note segment pertains to reflecting on and monitoring one's own learning process, knowledge, and experience with VPA.	so far the water samples that I have collected there is only one water sample that really stands out to me .
	Other	Note segment does not belong to any of the other categories (i.e., Reproduction, Elaboration, Metacognitive).	all bees are starving
Source of Note	Kiosk	Note segment contains information from research kiosk pages.	pesticides can cause mutations including extra limbs in frogs
	Test	Note segment contains information that could be traced to the laboratory test results.	water test : pH 4.5 , atrazine
	Observation	Note segment contains information based on what students observed in the virtual environment.	yellow tadpole : smaller than normal , short tail
	Dialogue	Note segment contains information from conversation with NPCs in VPA.	Another nam [sic] says that pesticides are the reason because 'he' sprays his fields with imidacloprid [sic].
	Combination	Note segment involves coordinating and integrating pieces of information from multiple disparate sources from the other categories (i.e., Kiosk, Test, Observation, Dialogue).	Internet Kiosk says pesticide (such as atrazine , which someone accused Garcia of using) can cause extra limbs to appear in frogs .

	Unknown	Note segment contains information whose source could not be identified.	i think the frog is an alien frog.
Hypothesis / Conclusion	Hypothesis	Note segment proposes a possible final hypothetical claim and generates a hypothesis about the possible causal factors (e.g., pesticides, pollution, parasites, genetic mutation, aliens) leading to the mutation of the six-legged frog or the death of the local bee population.	I think that the reason why the frong [sic] was abnormal and had six legs was because the water and pestisides [sic] in the water
	Conclusion	Note segment pertains to forming and drawing a conclusion from data that students collected (e.g., test results, kiosk pages, observation, dialogue, etc.).	Red bee is infected by parasites (Varroa Mites) as it has SMALL BROWN OR RED SPOTS AND STUBBY WINGS .
	Other	Note segment does not belong to Hypothesis or Conclusion.	frog has really low white blood

In addition, I listed all pieces of meaningful information that are presented to students in each VPA scenario (n = 121 in bee and n = 127 in frog), and then coded each segment in terms of whether any of the information was recorded in the segment. For instance, if a sentence segment mentioned that the six-legged frog is smaller than normal, the sentence would be given the corresponding code for this piece of information ("Observation_Six_Legged_Frog_1"). Note that one segment could be assigned multiple or zero *information* codes.

Two coders independently coded all note segments from a random 10% sample of students (among those who ever took notes) in the frog scenario. Cohen's (1960) kappa showed substantial inter-rater agreement was achieved for the *type of note* (κ = .81), the *source of note* (κ = .90), and the *information in note* (κ = .91). Results for *Hypothesis*/*Conclusion* (κ = .74) showed the need for further refinement, so definitions of each category in this scheme were further clarified in order to improve the reliability. Two rounds of coding of notes from an additional 10% of sample participants were conducted and a significantly improved agreement was achieved for *Hypothesis*/*Conclusion* (κ = .90). Discrepancies in final ratings in these random samples were resolved by discussion between the raters. Once the acceptable inter-rater agreement was established, the remaining note segments were then coded by one coder.

After all segments were coded, quantitative measures based on these categories were calculated for each note-taker (students who took notes) and used in later analysis. For example, the frequency of each code (e.g., *content reproduction*, *content elaboration*, etc.) was calculated for each note-taker, and each coding scheme, in each scenario. In addition, I computed the number of aggregated labels across coding schemes (e.g., segments coded as *content reproduction* from the research *kiosk, content elaboration* from field *observation*, etc.). In cases where a segment combined information from multiple disparate sources (e.g., *dialogue* and *test*), I counted this note as both a *combination* segment and as the specific categories they belonged to, when calculating these measures.

Three measures were developed based on the *information* codes. First, I counted the number of unique values of information recorded in each student's note. For example, if a piece of information (e.g., pH level of the control water is 7.5) was mentioned in at least one segment of the student's note, it received one point; otherwise 0 point was assigned for the information. Therefore, the quantity of unique information represents the amount of information presented in the environment that was noted in notepad. This measure complements the count of sentence segments measure by distinguishing students who wrote multiple pieces of meaningful information in one sentence from note-takers whose multiple sentences repeatedly mentioned the same piece of information. Second, based on the information code, I evaluated whether each piece of CVS evidence necessary to test the correct causal hypothesis and all possible hypotheses was recorded in the student's note or not. The total amount of CVS CFC evidence and CVS evidence recorded in notes (i.e., *CVS CFC-data notes* and *CVS-data notes*) comprise relevant measures of note content. They correspond to the measures on students' use of the control of variables strategy during their inquiry behaviors (CVS CFC-data score and CVS-data score). As with that analysis, recording CVS-data notes and CVS CFC-data notes does not necessarily mean that the students engaged in CVS. Instead, it indicates that information that is necessary for students to apply CVS to test hypotheses had been noted in the digital notepad.

The purpose of analysis three on note-taking as SRL strategy is two folds. First, I seek to explore whether the quantity of taking notes and the quantity of reviewing notes, which comprise the two fundamental functions of note-taking, and the content of notes taken by students, are associated with success on science inquiry task within Virtual Performance Assessments. Second, I aim to examine the development of note-taking and note-reviewing strategies in the environment.

Multilevel analyses were conducted to investigate the relationships between the meaningful features related to note-taking/reviewing quantity and note content distilled from the log files and measures of student success in the virtual environment (i.e., each student's CFC, ISE, CVS-data, and CVS CFC-data scores). Specifically, three-level logistic regression models and three-level regression models were fitted with students in each scenario nested within classes, and classes nested within teachers. In the three-level logistic regression models, the dependent variable is the student's CFC score, and each individual feature related to note-taking/reviewing quantity or note content serves as the single level-one predictor variable in each model. These three-level logistic regressions were conducted for each feature to determine the relationship between the note-taking/reviewing quantity or note content and student success on identifying a correct final claim after controlling for class- and teacher-level variability in each scenario.

Similarly, three-level regression models were fitted for all pairs of relationships between ISE/CVS-data /CVS CFC-data performance and individual features, with students in each scenario nested within classes, and classes nested within teachers. Student performance is the dependent variable and each individual feature is the level-one predictor variable in these models. Benjamini and Hochberg's posthoc control method was used to control for conducting multiple statistical analyses.

Analysis 3.1: Note-Taking as SRL Strategy and Science Inquiry Performance

In the frog scenario, 1,178 students spontaneously opened the notepad to take notes at least once (i.e., note-takers), and 807 students did not open the notepad to take notes at all (i.e., non-note-takers). In the bee scenario, 1,172 students opened the notepad to take notes at least once and 849 students did not access the notepad at all. Results revealed that the note-takers achieved a significantly higher average CFC score in the frog scenario (frog: *Ms* = 34% and 24%, *z* = 4.40, *p* < .001; bee: *Ms* = 30% and 27%, *z* = 1.04, *p* = .300), and significantly higher ISE score (frog: *Ms* = 54 and 44, *t*(1979) = 7.82, *p* < .001; bee: *Ms* = 49 and 43, *t*(2000) = 4.67, *p* < .001), CVS CFC-data score (frog: *Ms* = 2.45 and 1.85, *t*(1922) = 8.93, *p* < .001; bee: *Ms* = 2.40 and 1.73, *t*(2008) = 9.89, *p* < .001), and CVS-data score (frog: *Ms* = 12.27 and 8.98, *t*(1937) = 9.28, *p* < .001; bee: *Ms* = 12.06 and 8.59, *t*(2015) = 9.62, *p* < .001) than non-note-takers in both scenarios.

Quantity of Note-Taking/Reviewing Behavior and Performance

Three-level regression results indicated that measures of overall notepad use quantity (e.g., frequency of notepad access, time on notepad) among note-takers were significantly positively associated with science inquiry performance in the frog scenario when controlling for the class-level and teacher-level variability (statistics are reported in Table 9). For example, opening the notepad more frequently was positively associated with CFC performance ($e^B = 1.02$, $z = 5.08$, $p < .001$, adjusted $\alpha = .006$), ISE score ($\beta = .23$, $t(1142) = 7.87$, $p < .001$, adjusted $\alpha = .004$), CVS-data score ($\beta = .22$, $t(1132) = 7.34$, $p < .001$, adjusted $\alpha = .001$), and CVS CFC-data score ($\beta = .21$, $t(1100) = 7.14$, $p < .001$, adjusted $\alpha = .001$). These associations suggest that the more frequently students accessed the notepad and the more time they devoted to using the notepad, the better their average science inquiry performance was in the frog scenario.

Table 9

Three-level logistic regression of student CFC performance on each feature related to note-taking/reviewing quantity, and three-level regression of student ISE, CVS-data, and CVS CFC-data score on each of these features in the frog scenario

DV	**CFC**			**ISE**			**CVS-data**			**CVS CFC-data**		
Feature	**B**	**SE B**	**e^B**	**B**	**SE B**	**β**	**B**	**SE B**	**β**	**B**	**SE B**	**β**
Notepad access frequency	.02	.005	1.02*	.35	.04	.23*	.10	.01	.22 *	.02	.003	.21 *
Notepad time	.06	.01	1.06*	.73	.13	.16*	.22	.04	.17 *	.04	.01	.16 *
Percent of time on notepad	−.50	.34	.60	−.63	3.28	−.01	3.59	.98	.11 *	.58	.19	.09 *
Word count in note	.004	.001	1.00*	.06	.01	.15*	.01	.003	.11 *	.002	.0007	.10 *
Segment count in note	.04	.01	1.04*	.65	.09	.21*	.17	.03	.18 *	.03	.01	.19 *
Note-taking frequency	.03	.01	1.03*	.42	.06	.20*	.11	.02	.19 *	.02	.003	.18 *
Note-reviewing frequency	.06	.01	1.06*	.84	.12	.20*	.24	.04	.20 *	.05	.01	.20 *
Percent note-taking actions	−.17	.35	.84	−6.54	3.39	−.05	−2.27	1.02	−.06	−.51	.20	−.07 *
Percent note-reviewing actions	.17	.35	1.19	6.54	3.39	.05	2.27	1.02	.06	.51	.20	.07 *
Note-taking duration	.07	.02	1.07*	.81	.16	.14*	.26	.05	.16 *	.04	.01	.14 *
Note-reviewing duration	.14	.05	1.15*	1.68	.39	.12*	.49	.12	.12 *	.10	.02	.13 *
Avg note-taking duration	−.22	.18	.80	−6.36	1.55	−.12*	−1.81	.46	−.11 *	−.35	.09	−.11 *
Avg note-reviewing duration	.73	.35	2.08	7.71	3.08	.07*	1.19	.93	.04	.25	.18	.04
Note-taking to notepad time	−1.18	.48	.31*	−19.76	4.77	−.12*	−3.55	1.44	−.07 *	−.86	.28	−.09 *
Note-reviewing to notepad time	1.18	.48	3.26*	19.76	4.77	.12*	3.55	1.44	.07 *	.86	.28	.09 *

Note. For three-level logistic regression results, logistic coefficient of the predictor (B), standard error associated with the coefficient (SE B), and odds ratio for the predictor (e^B) are reported. For three-level regression results, coefficient of the predictor (B), standard error associated with the coefficient (SE B), and standardized coefficient (β) are reported. Statistically significant results after Benjamini and Hochberg's control are marked with *.

Results distinguishing between taking notes and reviewing notes, which comprise the two basic functions of note-taking (Di Vesta & Gray, 1972), indicated that taking notes frequently and spending more time in total on taking notes were all positively associated with science inquiry performance (e.g., CFC, ISE, CVS-data, CVS CFC-data) in the frog scenario (statistics reported in Table 9). In addition, the quantity of notes (i.e., the number of words in notes, the number of sentence segments in notes) was significantly positively related to science inquiry performance in the frog scenario.

Similarly, the more frequently students reviewed notes and the more time they spent reviewing notes (both in terms of duration and percentage), the more likely that they made a correct final claim about the causal factor leading to the frog mutation, and the better they performed in justifying the claim with supporting evidence and using CVS to test the hypotheses in the frog scenario. For example, reviewing notes frequently was positively associated with CFC score ($e^B = 1.06$, $z = 4.39$, $p < .001$, adjusted $\alpha = .008$), ISE score ($\beta = .20$, $t(1170) = 6.97$, $p < .001$, adjusted $\alpha = .002$), CVS-data score ($\beta = .20$, $t(1167) = 6.75$, $p < .001$, adjusted $\alpha = .003$), and CVS CFC-data score ($\beta = .20$, $t(1146) = 6.90$, $p < .001$, adjusted $\alpha = .003$). The percentage of time spent on reviewing notes (relative to the time devoted to taking notes) was also positively associated with science inquiry performance in the frog scenario. In other words, note-takers who spent a larger proportion of their time within the notepad reviewing their notes tended to be more likely to make a correct final claim ($e^B = 3.26$, $z = 2.48$, $p = .013$, adjusted $\alpha = .019$), justified their final claim with more supporting causal evidence ($\beta = .12$, $t(1165) = 4.14$, $p < .001$, adjusted $\alpha = .010$), and applied CVS more often to test all hypotheses ($\beta = .10$, $t(1175) = 3.36$, $p < .001$, adjusted $\alpha = .014$) and the correct final claim ($\beta = .08$, $t(1165) = 2.66$, $p = .008$, adjusted $\alpha = .017$) in the frog scenario.

However, very different results were obtained for the bee scenario (see Table 10). In this scenario, only the number of sentences recorded in notes was positively associated with the quantity of CVS evidence (β = .08, t(1160) = 2.67, p = .008, adjusted α = .017) and CVS CFC evidence (β = .08, t(1158) = 2.70, p = .007, adjusted α = .017) collected by students. The quantity of general notepad access and note-taking/reviewing behaviors were not significant predictors of science inquiry performance in the bee scenario. However, the frequency of reviewing notes and the percentage of time spent on reviewing notes were negatively associated with CFC performance (e^B = 0.97, z = −2.34, p = .020, adjusted α = .021; e^B = .27, z = −2.39, p = .017, adjusted α = .020).

Table 10

Three-level logistic regression of student CFC performance on each feature related to note-taking/reviewing quantity, and three-level regression of student ISE, CVS-data, and CVS CFC-data score on each of these features in the bee scenario

DV	**CFC**			**ISE**			**CVS-data**			**CVS CFC-data**		
Feature	**B**	**SE B**	**e^B**	**B**	**SE B**	**β**	**B**	**SE B**	**β**	**B**	**SE B**	**β**
Notepad access frequency	−.007	.005	.99	.04	.04	.03	.03	.01	.06	.01	.003	.06
Notepad time	−.004	.01	1.00	.21	.13	.05	.06	.04	.04	.01	.01	.04
Percent of time on notepad	−.73	.64	.48	−.32	5.82	−.002	−3.66	1.95	−.05	−.63	.36	−.05
Word count in note	.001	.001	1.00	.02	.01	.05	.01	.004	.05	.001	.0007	.05
Segment count in note	−.01	.01	.99	.13	.08	.05	.07	.03	.08 *	.01	.01	.08 *
Note-taking frequency	−.01	.01	.99	.05	.06	.03	.04	.02	.05	.01	.004	.06
Note-reviewing frequency	−.03	.01	.97*	.06	.12	.02	.04	.04	.03	.01	.01	.03
Percent note-taking actions	.62	.35	1.86	−.88	3.19	−.01	−1.69	1.07	−.05	−.31	.20	−.05
Percent note-reviewing actions	−.62	.35	.54	.88	3.19	.01	1.69	1.07	.05	.31	.20	.05
Note-taking duration	<.001	.02	1.00	.20	.16	.04	.07	.05	.04	.01	.01	.05
Note-reviewing duration	−.07	.05	.93	.28	.47	.02	−.13	.16	−.02	−.03	.03	−.03
Avg note-taking duration	.07	.14	1.07	.28	1.32	.01	−.37	.44	−.02	−.03	.08	−.01
Avg note-reviewing duration	−.16	.32	.85	1.90	2.80	.02	−.71	.93	−.02	−.20	.17	−.03
Note-taking to notepad time	1.32	.55	3.74*	2.14	4.63	.01	.48	1.55	.01	.19	.29	.02
Note-reviewing to notepad time	−1.32	.55	.27*	−2.14	4.63	−.01	−.48	1.55	−.01	−.19	.29	−.02

Note Content and Performance

In addition to the analysis of notepad use behavior quantity, I also examined the relationship between the content of notes and student performance on the science inquiry task. For this analysis, I used the same data as discussed above, but, due to limitations in logging, seven students who deleted all of their notes before exiting the environment had to be excluded from each scenario, leaving 1,171 note-takers in the frog scenario and 1,165 note-takers in the bee scenario.

In the frog scenario, students recorded an average of 8% of the total information presented in VPA to the notepad (M = 10.65, SD = 10.93). Their notes included 0.58 pieces of CVS CFC evidence (SD = 0.74) and 2.30 pieces of CVS evidence (SD = 2.62) on average. In this scenario, an average of 69% of a student's note segments were verbatim copies or close paraphrases of the content presented in the environment (M = 6.74, SD = 6.99), an average of 20% of note segments were semantically elaborative notes that added new information or generated inferences (M = 1.36, SD = 2.14), and 2% of the segments were metacognitive notes. In the bee scenario, students recorded an average of 10.96 pieces of unique information (SD = 11.20) in the digital notepad. Their notes included 0.53 pieces of CVS CFC evidence (SD = 0.73) and 2.26 pieces of CVS evidence (SD = 2.66) on average. An average of 71% of student note segments were copies or paraphrases of content in the environment (M = 6.75, SD = 6.77), 21% involved content elaboration (M = 1.32, SD = 2.02), and an average of 2% contained reflective and metacognitive content. In both scenarios, a relatively large percentage of a student's segments were based on information from research kiosk pages (39% in frog, 37% in bee), followed by notes that could be traced to students' observation in the environment (26% in frog, 29% in bee) and notes from laboratory test results (22% in frog, 22% in bee). A relatively smaller proportion of note segments (2% in frog, 4% in bee) coordinated multiple sources of information. Specifically, most reproductive notes reproduced content from kiosk informational pages. Among the elaborative segments, students elaborated largely on observation and test results. An average of 11% of student notes from the frog scenario generated possible causal hypotheses related to the mutation of frog, and an average of 6% of student notes attempted to draw conclusions based on data. In the bee scenario, on average, 8% of student notes involved potential causal hypotheses, and 6% of student notes drew conclusions from data.

Table 11

Three-level logistic regressions between the number of different categories of segments in a student's note and their CFC performance and three-level regressions of student ISE, CVS-data, and CVS CFC-data score on note content in the frog scenario

Scenario	**Frog CFC**			**Frog ISE**			**Frog CVS-data**			**Frog CVS CFC-data**		
Note Content	**B**	**SE B**	**e^B**	**B**	**SE B**	***β***	**B**	**SE B**	***β***	**B**	**SE B**	***β***
Information	.03	.01	1.03 *	.43	.06	.20 *	.12	.02	.20 *	.02	<.01	.19 *
CVS CFC-data Notes	1.14	.11	3.12 *	1.63	.84	.34 *	2.53	.26	.27 *	.59	.05	.33 *
CVS-data Notes	.15	.03	1.16 *	2.00	.25	.23 *	.77	.07	.29 *	.15	.01	.29 *
Reproduction	.04	.01	1.04*	.63	.10	.19*	.15	.03	.16 *	.03	.01	.17 *
Elaboration	.08	.03	1.08*	1.39	.30	.13*	.33	.09	.10 *	.06	.02	.10 *
Metacognition	−.16	.16	.86	−1.71	1.00	−.05	.40	.30	.04	.05	.06	.02
Test	.05	.02	1.05*	.87	.20	.12*	.68	.06	.32 *	.12	.01	.30 *
Kiosk	.07	.01	1.08*	.97	.13	.22*	.12	.04	.09 *	.03	.01	.13 *
Observation	−.01	.02	.99	.10	.16	.02	−.05	.05	−.03	−.02	.01	−.05
Dialogue	−.07	.05	.94	−.67	.51	−.04	−.21	.15	−.04	−.05	.03	−.05
Combination	.11	.07	1.12	1.61	.60	.07*	.57	.18	.09 *	.11	.04	.09 *
Hypothesis	.19	.06	1.21*	2.29	.60	.10*	.48	.18	.07 *	.11	.04	.09 *
Draw Conclusion from Data	.16	.06	1.17*	3.35	.58	.16*	.91	.18	.15 *	.15	.03	.13 *
Reproduction of Test	.04	.03	1.04	.76	.25	.08*	.75	.07	.28 *	.14	.01	.27 *
Reproduction of Kiosk	.07	.01	1.07*	.95	.13	.20*	.11	.04	.08 *	.03	.01	.11 *
Reproduction of Observation	−.002	.02	1.00	.15	.18	.02	−.01	.05	−.01	−.01	.01	−.03
Reproduction of Dialogue	−.05	.05	.95	−.57	.53	−.03	−.18	.16	−.03	−.04	.03	−.04
Reproduction of Combination	.46	.58	1.58	6.68	5.94	.03	2.61	1.80	.04	.58	.35	.05
Elaboration on Test	.16	.05	1.18*	2.39	.50	.13*	1.16	.15	.22 *	.22	.03	.21 *
Elaboration on Kiosk	.12	.06	1.13	1.77	.57	.09*	.40	.17	.07 *	.07	.03	.06
Elaboration on Observation	−.04	.05	.96	−.04	.52	<.01	−.44	.16	−.08 *	−.08	.03	−.08 *
Elaboration on Dialogue	−.59	.47	.56	−5.68	3.51	−.05	−2.39	1.06	−.06	−.53	.20	−.08 *

Elaboration on Combination	.11	.07	1.12	1.58	.61	.07*	.56	.18	.09 *	.11	.04	.09 *
Elaboration Test Hyp	.29	.13	1.33	4.18	1.30	.09*	2.27	.39	.17 *	.43	.07	.16 *
Elaboration Kiosk Hyp	.39	.13	1.47*	4.86	1.23	.11*	.79	.38	.06	.20	.07	.08 *
Elaboration Observation Hyp	−.04	.17	.96	−1.00	1.66	−.02	−1.50	.50	−.09 *	−.23	.10	−.07 *
Elaboration Dialogue Hyp	−1.42	.98	.24	−7.57	5.41	−.04	−2.85	1.64	−.05	−.64	.32	−.06
Elaboration Combination Hyp	.26	.21	1.29	3.55	2.10	.05	1.39	.63	.06	.32	.12	.07 *
Elaboration Test Conc	.23	.09	1.25	4.15	.93	.12*	1.65	.28	.17 *	.30	.05	.16 *
Elaboration Kiosk Conc	.07	.16	1.07	2.24	1.62	.04	.82	.49	.05	.11	.09	.03
Elaboration Observation Conc	.01	.10	1.01	1.99	.98	.06	.12	.30	.01	.01	.06	<.01
Elaboration Dialogue Conc	−17.38[a]	80.95	<.01	−12.51	15.70	−.02	−6.78	4.74	−.04	−1.38	.91	−.04
Elaboration Combination Conc	.18	.14	1.19	3.56	1.47	.07*	1.25	.44	.08 *	.26	.08	.09 *

Note. Significant results after post-hoc controls are marked with *. Extreme values in *a* because there are few cases of Elaborative Conclusion notes based on Dialogue.

As in the previous section, three-level logistic regressions and three-level regressions were conducted to examine the relationships between students' science inquiry performance and the count and percentage of each category of notes, using Benjamini and Hochberg's post-hoc control. Results for the frog scenario and the bee scenario are reported in Table 11 and Table 12 respectively. Overall, the associations between the percentage of various categories of notes and student performance were weaker than the associations between the frequency of the categories and performance. This was expected, considering that the percentage of types of information will not be informative if there are relatively few notes being taken in the first place. For example, a student whose notes included ten segments, with five of them from kiosk pages (50%) could be expected to learn more than a student who only encoded one note segment and the segment was from the kiosk (100%). Therefore, I focused on the results for frequency of features.

Table 12

Three-level logistic regressions between the number of different categories of segments in a student's note and their CFC performance and three-level regressions of student ISE, CVS-data, and CVS CFC-data score on note content in the bee scenario

Scenario	**Bee CFC**			**Bee ISE**			**Bee CVS-data**			**Bee CVS CFC-data**		

Note Content	**B**	**SE B**	**e^B**	**B**	**SE B**	***β***	**B**	**SE B**	***β***	**B**	**SE B**	***β***
Information	.005	.006	1.00	.19	.06	.10 *	.08	.02	.13 *	.02	<.01	.13 *
CVS CFC-data Notes	.60	.09	1.82 *	6.79	.83	.23 *	2.29	.28	.23 *	.47	.05	.26 *
CVS-data Notes	.04	.02	1.04	.82	.23	.10 *	.67	.08	.25 *	.12	.01	.25 *
Reproduction	−.01	.01	.99	.10	.09	.03	.06	.03	.05	.01	.01	.06
Elaboration	.05	.03	1.05	.77	.31	.07*	.23	.10	.07	.04	.02	.06
Metacognition	−.01	.11	.99	−.59	1.02	−.02	.62	.34	.05	.14	.06	.06
Test	.04	.02	1.04	.66	.20	.10*	.61	.06	.27 *	.11	.01	.25 *
Kiosk	−.02	.01	.98	.07	.12	.02	.05	.04	.04	.01	.01	.03
Observation	.001	.02	1.00	.10	.15	.02	−.11	.05	−.07 *	−.02	.01	−.06
Dialogue	−.01	.05	.99	−.09	.42	−.01	−.36	.14	−.07 *	−.06	.03	−.06
Combination	.15	.06	1.16*	2.11	.59	.10*	.50	.20	.07 *	.10	.04	.08 *
Hypothesis	.19	.06	1.21*	1.83	.60	.09*	.73	.20	.10 *	.12	.04	.09 *
Draw Conclusion from Data	−.03	.06	.97	.58	.55	.03	−.17	.18	−.03	−.02	.03	−.01
Reproduction of Test	.04	.02	1.04	.63	.24	.08*	.68	.08	.24 *	.12	.01	.23 *
Reproduction of Kiosk	−.03	.02	.97	.02	.13	<.01	.03	.04	.02	.01	.01	.02
Reproduction of Observation	−.01	.02	.99	.03	.17	<.01	−.13	.06	−.07 *	−.02	.01	−.06
Reproduction of Dialogue	−.001	.05	1.00	−.04	.44	<.01	−.39	.15	−.08 *	−.06	.03	−.06
Reproduction of Combination	.16	.12	1.17	2.32	1.13	.06	.43	.38	.03	.10	.07	.04
Elaboration on Test	.08	.06	1.08	2.11	.61	.10*	1.30	.20	.18 *	.22	.04	.17 *
Elaboration on Kiosk	.10	.06	1.11	1.23	.58	.06	.40	.19	.06	.07	.04	.06
Elaboration on Observation	.05	.05	1.05	.74	.48	.04	−.18	.16	−.03	−.02	.03	−.02
Elaboration on Dialogue	−.60	.48	.55	−2.32	2.95	−.02	−.46	.99	−.01	−.18	.18	−.03
Elaboration on Combination	.17	.08	1.18	2.28	.73	.09*	.58	.24	.07 *	.11	.05	.07 *
Elaboration Test Hyp	.32	.13	1.38*	4.96	1.25	.11*	2.35	.42	.16 *	.41	.08	.15 *
Elaboration Kiosk Hyp	.20	.11	1.22	1.59	1.09	.04	.76	.36	.06	.12	.07	.05
Elaboration Observation Hyp	.55	.15	1.73*	3.28	1.35	.07*	.22	.45	.01	.10	.08	.03
Elaboration Dialogue Hyp	−17.42[a]	181.02	<.01	−2.73	6.89	−.01	−1.05	2.30	−.01	−.23	.43	−.02
Elaboration Combination Hyp	.52	.18	1.68*	3.66	1.57	.07*	.94	.52	.05	.16	.10	.05
Elaboration Test Conc	−.07	.13	.93	1.52	1.20	.04	1.09	.40	.08 *	.18	.07	.07 *
Elaboration Kiosk Conc	.20	.12	1.22	2.02	1.17	.05	.13	.39	.01	.04	.07	.02
Elaboration Observation Conc	−.07	.08	.93	.44	.68	.02	−.50	.23	−.06	−.08	.04	−.05
Elaboration Dialogue Conc	−.22	1.18	.80	.13	10.36	<.01	−7.86	3.46	−.06	−1.72	.64	−.08 *
Elaboration Combination Conc	.16	.12	1.17	2.43	1.15	.06	.15	.38	.01	.04	.07	.02

Note. Significant results after post-hoc controls are marked with *.

Amount of unique information in notes and performance

The amount of unique information that was recorded in notes was positively associated with CFC performance in the frog scenario ($e^B = 1.03$, $z = 4.46$, $p < .001$, adjusted $\alpha = .009$), and was positively associated with ISE performance (frog: $\beta = .20$, $t(1164) = 7.01$, $p < .001$, adjusted $\alpha = .005$; bee: $\beta = .10$, $t(1149) = 3.40$, $p < .001$, adjusted $\alpha = .012$), CVS-data score (frog: $\beta = .20$, $t(1162) = 6.85$, $p < .001$, adjusted $\alpha = .005$; bee: $\beta = .13$, $t(1145) = 4.45$, $p < .001$, adjusted $\alpha = .009$), and CVS CFC-data score (frog: $\beta = .19$, $t(1154) = 6.51$, $p < .001$, adjusted $\alpha = .006$; bee: $\beta = .13$, $t(1147) = 4.40$, $p < .001$, adjusted $\alpha = .010$) in both scenarios. The more information that was recorded in notes, the higher their performance was on science inquiry tasks, and the more likely that they had collected the evidence that was necessary for CVS use and hypothesis testing.

Use of CVS in notes and performance

Both the amount of CVS evidence and CVS CFC evidence recorded in notes (i.e., *CVS-data notes* and *CVS CFC-data notes*) were associated with higher science inquiry performance. Specifically, the more information necessary for CVS to test the correct claim that were recorded in notes, the higher students' CFC performance (frog: $e^B = 3.12$, $z = 10.77$, $p < .001$, adjusted $\alpha = .003$; bee: $e^B = 1.82$, $z = 6.61$, $p < .001$, adjusted $\alpha = .005$) and ISE performance (frog: $\beta = .34$, $t(1166) = 12.69$, $p < .001$, adjusted $\alpha < .001$; bee: $\beta = .23$, $t(1160) = 8.14$, $p < .001$, adjusted $\alpha = .004$), and the more frequently that they demonstrated the use of CVS in their behaviors to test all hypotheses (frog: $\beta = .27$, $t(1168) = 9.81$, $p < .001$, adjusted $\alpha = .003$; bee: $\beta = .23$, $t(1145) = 8.24$, $p < .001$, adjusted $\alpha = .003$) and the correct hypothesis (frog: $\beta = .33$, $t(1169) = 12.05$, $p < .001$, adjusted $\alpha = .001$; bee: $\beta = .26$, $t(1151) = 9.30$, $p < .001$, adjusted $\alpha = .002$). Similarly, writing more CVS-data notes was associated with better science inquiry performance in both scenarios.

Type of note and performance

In the frog scenario, the number of segments that involved direct reproduction of content presented in the environment was significantly positively associated with CFC, ISE, CVS-data, and CVS CFC-data scores (statistics are reported in Table 11) such that the more students engaged in content reproduction, the more successful they were in identifying a correct final claim, supporting their claim with evidence, and testing hypotheses using CVS. In the bee scenario, the number of content reproductive notes was not significantly associated with science inquiry performance (Table 12). In both the frog and bee scenarios, the more note segments where students elaborated on content presented in the environment and introduced new semantic information and ideas, the better their average science inquiry performance was.

Hypothesis/Conclusion notes and performance

Generating more hypotheses about the potential causal factors in notes was associated with a statistically significant increase in the CFC score (frog: $e^B = 1.21$, $z = 3.12$, $p = .002$, adjusted $\alpha = .014$; bee: $e^B = 1.21$, $z = 3.10$, $p = .002$, adjusted $\alpha = .014$), ISE score (frog: $\beta = .10$, $t(1129) = 3.79$, $p < .001$, adjusted $\alpha = .011$; bee: $\beta = .10$, $t(1148) = 3.06$, $p = .002$, adjusted $\alpha = .015$), CVS-data score (frog: $\beta = .07$, $t(1143) = 2.61$, $p = .009$, adjusted $\alpha = .018$; bee: $\beta = .10$, $t(1147) = 3.69$, $p < .001$, adjusted $\alpha = .011$), and CVS CFC-data score (frog: $\beta = .09$, $t(1147) = 3.04$, $p = .002$, adjusted $\alpha = .015$; bee: $\beta = .09$, $t(1148) = 3.27$, $p = .001$, adjusted $\alpha = .013$) in both scenarios. In addition, the quantity of notes where students drew conclusions from data was also significant predictors of student performance in the frog scenario.

Note source and performance

The number of segments based on information from laboratory test results was significantly associated with a higher likelihood of identifying the correct final claim in the frog scenario (e^B = 1.05, z = 2.49, p = .013, adjusted α = .019). It was also significantly positively associated with ISE (frog: β = .12, t(1157) = 4.30, p < .001, adjusted α = .010; bee: β = .10, t(1159) = 3.33, p < .001, adjusted α = .013), CVS-data (frog: β = .32, t(1166) = 11.62, p < .001, adjusted α < .001; bee: β = .27, t(1153) = 9.61, p < .001, adjusted α = .002), and CVS CFC-data (frog: β = .30, t(1168) = 11.02, p < .001, adjusted α = .002; bee: β = .25, t(1139) = 8.14, p < .001, adjusted α = .003) in both scenarios.

In addition to experiment notes, the number of sentences based on research kiosk was also positively associated with performance in the frog scenario (CFC: e^B = 1.08, z = 5.47, p < .001, adjusted α = .007; ISE: β = .22, t(1167) = 4.30, p < .001, adjusted α = .004; CVS-data: β = .09, t(1168) = 3.06, p = .002, adjusted α = .015; CVS CFC-data: β = .13, t(1151) = 4.25, p < .001, adjusted α = .010). The number of sentence segments based on information from the research kiosk was not significantly related to performance in the bee scenario.

In both scenarios, the number of sentences where students combined information from multiple sources was positively associated with performance. Specifically, the quantity of note segments where students copied or paraphrased content from various disparate sources was not significantly associated with performance on science inquiry tasks. However, the more students combined information from multiple sources and added new information and ideas to it (e.g., by generating inferences), the more successful students were at identifying supporting evidence for their final claim (frog: β = .07, t(1143) = 2.59, p = .010; bee: β = .09, t(1142) = 3.14, p = .002), and the better their performance on using CVS to test the correct final claim (frog: β = .09, t(1138) = 3.09, p = .002, adjusted α = .014; bee: β = .07, t(1142) = 2.35, p = .019, adjusted

α = .020) and all potential causal factors (frog: β = .09, t(1147) = 3.02, p = .003, adjusted α = .015; bee: β = .07, t(1142) = 2.39, p = .017, adjusted α = .020).

Discussion

Overall, results from this analysis indicated that the quantity of students' note-taking/reviewing behavior and specific contents of their notes tended to be positively associated with performance in the environment's frog scenario. First, the quantity of general notepad usage was significantly positively associated with student science inquiry performance in the frog scenario, such that the more frequently students opened the notepad and the more time spent on using the notepad, the better their science inquiry performance in the frog scenario. Note-takers outperformed non-note-takers on science inquiry tasks, suggesting that it is advantageous to self-initiate the note-taking process and make use of the digital notepad for fostering performance on science inquiry and problem-solving. These results were also consistent with the claims by Chi (2009) that active learners who engage in taking and reviewing notes are more successful in learning than passive learners who do not take/review notes, probably because the active action of using the digital notepad for note-taking/reviewing intensifies students' understanding of presented material and strengthens existing knowledge, which is more effective than passive processing of external information.

Thus, in the frog scenario, taking notes more frequently in the digital notepad, devoting more time to taking notes, and producing more notes (e.g., encoding more sentences or words in notes) were all associated with better performance on identifying the supporting evidence and collecting the data needed for CVS use. Taking notes more frequently and typing more notes on computers probably indicated that student attention to instructional content increased (Einstein et al., 1985), that more information was selected from the environment and transferred to text in notepad (Piolat et al., 2005), and that generative processing was involved and deeper-level mental representations of the instructional content were constructed (Bui et al., 2013; Piolat et al., 2005). These potentially help explain why taking notes in digital notepad alone was associated with better performance on science inquiry in the frog scenario. In turn, this finding also suggests that there are positive encoding benefits of note-taking and note quantity on performance, in open-ended learning environments as well as in previous research on classroom note-taking (Bretzing et al., 1987; Cohn et al., 1995; Kobayashi, 2005). It seems that taking notes in the OELE did not limit student exploration of the environment or impede meaningful learning and performance, unlike in Trevors et al. (2014) where note-taking in an OELE was found to interfere with deep learning and was detrimental to performance.

Further, reviewing notes more frequently and spending more time on note-reviewing episodes, which might indicate that students were retrieving the notes they had stored in the notepad, was associated with better science inquiry performance in the frog scenario. These results indicated that the crucial role of reviewing notes as external storage on performance found in the previous literature on paper-based note-taking (Kiewra et al., 1991; O'Donnell & Dansereau, 1993) was replicated in the frog scenario. It is worth noting the different context of note-rereviewing in this work than in earlier work: within the open-ended learning environment, note-reaccessing and note-reviewing occurred in tandem with note-taking to solve a science inquiry problem in real-time, whereas in most of the previous work these two occupied separate phases, with note-taking largely being completed before note-reaccess and note-review commenced. Results also revealed that a higher proportion of time within notepad distributed to note-reviewing episodes relative to note-taking was related to better performance on science inquiry in the frog scenario. This corresponded with previous findings that the external storage function of note-taking is relatively more important than the encoding function (Kiewra et al., 1991; Rickards & Friedman, 1978), indicating that reviewing notes as external storage seemed to be more crucial and valuable for performance than merely recording notes. The potential contributions of the external storage function to science inquiry performance in the frog scenario as suggested by the note-reviewing measures indicates that students should be encouraged to review notes frequently in conjunction with taking notes and be provided sufficient time and opportunities to reaccess and review notes to ensure optimal inquiry performance in the frog scenario.

However, there was no significant relationship between the quantity of taking or reviewing notes and science inquiry performance in the bee scenario except that the number of sentence segments recorded in notes was positively associated with CVS measures. In addition, the proportion of time spent on reviewing notes was negatively associated with performance. It is still unclear why the results did not generalize to the bee scenario. In order to understand why differences were found in the relationships between note-taking/reviewing and performance between the frog and the bee scenarios, it is important to understand what kinds of notes taken by students were important in these scenarios.

The measures on note content showing the strongest relationships with science inquiry performance included CVS CFC-data notes, CVS-data notes, and the quantity of unique information recorded in notes in both the frog scenario and the bee scenario. In both scenarios, the number of pieces of unique information recorded in notes was positively associated with science inquiry performance. In other words, the more information presented in the environment that was recorded in notes, the better students performed in identifying the correct causal claim, selecting supporting evidence for the final claim, and using the control of variables strategy. The positive relationship between the number of idea units recorded in paper-based notes and student performance found in lecture-based contexts (Peverly et al., 2007) also holds true in the open-ended computer-based learning environment for science inquiry. In addition, the more CVS evidence and CVS CFC evidence that were recorded in notes, the higher students' science inquiry performance was. These results suggested the importance of recording the controlled comparisons in notes beyond simply collecting the information.

In the *type of note* coding scheme, content reproductive note segments represent what Chi (2009) refers to in her ICAP framework as active learning, contrasted with passive learning where students do not take notes when they access representations. An additional category, taking content elaborative notes, is conceptualized as constructive learning as students connect new knowledge and information with existing knowledge, generate inferences, and infer patterns and conclusions from presented content. Chi proposed in her review that constructive learning is generally superior to active learning. Overall, a majority of the notes taken by students were verbatim copies or close paraphrases of content presented in the environment, around 20% of the note segments involved introduction of new ideas and information through elaboration, and a very small proportion of notes were metacognitive. This was consistent with previous findings in classrooms that reformulated notes were rarer than verbatim/paraphrased notes (Boch & Piolat, 2005; Bretzing & Kulhavy, 1981).

According to the results, the more content reproductive note segments taken by students, the better student science inquiry performance in the frog scenario. The relatively shallow level of processing that entails copying or paraphrasing content (as opposed to deeply processing the information by making inferences) was associated with science inquiry success in the frog scenario. These results contradicted previous findings that verbatim or reproductive note-taking was likely to limit exploration of the open-ended learning environment and exposure to relevant information, interfere with deep learning, and thus was negatively related with performance (Trevors et al., 2014). Although lacking deep processing, it is likely that the pure process of copying or paraphrasing content from the environment to the digital notepad without much alteration still strengthened memory for knowledge, reduced cognitive load, increased the probability of activating relevant prior knowledge, hence leading to better performance on the science inquiry tasks. In addition, it is also possible that the review of the reproduced notes ensured the fidelity of the content, and that the students with more reproductive notes produced more complete notes, which past work has shown to be related to good learning performance (Carter & Van Matre, 1975; Cohn et al., 1995). However, this positive relationship was once again not replicated in the bee scenario.

Furthermore, the more notes students took that entailed deep processing of content presented in the open-ended learning environment and the introduction of new semantic information and ideas, the better they built causal explanations in both scenarios. This was consistent with previous research that constructive learning strategies such as elaboration lead to superior learning outcomes than active and passive learning strategies (Chi, 2009). In general, generative note-taking entails increased mental effort, construction of deeper mental representations, and a higher level of engagement in problem solving than shallower processing such as verbatim copying, thereby leading to better performance (Slotte & Lonka, 1999; Trafton & Trickett, 2001).

In addition to the type of notes, some sources of content seemed to be associated with better outcomes than others. For example, results suggested that the quantity of note segments that could be traced to the research kiosk was positively associated with performance in the frog scenario, and recording more notes of the information from research kiosk was associated with improved performance, probably because it added to the understanding and memory for domain-specific declarative knowledge presented in the kiosk and facilitated construction of a solid knowledge base.

Similarly, the quantity of notes from laboratory test results was positively associated with science inquiry performance in both scenarios, even after controlling for the frequency of viewing test results itself. Therefore, beyond merely reading test results, students should also be encouraged to take notes on the test results, which would probably promote the understanding and interpretation of the results, and help students realize the connections between the test results and the problems to be solved.

In contrast, the quantity of note segments based on dialogue with NPCs and the quantity of note segments from field observation were not significant predictors of science inquiry performance. Compared to information from kiosk and laboratory tests, information from field observation and dialogue with NPCs is more salient but less reliable. For example, farmers in the virtual environment would express personal thoughts and opinions that were not dependent on scientific evidence, and which often contradicted each other and were of low scientific value. Therefore, relying too much on this unscientific information and taking notes of it was not beneficial for learning. Instead, students should learn to think critically and act like a scientist, relying more on scientific facts by looking up information from past research, and conducting controlled tests and analyzing test results for evidence, to distinguish relevant information from irrelevant information (Kuhn, 1999).

Elaborating on information collected from multiple sources in the environment was more strongly associated with performance than merely reproducing combined information. That is, simply putting information from various sources together in notes was not sufficient. Students also needed to elaborate on the internal connections between the noted information to achieve the best science inquiry performance. This finding suggested the importance of organizing and synthesizing information from disparate sources in notes and reconstructing internal connections across various categories of information for science inquiry performance in the environment.

Generating more hypotheses or drawing more conclusions in notes, which also reflects constructive learning (Chi, 2009), was positively associated with performance. This echoes McQuiggan et al.'s (2008) finding that high-performing students tended to generate hypothesis in notes within an open-ended learning environment. When teaching about note-taking strategies in OELEs, students could be taught to think more deeply about the content and construct hypotheses and conclusions in their notes to assist with science inquiry.

Differences between frog and bee results

The pattern of results seen in this paper was markedly different between the frog scenario and the bee scenario, two scenarios designed with the original goal of being highly similar. While there were many positive associations between measures on the quantity of note-taking/reviewing and content of notes and science inquiry performance in the frog scenario, in the bee scenario, only the quantity of sentence segments and specific types of notes seemed to be positively associated with differences in performance. Content elaboration notes (especially from the tests and combined content), notes from tests and combined sources, and elaborative hypothesis notes (especially based on tests, observation, and combined sources) were positively associated with science inquiry performance in the bee scenario.

I postulate that the differences in results between the two scenarios were most likely caused by the differences in the design of the two learning contexts, despite similar design goals.

First, I hypothesize that there are aspects in the design of the open-ended learning environment that make it more difficult for students to infer and justify the causal factors in the bee scenario than in the frog scenario, as indicated by the relatively lower average performance in the bee scenario than the frog scenario. Meanwhile, students spent significantly more time in the frog scenario than in the bee scenario (*M* = 30 min. 56 sec., *SD* = 14 min. 24 sec. vs. *M* = 27 min. 43 sec., *SD* = 11 min. 56 sec.), *t*(2402) = 5.36, *p* < .001. That is, students tended to spend less time in conducting scientific inquiry in the bee scenario and their performance on selecting supporting evidence was lower in this scenario than in the frog scenario. It is possible that the difference was due to the fact that students were more familiar with the concepts and terms used in the frog scenario (e.g., water sample, blood test, pH level, etc.) compared to those in the bee scenario (e.g., larva test, nectar sample), or that the design of the evidence and counter-evidence associated with different claims in the two scenarios were different in terms of complexity. Accordingly, it is possible that more cognitive effort is required to solve the scientific problems in the bee scenario, while students did not distribute sufficient time to the inquiry and problem-solving process in this scenario. On the other hand, students engaged in a similar amount of note-taking in both scenarios, as indicated by the quantitative and content measures of note-taking. With note-taking occupying a similar amount of cognitive effort in the two scenarios, students might not have sufficient working memory space to attend to the scientific inquiry and self-regulated learning in the bee scenario, if it demanded more effort than the frog scenario. Consequently, the effects of note-taking and note-reviewing on science inquiry performance were limited in the bee scenario as compared to the frog scenario. Second, the higher amount of time devoted to the frog scenario might suggest that students were more motivated in this scenario, considering the similar amount of information presented in the two environments. In the frog scenario, students were supposed to find which factor had caused the frog to grow six legs. In the bee scenario, they had to figure out what was causing the bees to die. It is possible that the topic and concepts related to a frog growing six legs were more concrete and interesting to middle school students than the topic and concepts involved in bee death. Motivation has been found to be related to note-taking (Moos, 2009). The different results for the two scenarios may, therefore, be related to a difference in motivation. Third, I posit that different levels of cognitive processing are required to solve problems in the two scenarios. In the bee scenario, the scientific problem is slightly more abstract and difficult, and only deep-level thinking and cognitive processing, which is more reflective of constructive learning, leads to identification of the correct final conclusion and justification of the claim with evidence. By contrast, probably because the frog scenario is relatively easier and less complex, both relatively superficial cognitive processing (e.g., through verbatim copying or closely paraphrasing information presented in the environment) and deeper-level elaboration in notes were beneficial for subsequent learning performance in this scenario. Accessing, understanding, recording, and reviewing more facts (e.g., research information), without necessarily making inferences and elaborating on them, will assist with science inquiry performance in the frog scenario. This would explain why both reproductive notes and elaborative notes were associated with science inquiry performance in the frog scenario, while mainly elaborative notes that entailed constructive learning was related to performance in the bee scenario. This difference was not intended in the original design and indicates how difficult it is to generate truly isomorphic problems in complex learning contexts such as Virtual Performance Assessments. Fourth, I speculate that the types of knowledge and information that are crucial in the two scenarios are different, leading to different results in these scenarios. It seems that the design of the environment’s bee scenario makes declarative knowledge obtained from the research kiosk less crucial for problem-solving than in the frog scenario. More specifically, information from the research kiosk is important for identification of parasites as cause of the frog mutation and justification of this claim, while the research kiosk information in the bee scenario is less essential for successful science inquiry. Correspondingly, note from kiosk pages was only positively related to performance in frog scenario. Considering the importance of kiosk information in the frog scenario, recording and possibly reviewing the research information strengthens students' declarative knowledge, thereby fostering performance. Correspondingly, the relative lower importance of research kiosk information in the bee scenario compared to the frog scenario might also explain partially why reviewing notes in the notepad was positively associated with performance in the frog scenario, but was not significantly associated with performance in the bee scenario. This hypothesis is also consistent with our result that reproducing kiosk information in notes was not significantly related to performance in the bee scenario, and the result that the frequency of reproductive notes was overall not a significant predictor of performance in this scenario.

Analysis 3.2: Development of Note-Taking/Reviewing Strategy in VPA

The second part of analysis 3 examines whether there are consistent changes in the quantity of note-taking and note-reviewing activities executed by students and the content of notes taken between the two VPA scenarios the student completed. In three-level models, each meaningful feature related to the SRL strategy – quantity of note-taking/note-reviewing or content of notes – served as the dependent variable. The student's previous experience with VPA and gender as well as their interaction were the predictor variables in each model.

Quantity of Note-Taking/Reviewing Behaviors

In this section, I examine whether there were consistent changes in the quantity of note-taking and note-reviewing activities executed by note-takers as they transitioned from one VPA scenario to another. Descriptive statistics of the variables on the note-taking/reviewing quantity for the four groups of students are reported in Table 13.

Table 13

Descriptive statistics (means with standard deviations in parentheses) of the features related to note-taking/reviewing quantity for female first-time users (F-1), female second-time users (F-2), male first-time users (M-1), and male second-time users (M-2) in each scenario

Scenario	**Feature**	**F-1**	**F-2**	**M-1**	**M-2**
Frog	Notepad access frequency	17.77 (15.76)	21.06 (18.37)	10.97 (10.17)	13.69 (13.68)
	Notepad time	5.41 (4.86)	6.91 (7.10)	3.72 (3.61)	4.31 (4.01)
	Percent of time on notepad	0.21 (0.23)	0.23 (0.22)	0.14 (0.14)	0.16 (0.17)
	Word count in note	66.51 (60.82)	76.27 (71.69)	42.77 (41.77)	48.88 (49.97)
	Segment count in note	9.07 (7.71)	11.14 (8.95)	6.19 (5.80)	7.25 (6.68)
	Note-taking frequency	12.60 (11.24)	15.71 (13.63)	7.76 (7.61)	10.46 (10.33)
	Note-reviewing frequency	5.11 (6.05)	5.57 (6.73)	3.23 (3.65)	3.57 (4.61)
	Percent note-taking actions	0.74 (0.19)	0.75 (0.18)	0.71 (0.21)	0.77 (0.20)
	Percent note-reviewing actions	0.26 (0.19)	0.25 (0.18)	0.29 (0.21)	0.23 (0.20)
	Note-taking duration	4.48 (3.73)	5.82 (5.64)	3.24 (3.11)	3.79 (3.43)
	Note-reviewing duration	0.92 (1.84)	1.09 (2.28)	0.49 (1.10)	0.52 (1.07)
	Avg note-taking duration	0.46 (0.55)	0.45 (0.39)	0.50 (0.34)	0.42 (0.22)
	Avg note-reviewing duration	0.13 (0.18)	0.14 (0.16)	0.12 (0.30)	0.10 (0.15)
	Note-taking to notepad time	0.87 (0.14)	0.87 (0.14)	0.89 (0.14)	0.90 (0.12)
	Note-reviewing to notepad time	0.13 (0.14)	0.13 (0.14)	0.11 (0.14)	0.10 (0.12)
Bee	Notepad access frequency	16.75 (14.23)	21.01 (17.54)	11.68 (10.89)	12.93 (14.02)
	Notepad time	5.22 (4.40)	6.54 (6.12)	3.63 (3.36)	4.18 (4.47)
	Percent of time on notepad	0.15 (0.10)	0.19 (0.11)	0.11 (0.09)	0.14 (0.11)
	Word count in note	59.72 (57.14)	76.73 (72.82)	39.96 (42.61)	47.98 (49.57)
	Segment count in note	8.63 (6.88)	11.04 (9.08)	5.95 (5.67)	7.24 (7.12)
	Note-taking frequency	11.97 (10.65)	15.85 (13.07)	8.18 (7.93)	9.81 (10.31)
	Note-reviewing frequency	4.77 (5.75)	5.24 (6.10)	3.43 (4.03)	3.31 (4.41)
	Percent note-taking actions	0.74 (0.20)	0.77 (0.17)	0.72 (0.21)	0.77 (0.18)
	Percent note-reviewing actions	0.26 (0.20)	0.23 (0.17)	0.28 (0.21)	0.23 (0.18)
	Note-taking duration	4.40 (3.71)	5.64 (5.08)	3.19 (3.03)	3.80 (3.58)
	Note-reviewing duration	0.81 (1.48)	0.93 (1.57)	0.45 (0.84)	0.55 (1.22)
	Avg note-taking duration	0.48 (0.46)	0.43 (0.40)	0.50 (0.51)	0.50 (0.49)
	Avg note-reviewing duration	0.14 (0.23)	0.15 (0.25)	0.10 (0.12)	0.11 (0.25)
	Note-taking to notepad time	0.87 (0.14)	0.88 (0.13)	0.89 (0.13)	0.91 (0.12)
	Note-reviewing to notepad time	0.13 (0.14)	0.12 (0.13)	0.11 (0.13)	0.09 (0.12)

Multilevel modeling results on the relationship between experience and gender towards these measures after applying Benjamini and Hochberg’s post-hoc control method are presented in Table 14. Overall, none of the interaction terms was statistically significant. Therefore, I will report the main effects in the following sections.

Table 14

Relationship between experience with VPA and gender towards features related to note-taking/reviewing quantity in each scenario

	Frog						**Bee**					
	Experience		**Gender**		**Experience × Gender**		**Experience**		**Gender**		**Experience × Gender**	
Feature	**B (SE)**	**t**	**B (SE)**	**t**	**B (SE)**	**t**	**B (SE)**	**t**	**B (SE)**	**t**	**B (SE)**	**t**
Notepad access frequency	3.48 (1.10)	3.15*	−5.19 (1.05)	−4.94*	−1.58 (1.71)	−0.93	3.13 (1.06)	2.94*	−4.50 (1.04)	−4.34*	−2.86 (1.70)	−1.69
Notepad time	1.47 (0.38)	3.89*	−1.12 (0.36)	−3.12*	−1.12 (0.58)	−1.92	1.04 (0.35)	2.97*	−1.30 (0.34)	−3.82*	−0.77 (0.56)	−1.38
Percent of time on notepad	0.02 (0.02)	1.06	−0.06 (0.01)	−4.37*	−0.002 (0.02)	−0.08	0.04 (0.01)	5.36*	−0.03 (0.01)	−4.24*	−0.02 (0.01)	−1.60
Word count in note	10.79 (4.38)	2.46*	−18.17 (4.17)	−4.36*	−7.25 (6.77)	−1.07	14.79 (4.28)	3.45*	−16.27 (4.19)	−3.89*	−9.12 (6.85)	−1.33
Segment count in note	2.19 (0.56)	3.89*	−2.04 (0.53)	−3.82*	−1.55 (0.87)	−1.78	2.02 (0.54)	3.72*	−2.26 (0.53)	−4.26*	−1.19 (0.87)	−1.37
Note-taking frequency	3.28 (0.82)	4.01*	−3.8 (0.78)	−4.88*	−1.17 (1.26)	−0.93	3.21 (0.80)	4.03*	−3.53 (0.78)	−4.55*	−2.23 (1.27)	−1.75
Note-reviewing frequency	0.48 (0.42)	1.14	−1.38 (0.40)	−3.50*	−0.38 (0.64)	−0.59	0.16 (0.39)	0.40	−1.12 (0.39)	−2.90*	−0.44 (0.63)	−0.70
Percent note-taking actions	0.01 (0.02)	0.92	−0.03 (0.01)	−1.83	0.04 (0.02)	1.64	0.03 (0.01)	2.35	−0.02 (0.01)	−1.41	0.01 (0.02)	0.42
Percent note-reviewing actions	−0.01 (0.02)	−0.92	0.03 (0.01)	1.83	−0.04 (0.02)	−1.64	−0.03 (0.01)	−2.35	0.02 (0.01)	1.41	−0.01 (0.02)	−0.42
Note-taking duration	1.32 (0.31)	4.29*	−0.88 (0.29)	−3.00*	−0.95 (0.47)	−2.01	1.03 (0.29)	3.50*	−1.04 (0.29)	−3.60*	−0.65 (0.47)	−1.37
Note-reviewing duration	0.17 (0.13)	1.28	−0.28 (0.12)	−2.27	−0.17 (0.20)	−0.84	0.08 (0.10)	0.76	−0.28 (0.10)	−2.82*	0.01 (0.16)	0.07
Avg note-taking duration	−0.02 (0.03)	−0.48	0.03 (0.03)	0.81	−0.05 (0.05)	−1.03	−0.04 (0.04)	−1.09	0.02 (0.03)	0.57	0.04 (0.06)	0.73
Avg note-reviewing duration	0.01 (0.02)	0.81	−0.01 (0.02)	−0.45	−0.04 (0.03)	−1.46	0.01 (0.02)	0.77	−0.04 (0.02)	−2.31	0.01 (0.03)	0.22
Note-taking to notepad time	0.004 (0.01)	0.37	0.01 (0.01)	1.05	0.01 (0.02)	0.69	0.01 (0.01)	0.89	0.01 (0.01)	1.08	0.01 (0.02)	0.58
Note-reviewing to notepad time	−0.004 (0.01)	−0.37	−0.01 (0.01)	−1.05	−0.01 (0.02)	−0.69	−0.01 (0.01)	−0.89	−0.01 (0.01)	−1.08	−0.01 (0.02)	−0.58

Note. Coefficient of the predictor (B), standard error associated with the coefficient (SE B), and t-statistics (*t*) are reported for each term (experience, gender, and experience × gender). Statistically significant results after Benjamini and Hochberg’s control are marked with *.

In both the frog scenario and the bee scenario, students who used VPA for the second time accessed the notepad significantly more frequently (frog: *Ms* = 18.11 and 14.85, $t(1151) = 3.15$, $p = .002$, adjusted $\alpha = .013$; bee: *Ms* = 18.16 and 14.48, $t(1154) = 2.94$, $p = .003$, adjusted $\alpha = .015$) and spent more time in the notepad (*Ms* = 5 minutes 52 seconds and 4 minutes 41 seconds, $t(1145) = 3.89$, $p < .001$, adjusted $\alpha = .007$; bee: *Ms* = 5 minutes 42 seconds and 4 minutes 30 seconds, $t(1158) = 2.97$, $p = .003$, adjusted $\alpha = .015$) on average than students who were new to the VPA environment. Furthermore, significant gender effect was found with the females accessing the notepad significantly more frequently (frog: *Ms* = 19.01 and 11.92, $t(1158) = -4.94$, $p < .001$, adjusted $\alpha = .001$; bee: *Ms* = 18.56 and 12.10, $t(1150) = -4.34$, $p < .001$, adjusted $\alpha = .004$) and spending more time in the notepad (*Ms* = 5 minutes 58 seconds and 3 minutes 56 seconds, $t(1151) = -3.12$, $p = .002$, adjusted $\alpha = .014$; bee: *Ms* = 5 minutes 47 seconds and 3 minutes 49 seconds, $t(1155) = -3.82$, $p < .001$, adjusted $\alpha = .009$) than male students.

Further analysis that distinguished note-taking activities from note-reviewing activities revealed consistent differences between first-time and second-time users in both scenarios. Among the note-takers, the second-time users opened the notepad to take notes more frequently than the first-time users in both the frog scenario (*M*s = 13.60 and 10.52, $t(1151) = 4.01$, $p < .001$, adjusted $\alpha = .007$) and the bee scenario (*Ms* = 13.72 and 10.27, $t(1156) = 4.03$, $p = .006$, adjusted $\alpha = .006$). Second-time users also devoted significantly more time to taking notes in the digital notepad than first-time users in both scenarios (frog: *M*s = 5 min. and 3 min., 57 sec., $t(1146) = 4.29$, $p < .001$, adjusted $\alpha = .004$; bee: *M*s = 5 min. and 3 min., 51 sec., $t(1159) = 3.50$, $p < .001$, adjusted $\alpha = .011$). In addition, notes recorded by the second-time users were comprised of significantly more words (frog: *M*s = 65.30 and 56.33, $t(1157) = 2.46$,

$p = .014$, adjusted $\alpha = .017$; bee: Ms = 66.60 and 50.88, $t(1157) = 3.45$, $p < .001$, adjusted $\alpha = .012$) and more sentences (frog: Ms = 9.65 and 7.88, $t(1159) = 3.89$, $p < .001$, adjusted $\alpha = .008$; bee: Ms = 9.72 and 7.49, $t(1157) = 3.72$, $p < .001$, adjusted $\alpha = .010$) on average than notes recorded by their first-time user counterparts. In both scenarios, female students took notes more frequently (frog: Ms = 13.77 and 8.71, $t(1158) = -4.88$, $p < .001$, adjusted $\alpha = .002$; bee: Ms = 13.62 and 8.72, $t(1152) = -4.55$, $p < .001$, adjusted $\alpha = .002$), spent more time taking notes (frog: Ms = 4 minutes 59 seconds and 3 minutes 26 seconds, $t(1154) = -3.00$, $p = .003$, adjusted $\alpha = .014$; bee: Ms = 4 minutes 56 seconds and 3 minutes 23 seconds, $t(1156) = -3.60$, $p < .001$, adjusted $\alpha = .010$), and wrote more words (frog: Ms = 70.20 and 44.91, $t(1165) = -4.36$, $p < .001$, adjusted $\alpha = .003$; bee: Ms = 66.96 and 42.63, $t(1158) = -3.89$, $p < .001$, adjusted $\alpha = .008$) and sentences (frog: Ms = 9.85 and 6.56, $t(1165) = -3.82$, $p < .001$, adjusted $\alpha = .009$; bee: Ms = 9.65 and 6.38, $t(1153) = -4.26$, $p < .001$, adjusted $\alpha = .005$) than male note-takers.

Although the second-time users recorded a higher quantity of notes than the first-time users, they did not review notes significantly more frequently than the first-time users in either scenario (frog: Ms = 4.77 and 4.30, $t(1164) = 1.14$, $p = .253$, adjusted $\alpha = .030$; bee: Ms = 4.56 and 4.17, $t(1156) = .40$, $p = .692$, adjusted $\alpha = .046$). Likewise, note-takers from the two groups spent a similar amount of time reviewing their notes (frog: Ms = 52 sec. and 44 sec., $t(1165) = 1.28$, $p = .199$, adjusted $\alpha = .029$; bee: Ms = 48 sec. and 39 sec., $t(1159) = .76$, $p = .449$, adjusted $\alpha = .039$). However, females again reviewed notes significantly more frequently (frog: Ms = 5.28 and 3.35, $t(1168) = -3.50$, $p < .001$, adjusted $\alpha = .012$; bee: Ms = 4.97 and 3.39, $t(1159) = -2.90$, $p = .004$, adjusted $\alpha = .016$) and spent marginally significantly or significantly more time reviewing notes than males (frog: Ms = 59 seconds and

30 seconds, *t*(1170) = −2.27, *p* = .023, adjusted *α* = .019; bee: *M*s = 52 seconds and 29 seconds, *t*(1162) = −2.82, *p* = .005, adjusted *α* = .017).

Note Content

Considering that students became more frequent note-takers and took a greater quantity of notes as they became experienced in using the VPA, it would be useful to further explore how the content of notes taken by students in VPA developed over time. For example, which type of notes did the second-time users and female students record more than the first-time users and male students, and how did the content and quality of notes differ across the groups? Results on the comparisons of note content across different groups of note-takers are reported in Table 15 and Table 16. Again, none of the interaction terms were statistically significant in either scenario.

Table 15

Descriptive statistics (means with standard deviations in parentheses) of the features related to note content for female first-time users (F-1), female second-time users (F-2), male first-time users (M-1), and male second-time users (M-2) by scenario

Scenario	**Note Content**	**F-1**	**F-2**	**M-1**	**M-2**
Frog	Information	11.45 (10.67)	14.34 (12.99)	7.33 (8.35)	9.28 (10.42)
	CVS CFC-data notes	0.60 (0.74)	0.71 (0.78)	0.50 (0.72)	0.49 (0.72)
	CVS-data notes	2.51 (2.59)	2.67 (2.91)	2.01 (2.44)	1.78 (2.41)
	Reproduction	7.35 (7.19)	9.19 (8.09)	4.54 (5.23)	5.66 (6.16)
	Elaboration	1.46 (2.12)	1.54 (2.72)	1.14 (1.65)	1.28 (1.98)
	Metacognition	0.08 (0.49)	0.06 (0.29)	0.12 (0.48)	0.14 (1.33)
	Test	2.15 (3.31)	1.78 (3.52)	1.78 (3.05)	1.53 (2.86)
	Kiosk	4.00 (5.42)	5.49 (6.62)	1.99 (3.16)	2.77 (4.23)
	Observation	2.47 (3.94)	3.51 (5.32)	1.74 (3.15)	2.50 (4.14)
	Dialogue	0.35 (1.43)	0.35 (1.61)	0.23 (1.01)	0.10 (0.66)
	Combination	0.25 (0.87)	0.38 (1.87)	0.12 (0.39)	0.19 (0.67)
	Hypothesis	0.51 (1.14)	0.53 (1.05)	0.50 (0.96)	0.54 (1.09)
	Draw Conclusion from Data	0.48 (1.05)	0.62 (1.30)	0.43 (1.02)	0.43 (1.07)
	Reproduction of Test	1.59 (2.70)	1.22 (2.58)	1.35 (2.54)	1.12 (2.28)
	Reproduction of Kiosk	3.61 (5.27)	4.91 (6.25)	1.73 (3.00)	2.50 (4.01)
	Reproduction of Observation	1.84 (3.39)	2.74 (4.67)	1.23 (2.61)	1.95 (3.61)
	Reproduction of Dialogue	0.30 (1.33)	0.33 (1.59)	0.22 (1.01)	0.09 (0.61)
	Reproduction of Combination	0.01 (0.10)	0.01 (0.11)	0.01 (0.08)	0.02 (0.17)

	Elaboration on Test	0.53 (1.18)	0.56 (1.82)	0.42 (1.06)	0.40 (0.97)
	Elaboration on Kiosk	0.37 (1.08)	0.55 (1.76)	0.21 (0.57)	0.27 (0.81)
	Elaboration on Observation	0.62 (1.25)	0.71 (1.43)	0.48 (1.03)	0.56 (1.32)
	Elaboration on Dialogue	0.04 (0.29)	0.01 (0.09)	0.01 (0.08)	0.01 (0.11)
	Elaboration on Combination	0.24 (0.87)	0.36 (1.85)	0.11 (0.38)	0.18 (0.61)
Bee	Information	11.28 (10.32)	15.32 (13.51)	7.31 (8.13)	9.43 (11.12)
	CVS CFC-data notes	0.43 (0.63)	0.78 (0.88)	0.40 (0.59)	0.56 (0.76)
	CVS-data notes	2.13 (2.39)	3.04 (3.23)	1.80 (2.29)	2.04 (2.50)
	Reproduction	7.17 (6.27)	9.28 (8.10)	4.65 (5.14)	5.26 (6.52)
	Elaboration	1.26 (1.96)	1.46 (2.18)	1.04 (1.66)	1.72 (2.41)
	Metacognition	0.05 (0.35)	0.06 (0.33)	0.13 (0.96)	0.08 (0.50)
	Test	1.62 (2.94)	2.47 (3.87)	1.56 (2.47)	1.74 (2.98)
	Kiosk	3.60 (5.07)	5.12 (6.21)	2.00 (3.64)	2.94 (4.56)
	Observation	2.89 (4.35)	3.32 (4.78)	2.10 (3.57)	2.26 (3.56)
	Dialogue	0.48 (1.74)	0.32 (1.74)	0.17 (0.86)	0.21 (0.88)
	Combination	0.28 (1.01)	0.48 (1.18)	0.22 (0.70)	0.42 (1.33)
	Hypothesis	0.34 (0.85)	0.47 (1.10)	0.34 (0.90)	0.70 (1.34)
	Draw Conclusion from Data	0.42 (1.17)	0.62 (1.23)	0.27 (0.80)	0.61 (1.26)
	Reproduction of Test	1.26 (2.45)	1.94 (3.19)	1.18 (2.10)	1.20 (2.25)
	Reproduction of Kiosk	3.26 (4.84)	4.56 (5.76)	1.72 (3.45)	2.41 (4.20)
	Reproduction of Observation	2.28 (3.76)	2.62 (4.13)	1.65 (3.20)	1.53 (3.01)
	Reproduction of Dialogue	0.46 (1.71)	0.30 (1.67)	0.13 (0.67)	0.20 (0.85)
	Reproduction of Combination	0.11 (0.65)	0.14 (0.55)	0.06 (0.34)	0.11 (0.53)
	Elaboration on Test	0.34 (0.90)	0.52 (1.12)	0.36 (0.80)	0.51 (1.33)
	Elaboration on Kiosk	0.31 (0.90)	0.46 (1.14)	0.25 (0.73)	0.52 (1.63)
	Elaboration on Observation	0.58 (1.28)	0.68 (1.38)	0.42 (1.09)	0.70 (1.43)
	Elaboration on Dialogue	0.02 (0.16)	0.02 (0.16)	0.04 (0.31)	0.01 (0.11)
	Elaboration on Combination	0.17 (0.64)	0.34 (1.01)	0.16 (0.60)	0.31 (1.22)

Amount of unique information recorded in notes

In both scenarios, the second-time users recorded a significantly greater quantity of meaningful information than the first-time users (see Table 15 and Table 16). In the frog scenario, the first-time users recorded an average of 9.68 pieces of information presented in the environment while the second-time users recorded an average of 12.32 pieces of information. The difference was statistically significant, *t*(1153) = 3.77, *p* < .001, adjusted *α* = .004. A similar pattern was found in the bee scenario, where the first-time users noted significantly more information that was presented in the environment than the second-time users (*M*s = 13.24 and 9.51, *t*(1151) = 4.14, *p* < .001, adjusted *α* = .003).

Similarly, females covered significantly more information in their notes than their male counterparts (frog: *M*s = 12.55 and 8.01, *t*(1159) = −3.99, *p* < .001, adjusted *α* = .004; bee: *Ms* = 13.00 and 8.02, *t*(1148) = −4.44, *p* < .001, adjusted *α* = .001).

Table 16

Relationship between experience with VPA and gender towards features related to note content in each scenario

	Frog						**Bee**					
	Experience		**Gender**		**Experience × Gender**		**Experience**		**Gender**		**Experience × Gender**	
Note Content	**B (SE)**	**t**	**B (SE)**	**t**	**B (SE)**	**t**	**B (SE)**	**t**	**B (SE)**	**t**	**B (SE)**	**t**
Information	3.09 (0.82)	3.77 *	−3.10 (0.78)	−3.99 *	−1.68 (1.27)	−1.33	3.38 (0.82)	4.14 *	−3.54 (0.80)	−4.44 *	−1.86 (1.30)	−1.42
CVS CFC-data notes	0.13 (0.06)	2.24	−0.06 (0.05)	−1.17	−0.15 (0.09)	−1.66	0.33 (0.05)	6.00 *	−0.02 (0.05)	−0.39	−0.19 (0.09)	−2.19
CVS-data notes	0.25 (0.20)	1.26	−0.31 (0.19)	−1.63	−0.55 (0.31)	−1.76	0.73 (0.20)	3.70 *	−0.27 (0.19)	−1.38	−0.63 (0.32)	−2.00
Reproduction	1.92 (0.52)	3.69 *	−2.15 (0.49)	−4.36 *	−1.12 (0.81)	−1.39	1.80 (0.49)	3.64 *	−2.24 (0.48)	−4.65 *	−1.59 (0.79)	−2.02
Elaboration	0.12 (0.17)	0.74	−0.23 (0.16)	−1.47	−0.01 (0.26)	−0.04	0.14 (0.15)	0.89	−0.13 (0.15)	−0.85	0.51 (0.25)	2.07
Metacognition	−0.02 (0.05)	−0.46	0.04 (0.05)	0.80	0.04 (0.08)	0.56	0.01 (0.05)	0.27	0.08 (0.05)	1.78	−0.06 (0.07)	−0.77
Test	−0.34 (0.25)	−1.33	−0.29 (0.24)	−1.21	0.04 (0.39)	0.10	0.77 (0.24)	3.23 *	−0.04 (0.23)	−0.17	−0.64 (0.38)	−1.68
Kiosk	1.53 (0.39)	3.95 *	−1.54 (0.37)	−4.18 *	−0.93 (0.60)	−1.55	1.38 (0.38)	3.66 *	−1.42 (0.37)	−3.87 *	−0.60 (0.60)	−1.01
Observation	1.10 (0.33)	3.37 *	−0.60 (0.31)	−1.96	−0.41 (0.51)	−0.80	0.36 (0.32)	1.12	−0.71 (0.31)	−2.25	−0.27 (0.52)	−0.52
Dialogue	<0.01 (0.10)	0.04	−0.09 (0.10)	−0.97	−0.14 (0.16)	−0.86	−0.18 (0.11)	−1.66	−0.26 (0.11)	−2.41	0.20 (0.18)	1.15
Combination	0.13 (0.08)	1.57	−0.11 (0.08)	−1.41	−0.06 (0.13)	−0.48	0.18 (0.08)	2.28	−0.02 (0.08)	−0.20	0.01 (0.13)	0.08
Hypothesis	0.01 (0.08)	0.17	−0.02 (0.08)	−0.22	0.03 (0.13)	0.22	0.13 (0.08)	1.70	<0.01 (0.08)	0.05	0.22 (0.12)	1.80
Draw Conclusion from Data	0.14 (0.09)	1.57	−0.04 (0.08)	−0.46	−0.15 (0.14)	−1.08	0.16 (0.09)	1.86	−0.10 (0.08)	−1.19	0.17 (0.14)	1.28
Reproduction of Test	−0.36 (0.20)	−1.77	−0.20 (0.19)	−1.04	0.10 (0.32)	0.33	0.61 (0.20)	3.14 *	−0.09 (0.19)	−0.47	−0.62 (0.31)	−1.98
Reproduction of Kiosk	1.32 (0.37)	3.55 *	−1.47 (0.35)	−4.15 *	−0.73 (0.58)	−1.25	1.17 (0.35)	3.30 *	−1.40 (0.35)	−4.05 *	−0.64 (0.57)	−1.14
Reproduction of Observation	0.96 (0.28)	3.41 *	−0.49 (0.27)	−1.84	−0.30 (0.44)	−0.70	0.30 (0.28)	1.07	−0.57 (0.27)	−2.08	−0.48 (0.45)	−1.08
Reproduction of Dialogue	0.03 (0.10)	0.28	−0.06 (0.09)	−0.63	−0.16 (0.15)	−1.08	−0.18 (0.11)	−1.70	−0.28 (0.10)	−2.71 *	0.23 (0.17)	1.36
Reproduction of Combination	<0.01 (0.01)	0.31	<0.01 (0.01)	−0.38	0.01 (0.01)	0.65	0.03 (0.04)	0.73	−0.05 (0.04)	−1.18	0.02 (0.07)	0.24
Elaboration on Test	0.05 (0.10)	0.49	−0.08 (0.10)	−0.86	−0.07 (0.16)	−0.43	0.15 (0.08)	1.97	0.05 (0.08)	0.60	−0.02 (0.12)	−0.14
Elaboration on Kiosk	0.19 (0.09)	2.06	−0.15 (0.08)	−1.80	−0.13 (0.14)	−0.93	0.13 (0.08)	1.64	−0.04 (0.08)	−0.52	0.14 (0.13)	1.08
Elaboration on Observation	0.09 (0.10)	0.92	−0.13 (0.09)	−1.43	−0.02 (0.15)	−0.13	0.07 (0.10)	0.73	−0.12 (0.10)	−1.27	0.20 (0.16)	1.24
Elaboration on Dialogue	−0.03 (0.01)	−2.17	−0.03 (0.01)	−2.16	0.04 (0.02)	1.63	<0.01 (0.02)	−0.01	0.02 (0.02)	1.15	−0.03 (0.03)	−1.04
Elaboration on Combination	0.13 (0.08)	1.55	−0.11 (0.08)	−1.38	−0.07 (0.13)	−0.55	0.16 (0.06)	2.44	0.02 (0.06)	0.26	−0.01 (0.10)	−0.09

CVS-data notes

In the bee scenario, second-time users recorded significantly more information that is necessary for them to use the control of variables strategy to test the correct final claim (i.e., radiation) (*Ms* = .71 and .42, *t*(1156) = 6.00, *p* < .001, adjusted *α* < .001). Second-time users also outperformed first-time users in recording significantly more information for CVS to test all potential hypotheses (*Ms* = 3.04 and 2.13, *t*(1152) = 3.70, *p* < .001, adjusted *α* = .005). However, this pattern was not replicated in the frog scenario, where the first-time users and the second-time users did not record a significantly different number of CVS CFC-data notes (*Ms* = .55 and .62, *t*(1162) = 2.24, *p* = .026, adjusted *α* = .011) or CVS-data notes (*Ms* = 2.29 and 2.32, *t*(1155) = 1.26, *p* = .209, adjusted *α* = .026).

Content reproduction and content elaboration

According to the results, the higher quantity of notes for second-time users compared to first-time users and for female students compared to male students was highly driven by the differences in the content reproductive notes. In both scenarios, the second-time users recorded significantly more sentence segments that were verbatim copies or close paraphrases of the content presented in the VPA environment than the first-time users (frog: *Ms* = 7.78 and 6.14, *t*(1155) = 3.69, *p* < .001, adjusted *α* = .005; bee: *Ms* = 7.86 and 6.05, *t*(1151) = 3.64, *p* = .003, adjusted *α* = .006). That is, students with previous experience using the other VPA scenario tended to reproduce more content presented in the learning environment into notes in the digital notepad without adding new semantic information or ideas than students who were newly exposed to the environment. Similarly, female note-takers recorded significantly more content reproductive sentence segments than male note-takers (frog: *Ms* = 8.05 and 4.93,

$t(1160) = −4.36$, $p < .001$, adjusted $\alpha = .002$; bee: $Ms = 8.07$ and 4.86, $t(1149) = −4.65$, $p < .001$, adjusted $\alpha = .001$).

No significant main effects for experience or gender were found for content elaborative notes in either the frog scenario or the bee scenario. According to Chi's (2009) Interactive-Constructive-Active-Passive (ICAP) framework, elaborative and generative note-taking is a constructive learning activity that involves deep cognitive processing, and it predicts superior academic achievement than note-taking that involves relatively shallower level of processing such as verbatim copying, though verbatim copying still constitutes an active learning activity (Armbruster, 2009). That is, previous experience in completing the other VPA scenario seemed to have only led students to copy or paraphrase more information in notes, but did not prompt students to go beyond the superficial meaning of the instructional content and process the information deeply. Similarly, female students did not show a different quantity of content elaborative notes and level of cognitive processing involved in note-taking than male note-takers.

Source of note content

Comparison of the source of note content between the two groups of students revealed main effects for both experience and gender. In both scenarios, the second-time users recorded more sentences based on research information from the kiosk than the first-time users (frog: $Ms = 4.41$ and 3.13, $t(1159) = 3.95$, $p < .001$, adjusted $\alpha = .004$; bee: $Ms = 4.35$ and 2.89, $t(1154) = 3.66$, $p < .001$, adjusted $\alpha = .005$). These results were in line with our previous finding that second-time users were more likely to access the notepad after reading kiosk pages. Accordingly, the second-time users tended to make use of the digital notepad to verbatim copy or paraphrase information from the research kiosk more than the first-time users (frog: $Ms = 3.95$ and 2.80, $t(1160) = 3.55$, $p < .001$, adjusted $\alpha = .006$; bee: $Ms = 3.80$ and 2.57, $t(1154) = 3.30$,

p < .001, adjusted α = .007). Females also wrote more sentences on kiosk information (frog: *Ms* = 4.56 and 2.26, t(1164) = −4.18, p < .001, adjusted α = .002; bee: *Ms* = 4.25 and 2.31, t(1153) = −3.87, p < .001, adjusted α = .004), and reproduced more content from the kiosk than males (frog: *Ms* = 4.56 and 2.26, t(1164) = −4.15, p < .001, adjusted α = .002; bee: *Ms* = 4.10 and 2.00, t(1154) = −4.05, p < .001, adjusted α = .003).

Beyond taking more notes on kiosk research information, second-time users also took more notes from other sources than first-time users. For example, in the frog scenario, the second-time users took more notes that were based on observations than the first-time users (*Ms* = 3.11 and 2.16, t(1166) = 3.37, p < .001, adjusted α = .007), whereas in the bee scenario, the second-time users recorded more sentences based on laboratory experiment results than the first-time users (*Ms* = 2.21 and 1.59, t(1160) = 3.23, p = .001, adjusted α = .007).

In the bee scenario, the second-time users also recorded marginally significantly more sentences where they elaborated on information from multiple sources than the first-time users (*Ms* = 0.34 and 0.17, t(1161) = 2.44, p = .015, adjusted α = .009).

Hypothesis/Conclusion

No significant main effects for gender or experience were found for notes that involve hypothesis making or drawing conclusions, which were found to be positively associated with science inquiry performance in VPA.

Discussion

This analysis examined the development of note-taking and note-reviewing, which are important self-regulatory strategies, from multiple perspectives and explored the role of gender in the development process. Results on the evolution of the quantity of note-taking/reviewing behaviors and note content further suggested the more frequent utilization of these learning strategies by the second-time users than the first-time users.

To begin with, while using VPA, note-takers increasingly made use of the digital notepad to take notes. In both scenarios, note-takers with previous experience in the other VPA scenario tended to engage in a significantly higher frequency of note-taking activities, spend significantly more time on taking notes in the notepad, and record significantly more words and sentences in the notes than their counterparts who were exposed to VPA for the first time. All these measures were positively associated with science inquiry performance in the frog scenario. However, a short session of using one VPA scenario was not sufficient to change students' note-reviewing patterns in the other scenario, which has been shown to be relatively more important for learning than the encoding benefits in the frog scenario. Second-time users were not more likely to review notes more frequently or spend more time reviewing notes than note-takers who used VPA for the first time.

More information was transferred from the VPA environment and encoded as notes in notepads, and more complete notes were produced by the second-time users, potentially strengthening their understanding and mental representations of the instructional content. Investigation of the content of notes taken by students indicated that the second-time users tended to reproduce instructional content presented in VPA more than the first-time users. Particularly, they were more likely to copy or paraphrase research information from kiosk pages, which could potentially facilitate construction of a solid knowledge base. Probably due to the differences in the content of the two scenarios, second-time users also encoded a higher quantity of notes related to observations in the frog scenario and test results in the bee scenario.

In the bee scenario, note-takers who used VPA for the second time engaged in generative note-taking and built internal connections between information obtained from various sources more than note-takers who used VPA for the first time. Elaboration on combined content corresponds to constructive learning, leading to deeper-level mental representations of the instructional content (Bui et al., 2013) and may have led to the better performance seen in the bee scenario on CFC (for both males and females) and ISE (for females) for the second-time users. However, this pattern was not replicated as students transitioned from the bee scenario to the frog scenario.

In the bee scenario, students with previous experience also tended to record more evidence necessary for the application of the control of variables strategy to test hypotheses than the first-time users. The amount of CVS evidence recorded in notes has been found to be crucial for science inquiry performance in both scenarios. That is, despite the fact that they did not collect more CVS evidence, students who used VPA for the second time were more opportunistic and tended to record more information important for problem-solving and CVS use in notes in the bee scenario.

It is worth mentioning that the development of note-taking strategies across scenarios was consistent for male and female note-takers. Both male and female note-takers seemed to learn to make better use of the note-taking strategies as they gained experience with using VPA.

In addition to the development of the note-taking/reviewing strategies over time, gender-related differences were also found in the quantity of note-taking/reviewing behaviors and the content of notes. Females seemed to show advantages on both functions of note-taking: taking notes and reviewing notes. They not only took notes and reviewed notes more frequently, but also captured more information in notes through reproducing the instructional content presented in VPA, especially from research kiosk. These findings are consistent with previous research, which suggests that females were better note-takers than males and took more notes in quantity over classroom lectures (Reddington et al., 2015). The relatively more sophisticated self-regulatory behaviors and note-taking/reviewing strategies seen for the female students over male students could potentially help explain their improvement of performance on science inquiry tasks such as CFC and ISE across scenarios.

However, gender was not related to generative note-taking, where a deeper level of cognitive processing is required. Males and females did not differ significantly in the number of content elaborative notes they took. Their notes also contained a similar number of segments where they constructed internal connections between multiple sources, generated hypotheses and drew conclusions, all of which were found to be related to better learning outcomes. This is again consistent with previous findings that females tended to copy information verbatim more often than males in lecture-based settings (Maddox & Hoole, 1975).

CHAPTER VIII.

GENERAL DISCUSSION

This dissertation studies self-regulated learning (SRL) as a dynamic process and traces the development of self-regulatory skills in an open-ended virtual environment for middle school science named Virtual Performance Assessments (VPA). Specifically, I focus on studying how the key processes and strategies of self-regulated learning develop in VPA and whether male and female students develop SRL skills differently. Combining educational data mining techniques such as sequential pattern mining and feature engineering with multilevel analysis, this dissertation involved three analyses to study the relationship between experience with VPA and students' gender on self-regulatory skills. Analysis 1 studied the development of science inquiry expertise across the course of using two VPA scenarios and the role of gender in the development process. Science inquiry is closely related to SRL (Sabourin, Mott, et al., 2013). Therefore, information on the development of science inquiry skills sheds light on the development of self-regulatory skills in the environment. Analysis 2 of this dissertation examined students' behaviors and strategies that were representative of self-regulatory processes in Winne and Hadwin's SRL model (Winne, 2011; Winne & Hadwin, 1998, 2009) by applying differential pattern mining on students' interaction log data. Differences in the behavioral patterns executed by first-time versus second-time users and by female versus male students provide insights into the development of self-regulatory behaviors and the potential gender differences in SRL processes. Lastly, Analysis 3 focused on examining students' note-taking and note-reviewing strategies, which is an important component of self-regulated learning. Measures representing both the quantity of note-taking and note-reviewing behaviors and the content of notes were generated from log data and human coding of note content, and their development across VPA scenarios for male and female students was explored.

Summary

Results from Analysis 1 showed a differential effect of experience with VPA on science inquiry performance by gender. In general, female students who had previously used the other VPA scenario demonstrated better performance on identifying a correct final claim and justifying the final claim with supporting evidence than female first-time users in both the frog scenario and the bee scenario. Nevertheless, male students showed similar or even lower performance on science inquiry tasks as they used VPA for the second time. Only the CFC performance for male second-time users in the bee scenario was significantly higher than that for male first-time users. It is still unclear why female students' inquiry skills improved over time within VPA while male students did not seem to improve their science inquiry performance except for the CFC score in the bee scenario. One possible explanation was that male students were enthusiastic about VPA as it was first introduced to classrooms due to the novelty effect, while the initial enthusiasm declined as they used it for the second time. However, it is still not clear why the results did not replicate for females, whose CFC and ISE performance improved over the use of VPA.

The lack of improvement in the quantity of CVS data collected by both male and female students was unexpected. Among males in the bee scenario, the second-time users even showed lower CVS-data and CVS CFC-data performance than the first-time users. In addition to the novelty effect, this could also be attributed to the limitations of the CVS measures. As pointed out in Chapter V, the lack of increase in the quantity of CVS evidence and CVS CFC evidence collected by students across the course of using VPA does not necessarily mean that students actually applied CVS as frequently or less frequently when they used VPA for the second time.

Despite the lack of increase in the CVS evidence and CVS CFC evidence collected by students, both male and female second-time users recorded more CVS evidence and CVS CFC evidence in their notes than first-time users in the bee scenario. This indicated that the second-time users saw a higher level of importance in the controlled comparisons and recorded more of them in their notes. As such, science inquiry should not be evaluated solely by CVS-relevant scores. Instead, we should combine these measures with results on other measures such as CFC and ISE.

Among the first-time users in both the frog scenario and the bee scenario, the males collected a significantly higher quantity of CVS CFC evidence and CVS evidence than the females. This is consistent with previous findings that males generally surpass females in motivation in science (X. Chen & Weko, 2009; Cunningham et al., 2015; Curran & Kellogg, 2016; Halpern, 2004; Mullis et al., 2000; National Center for Education Statistics, 2016; Neuschmidt et al., 2008; Quinn & Cooc, 2015; Reilly et al., 2015). However, this gender difference disappeared in the bee scenario when students used VPA for the second time, as the male second-time users collected a significantly lower quantity of CVS / CVS CFC evidence than their first-time user counterparts. Despite the more evidence that male first-time users collected for CVS use than female first-time users, there was no gender difference in their performance on identifying supporting evidence in both scenarios between male first-time users and female first-time users. However, a gender difference favoring females emerged for ISE performance as students used VPA for the second time. This result conflicts with the higher achievement found in science for males than females in previous literature (Cunningham et al., 2015; National Center for Education Statistics, 2016), and could be attributed to the lack of improvement in performance for males possibly due to the novelty effect.

In order to better understand the gender difference in the development of science inquiry skills, students' self-regulatory behavioral patterns and learning strategies were mapped to the various phases in SRL framework and were compared in Analysis 2. Despite the mixed results on science inquiry performance, results from Analysis 2 indicated that both male and female students gained skills in regulating their inquiry behaviors and adopted more successful self-regulatory strategies as they used VPA. As such, after just a half hour completing the first scenario, students demonstrated more expert-like SRL behaviors in their second scenario — they executed note-taking strategies more often, and were more opportunistic in using resources and exploited more available sources of information (e.g., laboratory test results, research information) to help them solve inquiry problems than the first-time users (Gilhooly et al., 1997). For instance, the second-time users were more likely to access the digital notepad after reading research information or running and viewing experiment results, probably to record the information they think was important, or to review notes to help them build connections with the information they just obtained from kiosk or tests. These notes were later accessed more frequently by the second-time users before they submitted the final claim, possibly to review the information that was recorded to assist them in making the final decision. As they became more experienced, students also tended to read research information more frequently after viewing laboratory test results, possibly to interpret the results using the domain-specific information presented to them in the kiosk. VPA also enabled students to develop skills in self-monitoring and self-assessment, by stimulating students to make better use of their notes taken during learning to monitor and reflect on their learning and solutions. In other words, students' skills in applying learning strategies and monitoring their own learning, both of which are crucial phases in SRL models, developed over the use of VPA. In contrast, the first-time users generally executed longer behavioral patterns comprised of exploratory behaviors such as talking with NPCs, manipulating objects, and collecting data, as compared to the second-time users. This, again, might be attributed to the novelty effect (cf. Kubota & Olstad, 1991). That is, the higher attention of the first-time users resulted in higher interest and efforts in exploring the new learning environment than students who were more experienced with VPA.

The development of SRL behavioral patterns was mostly consistent among female and male learners (i.e., the interaction between experience and gender was not significant). However, females showed consistent advantages in making use of the notepad to record information from the research kiosk pages and experiment results, to review notes in order to identify a final claim, and to monitor and evaluate the claims they just submitted regarding frog mutation or bee deaths. These behaviors could again be mapped to the application of learning strategies and monitoring phases of Winne and Hadwin’s (2009) SRL framework, suggesting that females were more self-regulated learners and demonstrated more sophisticated self-regulatory behaviors than male students. This corresponds to previous literature that female students reported themselves as using self-regulatory strategies more often than males (Lee, 2002; Matthews et al., 2009; Pajares, 2002; Yukselturk & Top, 2013; B. J. Zimmerman & Martinez-Pons, 1990). On the other hand, male students tended to show patterns of exploratory behaviors such as data collection and conducting laboratory experiments, which is consistent with our previous findings that males generally collected more evidence for the use of CVS than females among the first-time users.

As results from Analysis 2 suggested the development of student behavior patterns related to note-taking and note-reviewing, Analysis 3 further focused on examining note-taking and note-reviewing strategies from multiple perspectives. Note-taking, an important self-regulatory strategy, has been studied mostly in classroom settings on undergraduate students where learners take notes of lectures by hand. Few studies have studied note-taking in the context of open-ended learning environments for middle school students, and to the best of my knowledge, no study has examined the development of note-taking strategies and gender difference in the development in OELEs. First, I investigated the relationship between note-taking/reviewing and science inquiry performance in VPA. However, different results were found in this analysis between the frog scenario and the bee scenario. In the frog scenario, the quantity of note-taking and quantity of note-reviewing were both significantly positively associated with science inquiry performance, suggesting that the benefits of both taking notes in digital notepad (encoding function) and reviewing notes (external storage function) on facilitating science inquiry performance within the environment. That is, the two functions of note-taking seemed to extend beyond traditional simple learning measures, to boosting performance on complex science inquiry tasks in the open-ended learning environment. These results corresponded to Chi's claim that active note-taking is superior to passive learning, and contradicted previous research showing that taking notes in computer-based notepad and the quantity of note-taking in computers was negatively associated with performance because of the cognitive overload imposed by OELEs (e.g., Trevors et al., 2014). On the other hand, only the number of sentences recorded in notes was positively associated with performance in the bee scenario. In order to understand why the results in the frog scenario did not generalize to the bee scenario, it is important to understand what kinds of notes taken by students were important in these scenarios.

Examination of note content indicated that the amount of unique information recorded in notes was positively associated with science inquiry performance in both scenarios. The more evidence from controlled comparisons that were recorded in notes for the use of the control of variables strategy, the better student science inquiry performance was. Constructive learning (e.g., through making inferences in notes, combining disparate sources of information in the environment, and hypothesizing or drawing conclusions in notes) was also related to better performance, in line with Chi's ICAP framework. However, the number of sentences that involved a relatively shallower level of cognitive processing (e.g., content reproductive notes, reproduction of research kiosk information) was only associated with higher science inquiry performance in the frog scenario. I postulate that the differences in results between the two scenarios were most likely caused by the differences in the design of the two learning contexts, despite similar design goals. For example, the scientific problem in the bee scenario is slightly more abstract and difficult, and only deep-level thinking and cognitive processing, which is reflective of constructive learning, leads to identification of the correct final conclusion and justification of the claim with evidence. By contrast, probably because the frog scenario is relatively easier and less complex, both relatively superficial cognitive processing and deeper-level elaboration in notes were beneficial for subsequent science inquiry performance in this scenario. Accessing, understanding, recording, and reviewing more facts (e.g., research information), without necessarily making inferences and elaborating on them, will assist with science inquiry performance in the frog scenario. In addition, the differences in the types of knowledge and information that are crucial in the two scenarios might also lead to the different results in these scenarios. These differences were not intended in the original design and indicate how difficult it is to generate truly isomorphic problems in complex learning contexts such as Virtual Performance Assessments.

Results on the development of note-taking and reviewing strategies further affirmed that experience with the open-ended learning environment prepared students to adopt more efficient note-taking strategies to assist their self-regulated learning. They gradually learned to take notes more frequently and spend more time taking notes. However, using VPA was not sufficient to change note-takers' note-reviewing behaviors, although it has been found to be positively related to performance in the frog scenario. The second-time users also tended to take more complete notes that are comprised of a higher quantity of unique meaningful information and reproduce more important domain-specific knowledge information from kiosk research pages, behaviors that have been previously found to promote inquiry performance. Although the second-time users collected a similar or lower amount of CVS evidence necessary to test hypotheses than the first-time users, they seemed to be better at recognizing the importance of these information and noting them down in the digital notepad for later review and problem-solving in the bee scenario. Prior experience with using VPA also stimulated students to reproduce more content through verbatim copying or close paraphrasing. These results were consistent with findings from Analysis 2 and indicated that students gradually learned to be better note-takers who engaged in note-taking more frequently and took more complete notes.

However, a half hour using VPA was not sufficient to stimulate students to engage in deeper-level cognitive processing and content elaboration during note-taking through such techniques as generating inferences, constructing connections between information from various sources, and generating hypotheses and conclusions in notes.

Consistent with the previous findings from Analysis 2 on gender difference, Analysis 3 showed marked gender difference in both the quantity of note-taking/reviewing behaviors and the content of notes. Female note-takers accessed the notepad more frequently for both note-taking and note-reviewing purposes, captured a higher quantity of unique information from the environment, and reproduced more content from the research kiosk than male note-takers. These results indicated that the previous findings of advantages for females on paper-based note-taking in lecture-based settings transferred to the computer-based note-taking in the open-ended learning environment for science inquiry. The higher level of involvement in note-taking (represented by the higher frequency and more time the female students engaged in reviewing notes and the higher quantity of information recorded in notes by females over males) could possibly explain why the science inquiry performance improved within VPA for female students while the pattern was not replicated for males. The more information the female students recorded from the environment, especially those on the research kiosk, might have added to the domain-specific knowledge base of the female students, which is crucial for problem solving in VPA.

Implications

Theoretical Implications

First, the results from this dissertation contribute to the existing SRL literature by providing insights into how self-regulatory skills and strategies develop within OELEs for science learning among middle school students, and the role of gender in this process. Measures are developed based on students' performance, behaviors and strategies, and are mapped to the various processes in Winne and Hadwin's (2009) SRL framework. The results shed light on how each self-regulatory process and strategy developed over the use of the open-ended learning environment.

This research extends the existing note-taking literature by examining note-taking within an open-ended learning environment for middle school students in authentic classroom settings. Therefore, this study tests the robustness and generalizability of a broad list of findings from traditional research on paper-based note-taking in lectures and adds to the literature on computer-based note-taking in OELEs for science inquiry, which comprises a common learning activity nowadays. For example, the analysis on the relationship between note-taking in VPA and science inquiry performance extends the existing note-taking and SRL literature by examining the correspondence of various aspects of note-taking, including the quantity of note-taking/reviewing behaviors and the content of notes, with multiple measures of science inquiry performance in VPA, such as identifying supporting causal evidence and using the control of variables strategy. The existing coding scheme developed by Trevors and colleagues (2014) was revised and enriched to enable analysis of the content of notes more comprehensively than in past work in open-ended learning environments. In addition to the traditional measures of note content such as the quantity of information recorded in notes, I also evaluated the quantity of data needed for the control of variables strategy that was recorded in notes, the number of sentences where students generated hypotheses, and the source of note content, etc. This research also contributes to the note-taking literature by studying the *development* of note-taking and note-reviewing strategies comprehensively for middle school students in the open-ended learning environment. For instance, to my knowledge, this dissertation represents the first study that examines the quantity of note-reviewing behavior and its development over time. In addition to the quantity of note-reviewing behavior, the timing of note-taking and note-reviewing behaviors and its development was also studied through sequential pattern mining, which has rarely been explored in previous literature. Analysis on the development of both the quantity and content of notes is informative on how middle school students gradually developed effective note-taking and note-reviewing strategies in the open-ended learning environment.

Additionally, this dissertation contributes to the literature on gender gaps in science learning by exploring the role of gender in the development of self-regulatory skills. Results from this research provide insights into the gender differences in science inquiry skills, SRL behaviors, and note-taking and note-reviewing strategies in open-ended learning environments.

This aspect of the dissertation also makes methodological contributions; it shows the value of analyzing the rich log files from open-ended learning environments, and combining a data-driven approach such as educational data mining methods with traditional statistical analysis to study self-regulated learning. Specifically, this dissertation applied sequential pattern mining to identify behavioral patterns that represent the various phases of SRL and developed quantitative features that represent both the quantity and content of notes from log data produced by around 2,000 students. Multilevel analysis was then conducted to compare these measures between first-time users and second-time users, enabling a comprehensive analysis of the development of self-regulatory behaviors and strategies in this environment. The rich action log data allows me to study the development of SRL unobtrusively in real time at a fine-grained level. These analyses on the development of SRL in VPA, which was originally designed to assess middle school students' science inquiry skills rather than self-regulatory skills, also show the potential of applying educational data mining techniques to analyze data for purposes beyond what the system was designed for.

Empirical Implications

This dissertation on the development of SRL skills within OELEs also provides implications for the instructional design of open-ended learning environments such as virtual environments that assess science inquiry and learning of ill-structured science topics. VPA, an open-ended virtual environment without any scaffolding embedded, has been shown to foster self-regulated learning in this study. Furthermore, researchers have argued for the effectiveness of scaffolds in open-ended computer-based learning environments (Azevedo, 2005; Quintana et al., 2004; Segedy et al., 2015). An increasing number of personalized learning environments now include various types of support for students in developing SRL skills, including giving regular reports about whether students are demonstrating SRL (Arroyo et al., 2007) and providing immediate feedback when students demonstrate behaviors associated with poorer SRL (Roll, Aleven, McLaren, & Koedinger, 2007). Therefore, this dissertation may be of value to educational practice by providing insights into detecting the development of SRL in real time and designing adaptive scaffolding to further invoke self-regulatory behaviors and strategies in OELEs. In the following paragraphs, I will discuss the implications of the results for designing future open-ended virtual environments to facilitate personalized learning, science inquiry, and self-regulated learning.

To begin with, results from Analysis 1 indicated that students' use of the control of variables strategy did not improve over the use of VPA, although female students' ability in identifying the correct final claim and selecting supporting causal evidence improved. The lack of improvement in CVS skills might be due to the limitation of the CVS measures. However, it could also be possibly caused by the novelty effect and the fact that VPA did not explicitly teach CVS. Therefore, scaffolding could be designed and embedded in VPA to stimulate students to collect data relevant to the research questions (e.g., CVS evidence) in their inquiry process. For instance, prompts could be used to remind students to also conduct experiments on the six-legged frog, and more importantly to interpret and compare the test results if a student was found to only run experiments on the frogs from the four virtual farms.

Sequential pattern mining helped us identify a list of behavior patterns in VPA that mapped to various SRL phases and explore how they developed over time. Although I was able to detect behavior patterns related to understanding task definitions, tactic execution of learning strategies, and self-monitoring, little information was obtained regarding the goal setting and planning mechanism in the SRL cycle. Did students make plans to accomplish their tasks, and how did they execute and adaptively change their plans? How detailed and practical were their goals? As the planning process was not represented in the behavior sequences or notes, we do not have insights into whether VPA promoted students to better plan their inquiry and problem solving process. In turn, this would hinder our ability to provide adaptive scaffolding for goal setting and planning to students who are in need. To better evaluate and facilitate this process, online prompts and scaffolding could be implemented to enable students to set meaningful learning goals and subgoals, explicitly list their plans in the notepad after being introduced to VPA, and evaluate and adapt their plans in real time. For example, students whose plans were too general according to natural language processing results could be prompted to create more practical subgoals (e.g., list their subgoals on tool usage, data collection, and data analysis).

As students used the system for the second time in the frog scenario, they were less likely to take notes of their ultimate tasks in the notepad than students who used VPA for the first time. This might suggest that students were familiar with what they were supposed to do and did not need to record their overall goal the second time they used VPA. Given that understanding task definition is a key SRL mechanism, guiding questions could be used to evaluate student understanding of their tasks, direct student attention to their tasks and lead them to take notes of it when students' behaviors showed evidence of confusion or signs of being at a loss about what they should do (e.g., indicated by long pauses or repeated meaningless actions (Sabourin, Rowe, Mott, & Lester, 2011, 2013)). Implementing these prompts to ensure that users have a good understanding of their tasks is especially meaningful when students were exposed to VPA for the first time and not sure about what they should achieve.

In this research, the self-monitoring process was mainly deployed by students during the final assessment stage, where students reviewed notes or read kiosk pages to self-evaluate their final claims. However, the system could provide scaffolds and feedback to encourage students to engage in monitoring activities throughout the learning and scientific inquiry process. For example, VPA could periodically prompt students to report their self-evaluation of knowledge (e.g., how much they feel that they have understood the content presented in the environment) and their judgment of learning and adequacy of information collected for problem solving, enable students to mark their goals and subgoals as accomplished or incomplete, and display their progress toward the goals to students so that they could monitor their learning (Azevedo, 2005). Adaptive scaffolds could be provided based on students' self-reports as well as their behavior patterns on self-monitoring and self-evaluation.

This dissertation's findings also illuminate the instructional design of scaffolds to improve student utilization of learning strategies such as note-taking and note-reviewing. Students' use of note-taking and note-reviewing strategies could be scaffolded by embedding prompts related to the notepad. For instance, students can be encouraged by computer agents in real time to access notepad to take notes more frequently and type more notes in order to promote their understanding and learning if the system detects low notepad access or a low word count in notes. Prompting students to reaccess notes as external storage frequently and encouraging them to spend sufficient time reviewing notes in the digital notepad might also be beneficial for academic success. As recording unique information and recording the data necessary for CVS use in notes are important to science inquiry performance, students should be encouraged to take these types of notes. For example, as students read results of potential controlled comparisons, scaffolding could be embedded to guide students to conduct CVS, link and compare the results, and record the results from controlled comparisons in notes. If a low frequency of notepad access is recorded after reading kiosk pages or running experiments, appropriate cues or prompts can be provided to encourage students to take notes of these contents that are crucial for problem solving. This is especially true in the frog scenario, as students did not record more sentences on experiment results as they used VPA for the second time, whereas notes on experiments were positively associated with science inquiry performance in this scenario. Such prompts may be less necessary when students access information of lower-importance, such as talking with NPCs, in order to avoid encouraging less effective note-taking strategies.

Similarly, these analyses provide insights into designing scaffolds to foster *generative* note-taking in VPA by encouraging the use of strategies such as connecting, generating inferences, and hypothesizing. For example, as behavior patterns where students access information from two different sources (e.g., reading kiosk page followed by viewing test results, or viewing both genetic test results and blood test results) are identified, the system could prompt students to go beyond verbatim copying or closely paraphrasing the content, and to delve deeper into the underlying meanings of the information and construct connections to interpret the test results based on the information in the research page, or compare the results from two tests. Students also should be prompted to generate hypotheses and conclusions from the data they collected, which they failed to improve across the two scenarios of VPA.

In addition, findings on gender-related differences in the development of self-regulated learning are practically meaningful to educators and instructional designers in designing personalized scaffolding to support the development of self-regulatory strategies and processes for both males and female learners. In this study (and other studies on note-taking in traditional classroom settings), males were found to take notes and review notes less frequently than females, and their notes included a lower amount of meaningful information, especially the research information on the kiosk. Considering the importance of note-taking, it could be beneficial to prompt male students to access the notepad more frequently, both for note-taking and note-reviewing purposes, and also remind them to record more information presented in the environment. Despite the fact that females engaged in more note-taking/reviewing and took more reproductive notes, they did not engage in a deeper level of cognitive processing and generative note-taking than their male counterparts. Therefore, instructional designers should embed scaffolding to stimulate constructive learning for both male and female students. Students should be asked to elaborate on certain data sources (e.g., connect content on lab tests with existing knowledge) in the environment, generate more hypothesis, and draw more conclusions from data. These findings will also have implications for closing gender gaps in self-regulated learning and science education.

The personalized scaffolds proposed above are meant to prompt and support students' self-regulatory processes and strategies in VPA in real time. They would be embedded in a broader design where they were introduced because of evidence of student need, and then gradually faded as the student demonstrated the relevant skills (as in Roll et al., 2007), so that students would emerge from their experience using the system with more generalizable self-regulated learning skill. Through introducing adaptive scaffolds that fade as the student demonstrates skill, it may be possible to enhance students' self-regulated learning in open-ended virtual assessments, benefitting not just their performance on the assessments, but what they take away from the experience.

Limitations and Future Directions

Although this study produced many findings, several findings were inconsistent between the two scenarios studied. For example, many of the positive relationships between note-taking/reviewing (especially the quantity of note-taking/reviewing behaviors) and science inquiry performance found in the frog scenario were not replicated in the bee scenario, which was highly structurally similar as the frog scenario. As a result, many of the findings were only partially supported. Further investigation should be conducted with content experts and instructional designers to examine whether it was the design of the bee scenario or other elements specific to the bee scenario that caused the different results. Investigating the relationship between the difficulty level of instructional content, cognitive load, and the effectiveness of note-taking/reviewing would also be meaningful to understand the different results. This would also provide insights into the instructional design of open-ended learning environments to maximize the benefits of digital note-taking. In addition, improvement of performance on identifying a correct final claim (CFC) and identifying the supporting evidence for their claim (ISE) was found for the female learners while students' performance on the use of CVS did not improve. Future studies should prompt students to explain their problem-solving strategies in VPA or collect think-aloud data to better understand whether students conducted controlled comparisons for problem solving in VPA or not. Furthermore, different results on the development of science inquiry expertise were obtained for male and female students. Female learners' performance on identifying a correct final claim and justifying the claim with supporting evidence improved over the use of VPA whereas male learners' science inquiry performance did not improve in general. I hypothesize that the difference in the development of science inquiry could be caused by the novelty effect for male students, or by the higher involvement in self-regulatory behaviors such as note-taking and the more information recorded in notes by females. Future research should be conducted to test these hypotheses. For instance, studies could be conducted to investigate whether prompting male students to take more notes would improve their science inquiry performance in the open-ended learning environment as they use the system the second time or not.

Additionally, replication and extension of these results should be conducted to validate the generalizability of the results across platforms, domain topics, populations, and types of tasks. For example, the comparison conducted here involved virtual scenarios within the same VPA architecture. The fact that the two scenarios were highly structurally similar might have facilitated the development and later demonstration of self-regulatory skills. Future work may involve exploring whether the development of SRL skills is robust or not. For instance, it is worth studying how SRL skills develop from VPA to assessments outside the system (e.g., other computer-based learning environments with different domain and interaction design). In addition, the current study examined the development of SRL skills across two VPA scenarios, each of which lasted for approximately 30 minutes. Future studies could explore how science inquiry performance, self-regulatory behavioral patterns, and computer-based note-taking/reviewing strategies develop over a long term, and how they are related to delayed learning outcomes and robust learning. Furthermore, the present study examined the development of self-regulatory skills and note-taking and note-reviewing strategies for middle school students, while most of the previous research on note-taking has focused on older populations (e.g., undergraduates and adults). Research has shown that even undergraduate learners have difficulties in applying self-regulatory strategies such as note-taking strategies effectively and may need additional scaffolding (Kiewra & Fletcher, 1984; Moos, 2009; Peverly et al., 2003; Piolat et al., 2005). As such, it is worth asking whether the results obtained among middle school students, who typically exhibit less sophisticated note-taking and self-regulated learning skills, also apply to older adults. Therefore, further analyses testing whether these results transfer to other science open-ended learning environments, for older and more proficient learners, on a variety of learning tasks, would be extremely informative and may help us understand whether younger students need different support for self-regulated learning than older students.

Another possible limitation of the study is the validity of measures such as note-reviewing frequency and CVS performance. In this research, note-reviewing was defined as opening the notepad without adding or changing the content of the notes, as an indicator of the external storage function. It is worth asking how closely these findings link to the broader behavior of note-reviewing. Although reaccessing notes is the first step for note-reviewing, it does not necessarily mean that students were reviewing notes. For example, a student might rapidly open and close the notepad without spending time reading the notes. Alternate operationalizations, such as opening the notepad for a minimum amount of time, could be considered. It is also possible that a student also reviewed notes before or after entering information in the notepad while the notepad remained open. Future analysis could combine log data with eye-tracking data to provide a more valid and reliable measure of the frequency of reviewing notes that excludes actions where notes were accessed but not reviewed. There are also limitations on the measures related to CVS, such as CVS-data score and CVS CFC-data score. As mentioned in Chapter V, students might not necessarily use the control of variables strategy even if they collected the data necessary for CVS use.

Furthermore, the online notepad provided to learners in this study is a plain text editor where students can enter any text. No additional features of existing popular note-taking applications such as collaborative note-taking, annotation, providing skeletal outlines for note-taking, and creating hierarchical lists, have been embedded within the notepad (Bauer & Koedinger, 2006; Kauffman, Zhao, & Yang, 2011). In addition, students cannot create graphs in the notepad, which could be easily achieved in paper-based note-taking. Potential future research includes exploring the effects of enabling these features on note-taking and performance. For instance, Chi's (2009) ICAP framework indicated that collaborative learning activities are superior to constructive, active, or passive learning activities. Embedding the collaborative note-taking feature in notepad would enable us to further test Chi's ICAP framework and compare collaborative note-taking with constructive note-taking and active note-taking.

In this work, different behavioral patterns during the scientific investigation process were detected for male and female learners. While male students collected more data and conducted more experiments, female students recorded more information that are important in the environment into their notes and tended to make more sophisticated use of learning strategies and resources to interpret the laboratory test results. Future work could potentially involve building machine-learned predictive models of a student's gender using students' behavioral patterns and other meaningful features on the inquiry behaviors. These models would provide us with a better understanding of gender differences in science inquiry and self-regulated learning.

Finally, research has suggested that motivation is an important component of self-regulated learning (Moos, 2009; Moos & Azevedo, 2008a; B. J. Zimmerman, 2008). Motivation also plays an important role in why students take notes and how notes influence learning (Moos, 2009). Future research should also include examining the development of motivation as a component of self-regulated learning in open-ended learning environments such as VPA and the role of gender in this process. This research would potentially add to the current dissertation study and provide a more comprehensive understanding of how self-regulatory skills develop in open-ended learning environments for both male and female learners.

REFERENCES

Agrawal, R., & Srikant, R. (1995). Mining sequential patterns. *Proceedings of the 11th IEEE International Conference on Data Engineering*, 3-14.

Aleven, V., Roll, I., McLaren, B. M., & Koedinger, K. (2010). Automated, unobtrusive, action-by-action assessment of self-regulation during learning with an intelligent tutoring system. *Educational Psychologist, 45*(4), 224–233.

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Education Psychology, 103*(1), 1-18.

Annis, L. F., & Annis, D. B. (1987). *Does practice make perfect? The effects of repetition on student learning*. Paper presented at the the Annual Meeting of the American Educational Research Association, Washington, D.C.

Armbruster, B. B. (2009). Taking notes from lectures. In R. F. Flippo & D. C. Caverly (Eds.), *Handbook of college reading and study strategy research* (pp. 220-248). New York, NY: Routledge.

Arroyo, I., Ferguson, K., Johns, J., Dragon, T., Meheranian, H., Fisher, D., . . . Woolf, B. P. (2007). Repairing disengagement with non-invasive interventions. In *Proceedings of the 13th International Conference on Artificial Intelligence in Education* (pp. 195-202).

Azevedo, R. (2005). Using hypermedia as a metacognitive tool for enhancing student learning? The role of self-regulated learning. *Educational Psychologist, 40*(4), 199-209.

Baker, D. (2002). Where is gender and equity in science education? *Journal of Research in Science Teaching, 39*(8), 659 - 663.

Baker, R. S., Clarke-Midura, J., & Ocumpaugh, J. (2016). Towards general models of effective science inquiry in virtual performance assessments. *Journal of Computer Assisted Learning, 32*(3), 267-280.

Baker, R. S., & Yacef, K. (2009). The state of Educational Data Mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1), 3-17.

Basol, G., & Balgalmis, E. (2016). A multivariate investigation of gender differences in the number of online tests received-checking for perceived self-regulation. *Computers in Human Behavior, 58*, 388-397.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi:10.18637/jss.v067.i01

Bauer, A. (2008). *Designing note-taking interfaces for learning*. Carnegie Mellon University. Pittsburgh, PA.

Bauer, A., & Koedinger, K. (2006). Pasting and encoding: Note-taking in online courses. In *Proceedings of the Sixth IEEE International Conference on Advanced Learning Technologies (ICALT'06)* (pp. 789-793). Kerkrade, The Netherlands: IEEE Computer Society.

Bazaldua, D. A. L., Baker, R. S., & San Pedro, M. O. Z. (2014). Comparing expert and metric-based assessments of association rule interestingness. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 44-51).

Benbow, C. P., & Arjmand, O. (1990). Predictors of high academic achievement in mathematics and science by mathematically talented students: A longitudinal study. *Journal of Educational Psychology, 82*(3), 430–441.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289-300.

Boch, F., & Piolat, A. (2005). Note taking and learning: A summary of research. *The WAC Journal, 16*, 101-113.

Boekaerts, M., Pintrich, P. R., & Zeidner, M. (2000). *Handbook of self-regulation.* San Diego, CA: Academic Press.

Bonner, J. M., & Holliday, W. G. (2006). How college science students engage in note-taking strategies. *Journal of Research in Science Teaching, 43*(8), 786–818.

Bouchet, F., Azevedo, R., Kinnebrew, J. S., & Biswas, G. (2012). Identifying students' characteristic learning behaviors in an Intelligent Tutoring System fostering self-regulated learning. In K. Yacef, O. Zaïane, A. Hershkovitz, M. Yudelson, & J. Stamper (Eds.), *Proceedings of the 5th International Conference on Educational Data Mining* (pp. 65-72). Chania, Greece: International Educational Data Mining Society.

Bouchet, F., Harley, J. M., Trevors, G. J., & Azevedo, R. (2013). Clustering and profiling students according to their interactions with an intelligent tutoring system fostering self-regulated learning. *Journal of Educational Data Mining, 5*(1), 104-146.

Bretzing, B. H., & Kulhavy, R. W. (1979). Notetaking and depth of processing. *Contemporary Educational Psychology, 4*(2), 145-153.

Bretzing, B. H., & Kulhavy, R. W. (1981). Note-taking and passage style. *Journal of Educational Psychology, 73*(2), 242-250.

Bretzing, B. H., Kulhavy, R. W., & Caterino, L. C. (1987). Notetaking by junior high students. *The Journal of Educational Research, 80*(6), 359-362.

Britner, S. L. (2008). Motivation in high school science students: A comparison of gender differences in life, physical, and earth science classes. *Journal of Research in Science Teaching, 45*(8), 955 - 970.

Bromage, B. K., & Mayer, R. E. (1986). Quantitative and qualitative effects of repetition on learning from technical text. *Journal of Educational Psychology, 78*(4), 271-278.

Brown, C. M. (1988). Comparison of typing and handwriting in “two-finger typists". *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 32*(5), 381–385.

Bui, D. C., Myerson, J., & Hale, S. (2013). Note-taking with computers: Exploring alternative strategies for improved recall. *Journal of Educational Psychology, 105*(2), 299-309.

Carrier, C. A., Williams, M. D., & Dalgaard, B. R. (1988). College students' perceptions of notetaking and their relationship to selected learner characteristics and course achievement. *Research in Higher Education, 28*(3), 223-239.

Carter, J. F., & Van Matre, N. H. (1975). Note taking versus note having. *Journal of Educational Psychology, 67*(6), 900-904.

Chen, X., & Weko, T. (2009) Students who study science, technology, engineering, and math (STEM) in postsecondary education (NCES 2009–161). In. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development, 70*(5), 1098-1120.

Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The journal of the learning sciences, 6*(3), 271-315.

Chi, M. T. H. (2009). Active ‐ constructive ‐ interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1*(1), 73-105.

Clark, R. E. (1983). Reconsidering research on learning from media. *Review of Educational Research, 53*(4), 445-459.

Clarke-Midura, J., & Dede, C. (2010). Assessment, technology, and change. *Journal of Research on Technology in Education, 42*(3), 309-328.

Clarke-Midura, J., Dede, C., & Norton, J. (2011). Next generation assessments for measuring complex learning in science. In *The Road Ahead for State Assessments*. MA: Rennie Center for Education Research & Policy.

Clarke-Midura, J., McCall, M., & Dede, C. (2012, February). *Designing virtual performance assessments.* Paper presented at the meeting of the American Association for the Advancement of Science, Vancouver, Canada.

Clarke-Midura, J., & Yudelson, M. V. (2013). *Towards Identifying Students' Causal Reasoning Using Machine Learning.* Paper presented at the Proceedings of the 16th International Conference on Artificial Intelligence in Education (AIED 2013), Berlin, Heidelberg.

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

Cohn, E., Cohn, S., & Bradley, J. (1995). Notetaking, working memory, and learning in principles of economics. *The Journal of Economic Education, 26*(4), 291-307.

Conway, M. A., & Gathercole, S. E. (1990). Writing and long-term memory: Evidence for a "translation" hypothesis. *The Quarterly Journal of Experimental Psychology Section A, 42*(3), 513-527.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 671-684.

Crawford, C. C. (1925). Some experimental studies on the results of college note-taking. *The Journal of Educational Research, 12*(5), 379-386.

Crooks, S. M., White, D. R., & Barnard, L. (2007). Factors influencing the effectiveness of note taking on computer-based graphic organizers. *Journal of Educational Computing Research, 37*(4), 369-391.

Cuban, L. (1986). *Teachers and Machines: The Classroom Use of Technology since 1920*. New York, NY: Teachers College Press.

Cunningham, B. C., Hoyer, K. M., & Sparks, D. (2015). *Gender differences in Science, Technology, Engineering, and Mathematics (STEM) interest, credits earned, and NAEP performance in the 12th grade (NCES 2015-075)*. Washington, DC: National Center for Education Statistics.

Curran, F. C., & Kellogg, A. T. (2016). Understanding science achievement gaps by race/ethnicity and gender in kindergarten and first grade. *Educational Researcher, 45*(5), 273-282.

Di Vesta, F. J., & Gray, G. S. (1972). Listening and notetaking. *Journal of Educational Psychology, 63*(1), 8-14.

Einstein, G. O., Morris, J., & Smith, S. (1985). Note-taking, individual differences, and memory for lecture information. *Journal of Educational Psychology, 77*(5), 522-532.

Else-Quest, N. M., Mineo, C. C., & Higgins, A. (2013). Math and science attitudes and achievement at the intersection of gnder and ehnicity. *Psychology of Women Quarterly, 37*(3), 293-309.

English, H. B., Welborn, E. L., & Killian, C. D. (1934). Studies in Substance Memorization. *Journal of General Psychology, 11*, 233-260.

Erwin, L., & Maurutto, P. (1998). Beyond access: considering gender deficits in science education. *Gender and Education, 10*(1), 51–69.

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin, 116*(3), 429–456.

Fisher, J. L., & Harris, M. B. (1973). Effect of note taking and review on recall. *Journal of Educational Psychology, 65*(3), 321-325.

Foos, P. W., Mora, J. J., & Tkacz, S. (1994). Student study techniques and the generation effect. *Journal of Educational Psychology, 86*(4), 567-576.

Forsterlee, L., & Horowitz, I. A. (1997). Enhancing juror competence in a complex trial. *Applied Cognitive Psychology, 11*(4), 305–319.

Forsterlee, L., Kent, L., & Horowitz, I. A. (2005). The cognitive effects of jury aids on decision‐making in complex civil litigation. *Applied Cognitive Psychology, 19*(7), 867-884.

Gilhooly, K. J., McGeorge, P., Hunter, J., Rawles, J. M., Kirby, I. K., Green, C., & Wynn, V. (1997). Biomedical knowledge in diagnostic thinking: The case of electrocardiogram (ECG) interpretation. *European Journal of Cognitive Psychology, 9*(2), 199-223.

Gobert, J. D., & Koedinger, K. (2011). *Using Model-Tracing to Conduct Performance Assessment of Students' Inquiry Skills within a Microworld*. Paper presented at the Paper presented at the Society for Research on Educational Effectiveness (SREE).

Gobert, J. D., Sao Pedro, M. A., Baker, R. S. J. d., Toto, E., & Montalvo, O. (2012). Leveraging educational data mining for real time performance assessment of scientific inquiry skills within microworlds. *Journal of Educational Data Mining, 4*(1), 111-143.

Greene, J. A., & Azevedo, R. (2007). A theoretical review of Winne and Hadwin’s model of self-regulated learning: New perspectives and directions. *Review of Educational Research, 77*(3), 334-372.

Greene, J. A., Moos, D. C., Azevedo, R., & Winters, F. I. (2008). Exploring differences between gifted and grade-level students’ use of self-regulatory learning processes with hypermedia. *Computers & Education, 50*(3), 1069–1083.

Griffith, A. L. (2010). Persistence of women and minorities in STEM field majors: Is it the school that matters. *Economics of Education Review, 29*, 911–922.

Hahsler, M., Gruen, B., & Hornik, K. (2005). Arules -- A computational environment for mining association rules and frequent item sets. *Journal of Statistical Software, 14*(15), 1-25.

Halpern, D. F. (1997). Sex differences in intelligence: Implications for education. *American Psychologist*, *52*(10), 1091–1102.

Halpern, D. F. (2004). A cognitive-process taxonomy for sex differences in cognitive abilities. *Current Directions in Psychological Science, 13*(4), 135-139.

Hannafin, M. J. (1995). Open-ended learning environments: Foundations, assumptions, and implications for automated design. In R. D. Tennyson & A. E. Barron (Eds.), *Automating Instructional Design: Computer-Based Development and Delivery Tools* (Vol. 140). Berlin, Heidelberg: Springer-Verlag

Hannafin, M. J., Land, S. M., & Oliver, K. (1999). Open learning environments. In C. M. Reigeluth (Ed.), *Instructional-Design Theories and Models: A New Paradigm of Instructional Theory* (Vol. II, pp. 115-140): Mahwah, N.J: Lawrence Erlbaum Associates.

Hartley, J., & Davies, I. K. (1978). Note-taking: A critical review. *Programmed Learning and Educational Technology*, *15*(3), 207-224.

Henk, W. A., & Stahl, N. A. (1984). *A meta-analysis of the effect of notetaking on learning from lecture*. Paper presented at the the 34th Annual Meeting of the National Reading Conference, St. Petersburg Beach, Florida.

Howe, M. J. A. (1970). Note-taking strategy, review, and long-term retention of verbal information. *The Journal of Educational Research, 63*(6), 285.

Igo, L. B., Bruning, R., & McCrudden, M. T. (2005). Exploring differences in students' copy-and-paste decision making and processing: A mixed-methods study. *Journal of Educational Psychology*, *97*(1), 103-116.

Igo, L. B., & Kiewra, K. A. (2007). How do high-achieving students approach web-based, copy and paste note taking? Selective pasting and related learning outcomes. *Journal of Advanced Academics, 18*(4), 512–529.

Jiang, Y., Paquette, L., Baker, R. S., & Clarke-Midura, J. (2015). Comparing novice and experienced students in Virtual Performance Assessments. *Proceedings of the 8th International Conference on Educational Data Mining*, 136-143.

Jovanovic, J., Solano-Flores, G., & Shavelson, R. J. (1994). Performance-based assessments: Will gender differences in science achievement be eliminated? *Education and Urban Society*, *26*(4), 352-366.

Kappe, R., & van der Flier, H. (2010). Using multiple and specific criteria to assess the predictive validity of the Big Five personality factors on academic performance. *Journal of Research in Personality, 44*(1), 142–145.

Kauffman, D. F., Zhao, R., & Yang, Y.-S. (2011). Effects of online note taking formats and self-monitoring prompts on learning from online text: Using technology to enhance self-regulated learning. *Contemporary Educational Psychology, 36*(4), 313–322.

Kay, R. H. (1992). An analysis of methods used to examine gender differences in computer-related behavior. *Journal of Educational Computing Research, 8*(3), 277-290.

Kay, R. H. (2008). Exploring gender differences in computer-related behaviour: Past, present, and future. In T. T. Kidd & I. Chen (Eds.), *Social Information Technology: Connecting Society and Cultural Issues* (pp. 12-30). Hershey, PA: IGI Global.

Kay, R. H., & Lauricella, S. (2011). Gender differences in the use of laptops in higher education: A formative analysis. *Journal of Educational Computing Research, 44*(3), 357-376.

Keller, J. M. (1999). Using the ARCS motivational process in computer-based instruction and distance education. *New Directions for Teaching and Learning, 1999*(78), 37-47.

Kiewra, K. A. (1984). Implications for notetaking based on relationships between notetaking variables and achievement measures. *Reading Improvement, 21*(2), 145-149.

Kiewra, K. A. (1985a). Investigating notetaking and review: A depth of processing alternative. *Educational Psychologist, 20*(1), 23-32.

Kiewra, K. A. (1985b). Students' note-taking behaviors and the efficacy of providing the instructor's notes for review. *Contemporary Educational Psychology*, *10*(4), 378-386.

Kiewra, K. A. (1989). A review of note-taking: The encoding-storage paradigm and beyond. *Educational Psychology Review, 1*(2), 147-172.

Kiewra, K. A. (2016). Note taking on trial: A legal application of note-taking research. *Educational Psychology Review, 28*(2), 377-384.

Kiewra, K. A., DuBois, N. F., Christian, D., McShane, A., Meyerhoffer, M., & Roskelley, D. (1991). Note-taking functions and techniques. *Journal of Educational Psychology, 83*(2), 240-245.

Kiewra, K. A., & Fletcher, H. J. (1984). The relationship between levels of notetaking and achievement. *Human Learning, 3*(4), 273-280.

Kinnebrew, J. S., Loretz, K. M., & Biswas, G. (2013). A contextualized, differential sequence mining method to derive students' learning behavior patterns. *Journal of Educational Data Mining, 5*(1), 190-219.

Kinnebrew, J. S., Segedy, J. R., & Biswas, G. (2014). Analyzing the temporal evolution of students' behaviors in open-ended learning environments. *Metacognition Learning, 9*(2), 187–215.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the Failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*(2), 75-86.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science, 15*(10), 661-667.

Kobayashi, K. (2005). What limits the encoding effect of note-taking? A meta-analytic examination. *Contemporary Educational Psychology, 30*(2), 242–262.

Kobayashi, K. (2006). Combined effects of note-taking/reviewing on learning and the enhancement through interventions: A meta-analytic review. *Educational Psychology Review, 26*(3), 459–477.

Kubota, C. A., & Olstad, R. G. (1991). Effects of novelty ‐ reducing preparation on exploratory behavior and cognitive learning in a science museum setting. *Journal of Research in Science Teaching, 28*(3), 225-234.

Kuhn, D. (1999). A developmental model of critical thinking. *Educational Researcher, 28*(2), 16-25+46.

Kuhn, D. (2010). What is scientific thinking and how does it develop? In U. Goswami (Ed.), *The Wiley-Blackwell Handbook of Childhood Cognitive Development* (pp. 497-523). Oxford, UK: Wiley ‐ Blackwell.

Kuhn, D., & Dean, D., Jr. (2005). Is developing scientific thinking all about learning to control variables? *Psychological Science, 16*(11), 866-870.

Kuhn, D., & Pease, M. (2008). What needs to develop in the development of inquiry skills? *Cognition and Instruction, 26*(4), 512-559.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). lmerTest: Tests in linear mixed effects models. *R package version 2.0-32*.

Land, S. M. (2000). Cognitive requirements for learning with open-ended learning environments. *Educational Technology Research and Development, 48*(3), 61-78.

Lee, I. S. (2002). Gender differences in self-regulated on-line learning strategies within Korea's university context. *Educational Technology Research and Development, 50*(1), 101-111.

Leelawong, K., & Biswas, G. (2008). Designing learning by teaching agents: The Betty's Brain system. *International Journal of Artificial Intelligence in Education, 18*(3), 181–208.

Maddox, H., & Hoole, E. (1975). Performance decrement in the lecture. *Educational Review, 28*, 17–30.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55-60).

Matthews, J. S., Ponitz, C. C., & Morrison, F. J. (2009). Early gender differences in self-regulation and academic achievement. *Journal of Educational Psychology, 101*(3), 689–704.

McCall, M., & Clarke-Midura, J. (2013, February). *Analysis of gaming for assessment.* Paper presented at the meeting of the Association of Test Publishers, Orlando, FL.

McElhaney, K. W., & Linn, M. C. (2010). Helping Students Make Controlled Experiments More Informative. In K. Gomez, L. Lyons, & J. Radinsky (Eds.), *Learning in the Disciplines: Proceedings of the 9th International Conference of the Learning Sciences (ICLS 2010)* (Vol. 1, pp. 786-793). Chicago, IL: International Society of the Learning Sciences.

McQuiggan, S. W., Goth, J., Ha, E., Rowe, J. P., & Lester, J. C. (2008). Student note-taking in narrative-centered learning environments: Individual differences and learning effects. In B. P. Woolf, E. Aïmeur, R. Nkambou, & S. Lajoie (Eds.), *Proceedings of the 9th International Conference on Intelligent Tutoring Systems (ITS 2008)* (pp. 510-519). Berlin, Heidelberg: Springer

Merceron, A., & Yacef, K. (2008). Interestingness Measures for Association Rules in Educational Data. In R. S. J. d. Baker, T. Barnes, & J. E. Beck (Eds.), *Proceedings of the first International Conference on Educational Data Mining* (pp. 57-66). Montreal, Canada.

Moos, D. C. (2009). Note-taking while learning hypermedia: Cognitive and motivational considerations. *Computers in Human Behavior*, *25*(5), 1120–1128.

Moos, D. C., & Azevedo, R. (2008a). Exploring the fluctuation of motivation and use of self-regulatory processes during learning with hypermedia. *Instructional Science, 36*(3), 203-231.

Moos, D. C., & Azevedo, R. (2008b). Self-regulated learning with hypermedia: The role of prior domain knowledge. *Contemporary Educational Psychology, 33*(2), 270–298.

Mueller, P. A., & Oppenheimer, D. M. (2014). The pen is mightier than the keyboard: Advantages of longhand over laptop note taking. *Psychological Science, 25*(6), 1159-1168.

Mullis, I. V. S., Martin, M. O., Fierros, E. G., Goldberg, A. L., & Stemler, S. E. (2000). *Gender differences in achievement: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: Boston College.

National Center for Education Statistics. (2016). The Nation's Report Card: 2015 Science at Grades 4, 8 and 12. Retrieved from https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2016162

National Research Council. (2011). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.

Neuschmidt, O., Barth, J., & Hastedt, D. (2008). Trends in gender differences in mathematics and science (TIMSS 1995–2003). *Studies in Educational Evaluation, 34*(2), 56–72.

Nye, P. A. (1978). Student variables in relation to notetaking during a lecture. *Programmed Learning and Educational Technology, 15*(3), 196–200.

O'Donnell, A., & Dansereau, D. F. (1993). Learning from lectures: Effects of cooperative review. *The Journal of Experimental Education, 61*(2), 116-125.

Pajares, F. (2002). Gender and perceived self-efficacy in self-regulated learning. *Theory Into Practice, 41*(2), 116-125.

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

Peck, K. L., & Hannafin, M. J. (1983). The effects of notetaking pretraining and the recording of notes on the retention of aural instruction. *The Journal of Educational Research, 77*(2), 100-107.

Peper, R. J., & Mayer, R. E. (1978). Note taking as a generative activity. *Journal of Educational Psychology, 70*(4), 514-522.

Peper, R. J., & Mayer, R. E. (1986). Generative effects of note-taking during science lectures. *Journal of Educational Psychology, 78*(1), 34-38.

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *BMJ: British Medical Journal, 316*(7139), 1236-1238.

Perry, N., Phillips, L., & Dowler, J. (2004). Examining features of tasks and their potential to promote self-regulated learning. *Teachers College Record, 106*(9), 1854-1878.

Peverly, S. T., Brobst, K. E., Graham, M., & Shaw, R. (2003). College adults are not good at self-regulation: A study on the relationship of self-regulation, note taking, and test taking. *Journal of Educational Psychology, 95*(2), 335-346.

Peverly, S. T., Ramaswamy, V., Brown, C., Sumowski, J., Alidoost, M., & Garner, J. (2007). What predicts skill in lecture note taking? *Journal of Educational Psychology, 99*(1), 167-180.

Peverly, S. T., & Wolf, A. D. (in press). Note-taking. In *Cambridge Handbook of Cognition and Education.* Cambridge, England: Cambridge University Press.

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

Pintrich, P. R., & Zusho, A. (2002). The development of academic self-regulation: The role of cognitive and motivational factors. In A. Wigfield & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 249–284). San Diego, CA: Academic Press.

Piolat, A., Olive, T., & Kellogg, R. (2005). Cognitive effort during note taking. *Applied Cognitive Psychology, 19*(3), 291-312.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*(2), 322–338.

Quinn, D. M., & Cooc, N. (2015). Science achievement gaps by gender and race/ethnicity in elementary and middle school: Trends and predictors. *Educational Researcher, 44*(6), 336-346.

Quintana, C., Reiser, B. J., Davis, E. A., Krajcik, J., Fretz, E., Duncan, R. G., . . . Soloway, E. (2004). A scaffolding design framework for software to support science inquiry. *The journal of the learning sciences, 13*(3), 337–386.

Reddington, L. A., Peverly, S. T., & Block, C. J. (2015). An examination of some of the cognitive and motivation variables related to gender differences in lecture note-taking. *Reading and Writing, 28*(8), 1155–1185.

Reilly, D., Neumann, D. L., & Andrews, G. (2015). Sex differences in mathematics and science achievement: A meta-analysis of National Assessment of Educational Progress assessments. *Journal of Educational Psychology, 107*(3), 645-662.

Richland, L. E., Bjork, R. A., Finley, J. R., & Linn, M. C. (2005). Linking cognitive science to education: Generation and interleaving effects. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society* (pp. 1850–1855). NJ: Erlbaum: Mahwah.

Rickards, J. P., & Friedman, F. (1978). The encoding versus the external storage hypothesis in note taking. *Contemporary Educational Psychology, 3*(2), 136-143.

Robinson, D. H., Katayama, A. D., Beth, A., Odom, S., Hsieh, Y., & Vanderveen, A. (2006). Increasing text comprehension and graphic note taking using a partial graphic organizer. *he Journal of Educational Research, 100*(2), 103-111.

Roll, I., Aleven, V., McLaren, B. M., & Koedinger, K. R. (2007). Designing for metacognition — Applying cognitive tutor principles to the tutoring of help seeking. *Metacognition and Learning, 2*(2–3), 125–140.

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.), *Lecture Notes in Computer Science: Artificial Intelligence in Education* (pp. 209–218). Berlin, Heidelberg: Springer.

Sabourin, J., Rowe, J. P., Mott, B. W., & Lester, J. C. (2011). When off-task is on-task: the affective role of off-task behavior in narrative-centered learning environments. In *Proceedings of the 2011 International Conference on Artificial Intelligence in Education (AIED 2011)* (pp. 534-536).

Sabourin, J., Rowe, J. P., Mott, B. W., & Lester, J. C. (2013). Considering alternate futures to classify off-task behavior as emotion self-regulation: A supervised learning approach. *Journal of Educational Data Mining, 5*(1), 9-38.

Sabourin, J., Shores, L. R., Mott, B. W., & Lester, J. C. (2012). Predicting student self-regulation 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* (pp. 470–475). Berlin, Heidelberg: Springer.

Sao Pedro, M. A., Baker, R. S., Gobert, J. D., Montalvo, O., & Nakama, A. (2013). Leveraging machine-learned detectors of systematic inquiry behavior to estimate and predict transfer of inquiry skill. *User Modeling and User-Adapted Interaction, 23*(1), 1-39.

Sao Pedro, M. A., Gobert, J. D., & Raziuddin, J. J. (2010). Comparing pedagogical approaches for the acquisition and long-term robustness of the control of variables strategy. In *Learning in the Disciplines: Proceedings of the 9th International Conference of the Learning Sciences (ICLS 2010)* (pp. 1024-1031). Chicago, IL: International Society of the Learning Sciences.

Sao Pedro, M. A., Jiang, Y., Paquette, L., Baker, R. S., & Gobert, J. D. (2014). Identifying transfer of inquiry skills across physical science simulations using educational data mining. In *Proceedings of the 11th International Conference of the Learning Sciences* (pp. 222-229).

Scalise, K., & Clarke-Midura, J. (2014, April). *mIRT-bayes as hybrid measurement model for technology-enhanced assessments.* Paper presented at the meeting of the National Council on Measurement in Education, Philadelphia, PA.

Schofield, J. W. (1995). *Computers and Classroom Culture*. New York, NY: Cambridge University Press.

Schraw, G. (2010). Measuring self-regulation in computer-based learning environments. *Educational Psychologist, 45*(4), 258–266.

Schraw, G., Crippen, K. J., & Hartley, K. (2006). Promoting self-regulation in science education: Metacognition as part of a broader perspective on learning. *Research in Science Education, 36*(1-2), 111–139.

Segedy, J. R., Kinnebrew, J. S., & Biswas, G. (2015). Using coherence analysis to characterize self ‐ regulated learning behaviours in open ‐ ended learning environments. *Journal of Learning Analytics, 2*(1), 13 ‐ 48.

Shores, L., Rowe, J., & Lester, J. (2011). Early prediction of cognitive tool use in narrative-centered learning environments. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence in Education* (pp. 320-327). Berlin, Heidelberg: Springer-Verlag.

Shute, V., Ventura, M., & Kim, Y. J. (2013). Assessment and learning of qualitative physics in Newton's Playground. *The Journal of Educational Research, 106*(6), 423-430.

Siler, S., Klahr, D., Magaro, C., Willows, K., & Mowery, D. (2010). Predictors of transfer of experimental design skills in elementary and middle school children. In (Ed.), . Part II, LNCS 6095, pp. . : . In J. K. V. Aleven, & J. Mostow (Ed.), *Proceedings of the Tenth International Conference on Intelligent Tutoring Systems (ITS 2010)* (pp. 198-208). Pittsburgh, PA: Springer.

Slotte, V., & Lonka, K. (1999). Review and process effects of spontaneous note-taking on text comprehension. *Contemporary Educational Psychology, 24*(1), 1-20.

Slotte, V., Lonka, K., & Lindblom-Ylänne, S. (2001). Study-strategy use in learning from text. Does gender make any difference? *Instructional Science, 29*(3), 255–272.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage Publications.

Taub, M., Azevedo, R., Bouchet, F., & Khosravifar, B. (2014). Can the use of cognitive and metacognitive self-regulated learning strategies be predicted by learners' levels of prior knowledge in hypermedia-learning environments? *Computers in Human Behavior, 39*, 356–367.

Taub, M., Azevedo, R., Bradbury, A. E., Millar, G. C., & Lester, J. (2017). Using sequence mining to reveal the efficiency in scientific reasoning during STEM learning with a game-based learning environment. *Learning and Instruction*, 1-11.

Thorley, C., Baxter, R. E., & Lorek, J. (2016). The impact of note taking style and note availability at retrieval on mock jurors' recall and recognition of trial information. *Memory, 24*(4), 560-574.

Trafton, J. G., & Trickett, S. B. (2001). Note-taking for self-explanation and problem solving. *Human–Computer Interaction, 16*(1), 1–38.

Trevors, G., Duffy, M., & Azevedo, R. (2014). Note-taking within MetaTutor: Interactions between an intelligent tutoring system and prior knowledge on note-taking and learning. *Educational Technology Research and Development, 62*(5), 507-528.

van der Graaf, J., Segers, E., & Verhoeven, L. (2015). Scientific reasoning abilities in kindergarten: dynamic assessment of the control of variables strategy. *Instructional Science, 43*(3), 381–400.

Weiss, I. R., Banilower, E. R., McMahon, K. C., & Smith, P. S. (2001). *Report of the 2000 national survey of science and mathematics education*. Retrieved from Horizon Research website: http://2000survey.horizon-research.com/reports/status.php

Whitley Jr., B. E. (1997). Gender differences in computer-related attitudes and behavior: A meta-analysis. *Computers in Human Behavior*, *13*(1), 1–22.

Williams, R. L., & Eggert, A. C. (2002). Notetaking in college classes: Student patterns and instructional strategies. *The Journal of General Education, 51*(3), 173-199.

Winne, P. H. (2011). A cognitive and metacognitive analysis of self-regulated learning. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 15-32). New York, NY: Routledge.

Winne, P. H., & Baker, R. S. (2013). The potentials of Educational Data Mining for researching metacognition, motivation and self-regulated learning. *Journal of Educational Data Mining*, *5*(1), 1-8.

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

Winne, P. H., & Hadwin, A. F. (2009). Studying as self-regulation. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in Educational Theory and Practice* (pp. 277-304). New York: Routledge.

Winne, P. H., & Jamieson-Noel, D. (2002). Exploring students' calibration of self reports about study tactics and achievement. *Contemporary Educational Psychology*, *27*(4), 551-572.

Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 531-566). San Diego, CA: Academic Press.

Winters, F. I., Greene, J. A., & Costich, C. M. (2008). Self-regulation of learning within computer-based learning environments: A critical analysis. *Educational Psychology Review, 20*(4), 429-444.

Wittrock, M. C. (1974). Learning as a generative process. *Educational Psychologist, 11*(2), 87-95.

Yukselturk, E., & Bulut, S. (2009). Gender differences in self-regulated online learning environment. *Educational Technology and Society, 12*(3), 12–22.

Yukselturk, E., & Top, E. (2013). Exploring the link among entry characteristics, participation behaviors and course outcomes of online learners: An examination of learner profile using cluster analysis. *British Journal of Educational Technology, 44*(5), 716–728.

Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist, 25*(1), 3-17.

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, M. Zeidner, & P. R. Pintrich (Eds.), *Handbook of self-regulated learning* (pp. 13–39). San Diego, CA: Academic Press.

Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal*, *45*(1), 166–183.

Zimmerman, B. J., & Martinez-Pons, M. (1990). Student differences in self-regulated learning: Relating grade, sex, and giftedness to self-efficacy and strategy use. *Journal of Educational Psychology, 82*(1), 51-59.

Zimmerman, B. J., & Schunk, D. H. (Eds.). (2001). *Self-regulated learning and academic achievement: Theoretical perspectives*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review, 20*, 99–149.

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review, 27*(2), 172–223.

APPENDIX. LIST OF CVS EVIDENCE AND CVS CFC EVIDENCE

Evidence that needs to be collected from controlled experiments/observations or kiosk pages for the application of the control of variables strategy (CVS) to test each hypothesis in each scenario. The CVS evidence necessary to test the correct final claim (CVS CFC) in each scenario were marked in bold.

Scenario	Evidence necessary for the use of CVS to test the hypothesis of …
Frog	**Parasites (correct claim)** • **Blood test on six-legged frog vs. red frog** • **Water test on control water vs. red water** • **Observation of six-legged frog vs. red frog** • **Research information from parasites kiosk page**
	Pesticides • Blood test on six-legged frog vs. yellow frog • Water test on control water vs. yellow water • Observation of six-legged frog vs. yellow frog • Research information from pesticides kiosk page
	Pollution • Blood test on six-legged frog vs. blue frog • Water test on control water vs. blue water • Observation of six-legged frog vs. blue frog • Research information from pollution kiosk page
	Radiation • Blood test on six-legged frog vs. green frog • Water test on control water vs. green water • Observation of six-legged frog vs. green frog • Genetic test on six-legged vs. green frog • Research information from radiation kiosk page
	Alien • Genetic test on six-legged frog vs. red frog • Genetic test on six-legged frog vs. yellow frog • Genetic test on six-legged frog vs. blue frog • Genetic test on six-legged frog vs. green frog • Research information from alien kiosk page
Bee	**Radiation (correct claim)** • **Protein test on dead bee vs. green bee** • **Nectar test on control nectar vs. green nectar** • **Observation of dead bee vs. green bee** • **Research information from radiation kiosk page**

Parasites

- Protein test on dead bee vs. red bee
- Nectar test on control nectar vs. red nectar
- Observation of dead bee vs. red bee
- Genetic test on dead bee vs. red bee
- Research information from parasites kiosk page

Pesticides

- Protein test on dead bee vs. yellow bee
- Nectar test on control nectar vs. yellow nectar
- Observation of dead bee vs. yellow bee
- Research information from pesticides kiosk page

Pollution

- Protein test on dead bee vs. blue bee
- Nectar test on control nectar vs. blue nectar
- Observation of dead bee vs. blue bee
- Research information from pollution kiosk page

Alien

- Genetic test on dead bee vs. red bee
- Genetic test on dead bee vs. yellow bee
- Genetic test on dead bee vs. blue bee
- Genetic test on dead bee vs. green bee
- Research information from alien kiosk page