How Immersive Virtual Environments Foster Self-Regulated Learning

Yang Jiang [a,*], Jody Clarke-Midura [b], Ryan S. Baker [c], Luc Paquette [d], Bryan Keller [a]

[a] Teachers College, Columbia University, United States

525 W 120th Street, New York, NY 10027, United States

yj2211@tc.columbia.edu

[b] Utah State University, United States

2830 Old Main Hill, Logan, UT 84322, United States

[c] University of Pennsylvania, United States

Graduate School of Education, 3700 Walnut St., Philadelphia, PA 19104, United States

[d] University of Illinois at Urbana–Champaign

1310 S 6th St., Champaign, Illinois 61820, United States

Abstract

Over the past decade, immersive virtual 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 chapter explores how the Immersive Virtual Assessments (IVAs), an immersive virtual environment designed to assess middle school students' science inquiry skills, fostered SRL. Our analyses combining educational data mining techniques with multilevel analysis indicated that students developed self-regulatory behaviors and strategies as they used IVAs. Experience with IVAs prepared students to adopt more efficient note-taking and note-reviewing strategies. Students also learned to exploit more available sources of information by taking and reviewing notes on them, in order to either solve inquiry problems, or to monitor their solutions.

Keywords: Immersive virtual assessments, self-regulated learning, note-taking, note-reviewing, self-regulatory behaviors and strategies, monitoring, science inquiry, instructional design, scaffolding

How Immersive Virtual Environments Foster Self-Regulated Learning

Introduction

Self-regulated learning (SRL) is important for academic success in various educational settings (Zimmerman & Schunk, 2001). 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, 2008). 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). One increasingly popular strategy for fostering SRL is to use personalized learning within computer-based environments (Azevedo, 2005). An increasing number of personalized learning environments now include various types of support for students in developing SRL skills, including both modeling those skills (Khachatryan et al., 2014), giving regular reports about whether students are demonstrating SRL (Arroyo et al., 2007), and even providing immediate feedback when students demonstrate behaviors associated with poorer SRL (Roll, Aleven, McLaren, & Koedinger, 2007). The challenge of open-ended learning environments such as immersive virtual 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, Kinnebrew, & Biswas, 2015). In the current study, we aim to explore how SRL manifests in an immersive virtual environment for middle school science, and how this environment can be enhanced to adapt to the needs and self-regulated learning of different learners.

Self-Regulated Learning

While researchers have developed many theoretical models of SRL (see Pintrich, 2000; 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, 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).

Winne and Hadwin's (1998) framework 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 (p. 278). 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. This framework offers a metacognitive view of SRL that integrates a more complex cognitive architecture (Greene & Azevedo, 2007; Winne, 2011), and has been adopted to study SRL in other open-ended learning environments (Moos, 2009; Moos & Azevedo, 2008). Given the interactive and open-ended nature of immersive virtual environments, this chapter applies Winne & Hadwin's model of SRL to the context of an immersive virtual environment.

SRL and Personalized Learning

Given the importance of SRL, it is crucial that schools and educators provide personalized learning experiences for learners to help them successfully regulate their cognition, metacognition, and learning with effective SRL strategies (Zimmerman & Schunk, 2001). Over the past decade, research has been conducted on designing personalized learning within computer-based learning

environments to prompt, support and enhance self-regulated learning (Azevedo, 2005). In this chapter, we aim to identify effective SRL behaviors and strategies in an immersive open-ended learning environment and examine how the environment fosters SRL. This research will provide implications for the instructional design of personalized learning environments to measure SRL behaviors in real time, assist and scaffold students' key self-regulatory processes and strategies, and prepare students as better self-regulated learners when these scaffolds fade out (Azevedo, 2005).

Student Behavior and SRL

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, & Lester, 2013). Therefore, researchers have studied students' observed actions and sequences of behaviors in computer-based learning environments to assess their SRL (Zimmerman, 2008). For instance, Aleven and colleagues (2010) evaluated learners' observed actions in an intelligent tutoring system to understand their use of help-seeking as an SRL strategy. 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; Zimmerman, 2008).

Sequential Pattern Mining (Agrawal & Srikant, 1995), a methodology that has been extensively used in Educational Data Mining (Baker & Yacef, 2009), has shown potential for discovering complicated patterns of SRL behaviors within open-ended learning environments. For example, Kinnebrew and colleagues (2014) applied differential pattern mining techniques, a form of sequential pattern mining where patterns over time are compared between different groups of individuals, to log data produced by students engaging in activities within Betty's Brain, an open-

ended learning environment for science learning. 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 concept map after taking quizzes 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 a virtual environment. In this chapter, we aim to apply sequential pattern mining to identify the action sequences that correspond to the four SRL processes in IVAs: task definitions, goal setting and planning, enacting study tactics and strategies, and metacognitively adapting study strategies.

Note-Taking as an SRL Strategy

Winne and Hadwin (1998) have identified the utilization of various learning strategies as a key component of their SRL framework. In immersive virtual 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, Duffy, & Azevedo, 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, very few studies have examined the role of note-taking strategy in immersive virtual environments or other online contexts.

Past research has shown that learners with different self-regulatory skills may invoke different note-taking/reviewing behaviors and show different patterns in the content of notes recorded (Trevors et al., 2014). In immersive 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 chapter studies whether learning science through immersive 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 developed 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 to exhibit a higher frequency of note-taking and note-reviewing, but also – and more importantly – to take notes that are of higher quality and involve a deeper level of cognitive processing.

The Present Study

In sum, the purpose of this chapter is to investigate how immersive virtual assessments foster self-regulated learning (SRL) and discuss how the learning processes and activities in the immersive virtual assessments map to various SRL processes. We explore these issues in the

context of the Immersive Virtual Assessment Project (IVA) (Clarke-Midura, McCall, & Dede, 2012). IVAs are 3D immersive virtual environments that have the look and feel of a videogame but are designed to assess middle school students' science inquiry skills *in situ.* The specific IVAs being studied were designed to be aligned with national standards for science education (National Research Council, 2011). Learning in the open-ended immersive virtual environments pose both opportunities and challenges for SRL. Two analyses were conducted to examine the effect of IVAs on SRL. In analysis 1, we applied sequential pattern mining on middle school students' action log data as they used IVAs within their science classes to track how their behaviors demonstrated SRL, and whether using IVAs promoted students' use of self-regulatory processes and strategies. In analysis 2, we employed feature engineering to explore the development of note-taking and note-reviewing strategies in IVAs, which are effective self-regulatory strategies for learning. We conclude with a discussion of the implications of our results for the instructional design of immersive virtual environments to facilitate personalized learning.

Immersive Virtual Assessments

The context for this research was 3-D Immersive Virtual Assessments (IVAs) designed as part of the Virtual Performance Assessment Project (Clarke-Midura et al., 2012). On the front end, students navigate their avatar around the virtual environment and solve scientific problems by making observations, gathering data, interacting with non-player characters (NPCs), reading kiosk informational pages for research, conducting virtual laboratory experiments, and taking notes. On the back end, students' actions are recorded automatically and unobtrusively in the form of process data (e.g., where they went and what they did in the immersive environment) as well as product data (e.g., student notes and final claims). The immersive assessments used in this study have been

validated in previous research to ensure that they are assessing the performance of science inquiry (Clarke-Midura et al., 2012; McCall & Clarke-Midura, 2013; Scalise & Clarke-Midura, 2014).

[Place Figure 1 approximately here]

This study uses data from two IVAs: the "frog scenario" (see Figure 1) and the "bee scenario." The two scenarios were designed as different forms of a test meant to assess the same inquiry practices in different contexts. Therefore, they have similar structure and mechanics, but the problems students are asked to solve have different content and surface features. In the frog scenario, students must determine why a frog has grown six legs. In the bee scenario, they must figure out why the bees are dying. Both scenarios have a similar look and feel. They are designed around a village that contains four farms, a lab, and an information kiosk. In both scenarios, students 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 factors is correct. Students can talk to NPCs from the farms who provide conflicting opinions, read informational pages about the five possible causal factors from a research kiosk, make observations at different farms, and conduct laboratory tests on samples they collect at the farms (e.g., frogs, tadpoles, water samples, bees, larvae, and nectar samples) in order to determine the correct answer (that parasites have caused the frog to grow six legs and radiation induced genetic mutation is causing the bees to die).

In order to keep track of their data, students have a digital notepad (Figure 2) that can be accessed at any time. This notepad was designed to not allow the student to copy and paste information (e.g., kiosk research pages, dialogue with NPCs, laboratory test results, observation, etc.). Instead, students 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.

[Place Figure 2 approximately here]

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 IVA's assessment of science inquiry skills for each student.

Importance of Immersive Virtual 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, 2008). The non-linearity and open-endedness of immersive virtual environments such as IVAs 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, 2008). In order to be successful in IVAs, students need to understand their inquiry tasks, set goals and make corresponding plans. Such plans include deciding where to go in the immersive 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 (1998) SRL framework: understanding task definition, goal setting and planning, enacting study tactics and strategies, and metacognitively adapting studying. In this chapter, we attempt to explore whether our immersive

virtual assessments help strengthen self-regulatory skills, and if so, how should educators design future environments to promote self-regulated learning.

Study on Development of SRL Skills Within IVAs

In order to illustrate how IVAs foster the development of skills in SRL, this chapter presents results from a study where more than 2,000 middle school students' behaviors were examined at a fine-grained level as they used Immersive Virtual Assessments (IVAs) in their science classes. Specifically, we applied a combination of educational data mining techniques (e.g., sequential pattern mining, feature engineering) and multilevel analysis on students' action log data and tracked how student behaviors demonstrated SRL, and how self-regulatory skills developed in IVAs.

Participants

The study presented in this chapter analyzed interaction log files produced by a total of 2,429 seventh and eighth grade students (12-14 years old) who used IVAs within their science classes at the end of the 2011-2012 school year. 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.

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. Prior to each assessment, students were shown a short introductory video that provided instructions on how to use the IVAs. 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.

Student actions, notes, and performance in the virtual assessments were automatically logged as they worked within each IVA scenario, and used for later analyses. On average, 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.). Overall, they generated a total of 381,331 actions within the frog scenario and a total of 396,760 actions within the bee scenario.

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 – novice users who were using IVA for the first time (*novice* group), and experienced users who had previously experienced the other IVA scenario (*experienced* group). Accordingly, among the 1,985 students who completed the frog scenario, 1,232 were novices (frog-novice) and 753 were experienced (frog-experienced). Among the students who completed the bee scenario, 1,198 were novices (bee-novice) and 825 were experienced (bee-experienced).

We explored students' development of self-regulatory skills while playing with IVAs by comparing the novice group and the experienced group. Specifically, two analyses were conducted to examine the development of SRL skills in IVAs. In analysis one, we applied differential pattern mining to compare the frequency of action sequences that were identified as representing self-regulatory processes and strategies between novice and experienced students in each scenario. In analysis two, we compared note-taking and note-reviewing behaviors, which are effective learning

strategies and important components of SRL, between the novice group and the experienced group within each IVA scenario.

Accordingly, two different types of measures related to SRL processes within IVAs were collected and developed for analyses: 1) Frequency metrics of action sequences related to different SRL processes and strategies, identified through sequential pattern mining; 2) Variables related to note-taking and note-reviewing, including purely quantitative measures based on actions involving IVA'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 sections.

In each analysis, multilevel modeling was conducted to investigate the potential differences between the novice group and the experienced group on these SRL-relevant measures. Multilevel models, also known as hierarchical linear 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 was adopted in the current study due to the hierarchical structure of our data, where the population consists of students nested within classes, and multiple classes that shared the same teacher. Specifically, 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 novice group and experienced

group on the frequency metrics of action sequences and the quantitative features on note-taking and note-reviewing. In these models, the dependent variable is each individual measure/feature related to SRL, while experience with IVAs (whether a student was a novice student (coded as 0) or had previous experience in the other IVA scenario (coded as 1)) serves as the single student-level predictor variable in each model. These three-level analyses, taking the hierarchy of data into consideration, enabled us to examine the relationship between student's experience with IVAs and their use of SRL processes and strategies after controlling for student- and teacher-level variability.

In this study, multilevel analyses were conducted for each 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, we controlled for false discovery rate by applying the Benjamini and Hochberg's (1995) post-hoc correction method.

Analysis 1: Behavior Pattern Analysis

In our previous work (Jiang, Paquette, Baker, & Clarke-Midura, 2015), we compared student performance on science inquiry tasks between the novice students and the experienced students within each IVA scenario, and found that experienced students showed significantly better performance on identifying the correct final claim (CFC) than novice students, in both the frog scenario and the bee scenario. Experienced students also outperformed novice students on designing causal explanations (DCE) in the bee scenario. The comparisons of student performance suggest that students are developing SRL skills across scenarios since SRL has been shown to be closely related to academic performance (Zimmerman, 1990), and the regulation of science inquiry performance is a crucial part of SRL (Pintrich & Zusho, 2002). In the current section, we aim to go beyond simply looking at whether previous experience in IVA improved student inquiry

performance, and instead delve into whether more experienced students used IVAs differently than less experienced students. Exploration of behaviors will enable us to better understand how students' self-regulatory behaviors and strategies develop over time. According to Pintrich and Zusho (2002), self-regulating one's behaviors and strategies are also key components of SRL.

We investigated patterns in behavior by applying sequential pattern mining to identify and compare the frequent sequential patterns of student actions between the two groups. Sequential pattern mining is a popular data mining technique that automatically identifies frequent temporal patterns of actions in the data (Agrawal & Srikant, 1995). It can be used to detect differentially frequent behavioral patterns of different groups of students (Kinnebrew, Loretz, & Biswas, 2013). An example sequential pattern in IVAs 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 contains 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. 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 IVA (*inspect*), saving objects to 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.

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., *view* → *read*). Arules was used to determine the most frequent two actions sequences by selecting the temporal associations of one specific action and a subsequent action with the highest support and confidence. In this study, sequential patterns of consecutive actions were selected with the cut-off thresholds of support = 0.0005 and confidence = 0.1. In the frog scenario, a total of 64 short sequential patterns (length = 2) were identified that met the minimum support and confidence constraints within the novice group; 61 patterns were identified within the experienced group. In the bee scenario, 66 patterns were

identified within the novice group; 63 were selected within the experienced group. 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 (Bazaldua, Baker, & San Pedro, 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*), we then identified a subset of sequential patterns that we believe corresponded to self-regulatory processes and strategies, and compared their frequency between the two groups.

To facilitate the comparison of the frequency metrics between the novice group and the experienced group, the support and confidence for each pattern were calculated separately for each student. Three-level regression tests that controlled for multiple comparisons with Benjamini and Hochberg (1995) corrections were then conducted to compare the metric values between the two groups in each scenario. Table 1 presents the comparison of the support and confidence metrics of frequent sequential patterns identified as reflective of self-regulatory processes and strategies that were found to be significantly different between the two groups.

[Place Table 1 approximately here]

Understanding Task Definition

One interesting behavior pattern with high interestingness was *help* → *note*, which we 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 IVAs. A window would pop up

to remind students of the ultimate goals they need to achieve when the button is clicked on. It seems that this pattern showed significantly higher support for the novice group than the experienced group in both scenarios (frog: *Ms* = .0007 and .0003, $t(1988) = -2.95$, $p = .003$; bee: *Ms* = .0007 and .0003, $t(1997) = -2.51$, $p = .012$). We speculate that this was due to the novelty effect (Jiang et al., 2015). That is, due to the increased attention and enthusiasm as the novel IVA 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 IVAs more frequently than experienced students, whereas previous experience in the other IVA scenario had familiarized experienced students 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.

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 experienced students than novice students (support: *Ms* = .0098 and .0063, $t(1979) = 4.77$, $p < .001$; confidence: *Ms* = .33 and .26, $t(1044) = 3.51$, $p < .001$). This pattern also showed significantly higher support for experienced students than novice students in the bee scenario (support: *Ms* = .0093 and .0064, $t(2012) = 2.90$, $p = .004$; confidence: *Ms* = .31 and .27, $t(1033) = 1.46$, $p = .143$). A similar pattern *read-MULT → note-MULT* also had significantly higher support and confidence for the experienced group than the novice group in the frog scenario (support: *Ms* = .0040 and .0029, $t(1982) = 2.59$, $p = .010$; confidence: *Ms* = .15 and .11, $t(1387) = 3.03$, $p = .002$). In the bee scenario, the support of this pattern was marginally significantly higher for the experienced group than the novice group (*Ms* = .0037 and .0027,

$t(2018) = 2.27$, $p = .023$, adjusted $\alpha = .015$). These results suggested that the experienced students were more likely to open notepad repeatedly to take or review notes after reading a research page (once or repeatedly). In other words, experienced students tended to show better 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 notepad most likely indicates more complete notes being encoded by users, further fostering student learning (Armbruster, 2009).

Similarly, experienced students were more likely to open the notepad to take or review notes after conducting laboratory experiments (*experiment → note-MULT)* or viewing test results (*look → note*). For the sequence *experiment → note-MULT*, the confidence for the experienced group was higher than that for the novice group in the bee scenario ($Ms = .16$ and .09, $t(1071) = 3.91$, $p < .001$). However, the confidence for this pattern in the frog scenario was not statistically significantly different ($Ms = .14$ and .12, $t(1142) = 1.48$, $p = .139$). For the pattern *look → note*, the confidence for the experienced group was marginally significantly higher than that for the novice group in the both scenarios (frog: $Ms = .12$ and .08, $t(1045) = 2.48$, $p = .013$, adjusted $\alpha = .013$; bee: $Ms = .09$ and .06, $t(954) = 2.30$, $p = .022$, adjusted $\alpha = .014$). That is, experienced students were more likely to access the notepad immediately after viewing the results of lab tests; they were also more likely to open the notepad repeatedly after running laboratory 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 IVAs.

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. These patterns appeared to have higher support for experienced students than novice students. In the frog scenario, the pattern *final claim → note* showed significantly higher support (but not confidence) for the experienced group than the novice group (support: *Ms* = .0019 and .0012, $t(1970) = 2.73$, $p = .006$; confidence: *Ms* = .26 and .22, $t(636) = 1.27$, $p = .205$). In the bee scenario, this pattern showed both significantly higher support and confidence for experienced students than novice students (support: *Ms* = .0015 and .0007, $t(2009) = 3.75$, $p < .001$; confidence: *Ms* = .13 and .08, $t(641) = 3.33$, $p < .001$). A similar pattern "*final claim → note-MULT*" also showed significantly higher support and confidence for experienced students in the frog scenario (support: *Ms* = .0014 and .0007, $t(1962) = 3.46$, $p < .001$; confidence: *Ms* = .24 and .16, $t(635) = 2.53$, $p = .012$). This finding indicated that experienced students were more likely than novice students 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.

On the other hand, mixed results were found for the patterns related to reading kiosk pages after submitting final claims. In the frog scenario, the pattern *final claim → read* showed significantly higher confidence for the novice group than the experienced group (*Ms* = .25 and .16, $t(639) = -2.78$, $p = .006$). By contrast, this pattern showed marginally significantly higher support for the experienced group than the novice group in the bee scenario (*Ms* = .0015 and .0009, $t(2013) = 2.42$, $p = .016$, adjusted $\alpha = .013$). The pattern *final claim → read-MULT* showed significantly higher confidence for the novice group than the experienced group in the bee scenario (*Ms* = .21 and .15, $t(637) = -2.98$, $p = .003$). This finding suggested that although experienced students were more likely to access the notepad after submitting final claims, novice students who were newly exposed to the IVA environment were sometimes more likely to read research information to check their answers after submitting a final claim than experienced students.

Discussion

In summary, our analysis of student behavior patterns within IVAs suggested that experience with learning in IVAs 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. Probably due to a novelty effect (cf. Kubota & Olstad, 1991), novice students who were introduced to IVA for the first time tended to access the help page that reminded them of their tasks and took notes on it more often than experienced students. This might suggest that students were more familiar with their tasks and did not need to record this information the second

time they used IVAs. On the other hand, experienced students generally showed better strategy usage than their novice counterparts during science inquiry. They tended to open the notepad more frequently after reading research information or running and viewing experiment results. Experienced students were also 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. On the contrary, novice students were more likely than experienced students to read kiosk pages, information that they had not organized into notes, after making their final claims.

Analysis 2: Development of Note-Taking/Reviewing Strategies in IVAs

Results from analysis 1 indicate that the experienced students were more likely to utilize the digital notepad for note-taking or note-reviewing purposes after reading research pages and test results presented in the virtual environment or after submitting their final claims than the novice students. Taking and reviewing notes are popular learning strategies that have been deemed as beneficial for academic success (Armbruster, 2009). Meanwhile, note-taking/reviewing strategies are critical elements of self-regulated learning (Azevedo, 2005; Moos, 2009). Therefore, we aim to further investigate how IVAs fostered the development of note-taking/reviewing strategies in analysis 2. Specifically, we examine the regulation of note-taking/reviewing strategies by comparing both the quantity of note-taking/reviewing behaviors and the content of notes taken, between novice students and experienced students, in each IVA scenario.

Quantity of Note-Taking/Reviewing Behaviors

In this section, we examine whether there were consistent changes in the quantity of note-taking and note-reviewing activities executed by students as they transitioned from one IVA scenario to another. Within IVAs, students could click on the digital notepad to take or review

notes. Measures representing the quantity of note-taking/reviewing behavior (see Table 2 for a description of the full set of measures) were developed and computed for each student based on the interaction logs and used in later analysis.

[Place Table 2 approximately here]

Results on the comparisons of note-taking/reviewing quantity between novice and experienced students after applying Benjamini and Hochberg's post-hoc control method are presented in Table 2. Overall, the novice students who used IVA for the first time and the experienced students who had previously used the other IVA scenario did not differ significantly on their average frequency of notepad access, the total amount of time spent in the notepad, or the proportion of total time in IVA that was distributed to the digital notepad in either the frog scenario or the bee scenario.

Despite the lack of significant differences in the quantitative measures of overall notepad access, further analysis that distinguished note-taking activities from note-reviewing activities revealed consistent differences between novice and experienced students in both scenarios. Among the note-takers, experienced students opened the notepad to take notes more frequently than novice students in both the frog scenario (*M*s = 13.60 and 10.52, $t(1158) = 4.51$, $p < .001$) and the bee scenario (*Ms* = 13.72 and 10.27, $t(1162) = 4.23$, $p < .001$). Experienced students also devoted significantly more time to taking notes in the digital notepad than novice students in both scenarios (frog: *M*s = 5 min. and 3 min., 57 sec., $t(1154) = 3.97$, $p < .001$; bee: *M*s = 5 min. and 3 min., 51 sec., $t(1166) = 3.77$, $p < .001$). In addition, notes recorded by experienced students were comprised of significantly more words (frog: *M*s = 65.30 and 56.33, $t(1162) = 2.42$, $p = .016$; bee: *M*s = 66.60 and 50.88, $t(1166) = 3.76$, $p < .001$) and more sentences (frog: *M*s = 9.65 and 7.88, $t(1159) = 3.69$,

$p < .001$; bee: Ms = 9.72 and 7.49, $t(1156) = 3.99$, $p < .001$) on average than notes recorded by their novice counterparts.

Although the experienced students recorded a higher quantity of notes than the novice students, they did not review notes significantly more frequently than novice students in either scenario (frog: Ms = 4.77 and 4.30, $t(1170) = 1.10$, $p = .272$; bee: Ms = 4.56 and 4.17, $t(1166) = .28$, $p = .776$). Likewise, note-takers from the two groups spent a similar total amount of time reviewing their notes (frog: Ms = 52 sec. and 44 sec., $t(1172) = 1.05$, $p = .294$; bee: Ms = 48 sec. and 39 sec., $t(1168) = 1.29$, $p = .196$).

Note Content

Considering that students became more frequent note-takers and took a greater quantity of notes as they became experienced in using the IVAs, it would be useful to further explore how the content of notes taken by students in IVAs developed over time. For example, which type of notes did the experienced students record more than the novice students, and how did the content and quality of notes differ between the two groups?

We followed the procedures recommended by Chi (1997) and Trevors et al. (2014) for the development of quantitative measures of note content. 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 by the first author, who adjusted inappropriate segmentation. 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 using three coding schemes: (1) The *type of note* coding scheme (verbatim or paraphrased content; Trevors et al., 2014), (2) The *source of note content* coding

scheme, and (3) The *hypothesis* or *conclusion* coding scheme. Details and examples of the coding schemes are shown in Table 3.

[Place Table 3 approximately here]

Two coders (i.e., the first and the second author) 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) and the *source of note* (κ = .90). 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 the first author.

After all segments were coded, quantitative measures based on these categories were calculated for each note-taker (e.g., student who took notes) and used in later analysis. For example, the number of each code (e.g., *content reproduction*, *content elaboration*, etc.) were calculated for each note-taker, and each coding scheme, in each scenario. In addition, we 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*), we counted this note as both a *combination* segment and as the specific categories they belonged to when calculating these measures.

[Place Table 4 approximately here]

Content reproduction and content elaboration. Comparisons of the number of note segments from each note content category between note-takers in the novice group and those in the experienced group are reported in Table 4. According to the results, the higher quantity of notes for experienced students compared to novice students was highly driven by the difference in the content reproductive notes. In both scenarios, the experienced students recorded significantly more sentence segments that were verbatim copies or close paraphrases of the content presented in the IVA environment than the novice students (frog: *Ms* = 7.78 and 6.14, $t(1161) = 3.73$, $p < .001$; bee: *Ms* = 7.86 and 6.05, $t(1157) = 3.52$, $p < .001$). That is, students with previous experience using the other IVA 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.

Mixed results were found for the content elaborative notes that involved a deeper level of cognitive processing. In the bee scenario, note-takers in the experienced group recorded significantly more note segments that entailed elaboration of instructional content presented in the environment than their novice counterparts (*Ms* = 1.55 and 1.16), $t(1163) = 2.72$, $p = .007$. 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, in the bee scenario, previous experience in completing the IVA frog scenario seemed to have not only led students to copy or paraphrase more information in notes, but may have also prompted students to go beyond the superficial meaning of the instructional content and process the information deeply, through such

techniques as generating inferences, identifying underlying patterns of data, constructing connections, self-questioning, and concept mapping. Engaging in constructive note-taking activity more for experienced students might also explain their superior performance in the bee scenario in both identifying correct final claims (CFC) and designing causal explanations (DCE) tasks, compared to novice students (Jiang et al., 2015). On the contrary, this trend was not replicated in the frog scenario, where no significant difference was found in the number of content elaborative note segments between novice students (M = 1.32, SD = 1.93) and experienced students (M = 1.44, SD = 2.45), $t(1167) = .99$, $p = .324$. Prior usage of the IVA bee scenario was not associated with differences in the quantity of content elaborative notes and level of cognitive processing involved in note-taking in the frog scenario.

Source of note content. Comparison of the source of note content between the two groups of students revealed differences between novice and experienced students. In both scenarios, experienced students recorded more sentences based on research information from the kiosk than novice students (frog: Ms = 4.41 and 3.13, $t(1164) = 3.94$, $p < .001$; bee: Ms = 4.35 and 2.89, $t(1161) = 4.28$, $p < .001$). This result was in line with our previous finding that experienced students were more likely to access the notepad after reading kiosk pages. Accordingly, experienced students tended to make use of the digital notepad to verbatim copy or paraphrase information from the research kiosk more than the novice students (frog: Ms = 3.95 and 2.80, $t(1165) = 3.68$, $p < .001$; bee: Ms = 3.80 and 2.57, $t(1161) = 3.75$, $p < .001$). The higher relative frequency of reading research information and taking notes on it, which might help experienced students interpret laboratory test results and facilitate the acquisition of domain-specific knowledge (Chen & Klahr, 1999), may have contributed to their higher inquiry performance than novice students.

Beyond taking more notes on kiosk research information, experienced students also took more notes than novices from other sources. For example, in the frog scenario, experienced students took more notes that were based on observations than novices (*Ms* = 3.11 and 2.16, $t(1166) = 3.80$, $p < .001$), whereas in the bee scenario, experienced students recorded more sentences based on laboratory experiment results than novices (*Ms* = 2.21 and 1.59, $t(1161) = 2.90$, $p = .004$).

Moreover, experienced students wrote nearly twice as many note segments that integrated information from multiple disparate sources than novice students in the bee scenario (*Ms* = .46 and .25), $t(1162) = 2.99$, $p = .003$. Previous experience in IVA appears to have led students to process information more deeply, to realize the connections between multiple pieces of information obtained from various sources, to organize and synthesize the information, and to construct more connections in notes in the bee scenario. Further examination of the combination notes indicated that most of the combination notes involved elaboration of combined content. Experienced students in the bee scenario produced nearly twice as many sentences as novices in which they elaborated on information combined from multiple sources (*Ms* = .33 and .17, $t(1161) = 3.03$, $p = .003$). This result also echoed our finding that experienced students took more elaborative notes than novice students in the bee scenario.

Hypothesis and conclusion notes. Experienced students also generated significantly more hypotheses related to the cause of the bee population death than their novice counterparts in the bee scenario (*Ms* = .55 and .34), $t(1146) = 3.49$, $p < .001$. Similarly, students who were using the IVAs for the second time also produced more sentences where they drew conclusions from data they collected (*Ms* = .62 and .36, $t(1163) = 3.43$, $p < .001$). Both hypothesis notes and conclusion notes are important components of content elaborative notes that involve constructive learning.

These differences were consistent with the higher quantity of content elaborative notes for experienced students in the bee scenario and affirmed the relatively deeper level of cognitive processing as students used IVAs for the second time. However, these differences were not replicated in the frog scenario.

Discussion

Our analysis on the differences in the evolution of the quantity of note-taking/reviewing behaviors and note content further suggested the more sophisticated utilization of these learning strategies by experienced students than novice students. To begin with, while using the IVAs, students increasingly made use of the digital notepad to take notes. In both scenarios, note-takers with previous experience in the other IVA 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 IVA for the first time. More information was transferred from the IVA environment and encoded as notes in notepads, and more complete notes were produced by experienced students, potentially strengthening their understanding and mental representations of the instructional content. However, a short session of using one IVA scenario was not sufficient to change students' note-reviewing patterns in the other scenario; experienced students were not more likely to review notes more frequently or spend more time reviewing notes.

Investigation of the content of notes taken by students indicated that the experienced students tended to reproduce instructional content presented in IVAs more than novice students. 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, experienced students also encoded a higher quantity

of notes in different sources depending on the scenario (observations in the frog scenario, and tests and combined sources in the bee scenario).

In the bee scenario, students with previous experience also tended to process information more deeply and engage in generative note-taking more than note-takers in the novice condition. Further analysis suggested that this behavior was driven largely by the experienced students' engagement in building internal connections between information obtained from various sources, generating hypotheses and inducing conclusions, and elaborating on the research information. These note contents all correspond to constructive learning, leading to deeper-level mental representations of the instructional content (Bui, Myerson, & Hale, 2013) and may have led to the better performance seen in the bee scenario on both CFC and DCE for experienced students. It seems that students' generative note-taking strategies developed as they transitioned from the frog scenario to the bee scenario, but not the other way around. We postulate that this difference was most likely caused by the differences in the content of the two learning contexts, despite similar design goals. It seems that it is slightly 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 DCE performance in the bee scenario than the frog scenario (Jiang et al., 2015). Therefore, the students' previous experience in the frog scenario, which was slightly easier and required a lower cognitive load, might have led students to engage in a deeper level of cognitive processing in the bee scenario than those who were introduced to IVA for the first time. However, previous experience in the more difficult bee scenario did not encourage experienced students in the frog scenario to delve deeper and generate more elaborative notes than novice students in the frog scenario.

Conclusion and Implications for Instructional Design

This chapter explores how student self-regulatory behaviors and strategies evolved within an immersive virtual environment for middle school science, combining educational data mining techniques such as sequential pattern mining and feature engineering with multilevel analysis. In conclusion, students gained skills in regulating their inquiry behaviors and adopted more successful self-regulatory strategies as they used IVAs. 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 novices (Gilhooly et al., 1997). The IVAs 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. Our analysis on note-taking and reviewing 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, take more complete notes, and reproduce more important domain-specific knowledge information from kiosk research pages, behaviors which have been previously found to promote inquiry performance. Particularly, students with previous experience in the IVA frog scenario engaged in deeper-level cognitive processing and content elaboration during note-taking in the bee scenario than students in the novice condition, through such techniques as generating inferences, constructing connections between information from various sources, and generating hypotheses and conclusions in notes. Altogether, the relatively more sophisticated self-regulatory behaviors and strategies seen within the experienced students over novice students could potentially help

explain their superior performance on science inquiry tasks such as identifying correct final claims and designing causal explanations to support their final claims.

Implications for Instructional Design of Immersive Virtual Assessments

Our results on the development of SRL skills within IVAs provide implications for the instructional design of immersive virtual environments that assess science inquiry and learning of ill-structured science topics. IVAs, an open-ended virtual environment without any scaffolding embedded, have shown to foster self-regulated learning in our study. Meanwhile, 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). Therefore, adaptive scaffolds have great potentials in further invoking self-regulatory behaviors and strategies in IVAs. In the following section, we discuss the implications of our results for designing future immersive virtual environments to facilitate personalized learning and self-regulated learning.

To begin with, sequential pattern mining helped us identify a list of behavior patterns in IVAs that mapped with various SRL phases and explore how they developed over time. Despite that we were 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? With the under-representation of the planning process by the behavior sequences or notes, we do not have insights about whether IVAs promoted students to better plan their inquiry and problem solving process, let alone providing adaptive scaffolding to deploy goal setting and planning to students who were in need. To better evaluate, emulate, 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 notepad after being introduced to IVAs, 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., illustrate their subgoals on tool usage, data collection, and data analysis).

As students used the system for the second time, they were less likely to access the help button and take notes of their ultimate tasks in the notepad than students who used IVAs for the first time. This might suggest that students were familiar with what they were supposed to do and did not need to record it the second time they used IVAs. 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). Implementing these prompts to ensure that users have a good understanding of their tasks is especially meaningful when students were exposed to IVAs for the first time and not sure about what they should achieve.

This chapter'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 low word count in notes. If 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 important contents that are crucial for problem solving. 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, our analyses provide insights on designing scaffolds to foster generative note-taking in IVAs 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.

In this study, 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, IVAs 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.

The personalized scaffolds proposed above are meant to prompt and support students' self-regulatory processes and strategies in IVAs 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 immersive virtual assessments, benefitting not just their performance on the assessments, but what they take away from the experience.

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.

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. *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, R. S. J. d., & 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.

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

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).

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.

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.

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

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.

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.

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.

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.

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

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.

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.

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.

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.

Khachatryan, G. A., Romashov, A. M., Khachatryan, A. R., Gaudino, S. J., Khachatryan, J. M., Guarian, K. R., & Yuga, N. V. (2014). Reasoning Mind Genie 2: An intelligent tutoring system as a vehicle for international transfer of instructional methods in mathematics. *International Journal of Artificial Intelligence in Education, 24*(3), 333-382.

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.

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., & 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*.

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

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.

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. (2008). Self-regulated learning with hypermedia: The role of prior domain knowledge. *Contemporary Educational Psychology, 33*(2), 270–298.

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

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.

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.

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.

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.

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.

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.

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

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.

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

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: Routledge.

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.

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

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

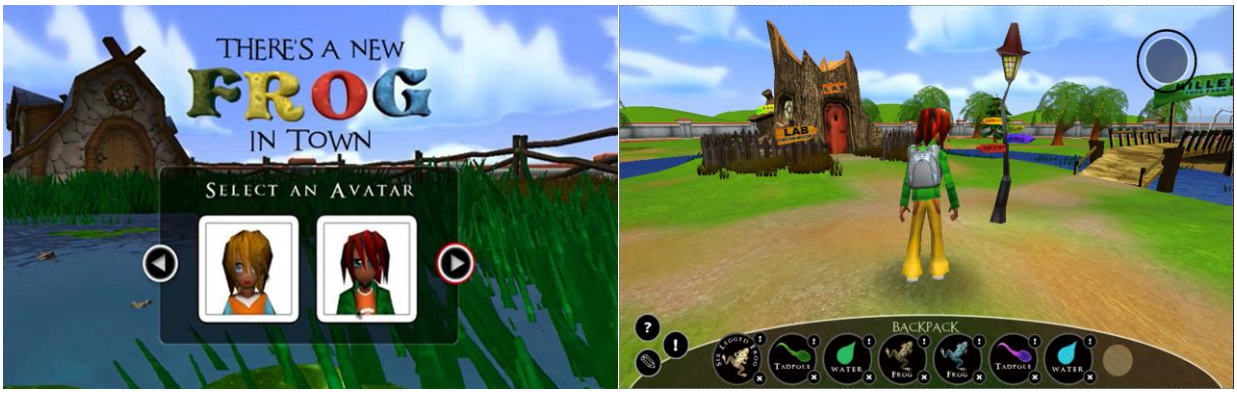

Figure 1. Screenshots of the IVA frog scenario.

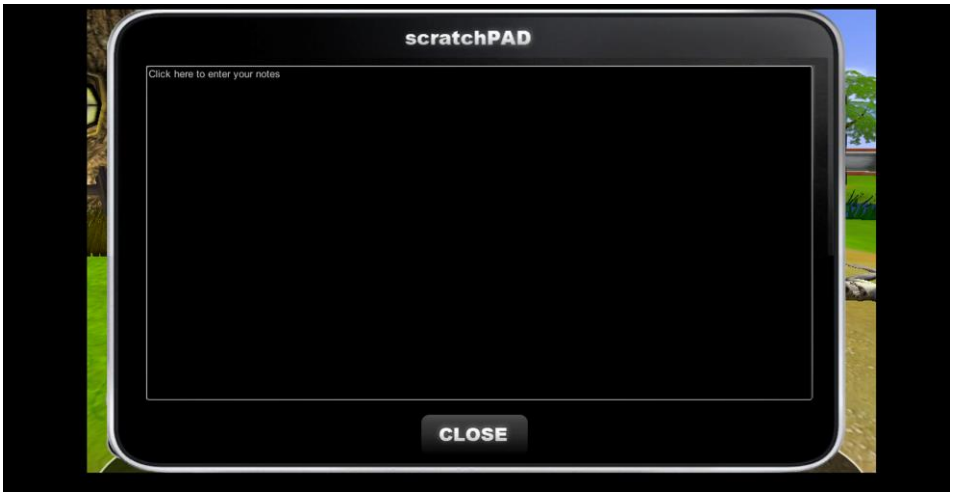

Figure 2. Screenshot of the digital notepad within IVA.

Table 1. Comparisons of the support and confidence of frequent sequential patterns related to self-regulatory processes and strategies between novice and experienced conditions in each scenario. Average support/confidence values and t-statistics from three-level regressions comparing differences in these values are reported for novice students (N) and experienced students (E) by scenario. Statistically significant results after Benjamini and Hochberg's post-hoc control are marked with *.

Pattern	Support in frog			Confidence in frog			Support in bee			Confidence in bee		
	N	E	t	N	E	t	N	E	t	N	E	t
help → note	.0007	.0003	−2.95*	.13	.13	−.03	.0007	.0003	−2.51*	.13	.17	1.14
read → note-MULT	.0063	.0098	4.77*	.26	.33	3.51*	.0064	.0093	2.90*	.27	.31	1.46
read-MULT → note-MULT	.0029	.0040	2.59*	.11	.15	3.03*	.0027	.0037	2.27	.11	.13	1.30
look → note	.0019	.0022	.90	.08	.12	2.48	.0012	.0019	2.19	.06	.09	2.30
experiment → note-MULT	.0031	.0030	−.42	.12	.14	1.48	.0025	.0031	1.47	.09	.16	3.91*
final claim → note	.0012	.0019	2.73*	.22	.26	1.27	.0007	.0015	3.75*	.08	.13	3.33*
final claim → note-MULT	.0007	.0014	3.46*	.16	.24	2.53*	.0007	.0010	1.56	.08	.09	.56
final claim → read	.0013	.0012	−.64	.25	.16	−2.78*	.0009	.0015	2.42	.10	.12	.86
final claim → read-MULT	.0020	.0024	.94	.37	.33	−.85	.0020	.0020	−.09	.21	.15	−2.98*

Table 2. Comparisons of features related to note-taking/reviewing quantity between novice students and experienced students in each scenario. Descriptive statistics (means with standard deviations in parentheses) of the features are reported for novice students (N) and experienced students (E) by scenario. T-statistics from three-level regressions comparing differences in note-taking/reviewing quantity by experience are reported. Statistically significant results after Benjamini and Hochberg's post-hoc control are marked with *. The first three features were computed for all students, and the remaining ones were computed for note-takers only.

Feature	Description	Frog-N	Frog-E	t	Bee -N	Bee -E	t
Notepad access frequency	Frequency of opening the notepad window	9.20 (12.98)	10.50 (15.63)	1.75	8.89 (12.24)	10.11 (15.34)	.52
Notepad time	Total amount of time in minutes that notepad was open	2.85 (4.14)	3.38 (5.50)	2.08	2.72 (3.82)	3.15 (5.09)	.74
Percent of time on notepad	Total amount of time on notepad divided by total time in IVA	11% (18%)	12% (19%)	.48	8% (10%)	10% (12%)	1.96
Word count in note	Number of words in note-taker's note	56.33 (54.74)	65.30 (65.22)	2.42*	50.88 (52.06)	66.60 (66.95)	3.76*
Segment count in note	Number of sentence segments in note-taker's note	7.88 (7.09)	9.65 (8.32)	3.69*	7.49 (6.49)	9.72 (8.62)	3.99*
Note-taking frequency	Frequency of note-taking actions	10.52 (10.13)	13.60 (12.66)	4.51*	10.27 (9.71)	13.72 (12.49)	4.23*
Note-reviewing frequency	Frequency of note-reviewing actions	4.30 (5.24)	4.77 (6.05)	1.10	4.17 (5.10)	4.56 (5.63)	.28
Percent note-taking actions	Frequency of note-taking divided by frequency of notepad access	73% (20%)	76% (19%)	2.58*	73% (20%)	77% (17%)	3.44*
Percent note-reviewing actions	Frequency of note-reviewing divided by frequency of notepad access	27% (20%)	24% (19%)	−2.58*	27% (20%)	23% (17%)	−3.44*
Note-taking duration	Total amount of time (in minutes) spent on taking notes	3.95 (3.53)	5.01 (4.97)	3.97*	3.86 (3.48)	4.99 (4.69)	3.77*
Note-reviewing duration	Total amount of time (in minutes) spent on reviewing notes	.74 (1.58)	.86 (1.91)	1.05	.65 (1.25)	.80 (1.47)	1.29
Avg note-taking duration	Average duration (in minutes) of a note-taking action	.48 (.47)	.44 (.33)	−1.48	.49 (.48)	.46 (.44)	−.92
Avg note-reviewing duration	Average duration (in minutes) of a note-reviewing action	.13 (.24)	.12 (.16)	−.11	.12 (.19)	.14 (.25)	1.40
Note-taking to notepad time	Ratio of time spent on note-taking actions and total time on notepad	88% (14%)	88% (13%)	1.01	88% (14%)	89% (13%)	1.44
Note-reviewing to notepad time	Ratio of time spent on note-reviewing actions and total time on notepad	12% (14%)	12% (13%)	−1.01	12% (14%)	11% (13%)	−1.44

Table 3. 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 IVA.	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 IVA.	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

Table 4. Comparisons of the number of various categories of segments in a student's note between note-takers in the novice group and the experienced group in each scenario. Descriptive statistics (means with standard deviations in parentheses) of the note content are reported for novice students (N) and experienced students (E) by scenario. T-statistics from three-level regressions comparing differences between the two groups are reported. Statistically significant results after Benjamini and Hochberg's post-hoc control are marked with *.

Note Segment	Frog-N	Frog-E	t	Bee -N	Bee -E	t
Reproduction	6.14 (6.57)	7.78 (7.57)	3.73*	6.05 (5.92)	7.86 (7.81)	3.52*
Elaboration	1.32 (1.93)	1.44 (2.45)	.99	1.16 (1.83)	1.55 (2.26)	2.72*
Metacognition	.10 (.49)	.09 (.87)	−.18	.08 (.69)	.07 (.40)	−.42
Test	1.99 (3.20)	1.68 (3.27)	−1.60	1.59 (2.74)	2.21 (3.59)	2.90*
Kiosk	3.13 (4.69)	4.41 (5.93)	3.94*	2.89 (4.55)	4.35 (5.77)	4.28*
Observation	2.16 (3.64)	3.11 (4.91)	3.80*	2.54 (4.04)	2.95 (4.41)	1.23
Dialogue	.30 (1.26)	.25 (1.32)	−.60	.34 (1.43)	.28 (1.49)	−1.03
Combination	.19 (.71)	.30 (1.51)	1.70	.25 (.89)	.46 (1.24)	2.99*
Hypothesis	.51 (1.06)	.53 (1.07)	.40	.34 (.87)	.55 (1.19)	3.49*
Draw Conclusion from Data	.46 (1.04)	.54 (1.21)	1.19	.36 (1.02)	.62 (1.24)	3.43*
Reproduction of Test	1.49 (2.63)	1.18 (2.46)	−2.00	1.22 (2.30)	1.68 (2.91)	2.60
Reproduction of Kiosk	2.80 (4.53)	3.95 (5.59)	3.68*	2.57 (4.34)	3.80 (5.36)	3.75*
Reproduction of Observation	1.58 (3.09)	2.42 (4.29)	3.93*	2.00 (3.53)	2.24 (3.80)	.76
Reproduction of Dialogue	.27 (1.20)	.23 (1.30)	−.48	.31 (1.35)	.26 (1.43)	−.90
Reproduction of Combination	.01 (.09)	.01 (.14)	.95	.09 (.53)	.13 (.54)	1.24
Elaboration on Test	.48 (1.13)	.50 (1.54)	.33	.35 (.86)	.51 (1.20)	2.35
Elaboration on Kiosk	.30 (.90)	.44 (1.47)	2.00	.28 (.83)	.48 (1.33)	2.94*
Elaboration on Observation	.56 (1.16)	.65 (1.38)	1.14	.51 (1.20)	.69 (1.40)	1.97
Elaboration on Dialogue	.03 (.22)	.01 (.10)	−1.42	.03 (.24)	.02 (.15)	−.90
Elaboration on Combination	.18 (.70)	.29 (1.49)	1.62	.17 (.62)	.33 (1.09)	3.03*