Comparing Novice and Experienced Students within Virtual Performance Assessments

Yang Jiang, Luc Paquette, Ryan S. Baker
Teachers College, Columbia University
525 W 120th Street
New York, NY 10027
yj2211@tc.columbia.edu
paquette@tc.columbia.edu
baker2@exchange.tc.columbia.edu

Jody Clarke-Midura
Utah State University
2830 Old Main Hill
Logan, UT 84322
jody.clarke@usu.edu

ABSTRACT

Inquiry skills are an important part of science education standards. There has been particular interest in verifying that these skills can transfer across domains and instructional contexts [4,15,16]. In this paper, we study transfer of inquiry skills, and the effects of prior practice of inquiry skills, using data from over 2000 middle school students using an open-ended immersive virtual environment called Virtual Performance Assessments (VPAs) that aims to assess science inquiry skills in multiple virtual scenarios. To this end, we assessed and compared student performance and behavior within VPA between two groups: novice students who had not used VPA previously, and experienced students who had previously completed a different VPA scenario. Our findings suggest that previous experience in a different scenario prepared students to transfer inquiry skills to a new one, leading these experienced students to be more successful at identifying a correct final conclusion to a scientific question, and at designing causal explanations about these conclusions, compared to novice students. On the other hand, a positive effect of novelty was found for motivation. To better understand these results, we examine the differences in student patterns of behavior over time, between novice and experienced students.

Keywords

Virtual environment, science inquiry, educational data mining, sequential pattern mining, transfer, novelty effect.

1. INTRODUCTION

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. As such, science inquiry skills have been a critical component of the K-12 science curriculum standards [18]. It is particularly crucial that students acquire inquiry skills which are not specific to a domain or instructional context, but which can transfer broadly, preparing students for using science and understanding science in their future schooling, and in their lives [4, 15, 16].

With the increasing popularity of online learning systems that engage learners in science inquiry activities [e.g., 7, 21], Educational Data Mining (EDM) techniques have proven effective in automatically assessing science inquiry skills. Sao Pedro et al. demonstrated that science inquiry skills can be assessed within online learning activities using EDM, predicting future performance not only within the same domain [21], but also across domains [22].

Many studies of student inquiry behavior have been conducted within open-ended online learning environments, such as virtual environments, in which users have the freedom to decide their own inquiry behaviors. This, combined with the fact that these open-ended environments are typically more loosely-scaffolded and coarser-grained than more tightly-scaffolded systems such as intelligent tutoring systems or simulations [e.g., 21], makes the assessment of science inquiry in these contexts challenging. Sequential Pattern Mining [1], a methodology that has been extensively used in EDM [23], has shown potential in discovering complicated patterns of learning behavior within open-ended learning environments. For example, Kinnebrew and colleagues [13] applied sequential pattern mining techniques to log data produced by students engaging in activities within Betty's Brain, an open-ended learning environment for science learning. This allowed them to study the differences in students' productive and unproductive learning behaviors by identifying frequent sequential patterns related to the use of concept maps and determining which sequential patterns were characteristic of high-performing students as compared to low-performing students. Differential pattern mining was also used by Sabourin and colleagues [20] to analyze the differences in inquiry behaviors utilized by learners depending on their level of self-regulation within a virtual environment. In another study, Gutierrez-Santos et al. [10] conducted analysis of student actions to detect repetitive sequences in an open-ended learning environment.

Another EDM approach that has proven useful for the study of inquiry behaviors in open-ended contexts involves in-depth analysis of features distilled from log data. For instance, Baker and Clarke-Midura [2] distilled a set of features related to inquiry behavior from log data in Virtual Performance Assessments (VPAs), an open-ended immersive virtual environment used in the current study, to develop predictive models of student success on two inquiry tasks. The current study combines both sequential pattern mining and analysis of features related to science inquiry to study transfer of inquiry skills. In doing so, we also analyze differences in inquiry behavior between novice students and experienced students.

The degree of student experience with an environment can also be hypothesized to have important impact on their inquiry. Clark [6] argued that 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., 8, 12, 24] indicated that students showed greater initial enthusiasm and motivation in classrooms when novel educational technologies were introduced. This enthusiasm gradually diminished as students were more familiar with the technologies and the initial novelty effect wore off. Therefore, in our study, we investigate whether relative novelty created by the introduction of a new 3D virtual environment will lead to differences in motivation and learning between novice students and experienced students. We also study the relationship between the potential novelty effect and inquiry skills.

To research these questions, we assess and compare student performance and behavior within VPA between two groups: novice students who had not used VPA previously, and more experienced students who had previously spent one class session completing a different VPA scenario. We compare student performance on two inquiry skills – identifying a correct final claim and designing causal explanations. We also compare student responses to a motivation survey between the two groups. Finally, we analyze the difference in student behavior between the two groups using differential sequence mining.

2. VIRTUAL PERFORMANCE ASSESSMENTS

This study was conducted within the context of Virtual Performance Assessments (VPAs; see http://vpa.gse.harvard.edu). VPAs are online 3D immersive virtual environments, designed using the Unity game development engine [26] that assess middle school students' science inquiry skills, in line with state and national standards for science content and inquiry processes. Within VPAs, whose interface is similar to that of video games, students engage in authentic inquiry activities and solve scientific problems by navigating around the virtual environment as an avatar, making observations, interacting with non-player characters (NPCs), gathering data, and conducting laboratory experiments. VPAs enable automated and non-intrusive collection of process data (logged actions and behaviors) and product data (student final claims), facilitating the capture and assessment of science inquiry *in situ*.

Multiple VPA assessment scenarios have been developed. In this study, two scenarios were used, the frog scenario and the bee scenario. In the frog scenario (see Figure 1), students are presented with a six-legged frog in the virtual environment and have to collect and reason through evidence to determine what is causing the frog's mutation, selecting from a set of possible causal factors including parasites (the correct causal explanation), pesticides, pollution, genetic mutation, and space aliens. In this scenario, students can talk with NPCs from four virtual farms who provide conflicting opinions, collect items such as frogs, tadpoles, and water samples at each farm, run laboratory experiments on water quality, frog blood and DNA, and read informational pages from a research kiosk. Once students think that they have sufficient data, they submit a final conclusion on the causal factor resulting in the mutation, and justify their final claim with supporting evidence. In the bee scenario, students must determine what causes the death of a local bee population. Similar to the frog scenario, they can talk with NPCs from four different farms, read informational pages at the research kiosk, and conduct tests (e.g., nectar test, protein test, genetic test) on the items they have collected at the farms (e.g., nectar samples, bees, larvae). By the end of the assessment, students choose a final claim about the cause of the bee deaths from possible hypotheses including genetic mutation (the correct causal factor), parasites, pesticides, pollution, and space aliens, and support their final claim with evidence. The activities in each VPA scenario are deliberately similar, allowing researchers to assess performance of the same inquiry practices in different contexts.

Figure 1. Screenshots of the VPA frog scenario.

3. DATA SET

Data for this study was composed of action logs produced by middle school students who used Virtual Performance Assessments within their science classes at the end of the 2011-2012 school year. A total of 2,431 students in grades 7-8 (12-14 years old) from 138 science classrooms (40 teachers) participated in this study. These students were from a diverse range of school districts in the Northeastern and Midwestern United States, and Western Canada. A total of 1,985 students completed the frog scenario and 2,023 students completed the bee scenario, with 1,579 students completing both scenarios. Overall, students completed 423,616 actions within the frog scenario and 396,863 actions within the bee scenario. They spent an average of 30 minutes and 47 seconds (*SD* = 14 minutes, 6 seconds) in the frog scenario and an average of 26 minutes and 5 seconds (*SD* = 12 minutes, 27 seconds) in the bee scenario.

The 2,431 students were randomly assigned to begin with either the frog scenario or the bee scenario. Two weeks later, they were assigned to complete the other scenario. Therefore, within each scenario, participants could be put into two groups – novice users who were using VPA for the first time (*novice* group), and experienced users who had previously experienced the other VPA scenario (*experienced* group). Accordingly, among the 1,985 students who completed the frog scenario, 1,232 completed the frog scenario as their first scenario (frog-novice) and 753 had previous experience in the bee scenario (frog-experienced). Among the students who completed the bee scenario, 1,198 students had no previous experience in the frog scenario (bee-novice), whereas 825 had previous experience in the frog scenario (bee-experienced). Student actions and performance in the virtual environment were logged as they worked within each VPA scenario and used for later analyses.

4. OVERALL ANALYSIS

In this section, we compare student performance on identifying a correct final claim and constructing causal explanations, the amount of time spent on VPA, and students' motivation level, between the novice group and the experienced group, within each VPA scenario.

4.1 CFC and DCE Performance

To explore the transfer of student science inquiry skills between scenarios, two measures of student performance within the VPAs were collected and compared between the two groups of students within each scenario: 1) the correctness of the student's final claim (CFC) on the cause of the six-legged frog or the death of the

bees; and 2) student's success in designing causal explanations (DCE) for why that claim is correct.

In each VPA, students submitted a final claim by choosing from five possible causal factors. A student's final claim was considered correct if the student concluded that the mutation of the six-legged frog was caused by parasites (correct causal factor), or that the bee deaths were caused by genetic mutation (correct choice). Otherwise, if the student selected the other potential hypotheses, the student's final claim was considered incorrect. Overall, 29.6% of students correctly concluded that parasites led the frog to have six legs, and 28.3% of students made a correct claim on what was killing the bee population. In this paper, a chi-square test was conducted to compare student CFC performance between the two groups in each scenario.

In the bee scenario, 34.8% of experienced students who had previously used the frog scenario identified correctly that genetic mutation was killing the bees, while 23.9% of novice students (*without* prior experience in the frog scenario) made the correct final conclusion. This difference was statistically significant according to a chi-square test, $\chi^2(1, N = 2023) = 28.67$, $p < .001$. Logistic regression results revealed that the odds of making a correct final claim for experienced students (0.533) was statistically significantly larger than the odds for novice students (0.314) by 70%. This suggested that the students transferred what they learned about how to make a correct final claim from the frog scenario to the bee scenario.

Similarly, in the frog scenario, a statistically significantly higher percentage of experienced students (33.2%) made a correct final claim than the percentage of novice students (27.5%) who made a correct conclusion, $\chi^2(1, N = 1985) = 7.45$, $p = .006$. Logistic regression results indicated that previous experience in the bee scenario significantly improved the odds of making a correct final claim in the frog scenario by 31.5% (odds = 0.378 for novice group and 0.497 for experienced group).

The DCE measure evaluates student ability in supporting final conclusions with evidence. By the end of the assessment in each scenario, students needed to select the evidence that supported their claims from the data they had collected within the virtual world and the results of laboratory tests they had conducted. They were then presented with all possible data (including data that the students did not collect/conduct) and asked to identify the evidence supporting their claim. In each VPA scenario, most evidence was consistent with the correct causal claim. However, for the incorrect claims, there was often evidence consistent with those claims along with counter-evidence that conclusively disproved those hypotheses. Therefore, even if students were unsuccessful in identifying the correct final conclusion, 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. Student success in selecting evidence and constructing causal explanations were aggregated into a single composite DCE measure that ranges from 0 to 100%, by averaging across the use of each piece of evidence. The mean DCE score for the frog scenario was 50.0% (*SD* = 23.3%), and the average DCE score for the bee scenario was 46.1% (*SD* = 21.4%). A two-tailed Mann-Whitney U test, a nonparametric alternative to t-test, was then conducted to compare student ability in designing causal explanations between the two groups in each scenario.

Results of the Mann-Whitney U test comparing the DCE score between the two groups in the bee scenario suggested that the experienced group had a significantly higher average DCE score (*M* = 48.9%, *SD* = 19.3%) than the novice group (*M* = 44.2%, *SD* = 23.8%), *U* = 453873, *Z* = -3.12, *p* = .002. Further analyses revealed that the difference in DCE performance was dependent on the correctness of final claims. Among students who made a correct final claim in the bee scenario, the experienced group achieved significantly higher DCE scores (*M* = 75.1%, *SD* = 18.3%) than the novice group (*M* = 68.1%, *SD* = 20.5%), *U* = 32448.5, *Z* = -4.34, *p* < .001. However, among students who did not make a correct final claim, the novice group showed higher DCE scores (*M* = 36.7%, *SD* = 11.2%) than the experienced group (*M* = 34.9%, *SD* = 11.4%), *U* = 223797, *Z* = -2.80, *p* = .005.

In the frog scenario, student performance in designing causal explanations for the novice group (*M* = 49.7%, *SD* = 22.7%) was not statistically significantly different from the experienced group (*M* = 50.6%, *SD* = 24.3%), *U* = 454398, *Z* = -.76, *p* = .446.

4.2 Time

As each VPA scenario logged the timing of each student starting and exiting the virtual environment, we also compared the total amount of time students spent within VPA recorded by the log data between the novice group and the experienced group, by employing one-way ANOVA.

An analysis of variance showed that, on average, novice students without previous experience in the frog scenario spent significantly more time in the bee scenario (*M* = 27 minutes, 43 seconds, *SD* = 11 minutes, 56 seconds) than experienced students who had used the frog scenario (*M* = 23 minutes, 43 seconds, *SD* = 12 minutes, 48 seconds), *F* (1, 2021) = 51.64, *p* < .001. On the other hand, the total amount of time spent in the frog scenario by novice students (*M* = 30 minutes, 56 seconds, *SD* = 14 minutes, 24 seconds) and experienced students (*M* = 30 minutes, 33 seconds, *SD* = 13 minutes, 35 seconds) was not statistically significantly different (*F* (1, 1983) = .36, *p* = .548).

4.3 Motivation

In this study, students completed an online motivation survey shortly after they finished the VPA assessment for each scenario. Student responses to the survey were analyzed to better understand the impact of experience with the environment on learning and motivation. The survey was adapted from the Intrinsic Motivation Inventory [IMI; 27] and the Player Experience of Need Satisfaction [PENS; 19] survey and was comprised of 27 six-point Likert-type items that aimed to measure seven components related to student motivation, autonomy, and in-game immersion: interest/enjoyment, perceived competence, effort/importance, pressure/tension, value/usefulness, presence/immersion, and autonomy. Items were slightly modified to fit the specific activity in this game-like environment. Student subscale scores were calculated by averaging across all items on each subscale. One-way ANOVA was applied to assess whether there were any systematic differences in student motivation between the novice group and the experienced group within each VPA scenario. Given the substantial number of statistical tests, we controlled for the proportion of false positives by applying Storey's q-value method [25] (calculated using the QVALUE package for R).

Analyses of motivational survey results (see Table 1) indicated that, on average, novice students scored significantly higher on the interest/enjoyment subscale than experienced students in both scenarios (*F*(1, 1800) = 50.02, *q* < .001 for the frog scenario; *F*(1, 1740) = 27.67, *q* < .001 for the bee scenario). Similarly, students

in the novice group had a significantly higher level of perceived effort invested to the VPA activity and perceived importance of the activity than students in the experienced group ($F(1, 1800) = 25.41$, $q < .001$ for the frog scenario; $F(1, 1740) = 18.94$, $q < .001$ for the bee scenario). Novice students also regarded the VPA activity as more useful and valuable than experienced students, $F(1, 1800) = 19.37$, $q < .001$ for the frog scenario; $F(1, 1740) = 4.66$, $q = .019$ for the bee scenario. Finally, novice students also had significantly higher presence/immersion, autonomy, and tension/pressure subscale scores than the experienced students, indicating that they were more immersed in the virtual environment, and felt a higher sense of autonomy and a higher level of tension/pressure than experienced students. These corresponded to previous findings on novelty effect [8, 12].

Table 1. Average subscale scores on the motivational survey (standard deviations in parentheses) by condition. Differences that are sig. after post-hoc controls ($q < 0.05$) are marked by *.

Subscale	Frog-N	Frog-E	*F (q)*	Bee-N	Bee-E	*F (q)*
int/enj	4.47 (1.32)	3.98 (1.55)	50.02* (<.001)	4.26 (1.42)	3.87 (1.56)	27.67* (<.001)
comp	4.28 (1.21)	4.23 (1.37)	0.73 (.213)	4.13 (1.27)	4.14 (1.37)	0.006 (.473)
eff/imp	4.38 (1.19)	4.06 (1.44)	25.41* (<.001)	4.21 (1.30)	3.91 (1.49)	18.94* (<.001)
val/use	4.07 (1.41)	3.74 (1.62)	19.37* (<.001)	3.84 (1.51)	3.67 (1.64)	4.66* (.019)
pres/ten	1.86 (1.25)	1.72 (1.39)	4.62* (.019)	1.85 (1.29)	1.69 (1.38)	5.86* (.011)
pres/imm	3.51 (1.36)	3.16 (1.53)	24.72* (<.001)	3.36 (1.42)	3.13 (1.53)	10.14* (.001)
auto	4.26 (1.29)	3.82 (1.55)	41.12* (<.001)	4.01 (1.41)	3.76 (1.56)	11.42* (.001)

Note. Frog-N = frog-novice, Frog-E = frog-experienced, Bee-N = bee-novice, Bee-E = bee-experienced. Int/enj=interest/enjoyment, comp=perceived competence, eff/imp=effort/importance, pres/ten=pressure/tension, val/use=value/usefulness, pres/imm=presence/immersion, auto= autonomy.

5. USAGE ANALYSIS

In the previous section, differences were found in motivation and learning outcomes between novice and experienced students. In the current section, we aim to go beyond just looking at whether previous experience in VPA improved student inquiry performance, and instead look into whether more experienced students used VPAs differently than less experienced students. For example, this will allow us to determine whether the higher success for experienced students within VPAs was related to the acquisition and transfer of science inquiry skills, or whether it was merely the result of increased familiarity and proficiency with using the system and tools than novice users.

We studied these questions by investigating the prevalence of specific behaviors between groups, and by applying sequential pattern mining to identify and compare the frequent sequential patterns of student actions between groups.

5.1 Comparing Behaviors Between Groups

In order to understand student behavior, and how it differed between groups, a set of 30 semantically meaningful features of student behavior thought to potentially differ between groups were distilled from raw interaction data and were compared between the novice and experienced groups in each scenario. These features were a subset of the 48 features that were used to build models predicting a student's CFC and DCE performance within the frog scenario in [2]. Examples of these features will be given in the following paragraphs.

After distilling the 30 features from each student's interaction logs, t-tests were conducted to compare the value of each feature between the experienced and novice groups, within each scenario. Storey's q-values [25] were calculated to control for multiple comparisons. Table 2 presents the average values of 10 features that strongly differentiated between groups.

According to the results, features representing the maximum or average fullness of a student's backpack in the frog scenario, both including repeats (e.g. picking up two green frogs counts as two objects), and not including repeats (e.g. two green frogs counts as one object), had significantly higher value for the novice group than the experienced group. Similar results were found in terms of the number of times a student went to the lab to run tests, the number of different (types of) non-sick frogs that the student took to the lab at the same time, the number of times that lab water was taken to the lab, and the percentage of time the student spent at farms to collect evidence in the frog scenario. Similarly, novice students in the bee scenario had higher values on all these features than experienced students. This suggested that novice students collected significantly more data for testing and spent a larger proportion of time on collecting evidence in farms than the experienced students in both scenarios. This finding was consistent with the higher motivation level of novice students (in both scenarios) and the longer time they spent working on VPA

Table 2. Comparisons of features between novice group and experienced group. Sig. differences ($q<0.05$) are marked by *.

Feature	Frog-N	Frog-E	t	q	Bee-N	Bee-E	t	q
The number of times student went to the lab	6.66	5.14	6.81	<.001*	16.37	12.71	8.97	<.001*
Maximum number of items (including repeats) in backpack	7.48	6.69	11.25	<.001*	6.03	4.76	11.57	<.001*
Maximum number of items (not including repeats) in backpack	7.45	6.65	11.68	<.001*	8.54	7.28	12.27	<.001*
Average number of items (including repeats) in backpack	4.77	4.02	11.39	<.001*	3.86	3.06	11.91	<.001*
Average number of items (not including repeats) in backpack	4.75	4.00	11.50	<.001*	6.17	5.14	11.61	<.001*
Number of times that lab water/nectar was taken to the lab	0.42	0.38	2.11	.022*	1.69	0.93	8.31	<.001*
Number of different (types of) non-sick frogs/bees student took to the lab at the same time	1.87	1.70	2.34	.014*	4.32	3.90	4.09	<.001*
How long, on average, did students spend reading information pages? (average per read)	15.28	17.17	-0.72	.146	11.93	13.93	-2.07	.027*
How long, in total, did student spend reading information page on correct hypothesis?	32.33	35.13	-0.70	.146	23.45	27.46	-2.20	.021*
Percentage of time student spent at farms	0.29	0.26	4.43	<.001*	0.34	0.31	5.46	<.001*

(in the bee scenario).

Despite the fact that the novices collected more data and spent more total time within the VPA bee scenario, they spent significantly less time on reading an information page at the research kiosk each time they accessed the page (M = 11.93 seconds, SD = 17.69 seconds) than experienced students (M = 13.93 seconds, SD = 25.48 seconds), t (2021) = -2.07, q = 0.027, *Cohen's D* = 0.15. In specific, experienced students spent more time in total reading the information page on the correct hypothesis – genetic mutation (M = 27.46 seconds, SD = 46.51 seconds) compared to novice students (M = 23.45 seconds, SD = 35.46 seconds), t (2021) = -2.20, q = 0.021, *Cohen's D* = 0.11. Gaining more information about the correct hypothesis might have contributed to the students' domain-specific knowledge base, which had been found to be crucial for problem solving and the development of expertise [5]. However, the corresponding pattern was not statistically significant for the frog scenario, probably due to higher standard deviations.

5.2 Sequential Pattern Mining

In this section, we investigate patterns in behavior by the two groups, over time. Prior to performing sequential pattern mining, detailed raw action 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*), talking with NPCs (*talk*), saving objects to backpack (*save*), discarding objects (*discard*), opening and reading informational pages at the research kiosks (*read*), running laboratory tests (*blood/protein test, water/nectar sample test, genetic test*), reviewing and looking at test results (*look*), 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, closing the scratchpad, and entering/exiting a specific area were filtered out from the raw interaction data. Second, as in [13], 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. This adjustment prevents frequent patterns from being overlooked merely due to differences in how many times the action is repeated. 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 [11] within the statistical software program R. Arules was used to determine the most frequent short sequences of two actions 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 51 short sequential patterns (length = 2) were identified that met the minimum support and confidence constraints within the novice group; 54 patterns were identified within the experienced group. In the bee scenario, 55 short sequential patterns met the minimum constraints within the novice group; 59 were selected within the experienced group. These patterns were similar across the 4 conditions, and most had support and confidence considerably higher than the threshold. They were then ordered according to their *Jaccard* similarity coefficient – a measure of the patterns' interestingness [17] that was found to be the most highly correlated with human judgments [3] – to find interesting sequential patterns. According to [3], lower *Jaccard* measures indicated higher interestingness.

To facilitate the comparison of the frequency measures between the novice group and the experienced group, the support and confidence for each pattern were calculated separately for each student. Mann-Whitney U tests that controlled for multiple comparisons were then conducted to compare the metric values between two groups in each scenario.

Table 3 presents the comparison of the support and confidence levels of 9 frequent sequential patterns with low *Jaccard* measure (indicating high interestingness) across conditions that were considered as meaningful due to the nature of the actions they contained. The sequential patterns with the lowest *Jaccard* included patterns related to making final claims (*final claim*) or starting to answer final questions (*start final questions*) and reading informational pages (*read*), such as "*final claim → read-MULT*", "*final claim → read*", "*read-MULT → final claim*", "*start final questions → read-MULT*", and "*start final questions → read*". These patterns indicated that students tended to review research and read informational pages as resources to assist with their decision-making before submitting a final claim, or that they used the research information to check and monitor the claims they had just made. All these 5 patterns appeared to have higher support for experienced students than novice students within each

Table 3. Comparison of the support and confidence of 9 frequent sequential patterns between novice and experienced conditions. Average support/confidence values, and post-hoc controlled sig. of tests are presented. Sig. differences (q<0.05) are marked by *.

	support			confidence			support			confidence		
Pattern	**Frog-N**	**Frog-E**	**q**	**Frog-N**	**Frog-E**	**q**	**Bee-N**	**Bee-E**	**q**	**Bee-N**	**Bee-E**	**q**
final claim → read-MULT	.0033	.0043	.420	.296	.313	.594	.0030	.0036	.619	.326	.298	.420
read-MULT → final claim	.0061	.0074	.584	.114	.109	.619	.0055	.0064	.675	.101	.109	.594
final claim → read	.0020	.0026	.675	.164	.158	.675	.0014	.0024	.018*	.142	.193	.107
start final questions→ read-MULT	.0046	.0047	.594	.282	.261	.594	.0044	.0049	.675	.274	.257	.594
start final questions → read	.0029	.0033	.682	.160	.167	.675	.0025	.0027	.675	.147	.142	.675
look-MULT → read-MULT	.0027	.0032	.718	.143	.176	.517	.0028	.0030	.594	.141	.189	.309
look → read	.0025	.0028	.711	.103	.142	.214	.0017	.0021	.675	.080	.107	.361
look → read-MULT	.0027	.0033	.675	.113	.158	.073	.0019	.0027	.594	.105	.155	.018*
look-MULT → read	.0021	.0021	.594	.104	.117	.675	.0021	.0017	.018*	.106	.101	.420

scenario, but most of the differences were not statistically significant. In the bee scenario, the pattern *final claim → read* showed significantly higher support and marginally significantly higher confidence for the experienced group than the novice group (for *support*, $Ms = 0.024$ and 0.014, $U = 474169.5$, $Z = -3.03$, $q = 0.018$; for *confidence*, $Ms = 0.193$ and 0.142, $U = 46833.5$, $Z = -2.32$, $q = 0.107$). This finding indicated that experienced students who had previously used the frog scenario were more likely to review research and read information, possibly to monitor their answers and reflect on previous steps [cf. 15], after submitting a final claim in the bee scenario than novice students. However, this trend was not replicated in the frog scenario (for *support*, $Ms = 0.0026$ and 0.0020, $U = 462294.5$, $Z = -.23$, $q = .675$; for *confidence*, $Ms = 0.158$ and 0.164, $U = 58423.5$, $Z = -.32$, $q = .675$).

Another four interesting sequential patterns corresponded to looking at laboratory test results (once or repeatedly), followed by reading informational pages (once or repeatedly) (i.e., *look-MULT → read-MULT, look → read, look → read-MULT, look-MULT → read*). For three out of the four patterns, both the support and the confidence for the experienced group were higher than those for the novice group in both scenarios. For the pattern *look → read-MULT*, the confidence for the experienced group was statistically significantly higher than that for the novice group in the bee scenario and marginally higher than confidence for the novice group in the frog scenario (in bee scenario, $Ms = 0.105$ and 0.155, $U = 94500.5$, $Z = -3.09$, $q = .018$; in frog scenario, $Ms = 0.113$ and 0.158, $U = 111697.5$, $Z = -2.53$, $q = .073$). That is, experienced students were more likely to read multiple research information pages on possible causal factors immediately after looking at the results of lab tests. 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 [9]. The higher relative frequency of reading research information, which might help experienced students interpret laboratory test results and facilitate the acquisition of domain-specific knowledge [4], might have contributed to their higher success on making correct final claims than novice students.

In addition to two-action patterns, a differential sequence mining technique developed by Kinnebrew and colleagues [13] was utilized for identifying longer sequential patterns (length > 2) that occurred with significantly different frequencies between the two 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 [13]. 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 sequential pattern mining algorithm [13]. 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 q-value post-hoc control [25] was applied to select significantly differentially frequent patterns.

The 25 most differentially frequent long patterns with at least three consecutive actions were identified in the frog scenario and the 32 differentially frequent long patterns were identified in the bee scenario by employing a cutoff s-support of 50% and a cutoff q-value of 0.05 for comparison of pattern usage between two groups. 14 out of the 25 long patterns in the frog scenario and 16 out of 32 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 novice group. 11 long patterns in frog scenario and 16 in the bee scenario were frequently used only by students in the novice group. All differentially frequent long patterns had a

Table 4. Top differentially frequent patterns between the novice group (nov) and the experienced group (exp).

Scenario	Pattern	s-support		i-support			Frequent
		nov	exp	nov	exp	q	
Frog	talk-MULT → inspect → save → inspect → save	0.58	0.36	0.78	0.45	<.001	nov
	talk-MULT → inspect → save → inspect	0.59	0.37	0.79	0.46	<.001	nov
	save → discard → inspect → save	0.53	0.36	0.74	0.48	<.001	nov
	inspect → save → discard → inspect	0.53	0.36	0.75	0.49	<.001	nov
	inspect → save → discard → inspect → save	0.53	0.36	0.74	0.48	<.001	nov
	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	nov
	talk-MULT → inspect → save → inspect → save → inspect → save	0.59	0.27	0.71	0.32	<.001	nov
	talk-MULT → inspect → save → inspect → save	0.74	0.45	0.99	0.57	<.001	nov
	talk-MULT → inspect → save → inspect	0.74	0.45	0.99	0.57	<.001	nov
	start assessment → talk-MULT → inspect	0.51	0.26	0.51	0.26	<.001	nov
	talk-MULT → inspect → save	0.85	0.62	1.30	0.87	<.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

higher s-support and a significantly higher average i-support for novice students than experienced students.

Table 4 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 novice group 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. It seemed that novice students who had not used VPA before executed more sequences comprised of exploratory behaviors such as talking with NPCs and collecting data, while more experienced students focused primarily on what was necessary to answer the core inquiry question.

6. DISCUSSION AND CONCLUSION

This paper investigates the transfer of student science inquiry skills across two Virtual Performance Assessment scenarios, and the impact of the novelty of the immersive virtual environment on motivation and learning. We do so by comparing performance and behaviors between novice students and experienced students. A novelty effect was found as novice students who engaged in VPA for the first time showed significantly higher scores on motivational survey subscales such as interest/enjoyment, effort/importance, pressure/tension, value/usefulness, presence/immersion, and autonomy than more experienced students. As these students were first introduced to the novel 3D virtual environment, the initial attraction and attention led to higher enjoyment, greater effort invested in the tasks, a higher sense of immersion and a higher sense of autonomy. These measures tended to decline when students became relatively experienced and familiar with the environment, consistent with previous findings on the novelty effect [8, 12]. Sequential pattern mining and comparison of overall behavior prevalence using student action log data indicated that novice students engaged in more exploratory behaviors -- they collected more data in the environment and had higher frequency of long sequences comprised of exploratory actions such as talking with NPCs, manipulating objects, and collecting data, as compared to more experienced students. This, again, might be attributed to the novelty effect [cf. 14]. That is, the higher attention of novice students resulted in higher interest and efforts in exploring the new learning environment than students who were more experienced with VPA.

However, another possibility is that the experienced students focused more on the goal at hand, than on the environment they were researching this issue on, leading to less exploration and more attention directly to the information most likely to be useful. This itself may reflect the fact that novelty is wearing off, but may be a positive aspect of the disappearance of the novelty effect. Indeed, despite the experienced students' relatively lower motivation and fewer exploratory behaviors, they outperformed the novice students in identifying a correct final claim in both scenarios and in designing causal explanations (in one scenario). Experienced students generally showed more effective problem solving. They not only tended to read research information pages more often immediately after submitting a final claim or reviewing laboratory test results, but also spent more time reading the information each time they accessed a new page. As such, even after just a half hour completing the first assessment, students demonstrated more expert-like science inquiry behaviors -- they made more use of the research information available as resources [cf. 9], in order to either interpret results, or to monitor and reflect on their final claims [cf. 15]. The information from the pages may also have added to the domain-specific knowledge base of experienced students, which have been found to be crucial for problem solving and expertise development [5]. This corresponds to earlier findings that the transfer of domain-general inquiry strategy has the potential to facilitate the acquisition of domain-specific knowledge [4]. In conclusion, the experienced students successfully consolidated and transferred science inquiry skills they had learned from the first scenario during the approximately 30-minute engagement to the second scenario.

The current study contributes to research on the assessment of the transfer of science inquiry skills by proposing the application of a combination of educational data mining techniques such as sequential pattern mining as supplements to the traditional analysis of success between conditions. One limitation of this study is that 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 transfer. Future work may involve exploring whether far transfer of science inquiry occurs from VPA to assessments outside the system (e.g., other computer-based learning environments with different domain and interaction design).

7. ACKNOWLEDGMENTS

The research presented here was supported by the Bill and Melinda Gates Foundation. We also thank Chris Dede for his support and suggestions.

REFERENCES

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

[2] Baker, R.S.J.d., Clarke-Midura, J. 2013. Predicting successful inquiry learning in a Virtual Performance Assessment for science. In *Proceedings of the 21st International Conference on User Modeling, Adaptation, and Personalization*, 203-214.

[3] Bazaldua, D. A. L., Baker, R. S., San Pedro, M. O. Z. 2014. Combining expert and metric-based assessments of association rule interestingness. In *Proceedings of the 7th International Conference on Educational Data Mining*, 44-51.

[4] 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.

[5] Chi, M. T. H., Glaser, R., & Rees, E. 1982. Expertise in problem solving. In *Advances in the Psychology of Human Intelligence*, R. Sternberg, Ed. Vol. 1, Erlbaum, Hillsdale, NJ, 7–76.

[6] Clark, R. E. 1983. Reconsidering research on learning from media. *Review of educational research*, *53*(4), 445-459.

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

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

[9] 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.

[10] Gutierrez-Santos, S., Mavrikis, M., & Magoulas, G. 2010. Sequence detection for adaptive feedback generation in an exploratory environment for mathematical generalisation. In *Artificial Intelligence: Methodology, Systems, and Applications*. Springer Berlin Heidelberg, 181-190.

[11] 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).

[12] 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.

[13] 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.

[14] 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.

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

[16] Kuhn, D., Schauble, L., & Garcia-Mila, M. 1992. Cross-domain development of scientific reasoning. *Cognition and Instruction*, 9, 285-327.

[17] Merceron, A., & Yacef, K. 2008. Interestingness measures for association rules in educational data. *Educational Data Mining*, 8, 57-66.

[18] National Research Council. 2011. *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. The National Academies Press, Washington, DC.

[19] Ryan, R., Rigby, C., & Przybylski, A. 2006. The motivational pull of video games: A self-determination theory approach. *Motivation & Emotion, 30(4),* 344-360.

[20] Sabourin, J., Mott, B., & Lester, J. 2013. Discovering behavior patterns of self-regulated learners in an inquiry-based learning environment. In *Artificial Intelligence in Education* (Jan. 2013), Springer Berlin, Heidelberg, 209-218.

[21] Sao Pedro, M., Baker, R., Gobert, J., 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*, 1-39.

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

[23] Scheuer, O., & McLaren, B. M. 2012. Educational data mining. In *Encyclopedia of the Sciences of Learning*. Springer US, 1075-1079.

[24] Schofield, J. W. 1995. *Computers and Classroom Culture*. Cambridge University Press, New York, NY.

[25] Storey J. 2002. A direct approach to false discovery rates. *J Roy. Stat. Soc.*, 64, 479-498.

[26] Unity Technologies. 2010. *Unity Game Engine*.

[27] University of Rochester. 2015. *Intrinsic Motivation Inventory.* Retrieved January 25, 2015, from http://www.selfdeterminationtheory.org/intrinsic-motivation-inventory/