JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Volume 10(1), 168–188. https://doi.org/10.18608/jla.2023.7681

Serious Game Analytics by Design: Feature Generation and Selection Using Game Telemetry and Game Metrics: Toward Predictive Model Construction

Wenyi Lu[1], Joe Griffin[2], Troy D. Sadler[3], James Laffey[4], Sean P. Goggins[5]

Abstract

The construction of prediction models reflecting players' learning performance in serious games currently faces various challenges for learning analytics. In this study, we design, implement, and field test a learning analytics system for a serious game, advancing the field by explicitly showing which in-game features correspond to differences in learner performance. We then deploy and test a system that provides instructors with clear signals regarding student learning and progress in the game, which instructors could depend upon for interventions. Within the study, we examined, coded, and filtered a substantial gameplay corpus, determining expertise in the game. Mission HydroSci (MHS) is a serious game that teaches middle-school students water science. Using our logging system, designed and implemented along with game design and development, we captured around 60 in-game features from the gameplay of 373 students who completed Unit 3 of MHS in its first field test. We tested eight hypotheses during the field test and presented this paper's results to participating teachers. Our findings reveal several features with statistical significance that will be critical for creating a validated prediction model. We discuss how this work will help future research establish a framework for designing analytics systems for serious games and advancing gaming design and analytics theory.

Notes for Practice

- Our study constructs a foundation for future development of learning prediction models and takes a small step forward in the domain of learning analytics in serious games. Although previous studies with pre- and post-tests have shown the potential for serious games to improve learning, comprehensive analytics is still inadequate. This motivates our development of a granular, learning-focused logging and analytic system, designed alongside the game, to create hypothesis-driven analysis. This study shows that features generated from our logging system distinguish the learning performance of expert and novice students.
- Our results illustrate that students with different expertise levels behave differently in specific ways, including size of the game area explored, task completion speed, in-game event sharing, and other key tasks.
- Instructors can conduct appropriate interventions based on the analytic results, which provide meaningful indicators of student progress.
- Designers can refine their game designs based on findings from the logging system to improve the gameplay flow and optimize learning experiences for students of different expertise levels.

Keywords

Serious games, learning analytics, game-based learning environment, logging system, trace data, educational data mining, performance measurement

Submitted: 31/01/2022 — **Accepted:** 27/09/2022 — **Published:** 05/03/2023

[1] *Email: wldh6@mail.missouri.edu Address: Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, Missouri, USA. ORCID ID: https://orcid.org/0000-0002-6449-3284*

[2] *Email: jggmr2@mail.mizzou.edu Address: School of Information Science and Learning Technologies, University of Missouri, Columbia, Missouri, USA. ORCID ID: https://orcid.org/0000-0002-4343-1041*

[3] *Email: tsadler@unc.edu Address: School of Education, University of North Carolina, Chapel Hill, North Carolina, USA. ORCID ID: https://orcid.org/0000-0002-9401-0300*

[4] *Email: jimlaffeymu@gmail.com Address: School of Information Science and Learning Technologies, University of Missouri, Columbia, Missouri, USA. ORCID ID: https://orcid.org/0000-0002-0434-4260*

Corresponding author [5] *Email: outdoors@acm.org Address: Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, Missouri, USA.*

1. Introduction

Serious games can help players improve motor and cognitive skills in a specific knowledge domain. The use of serious games for learning includes entertainment-oriented video games, which engage students in ways that complement the subject matter curriculum. If a game proves engaging and relevant to a curriculum, it may be able to increase student interest in a subject, which increases learning (K. Renninger, 2010; K. Renninger et al., 2011; K. A. Renninger, 2012). Given the potential of serious games to close learning gaps in STEM disciplines for students across various cultural and economic categories, connecting student assessments to gameplay is essential for understanding how and to what extent particular games aid learning. The first step toward in-game assessment is identifying activity patterns strongly correlated with in-game success based on curricular features designed into a specific game. Such features may also reveal hints that help researchers know under what circumstances student players will most likely abandon a serious game or fail in its execution.

Identifying appropriate in-game features to predict serious game success, particularly learning the material through the serious game, remains a persistent challenge (Fulmore, 2015; Clarke-Midura et al., 2015; Liu et al., 2016). Comparisons of pre-test and post-test student performance in other studies confirm that serious games can improve learning (Fridenson-Hayo et al., 2017; Chaves et al., 2015; Chittaro & Buttussi, 2015). The more closely a game connects to the curriculum it is designed for, the more likely it is that learning gains will be in evidence (Barzilai & Blau, 2014). However, what is missing from prior studies are analytical systems designed to make student learning gains visible to teachers through integration with the serious game, enabling learning-focused intervention.

There are two primary shortcomings in the connection between gaming analytics and teacher understanding of learning in prior studies. First, most analyzed logs are not designed alongside the game and do not provide explicit evidence of learning in gameplay. Instead, prior studies use a kind of technological side effect of the existence of logs. In analytics driven by technological side effects, the analytics produced are structurally disconnected from the learning-connected features of a game and, therefore, insufficient for prediction. Second, current logging systems capture high-level information, ignoring different game quests' properties and restricting prediction capabilities. Logging data not tuned to capture critical learning events is narrower and shallower than the data collection we designed into MHS. More commonly available incidental logs rely on machine learning and data mining to discern any value, often with disappointing results. In addition, narrow logs have a low likelihood of generating features or predictors that accurately reflect similarities in players' gameplay and distinguish between players who learned and those who did not.

One previous article presents an innovative way to measure students' learning behaviours under a game-based-learning virtual environment (Cutumisu et al., 2015). Their in-game learning assessment uses a learning theory framework called SRL (self-regulated learning). They designed and developed a simple game to test their hypothesis, asking students to create posters that satisfy customers' requirements and follow the design theories. In-game characters provided feedback to participants about their finished posters. Participants then chose to move on to the next design task or go back to revise the current poster. The paper showed that logging data, such as frequency of reading constructive feedback, time spent reading constructive feedback, and the resulting total edits, correlates with academic post-test scores and excellent in-game performance, measured by the number of posts with favourable evaluations. The authors concluded that these results fit within the SRL theory applied through the analysis of game-embedded assessments. However, this study used a simple game that did not require user navigation through a virtual space, requiring the performance of only one type of task.

To close the gap in gaming analytics effectiveness and applicability in more complex game environments, we designed a logging system for Mission HydroSci (MHS) alongside the game itself. Such intentional, learning-focused design is urgently required, alongside rigorous research of its implementation. Ambitious design goals like integration of analytics and gameplay face four intertwined challenges: (1) identifying and distinguishing between different levels of engagement with the game; (2) determining which game components make the most significant contributions to learning; (3) building gaming analytics that distinguishes between game skill and subject matter knowledge achievements (Caballero-Hernández et al., 2017); and (4) constructing gaming analytics for more complex, curriculum-centred serious games. These gaps emerge from software development routines surrounding the creation of game logs that are distinct from the development of analytics, resulting in an imperfect connection between what is logged by the game and the analytics that is most likely to be valid.

We applied a theory of transformational play (Barab et al., 2010) throughout the game design and development process for MHS and explicitly connected game log design decisions with game design choices and their intended learning outcomes. In this way, we operationalized the complexity of measuring player success through defined "gates" that serve as signals of

subject matter learning in MHS. Furthermore, MHS analytics identifies player interaction patterns that are antecedent to explicit learning achievements, increasing the potential of MHS to (a) provide player advice that nudges them toward success, (b) enhance clear indicators of gaming achievements, and (c) serve as potential implicit indicators of the types of productive failure (Kapur, 2008) that lead to learning.

We describe related, prior work in gaming analytics and present an overview of MHS and its core gaming analytics features in the rest of this paper. Next, we describe the present study and the eight hypotheses we test, using gaming analytics and player learning outcomes from 373 players in the first empirical evaluation of MHS's analytics system. Our findings show that our design can identify explicit learning achievements through game log analysis, classify antecedent behaviours associated with those achievements, identify implicit game behavioural patterns associated with gameplay expertise, and distinguish these from behavioural patterns associated with learning. We close by discussing the gaming analytics landscape and outlining our next steps.

2. Literature Review

Models linking patterns of player engagement in serious games to learning and performance outside the game context are missing in research on contemporary games for learning. Closing this gap will require sufficiently detailed logs of the choices made by players during a serious game and the subsequent use of these logs to construct computational models. Predicting player learning outcomes and distinguishing them from gaming achievements is a goal for the field that builds on a substantial series of work. This paper will discuss vital features that potentially become significant components of an interpretable validated prediction model. Detailed computational model construction procedures are part of our ongoing research and beyond the scope of this paper.

2.1 Gaming Analytics Validity Questions

Gaming analytics utility requires evaluations showing that what a teacher sees is valid. Validity indicators express confidence through probability measures about the relationship between these measures and a learning result. The operationalization of learning effectiveness indicators is critical for parents, teachers, and individuals who expect subject matter knowledge to result from serious gameplay (All et al., 2015). For example, research that builds video games with embedded, formative assessment designs seems likely to help evaluate whether or not persisting in video games transfers to persistence in other learning contexts (Ventura & Shute, 2013).

2.2 Semantics and Grammars of Behaviour

It is not the discrete click of a button that yields clues to the relationship between learning and gameplay but the assembly of interaction sequences of gameplay. We refer to these recurrent, predictive sequences as "event grammars" (Hauge et al., 2014). Identifying behavioural patterns with valid, valuable connections to gameplay skill and subject matter learning requires log data with at least a conceptual connection to the learning objectives (logical or physical connections are even better) and algorithms capable of identifying similarities within sequenced data. The log data and the algorithms that process it and summarize it ought to form a specific enough vocabulary and set of "event grammars" to withstand qualitative review (Serrano-Laguna et al., 2017). Fundamental analysis tools are too general to draw connections to the semantics behind log file vocabularies if event sequences are not "chunked" into these grammars, which serve as a secondary "end of the wire" processed signal. Ultimately, an analytic grammar of gaming will enable the development of real-time analytics for adaptive gaming, as well as personalized support (Hauge et al., 2014). Developing such a grammar will also enable the interoperability of gaming analytics across different games, genres, and platforms. These adaptations and generalizations are essential for a solid uptake of gaming analytics in authentic educational and training settings.

2.3 Analytics and Gaming Co-design Evolution

Gaming analytics systems can provide helpful feedback to developers of serious games, which will help them to evolve the game design in ways that maximize both learning and engagement (Buckingham Shum & Deakin Crick, 2016). In the history of gaming analytics, however, there remains a significant gap between high-level learning constructs and how those constructs, or facsimiles of them, are operationalized in learner activity log files. More visionary and aspirational analyses of the potential of gaming analytics ask future work to include pop-up alerts and warnings to be shown to the teacher during a game session in the class, more advanced and user-friendly visualizations, and multi-game studies of serious games aimed at classifying both the logged data and the feedback provided in dashboards (Fernández, 2016).

Designing the analytics system in parallel with the game itself also helps to solve common issues in serious games, such as the lack of a social bridge between the game realization stage (design, development, pedagogical support, testing, etc.) and the game dissemination stage (general adoption, coaching, marketing, etc.) or the game use stage (use of serious game support,

reading and interpreting serious game support) (Alvarez et al., 2017). Furthermore, with abundant data and analytics connected to learning goals, stakeholders in serious game projects can ensure communication between different stakeholders or counsels to help identify key problems during project development, testing, and dissemination.

To fully realize the potential of serious games, researchers will need to innovate and devise new training performance metrics and methods to (1) better measure human performance with serious games (tracing of in-game actions, inference of cognitive process, categorization of psychological profiles); (2) improve metrics and methods for the measurement of skills and cognitive abilities; (3) identify likely expert performance through pattern recognition and focus on distilling trainable aspects; (4) score and distill in-game user-generated data to produce actionable insights; and (5) transform analytics into prescriptive, actionable insights for the improvement of human performance (Loh et al., 2015).

2.4 Ensuring the Utility of Gaming Analytics Tools

Gaming analytics can move quickly from helpful cognitive aids for teachers and students to cognitive overload (van Leeuwen, 2015). Van Leeuwen's study (2015) contrasts the results of using learning analytics to support students and teachers simultaneously with cases where only the teacher can access learning analytics. Overload is more likely when students have access to the analytics. This condition quickly leads to teachers dashing around the room responding to content, game, and analytics questions triggered in many cases by the analytics. Further studies in real classrooms should examine whether teachers can use learning analytics in these situations by using the logged history of the collaboration to adapt to the students' requests for support.

Identifying students struggling in the game is likely to be the highest value for teachers, and significantly so if the alerts for students are correlated with learning outcomes derived from research and presented using probabilistic language (Martin et al., 2016). In addition, discerning patterns will require analytic indicators that separate deliberate student movement through a game from accidental or unskilled movement (Kang et al., 2017).

In prior literature, three main approaches to creating indicators apply algorithms to log data. The first approach supports design and development of serious games by mapping actual player progress with intended flow construction and curriculum implementation. Where the information logged is not aligned with the learning intentions of the game design, or the logs indicate that learning is not significant in the current game design, the game and its logging infrastructure co-evolve. The second approach is to design and build an easy-to-follow visualization dashboard for instructors, giving them better insight into student performance and opportunities to provide guidance. Finally, the third approach is to construct and validate a model to predict players' future behaviour or performance and then classify it into different categories for later pattern discovery.

Whether the analytics can provide helpful game design support or an appropriate dashboard for instructors to pinpoint the moments when students struggle is a function of the quality of presentation and the validity of the analytics model. Therefore, the top priorities for gaming analytics research are identifying the likely benefits and challenges for building comprehensive player performance and behaviour indicators and developing useful prediction models.

2.5 Limits of Learning Performance Measurement to Date

To date, measuring learning in serious game research does not clarify how extensive video game experience (often involving fast-paced action games) is related to performance improvements on tests of cognitive ability or subject matter knowledge (Towne et al., 2016). In their discussion of individual performance assessments through analytics in the "Space Fortress" game, Towne and colleagues (2016) extracted behavioural data from the game logs but were unable to collect specific process data such as verbal reports or logs contextualized by the game's structure. They were also not able to identify the mechanisms mediating superior performance. They argue that future studies should identify and measure the strategies of individual participants so it is possible to measure their effects on performance. It is also essential to develop new methods for discerning the role of strategy in individual performance. (Do players tend toward specific strategies based on personality or prior gaming experience? Do specific strategies fare better in some serious games?) Finally, they show advantages for future studies incorporating previous gaming experience, computer experience, familiarity with specific games, and motivation to triangulate learning outcome measures embodied in a learning analytics system.

3. Research Context

In the serious game we study, the system captures players' in-game behaviours, including explicit signals of learning progress. We plan to construct machine learning models with our log data to predict player performance to make probabilistic predictions of player success. MHS is a serious game whose main goal is to teach middle-school students about water systems. It has an integral and consistent storyline that elaborately shepherds the player through the water science curriculum. The player's role in the game is to examine a new planet's environment, especially water conditions, and solve water-related problems to establish the planet as a habitable human outpost. There are six units in the game. Unit 1 teaches players how to navigate the game and

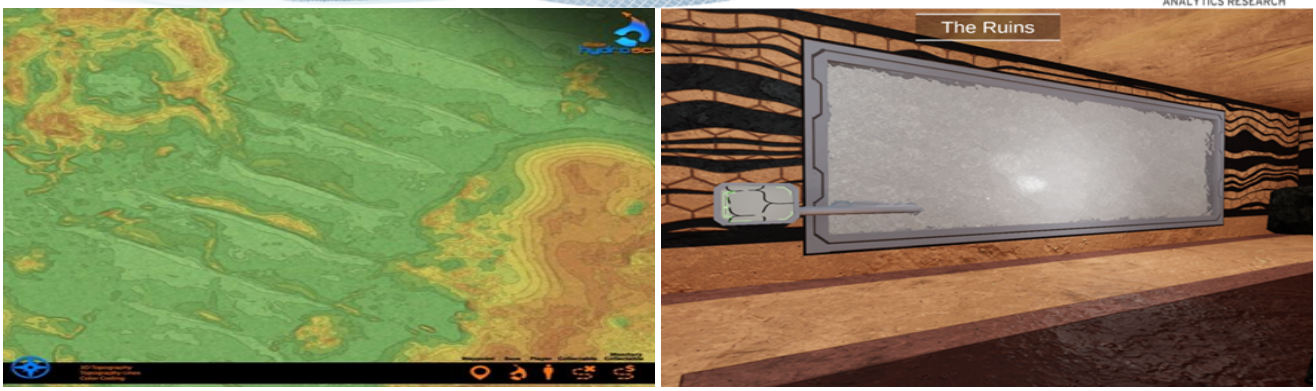

Figure 1. MHS Game*

*An overview of the terrain in Unit 4 is shown on the left. This is an aerial view the student/player can bring up to get a sense of where they are in the game during play. On the right is a window from Unit 4, which allows the student/player to see how water flows through groundwater systems.

familiarizes them with the argumentation system. Unit 2 builds player knowledge about the planet's topography and watershed characteristics. Unit 3 introduces more specific knowledge about the properties of water flow. Unit 4 leads players through a game that helps them understand groundwater systems. Unit 5 adds the atmospheric water movement cycle to the students' repertoire of knowledge. Finally, Unit 6 incorporates prediction games focused on spreading dissolved materials through a watershed, explaining water movement from the surface to the groundwater system and the movement between surface and atmospheric systems. Figure 1 shows screenshots from parts of MHS as an example.

Instead of purely involving game elements from a single game genre, MHS embraces game characteristics from various game types, including problem-solving and puzzle games, role-playing games, and adventure games. This design requires a sophisticated, learning goal–connected logging system to capture the explicit behavioural features that correlate with subject matter learning, and our analytics procedure applies to the evaluation of learning and gaming performance. By quantitatively proving the significant influence on learning and gaming performance of specific game features, researchers can diversify the serious game genre, which solves the deceptively tricky issue of defining the characteristics of a game (Boyle et al., 2016).

3.1 Logging System Design

We designed the logging system in conjunction with the game itself, which means that we are logging events that are specifically designed and related to players' active learning in the game. Our approach differs significantly from past approaches to gaming analytics, which largely rely on the logs made available by default from a technology platform (S. P. Goggins, Gallagher, et al., 2010; S. P. Goggins, Galyen, et al., 2010; S. Goggins et al., 2016). Our co-design approach to analytics builds on Carvalho and colleagues (2015), whose Activity Theory-based Model of Serious Games (ATMSG) describes how game designers can achieve higher learning outcomes through exhaustive trace data from multiple layers. The explicit connections between the pedagogical and game design of MHS and the logged events enable us to explore a set of hypotheses about player patterns of engagement more deeply than in prior work. The detailed logs in the data section and the complete technological architecture are available on GitHub[1].

Our research process incorporates game design, logging system design, prototyping, testing, and iteration of both subsystems over time. It follows a hermeneutic circle of evolution, as depicted in Figure 2.

4. Methods

4.1 Hypotheses

Our methods iteratively connect learning outcomes with actionable information discerned from logs co-designed with the game. This study evaluates eight specific hypotheses related to gameplay analytics in MHS. Each of these hypotheses connects to how we expect our logging system to help us make sense of and tease out learning outcomes within the game, though focused on identifying distinctions between experts and novices, as outlined in Loh and colleagues (2015). The potential log features emerge from our gameplay observations and the variables that seemed to correspond to higher degrees of success during a

[1] https://github.com/mission-hydro-sci

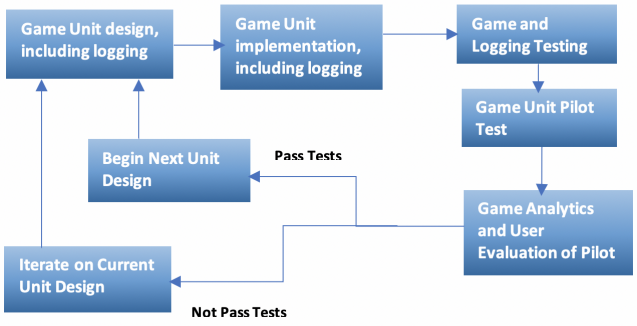

Figure 2. Illustration of the Hermeneutic Development Cycle for MHS Associated with the Game and Logging System

series of informal, preliminary field studies and unit tests. Our hypotheses focus on applying the principles associated with grammar of behaviour in the specific designed context of MHS.

- H1: The spatially calculated volume of the game exploration space differs between expert and novice students.
- H2: The speed of task completion differs between expert and novice students.
- H3: The number of different, specific event types (movement event, dialogue event, state update event, mission update event, trigger event, and others) performed by expert and novice students is different.
- H4: The frequencies of use of various in-game tools and the corresponding average duration of using these tools differ between expert and novice students.
- H5: The frequencies of interaction (throw, deliver) with key items (crate, sensor) of the game between expert and novice students are different.
- H6: In argumentation, the frequency of hovering for text guidance and the speed of reading each choice differ between expert and novice students.
- H7: Argument claim construction performance differs between expert and novice students.
- H8: Argument claim construction performance from the previous unit influences performance in a subsequent (current) unit.

The next subsection will introduce our data collection procedure, the original data set, and how we cleaned it. Also, we will illustrate how we distinguish expert and novice students, what features we generate to test the hypotheses mentioned above, and how we generate them from the original data set.

4.2 Data Collection and Processing

We analyzed gameplay logs from students coming from seven different local middle schools. They played MHS in a classroom with their teacher during the spring and early summer of 2018. The logging system has general features captured from all units and unit-specific features related to learning for each unit. This study is focused on Unit 3, because it is the first unit with significant logging complexity; essentially, it is presented following a training unit (Unit 1) and an orientation unit (Unit 2) and therefore contains some of the first complex learning arcs in MHS. Since the logging system records both player movement and passage through learning milestones in the game, our analysis includes approximately 1.2 million player action records from 373 students in our study after deleting incomplete player records. The composition of the logs, shown in Figure 3, grounds our presentation of research findings and demonstrates a strictly followed data provenance. These are the row-level raw records of player interaction.

In the final quest of Unit 3, students plant five seeds in pods set alongside the riverbanks. A congratulations dialogue box will appear if students plant a seed in the correct pod. Otherwise, a reminder dialogue box will show up telling students they made a mistake and revealing hints about correct pod locations. Based on our on-site gameplay observations and discussions with designers, we consider students who planted the last two seeds in correct pods as experts. To become experts, students must complete all primary quests before the final one, understand all the concepts of this unit's curriculum, and choose the

	ItemID	classId	buildType	installId	playerName	playerId	timestamp
1	5ab3de56a823890aeb055d54	75	FieldTest	1f36e0fb-e25b-41b9-a204-e15034a2aa20	MarcoAlejo	MarcoAlejo1f36e0fb-e25b-41b9-a204-e15034a2aa20	2018-03-22 11:47:55
2	5ab3de56a823890aeb055d55	75	FieldTest	1f36e0fb-e25b-41b9-a204-e15034a2aa20	MarcoAlejo	MarcoAlejo1f36e0fb-e25b-41b9-a204-e15034a2aa20	2018-03-22 11:47:55
3	5ab3de56a823890aeb055d56	75	FieldTest	1f36e0fb-e25b-41b9-a204-e15034a2aa20	MarcoAlejo	MarcoAlejo1f36e0fb-e25b-41b9-a204-e15034a2aa20	2018-03-22 11:48:12
4	5ab3de56a823890aeb055d57	75	FieldTest	1f36e0fb-e25b-41b9-a204-e15034a2aa20	MarcoAlejo	MarcoAlejo1f36e0fb-e25b-41b9-a204-e15034a2aa20	2018-03-22 11:48:12
5	5ab3de56a823890aeb055d58	75	FieldTest	1f36e0fb-e25b-41b9-a204-e15034a2aa20	MarcoAlejo	MarcoAlejo1f36e0fb-e25b-41b9-a204-e15034a2aa20	2018-03-22 11:48:12
6	5ab3de56a823890aeb055d59	75	FieldTest	1f36e0fb-e25b-41b9-a204-e15034a2aa20	MarcoAlejo	MarcoAlejo1f36e0fb-e25b-41b9-a204-e15034a2aa20	2018-03-22 11:48:12
7	5ab3de56a823890aeb055d5a	75	FieldTest	1f36e0fb-e25b-41b9-a204-e15034a2aa20	MarcoAlejo	MarcoAlejo1f36e0fb-e25b-41b9-a204-e15034a2aa20	2018-03-22 11:48:12
8	5ab3de56a823890aeb055d5b	75	FieldTest	1f36e0fb-e25b-41b9-a204-e15034a2aa20	MarcoAlejo	MarcoAlejo1f36e0fb-e25b-41b9-a204-e15034a2aa20	2018-03-22 11:48:12
9	5ab3de56a823890aeb055d5c	75	FieldTest	1f36e0fb-e25b-41b9-a204-e15034a2aa20	MarcoAlejo	MarcoAlejo1f36e0fb-e25b-41b9-a204-e15034a2aa20	2018-03-22 11:48:12
10	5ab3de56a823890aeb055d5d	75	FieldTest	1f36e0fb-e25b-41b9-a204-e15034a2aa20	MarcoAlejo	MarcoAlejo1f36e0fb-e25b-41b9-a204-e15034a2aa20	2018-03-22 11:48:12

	platform	sessionId	teacherId	type	unit	buildVersion	PlayerpostionX	PlayerpostionZ	PlayerpostionY
1	WindowsPlayer	674a7d64-676a-4f79-9a3e-48eb649c215d	Coslet	TriggerEvent	0	0.7.37	-3.52	-66.18	0.8179998
2	WindowsPlayer	674a7d64-676a-4f79-9a3e-48eb649c215d	Coslet	TriggerEvent	0	0.7.37	-3.52	-66.18	0.8179998
3	WindowsPlayer	674a7d64-676a-4f79-9a3e-48eb649c215d	Coslet	ArgumentationAddNodeEvent	0	0.7.37	-3.52	-66.18	0.8074051
4	WindowsPlayer	674a7d64-676a-4f79-9a3e-48eb649c215d	Coslet	ArgumentationAddNodeEvent	0	0.7.37	-3.52	-66.18	0.8074051
5	WindowsPlayer	674a7d64-676a-4f79-9a3e-48eb649c215d	Coslet	ArgumentationAddNodeEvent	0	0.7.37	-3.52	-66.18	0.8074051
6	WindowsPlayer	674a7d64-676a-4f79-9a3e-48eb649c215d	Coslet	ArgumentationAddNodeEvent	0	0.7.37	-3.52	-66.18	0.8074051
7	WindowsPlayer	674a7d64-676a-4f79-9a3e-48eb649c215d	Coslet	ArgumentationAddNodeEvent	0	0.7.37	-3.52	-66.18	0.8074051
8	WindowsPlayer	674a7d64-676a-4f79-9a3e-48eb649c215d	Coslet	ArgumentationAddNodeEvent	0	0.7.37	-3.52	-66.18	0.8074051
9	WindowsPlayer	5ccf60ec-a8fa-494c-80a8-9abf0728fa73	Coslet	QuestEvent	0	0.7.37	-3.52	-66.18	0.8074051
10	WindowsPlayer	5ccf60ec-a8fa-494c-80a8-9abf0728fa73	Coslet	QuestEvent	0	0.7.37	-3.52	-66.18	0.8074051

	CamerarotationX	CamerarotationZ	CamerarotationY	QuestTableF	TasktableF	Scenenames
1	-1.0000000	5.960464e-08	0		<NA>	U1,U1_Quest
2	-1.0000000	5.960464e-08	0		<NA>	U1,U1_Quest
3	-0.9991229	-4.187560e-02	0		<NA>	U1,U1_Quest
4	-0.9991229	-4.187560e-02	0		<NA>	U1,U1_Quest
5	-0.9991229	-4.187560e-02	0		<NA>	U1,U1_Quest
6	-0.9991229	-4.187560e-02	0		<NA>	U1,U1_Quest
7	-0.9991229	-4.187560e-02	0		<NA>	U1,U1_Quest
8	-0.9991229	-4.187560e-02	0		<NA>	U1,U1_Quest
9	-0.9991229	-4.187560e-02	0		<NA>	U1,U1_Quest
10	-0.9991229	-4.187560e-02	0		<NA>	U1,U1_Quest

Figure 3. List of 10 Log Records from the 1.2 Million Records Analyzed*

*Each row includes an ItemID for connecting to detailed information (e.g., dialogue names along with related content, item names that players interacted with, task completion status), an InstallID that identifies each computer, a unique playerID across all MHS installations, timestamp, operating system, sessionID, teacherID, event type, event unit of measure, and MHS build version. In addition, the three-dimensional space the player occupies includes camera rotation, which is where a player character is looking from that position, and is stored in the X, Y, and Z labelled fields at the end of the rows. The final column includes a scene name, which references a specific part of the unit design and can filter data and ascertain expected player behaviours within a session.

correct river branch for depositing pods. Accomplishing this requires dialogue with non-player characters (NPCs), and the player outcomes from the first three planting pods (before the fourth pod, students need to go back and forth between different riversides/banks to determine which side the pollutant materials flow along), which is why we use "successfully planting the last two seeds in correct pods" as a proxy for identifying "experts." Of the 373 students in our study, 24 planted the last two seeds in the correct locations, and 349 planted at least one seed in the wrong location in the last two trials. Based on the above-mentioned expert standard, we generated a categorical variable named "U3Performance" to distinguish between experts and novices.

4.2.1 Discretization

After some exploration analysis, we found that most features extracted from the logging system to test the eight hypotheses mentioned earlier contain high variation or numeric outliers, creating significant noise that imparts unstable factors to the testing results. To avoid statistical anomalies arising from non-standard probability distributions in both performance groups, we decided to discretize features containing massive outliers based on the variable distribution and then apply them as ingredients of the hypothesis testing procedure. Considering only 24 members involved in the expert group, we thought removing outliers directly from the dataset meant excluding potentially useful information from the expert group and decreasing the statistical power of hypothetical tests because of unbalanced group sizes.

During the discretization procedure, we first calculated minimum, first quartile, median, and third quartile values for each task's completion time to generate the speed feature. Secondly, we created four feature levels based on these summary statistics. These summary statistics place each player in a performance quartile, with the fastest players in quartile 1 and the slowest players in quartile 4 for each task.

4.2.2 The Size of the Explored Game Area

We reference players' in-game navigating trajectories to measure the space explored by each player. We hypothesize that expert students are more efficient with game navigation, finding the locations to complete a quest with a clear goal instead of walking back and forth in similar circles or aimlessly exploring places far away from correct walking tracks. Inspired by the in-game position data-processing method (Loh et al., 2016), we used a variable called "exploreAreaPercentage" to test this hypothesis.

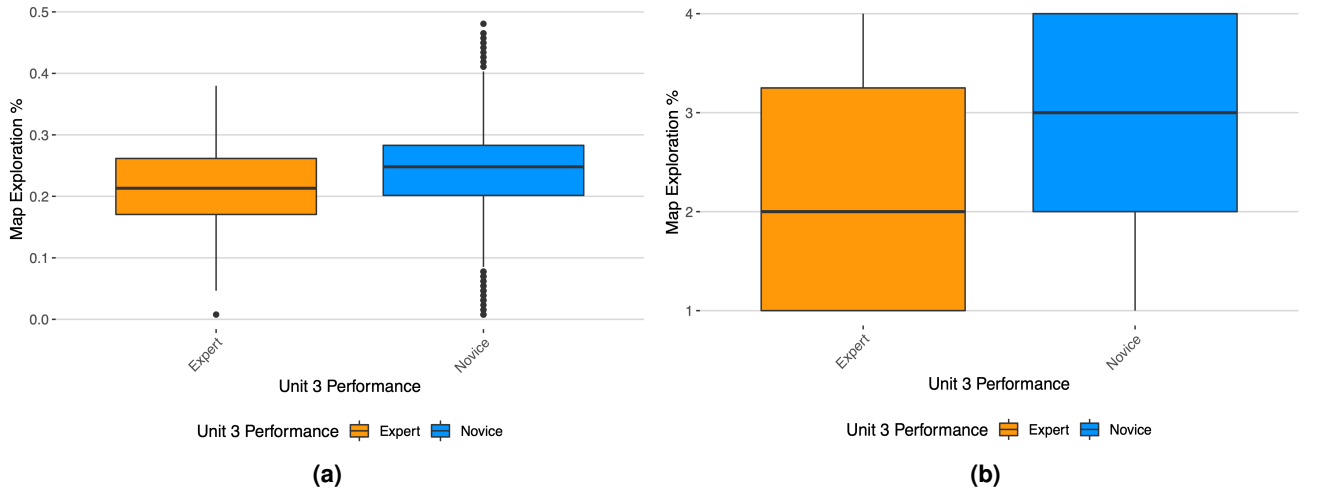

Figure 4. Map Exploration Size*

*Figure (a) shows the distribution of map exploration size in the original format. Figure (b) shows the map exploration size after discretization.

Like books, digital video, and other media, game as an emerging information carrier provides knowledge with immediate practice opportunities in a virtual environment and diversified ways of interacting where players learn things at their own pace. The game world contains detailed granularity of massive stories and abundant integrated curriculum knowledge as a basic game element. By exploring it, players can become familiar with game navigation and mechanics and better understand the knowledge blended with it. In addition, measuring players' game environment and exploring status offers researchers more opportunities to learn about players' gaming and learning procedures.

Our logging system provides deep, wide, and learning-connected behavioural data. This data includes, for example, measurement of the size of the explored game area in three dimensions, x, y, and z, along with information about the player character's directional gaze, called the *spatial axis*. Inspired by Loh and colleagues' work (2015), we split the whole game map into 10 by 10 by 10 grids. Each grid has a unique tag. Each tag contains a group of four numbers. The first number represents the row number, the second is the column number, the third is the height, and the fourth is the number of times a player steps into this grid. For example, (1, 1, 1, 1) means the player steps into the grid in row 1, column 1, and height 1 for the first time. We used E_T to represent the total number of unique tags in log data from field test 1 of Unit 3. E_i represents the unique tag player i has after finishing Unit 3. We utilize E_i/E_T to represent the size of the game environment player i has explored. Figure 4 shows the distributions related to variables representing the size of the explored game area.

4.2.3 Speed of Completing Tasks

This feature captures information to test the second hypothesis. This hypothesis focuses on how quickly players move through the game. We hypothesize that expert students will complete quests faster than others. Duration of task or quest completion was a powerful predictor or classifier in multiple academic articles (Loh & Sheng, 2014; Loh & Li, 2015; Loh et al., 2015; Palmquist & Crowley, 2007).

To generate an appropriate feature to test this hypothesis, we constructed a data set describing how long each player took to complete a specific task. Each column represents a specific task, and each row represents a player. Figure 5 visualizes the distributions of duration of task completion in minutes (a) and in discretized level format (b).

4.2.4 The Percentage Shares of Different Events

The third hypothesis tests differences in the types of events triggered by players in the game and evaluates them as a percentage of total events. The inspiration for this measurement comes from field observation reports during MHS field test 1. One of the reports shows that expert students tended to have smaller numbers of reminder dialogue boxes (e.g., remind students of the correct direction to go), which would reduce their total percentage of dialogue boxes compared to their entire pool of events. Event types such as movement and dialogue boxes differ significantly in hypothesis tests. We hypothesized that a subset of event types would vary widely between experts and novices.

As we mentioned in the data collection section, the MHS log data set has an event-type column to describe players' in-game actions with a specific timestamp. There are 10 categories of event types: triggering events, movement events, complete mission

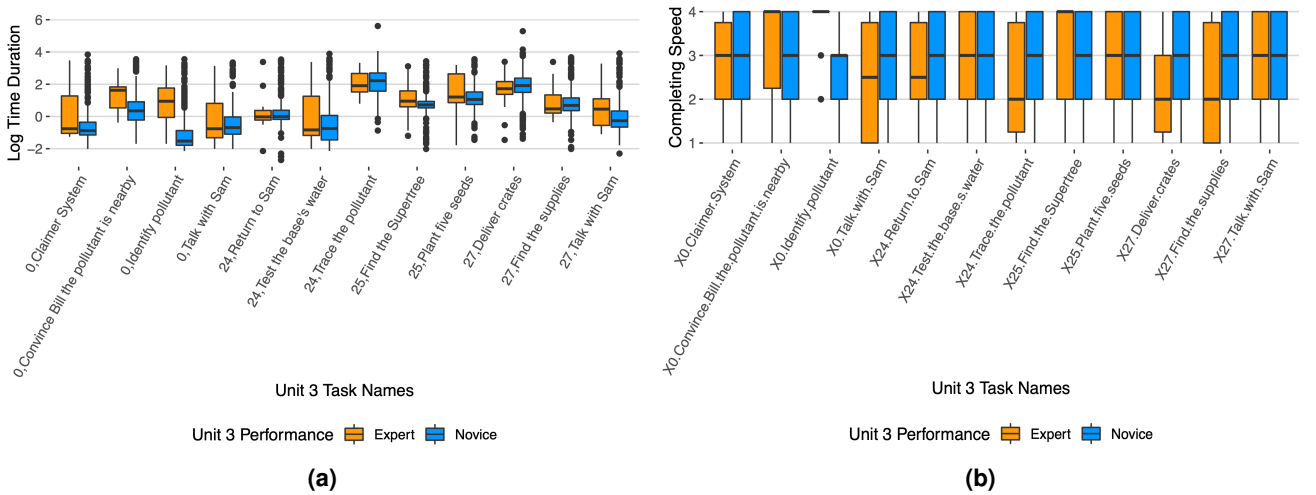

Figure 5. Duration of Task Completion*

*Figure (a) shows the distributions of log-transformed duration of task completion. Figure (b) shows the distributions of task completion duration after discretization.

events, statement update events, dialogue events, arf (an artificial intelligence with an avatar image that helps players in specific game points) related events, hotkey pressing events, jump events, argument-related events, and toggle board–related events. We used P_t/S_t, P_{mo}/S_t, P_{ms}/S_t, P_s/S_t, P_d/S_t, P_a/S_t, P_k/S_t, P_j/S_t, P_{ar}/S_t, P_{to}/S_t to represent the percentage shares of different events for later hypothesis testing. What each symbol represents is listed in Table 1.

Table 1. A Detailed Explanation of Different Percentage Shares of Events

Symbol	Represent content
S_t	Adding all tracing data records of a specific player
P_t	Sum of the number of events that happened when players interacted with in-game items
P_{mo}	Sum of the number of events that happened when players walked or ran in the game environment
P_{ms}	Sum of the number of events that happened when players accepted or completed a task or a quest
P_s	Sum of the number of events that happened when players entered a new game scene
P_d	Sum of the number of events that happened when players triggered out, clicked selection nodes, or closed a dialogue
P_a	Sum of the number of events that happened when players opened or closed the arf help menu, or selected choice buttons in the arf help menu
P_k	Sum of the number of events that happened when players pressed a hotkey button on the keyboard
P_j	Sum of the number of events that happened when players jumped into the game environment
P_{ar}	Sum of the number of events that happened when players opened the argument claimer system, dragged on or off an argument node, hovered on or off a node, succeeded in argument, failed in argument, or closed the argument claimer system
P_{to}	Sum of the number of events that happened when players utilized a skateboard to move in the game environment

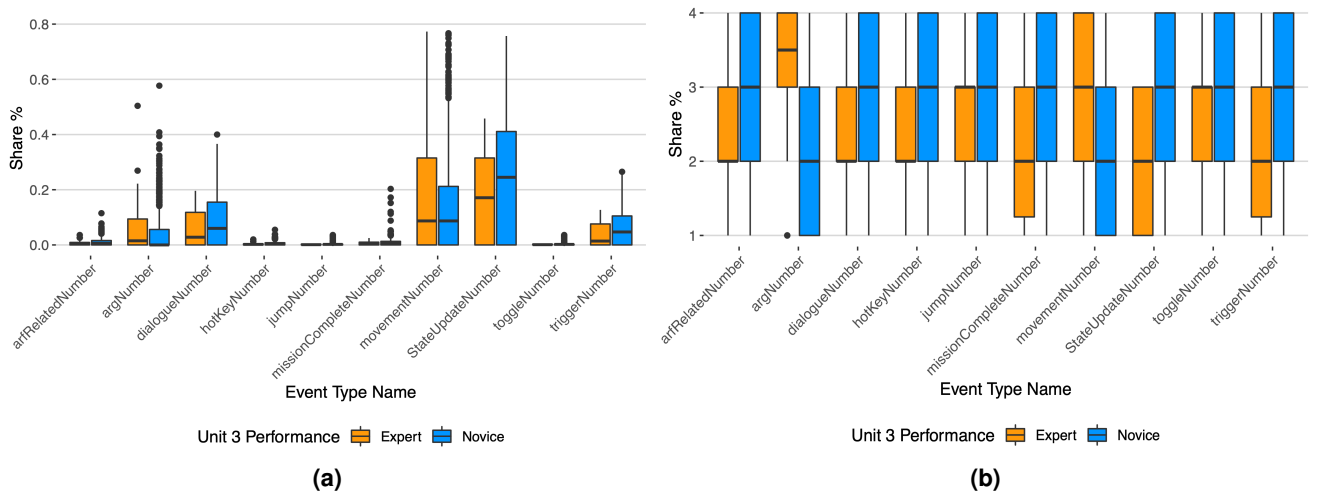

Figure 6. Event Type Shares*

*Figure (a) shows the distributions of event type shares in percentage format. Figure (b) shows the distributions of event type shares after discretization.

The feature names we have used to conduct hypothesis tests are "triggerNumber," "movementNumber," "missionCompleteNumber," "StateUpdateNumber," "dialogueNumber," "arfRelatedNumber," "hotKeyNumber," "toggleNumber," "jumpNumber," and "argNumber." Figure 6 shows the distributions of each event type before and after discretization.

4.2.5 Tool Use Status

The fourth hypothesis highlights the specific, available tools used by players. During the field test, we observed that expert students can better understand tool functions and can smoothly apply tools in appropriate situations. However, expert students also tend to complete side and main quests, which further skews their tool use in our hypothesis because of differences in tools required on side quests.

Seven types of in-game tools provide players with information and knowledge, helping them realize self-paced learning. These seven tools include a background information menu including character, storyline, and game context information; a chat log; crash diagnostics to explain ship damage status and repair strategies; a help menu; a game map; a quest menu to show game progress; and a menu displaying information about the player's current task.

Two features are generated from the tool-using status: the tool's frequency and the tool-using speed. We calculated how many times a player opened it and the average speed during their usage of each tool. The frequency tool calculates all open records for a specific type of tool. MHS calculates tool use speed in the same way as task completion speed. Finally, we have 12 features to measure players' in-game tool use: "BackingInfoMenuNode," "ChatLogNode," "CrashDiagnosticsMenuNode," "HelpMenuNode," "MapMenuNode," "Quest-MenuNode," "BackInfoSpeed," "ChatLogSpeed," "CrashDiagnosticsMenuSpeed," "helpMenuSpeed," "mapMenuSpeed," and "questMenuSpeed." The distributions for each variable in this section in original or discretized format are in Figure 7.

4.2.6 The Frequency of Triggering Key In-Game Objects

We test the fifth hypothesis using variables that indicate the success of delivering crates and tossing sensors into the correct river. As the expert versus novice section mentioned, we have four key quests to evaluate players' curriculum learning performance. Among them, the quests "crate delivery" and "tossing sensors to find pollution" require players to interact with in-game items. The logging system captures related measures and performance using four features: (1) "crateCorrectFreq," (2) "crateWrongFreq," (3) "tossCleanFreq," and (4) "tossPollutedFreq." The "crateCorrectFreq" feature is the frequency with which a player delivers crates into the correct river. The "crateWrongFreq" feature is the frequency with which a player delivers crates into the wrong river. The "tossCleanFreq" feature is the frequency with which a player tosses a sensor into the river with a polluted outcome. Finally, the "tossPollutedFreq" feature indicates the frequency with which a player tosses a sensor into the river with a clean outcome.

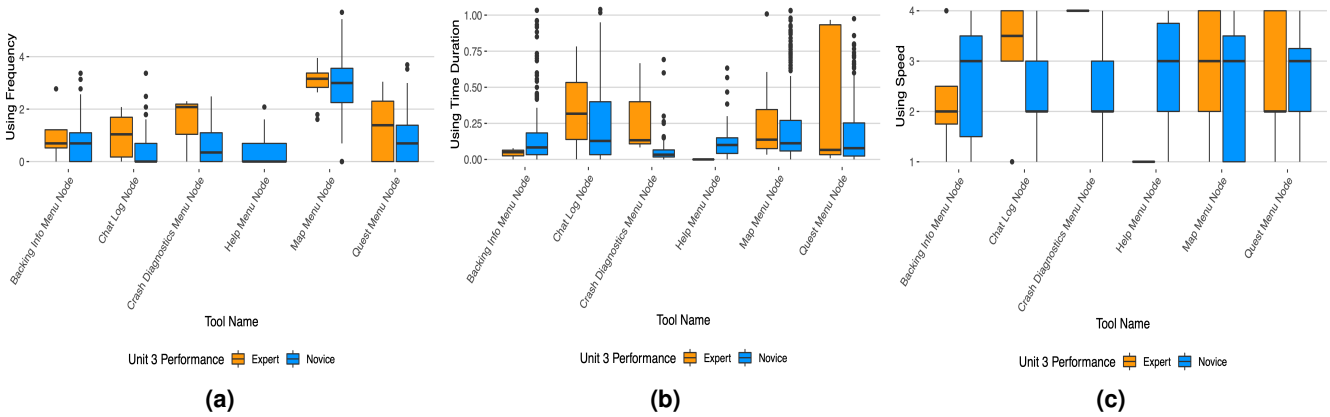

Figure 7. Frequency, Duration, and Speed Distributions for Different In-Game Tools*

*Figure (a) shows the distributions of the frequencies of using different in-game tools. Figure (b) shows the distributions of the average duration of using different in-game tools in minute format. Figure (c) shows the speed distributions of different in-game tools in discretized level format.

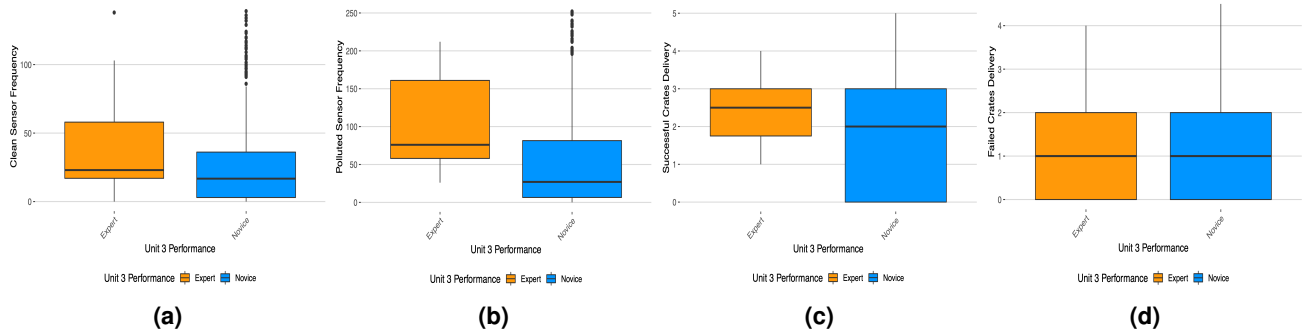

Figure 8. Sensor-Tossing and Crate-Delivery Frequency Distributions*

*Figure (a) shows the distributions of sensor-tossing frequency with clean outcomes. Figure (b) shows the distributions of sensor-tossing frequency with polluted outcomes. Figure (c) shows the frequency distributions of correct crate delivery. Figure (d) shows the frequency distributions of wrong crate delivery.

By calculating these features, we can get information about how many crates players deliver to the correct river and how many to the wrong one, and also get information about what path they follow to the tosses. These features may have a high potential to be a powerful classifier to identify expert players. Figure 8 shows the distributions of key item interaction frequencies.

4.2.7 Hovering Frequency and Corresponding Reading Speed per Choice Node within Argument System

Constructing a sound scientific argument is one of the key quests in Unit 3. During this quest, students must read nodes' pop-up text information to choose the correct combination of different argument components, including evidence, claim, and reason. In hypothesis 6, from our observation, expert and novice students behave differently, especially when they hover on choice nodes and read the pop-up contents. In our preliminary studies, we observed that students with an unusually high frequency of hovering on choice nodes or reading a node's content quite slowly are less likely to be experts. Therefore, we operationalized the following variables: the hovering frequency for each choice node, the summation of each node's hovering frequency, and the summation of the reading times for each choice node. The nodes containing pop-up text include "U3.Claim," "Site.A," "Site.B," "Resaoning.5," "Reasoning.2," "Reasoning.3," "Reasoning.4," "Reasoning.6," "Reasoning.7," and "Reasoning.1." The reading times are then discretized as reading speed rate, calculated in the same way as task completion speed. Figure 9 displays distributions of the variables included in testing hypothesis 6.

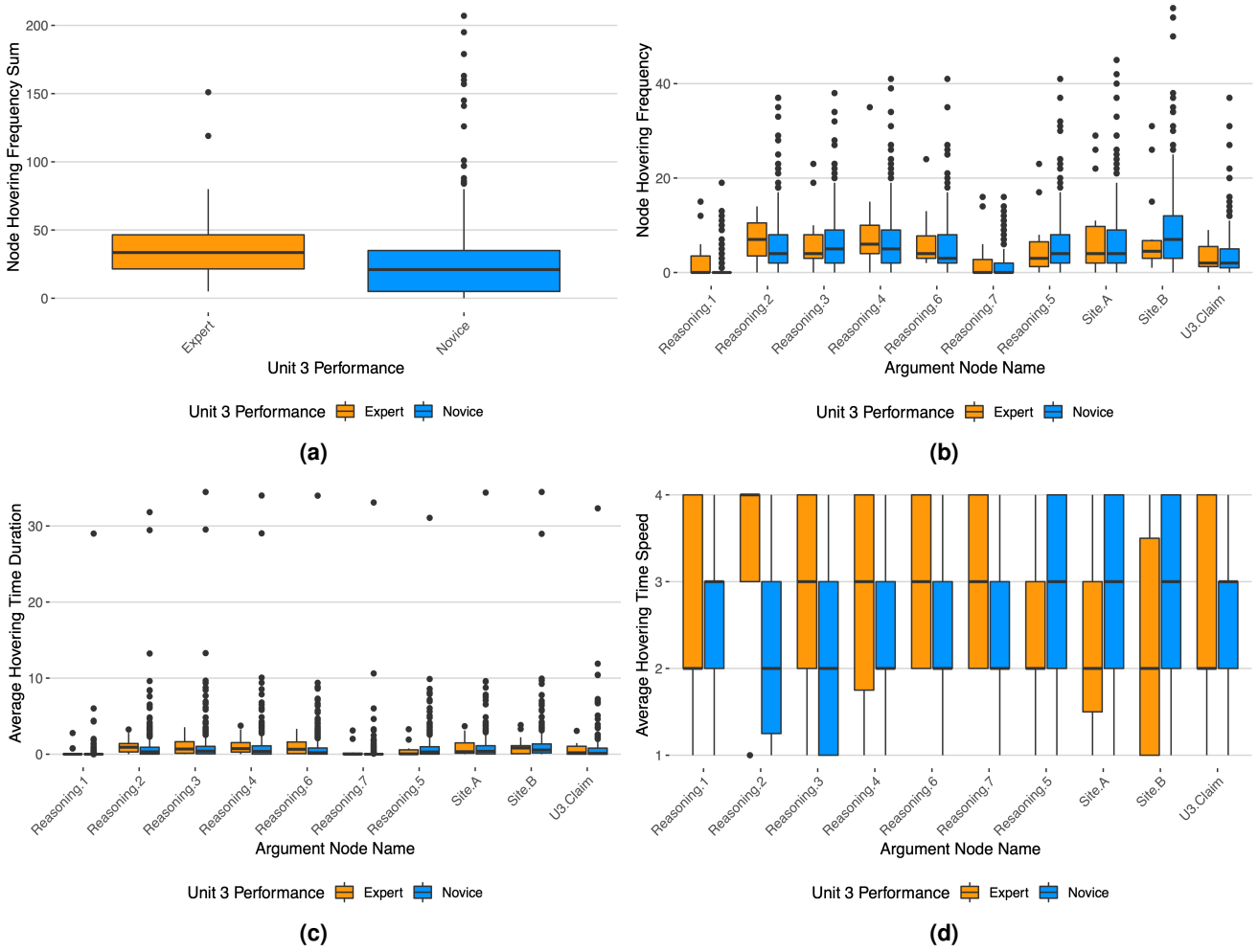

Figure 9. Variable Distributions for Testing Hypothesis 6*

*Figure (a) shows the distributions of the summation of nodes' hovering frequency. Figure (b) shows the distributions of each node's hovering frequency. Figure (c) shows the reading time for each argument node. Figure (d) shows each argument node's discretized reading time or reading speed rate.

4.2.8 Argument Performance

In hypothesis 7, we suppose that experts are more immediately and reliably successful at constructing an argument than novices in the current unit. In hypothesis 8, we assume expert students can more efficiently learn patterns and mechanics regarding argument construction from the previous unit's experience and apply the knowledge in the current unit than novice students. Because Unit 1 is a tutorial unit, which does not include a completed argument construction task, we only consider Unit 2 as the previous unit. To test the hypothesis, we operationalized "argumentLevel," which creates a synthetic measure of students' argumentation performance based on a calculation of their success and failure frequencies, as well as an indicator of whether or not they do ultimately construct a valid scientific argument. Features we used to evaluate students' argumentation performance include the number of times they submitted correct answers and the frequency of wrong answer submissions.

"ArgumentLevel" has four different levels. We describe each level's exact calculation method in Table 2. Using the same feature generation method, we evaluate students' argument construction performance in MHS Unit 2 with the name "U2argumentLevel." Figure 10 visualizes proportion shares of each argument performance level under different units, faceted by expertise groups.

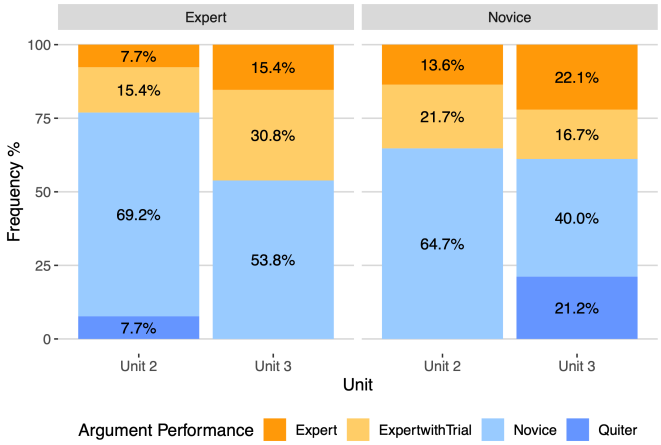

Figure 10. The Percentage Stacked Bar Plot Displaying Students' Argument Performance in Unit 2 and Unit 3

Table 2. A Detailed Explanation of Four Levels of Argument Performance

Level	Description
Quitter	Players in this level opened the argument claim system but had no successful submissions and more than 0 failure submissions.
Novice	Players in this level had at least one successful submission. They had more than three (median level) failure trials. (Students may just want to try what combination is correct instead of understanding the content first.)
ExpertwithTrial	Players in this level had at least one successful submission and less than or equal to three failure trials.
Expert	Players in this level had no failure submissions but had successful submissions.

5. Results

Our results validate six hypotheses out of eight with statistical significance, and the remaining two show promising, less statistically significant results. We explain our evaluation of our hypotheses and the salient statistical tests next, followed by a table of assertions about how to interpret our analysis of each hypothesis and a brief explanation of key statistical analysis choices.

5.1 Analysis and Statistical Tests

After processing the data and generating corresponding features, the next step is to select the appropriate statistical tests for our hypotheses. As is typical among field-gathered behavioural data, few variables follow a normal distribution. Also, considering that the dependent variable contains two independent groups, and the task is to construct hypothetical tests on ordinal or numeric features, we selected the non-parametric statistical test, the Mann-Whitney U test, also called the Wilcoxon rank-sum test, to examine our hypotheses. Specifically, the Mann-Whitney U test is selected because it is a non-parametric test that does not need a normal distribution assumption, it has a high tolerance for outliers, and it can deal with small sample sizes. The test's null hypothesis is that both populations' distribution functions are equal. The alternative hypothesis is that the distribution functions are not equal. Informally, we are testing to see if mean ranks, approximately equal to the median value, differ between groups. The Mann-Whitney U test is applied to all our numeric or ordinal features, such as features related to frequencies or counts.

For categorical features, such as features used in hypotheses 7 and 8, we conducted Fisher's exact test, which is applied when we want to test the null hypothesis that there is no association between the two included categorical features within a contingency table. This test performs well when dealing with small, unbalanced sample sizes, matching our data characteristics, making it more appropriate in our case than the other popular method, the Chi-square test, which requires comparatively large and balanced sample sizes.

However, there is a debate or uncertainty on how to define the data type after discretization. For example, we can treat discretized variables as ordinal or numeric. If we assume that each student belongs to a certain level after discretization, then

the levels are equally distributed. Alternatively, categorical variables work if we assume that each student belongs to a certain group and the differences between groups cannot be quantitatively measured. To handle this uncertainty, we used two methods to conduct the test for each discretized feature. One is the Mann-Whitney U test, if we assume that the discretized feature is numeric, and the other is Fisher's exact test, if we assume that it is categorical. We can validate and add statistical power to our conclusions by using both tests on discretized features.

5.2 Hypothesis Testing Outcomes

We show the results of the statistical tests in Table 3.

Table 3. Statistical Tests of Hypotheses with Significance and Descriptions.

Hypothesis	Significant variables	Test results	
		Mann-Whitney U	Fisher's exact
H1: Size of game area exploration	exploreAreaDiscretize	8281*	9.178*
H2: Task completion speed	X0.Convince.Bill.the.pollutant.is.nearby	1682.0**	12.473*
	X0.Identify.pollutant	1032.5***	33.251***
	X25.Find.the.Supertree	2102.5*	9.209.
	X27.Deliver.Crates	2074.5*	7.397.
H3: Event type percentage shares	argNumber	1978*	6.129
H4: Tool-using status	Help.Manu.Node.Freq	0***	NA
	Crash.Diagnostics.Menu.Node.Speed	31.5*	8.282*
H5: Key item interactions	TossPolluted	1850**	NA
	crateThrowSuccess	2263.5*	NA
H6: Argument node hovering status	Arg.Node.Sum.Freq	1987.5*	NA
	Reasoning.1.Freq	1880.5*	NA
	Reasoning.2.Speed	1153.5*	8.922*
H7: U3 argument performance	NA	NA	NA
H8: U2 argument performance	U2ArgPerformance	NA	33.983*

The first column, "Hypothesis," lists what hypothesis we are testing. The second column, "Significant variables," lists features that are tested out as statistically significant (p-value less than 0.05) by either one of the two testing methods. If no significant feature exists, NA is entered in the corresponding cell. The third column, "Test results," contains two sub-columns, "Mann-Whitney U" and "Fisher's exact." NA is entered in the cell under either of the following two conditions: (1) there are no significant features tested out by both methods under a certain hypothesis; (2) only one test method is applied to the corresponding significant feature because it is either a numeric feature or a categorical feature but not a discretized feature. The marks ".", "*", "**", and "***" following the testing result value represent the significance degree reflected by p-value. "." indicates that the p-value is less than 0.1 but larger than or equal to 0.05; "*" indicates that the p-value is less than 0.05 but larger than or equal to 0.01; "**" indicates that the p-value is less than 0.01 but larger than or equal to 0.001; "***" indicates that the p-value is less than 0.001.

5.2.1 H1: The Volume of Spatially Calculated Game Exploration Space Is Different between Expert Students and Novice Students

Under this hypothesis, we identified that expert students explored significantly fewer game areas than novice students. Combined with our on-site observations, we speculated that this phenomenon might be because expert students are more efficient than novice students in the following ways:

1. getting familiar with the game environment and mechanics,
2. figuring out the correct path to complete tasks by identifying useful information during gaming procedures, and
3. focusing on completing quests and ignoring distractions.

5.2.2 H2: The Speed of Task Completion Differs between Expert and Novice Students

For hypothesis 2, we found that (1) expert students are significantly slower than novice students at completing the tasks "Convince.Bill.the.pollutant.is.nearby," "Identify.pollutant," and "Find.the.Supertree." (2) Expert students are significantly faster than novice students at completing the task "Deliver.Crates." We can also observe that expert students complete the task

"Trace.the.pollutant" more quickly than novice students, which does not have statistical significance but reveals a potential pattern in distinguishing between experts and novices.

5.2.3 H3: The Number of Different, Specific Event Types (Movement Event, Dialogue Event, State Update Event, Mission Update Event, Trigger Event, and Others) Performed by Expert and Novice Students Is Different

In this section, we figured out that expert students' argument event type shares are significantly higher than those of novice students. Although having no statistical significance, the following findings contain promising patterns to distinguish students' performances: (1) Expert students have fewer event shares than novice students on triggering dialogues, using in-game tools, and interacting with in-game items. (2) Expert students have more shares than novice students on events of navigating within the game environment.

5.2.4 H4: The Frequencies of Using Various In-Game Tools and the Corresponding Average Duration of Using Those Tools Differ between Expert and Novice Students

Under this hypothesis, the results show that expert students did not use the "help" tool. Instead, they read the information within the tools of backing information and quest reviews faster than their novice peers. Regarding speed of tool use, expert students read information presented by the "Crash.Diagnostics" tool at a significantly slower rate than novice students. This tool provides background stories and information about the game world and hints on how to solve side-task puzzles with extra credits. Furthermore, we also identify potential patterns, although without statistical significance, that expert students more frequently use the chat log, map, and quest review tools than novice students.

5.2.5 H5: The Frequencies of Interaction (Deliver, Toss) with Key Items (Crate, Sensor) of the Game between Expert Students and Novice Students Are Different

By examining features within this hypothesis, we found that expert students delivered significantly more crates to the correct river than novice students. They also tossed significantly more sensors to test whether the soluble materials polluted the river area than their novice peers.

5.2.6 H6: In Argumentation, the Frequency of Hovering for Text Guidance and the Speed of Reading Each Choice Differ between Expert and Novice Students

In Unit 3's argument system, students are asked to select the appropriate reasoning to connect the existing claim and evidence. By investigating hypothesis 6, we identified that expert students had significantly higher total hovering counts on all choice nodes than novice students. Specifically, they hovered on the node "reasoning.1" significantly less but spent significantly more time reading the "reasoning.2" node than their novice peers. We also observed that expert students received useful information from the existing nodes "Site.A," "Site.B," and "U3.Claim" faster than novice students, which may indicate that they understood the problem more efficiently than novice students. Meanwhile, expert students are more quick than novice students to determine which reasoning nodes, such as "Reasoning.1" and "Reasoning.5," are the least appropriate for solving this problem and focus on investigating promising answers, reflected by spending more time on reading nodes "Reasoning.2," "Reasoning.3," "Reasoning.4," "Reasoning.6," and "Reasoning.7" than novice students.

5.2.7 H7: Argument Claim Construction Performance Differs between Expert and Novice Students

Although the test proves no significant results, we can observe from Figure 10 that all expert students submit at least one successful argument. In contrast, 21.2% of novice students did not submit a successful argument, which may indicate that students can gain knowledge if the game helps them complete the following tasks by submitting a successful argument.

5.2.8 H8: Argument Claims Construction Performance from the Previous Unit Influences Performance in a Subsequent (Current) Unit

We identified a significant association between students' performance in Unit 3 and their Unit 2 argument performance. Figure 10 reveals that the association is negative, in which 7.1% of expert students made no successful submission while all novice students made at least one successful submission within Unit 2's argument system. As well, compared to 35.3% of novice students who made successful argument submissions within three failures, only 23.1% of expert students constructed successful arguments within the same number of failed trials.

5.2.9 Summary of Results

In general, we identified that some features describing (1) the size of the explored game world, (2) task completion speeds, (3) game event shares, (4) tool use status, (5) key item interaction frequencies, (6) node hovering status within argument system, and (7) argument performance from the previous unit are significantly different between expert students and novice students with more than or equal to 95% confidence. Nevertheless, we also note that although argument performance from the current unit has no statistically significant difference between expert and novice students, Figure 10 reveals a trend that expert students

performed better than novice students on current argument construction. Hypothesis 1 shows that expert students focused on exploring areas relating to the paths to complete tasks instead of wandering around aimlessly or being easily distracted by environments, as some novice students did. For hypothesis 2, expert students spent much more time than their novice peers on tasks that provided new information or subject matter knowledge to learn carefully, but they completed tasks requiring applying knowledge significantly more quickly than novice students, which may reveal that expert students identify useful new information and learn knowledge more efficiently than novice students within the game environment. Our test of hypothesis 3 illustrates that experts' argument and movement event shares were significantly higher than those of novices, indicating that experts spent most of their time promoting game progress and constructing arguments, which might be caused by higher information load in argument-related tasks than in other tasks. In contrast, their help-seeking event shares, such as triggering dialogues to adjust players' failed actions or choices, referring to in-game tools, or interacting with items to familiarize themselves with the game environment or mechanics, are lower than those of their novice peers. By investigating hypothesis 4, we noticed that experts knew how to use tools effectively and efficiently in appropriate situations to assist themselves in completing tasks. They also understood information embedded within tools more quickly than novices, who may need to keep referring back to the "Help" menu to understand each tool's specific function. Within hypothesis 5, we figured out that expert students delivered significantly more crates to the correct river than novice students, which suggests that expert students are stronger in transferring knowledge within the game than their novice peers.

Furthermore, experts tossed more sensors into the river and found the polluted sources quicker than novices, indicating that they quickly mastered the sensor's function and applied it on approaching the destination, another sign of better knowledge learning and transferring skills. Hypothesis 6 reveals that expert students filtered out wrong answers and compared potentially right answers much faster than novices, who were more obtuse when figuring out the key problem to solve and which information warranted attention for building a solution. Although hypothesis 7 has no statistical difference, we identified a trend that expert students performed better than novice students. However, experts make mistakes and may perform poorly at the initial stages, but they can learn from failures, which is, to some extent, validated by hypothesis 8.

5.3 Extended Thoughts

Our hypotheses tests specifically prove that we can extract a set of variables that quantitatively correspond to in-game learning and gaming performance from the complex game world of MHS. Our findings suggest a clear and robust path for identifying in-game performance differences. We also identified users struggling as novices, in contrast with experts, which will prove useful in the future study of serious games and gaming analytics. Since our analysis specifically operationalizes log file variables we designed into the game to make learning visible, our analysis of those variables and statistical testing of their values together strengthen the worth of our findings.

We summarize a few specific observations based on the outcomes:

1. Expert students explored the MHS world to a lesser extent than novice students.
2. Expert students completed tasks where they needed to identify new information or acquire knowledge more slowly than novice students. However, expert students finished tasks requiring knowledge application much more quickly than their novice peers.
3. Expert students spent more time on harder tasks or tasks needing more information processing, a pattern novice students may not follow.
4. Expert students applied tools more effectively and efficiently than novice students, which can be reflected by assistant tool usage status or the interaction status of some task-specific items, like pollutant sensors.
5. Expert students made some wrong decisions during gameplay, which may in some cases inspire the design of MHS. For example, most expert students selected the wrong river on their first one or two attempts in the key quest of delivering crates to an NPC.

However, after reading the reminder dialogues, which inform students why their choice is wrong, expert students adapted immediately, while novices frequently made similar errors on their next attempt. We also noticed that expert students made more wrong decisions on planting seeds in pods near the riverbanks in the first three trials than novice students. The average count of mistakes made by experts is 2.45, while the same number for novices is 1.56, which has statistical significance with 95% confidence, tested by the Mann-Whitney U method. Furthermore, in the spirit of productive failure, we observed that students who were less successful in the argumentation task in Unit 2 were more likely to succeed in their argumentation in Unit 3. Our work quantitatively proves the importance of incorporating an effective negativity mechanism in game design, which is a supplement to the research of Gauthier and Jenkinson (2018), who applied qualitative methods to evaluate learning performance

gains from failure experiences in both serious games and a simulation environment. Finally, we observed that experts triggered fewer reminder dialogues but absorbed the embedded information within the dialogues better than novice students, making them make the same mistakes more than once much less than novice students. Considering the information-processing differences between expert and novice students, we speculate that designing different reminder dialogues depending on players' gaming performances (e.g., failure times) may help novices learn better in MHS.

When completing the argumentation quest, expert students also reflected some unique properties: (1) Most of them constructed a correct argument without making a large number of wrong submissions (a kind of "brute force" approach to finding the right answer). (2) Expert students also filtered out wrong choices (nodes) and ensured the final answer by carefully reading the potentially correct nodes' content, as shown by their reading time of nodes' pop-up text. Finally, (3) expert students generally did not select a common misconception as a part of the final answer, which contrasts with novice students.

6. Discussion

6.1 Implications for Logging System Design in Serious Games

In this study, we propose eight hypotheses based on prior work reviews and on-the-spot observation records (Laffey et al., n.d.) and obtain a summary of the outcome that features included in these hypotheses are vital to distinguish expert students and novice students with statistical significance. Our results provide a foundation for building prediction models that consistently and validly measure learning performance.

For developers who design and integrate logging systems for serious games, we recommend co-designing the game and logs to ensure a connection between learning goals and the gameplay. Even with a planned integration, we continuously integrated the thinking for the game and the analytics system. The design and implementation procedure we used is not a linear pipeline since the outcome of log system testing may trigger redesigns of the game and vice versa. For example, the logging system sometimes helps designers identify gameplay issues and adapt the game's design.

For the data analyst, the standard of students' in-game behaviour and interaction feature generation and organization is still emergent (Serrano-Laguna et al., 2017). Designing and generating measurement metrics is a function of each game's unique properties. However, our hypothesis-driven analytics outcome shows some general measurement features that could apply more widely for monitoring serious games and providing learning-centred analytics:

1. game world exploration sizes, calculated by trajectory data (moving positions along with time);
2. task completion speeds;
3. in-game tool usage status;
4. the triggering frequencies and reading status of text-enriched boxes, nodes, or dialogues, functioning to remind or deliver important information;
5. the interaction status with key items (boxes, switches, sensors, etc.), which are crucial to complete quests;
6. the percentage shares of the different event or action types.

It is also important to note that game use and logging characteristics differ in small ways from classroom to classroom. Therefore, we will develop learning performance prediction models to evaluate the prediction power of features mentioned above and their interaction effects with different classroom arrangements.

6.2 Key Contributions to Instructors and Designers Based on Analytic Results

Our findings show that our approach to designing the game alongside the logging system enables us to (a) identify explicit learning achievements through game log analysis, (b) classify antecedent behaviours associated with these achievements, and (c) identify implicit game behavioural patterns associated with gameplay expertise and distinguish them from behavioural patterns associated with learning, especially in terms of information-processing ability.

Students play serious games like MHS in regular classrooms or as supervised extracurricular activities. We are also developing the game for remote students and home-schooled students. One of the main goals of this article is to help instructors identify struggling students. Recognizing significant features to distinguish students' expertise within the game validated by hypothetical tests suggests promising, concrete analytical strategies for advancing gaming and learning analytics. Based on the outcome summary of our hypothesis-driven analytics, we think, for example, that instructors will be able to intervene based on signals focused on

1. identifying students who explored an unexpectedly wide range of the game world,
2. rate of quest or task completion below or above specific thresholds,
3. spending a small share of time on comparatively complex tasks or quests,
4. reading argument nodes aimlessly or wasting much time on wrong choices,
5. frequent argumentation errors without successful submission,
6. interaction of objects in the game as part of the learning goals,
7. dialogue-triggering frequency and corresponding reading speed, and
8. frequency of tool usage and corresponding speed.

Of course, how instructors choose to intervene in gameplay is a function of their course arrangement and curriculum pace. Using MHS and the design guidance from our lab, instructors can also check students' in-game behaviours and interactions on our tracing-data visualization website, revealing students' statements on the aforementioned in-game features.

We have five initial pieces of advice for designers on serious game design. First, since gaming time is restricted in the classroom, designers may need to set up a boundary for the game world, preventing aimless area exploration to some extent. To better guide students, especially those with novice expertise levels, designers could improve the graphic sophistication of or add more details to the game environment around the navigation paths to complete key tasks, which, to some extent, lure students into the correct trajectory. Second, inspired by some novice students who feel confused by different in-game tools, even with contextual help, we suggest that designers create mini-tasks as a separate "practice space" to help students better understand each tool's application scenarios. Third, it is crucial to implement timely feedback to provide actionable guidance for teachers and students, for example, providing precise feedback when students submit the wrong answers during argument construction so that they know what they might need to learn to modify their argument, or seeing where they made mistakes (e.g., selecting the wrong evidence node or reasoning and claim) and better understanding the knowledge embedded in the quest. Fourth, when students run ahead in the wrong direction or interact with task-related items incorrectly, these are queues where MHS players need more background or contextual information. Designers can offer reminders (e.g., dialogues to let students know they need to change their decisions or implicitly describe what the correct answer looks like or the right thing to do next). Doing this can help decrease students' frustration, improve game engagement, and increase learning analytics data sources in the game context. Finally, we discovered from our results that students learned from negative outcomes, so we advise designers to consider negative outcomes as potentially useful for learning.

Furthermore, unlike real-world experiments, experiments within the game environment have much fewer restrictions, significantly higher tolerance for failures, and extremely low costs and risks, making them an ideal place to practise skills and knowledge and test new ideas. To adequately apply these advantages to students' learning, designers should leave enough space and time to make mistakes, learn from failure feedback, practise knowledge under different circumstances, and test new ideas to optimize their learning outcomes. For example, although the time for a course is fixed, designers could implement several side-quests in multiple formats, with which students can practise and review as after-class assignments the same subject matter knowledge introduced in the main quest.

6.3 Limitations and Future Possibilities

In summary, we have generated useful student-game interaction variables and identified appropriate performance measurement features that could become ingredients to formulate a solid learning performance prediction model for our next step. However, it is important to note that our work has limitations and to provide specific guidance for future work. First, although we conducted hypothesis-driven analytics and received expected outcomes, substantial influence from each feature to distinguish expert students from others needs further investigation. For example, the main goal of this study is to conduct an initial field test on an unfinished version of MHS to see the playability in the classroom environment and the executability of the logging system. We lack important external sources of pre-post assessments on students' knowledge and skills, which casts doubt on our conclusion, such as the association between knowledge gains and the final task's performance, the influences of students' pre-equipped skills and knowledge on their final success, and the effectiveness of the game to teach expected knowledge and skills. These deficiencies motivated us to construct a validated and interpretable prediction model to investigate further how the identified in-game features in this study influence students' learning outcomes, with necessary external assessment results under a more rigorous experimental design.

Another limitation emerges from our use of event types as a category of distinguishing features. Our results show that some players' event-type percentage shares are important factors in identifying experts. However, we explored only part of the information given by event types, involving a percentage share of single event types other than interaction effects from multiple event types, ignoring event order and influence caused by transitions between different event types. In future work, we may construct a computational model to evaluate interaction effects and use techniques from the field of sequential data analytics to investigate what other critical information the event-type sequences can provide. Beyond the event types, we believe the intersections between features used in different hypotheses within our study imply amplified research potential in terms of the optimal gameplay flow related to serious games, sophisticated usability tests, and intelligent stealth assessment systems embedded within the game.

The third limitation comes from treating students with the same characteristics, such as personalities, demographic backgrounds, player types, and learning styles. Future data combinations describing the students' unique traits with behavioural log data will help us develop richer models and improve our model prediction ability. These improved models will also help researchers explore relationships between students' in-game behaviours and their characteristics to improve the learning experiences of different students by incorporating personalized designs into the game (Shute & Ventura, 2013).

7. Conclusion

This article analyzed log data to generate statistically significant in-game features that show potential for developing learning prediction models for future serious games. To evaluate how these features affect students' performances (positive, negative, and to what extent), we plan to construct an interpretable computational model combined with in-game log data and measurements from external sources and then discuss relationships between features for more granular analysis.

Looking across all six MHS units in future work, we will face scaling challenges. The number of in-game interaction features is huge because of the game's multiple curriculum units, numerous quests, and complex in-game interactions between players. Filtering features based on observation records or outside information to reduce data dimension poses some risk of information loss and model instability. Instead, we think unsupervised and deep learning techniques will prove efficient and effective, especially during pre-model construction procedures, on outlier smoothing and feature selection. Such techniques can decrease high-dimensional metrics for distinguishing relevant signals from noise (Graves et al., 2016), discover hidden variables, and identify important features that explain a specific pattern. Our experience designing logging to explicitly measure learning provides higher than average assurance that our application of less supervised modelling will perform significantly better than the commonplace application of the same algorithmically centred technologies on impoverished, narrow logfile data.

Declaration of Conflicting Interest

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

Funding

The work described herein is supported by the US Department of Education's Institute of Education Sciences (R305A150364) and Investing in Innovation (i3) program (U411C140081). The ideas expressed are those of our project team and do not necessarily reflect the views of the funders.

References

All, A., Nuñez Castellar, E. P., & Van Looy, J. (2015). Towards a conceptual framework for assessing the effectiveness of digital game-based learning. *Computers & Education*, *88*, 29–37. https://doi.org/10.1016/j.compedu.2015.04.012

Alvarez, J., Plantec, J.-Y., Vermeulen, M., & Kolski, C. (2017). RDU Model dedicated to evaluate needed counsels for Serious Game projects. *Computers & Education*, *114*, 38–56. https://doi.org/10.1016/j.compedu.2017.06.007

Barab, S. A., Gresalfi, M., & Ingram-Goble, A. (2010). Transformational play: Using games to position person, content, and context. *Educational Researcher*, *39*(7), 525–536. https://doi.org/10.3102/0013189X10386593

Barzilai, S., & Blau, I. (2014). Scaffolding game-based learning: Impact on learning achievements, perceived learning, and game experiences. *Computers & Education*, *70*, 65–79. https://doi.org/10.1016/j.compedu.2013.08.003

Boyle, E. A., Hainey, T., Connolly, T. M., Gray, G., Earp, J., Ott, M., Lim, T., Ninaus, M., Ribeiro, C., & Pereira, J. (2016). An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. *Computers & Education*, *94*, 178–192. https://doi.org/10.1016/j.compedu.2015.11.003

Buckingham Shum, S., & Deakin Crick, R. (2016). Learning analytics for 21st century competencies. *Journal of Learning Analytics*, *3*(2), 6–21. https://doi.org/10.18608/jla.2016.32.2

Caballero-Hernández, J. A., Palomo-Duarte, M., & Dodero, J. M. (2017). Skill assessment in learning experiences based on serious games: A systematic mapping study. *Computers & Education*, *113*, 42–60. https://doi.org/10.1016/j.compedu.2017.05.008

Carvalho, M. B., Bellotti, F., Berta, R., De Gloria, A., Sedano, C. I., Hauge, J. B., Hu, J., & Rauterberg, M. (2015). An activity theory-based model for serious games analysis and conceptual design. *Computers & Education*, *87*, 166–181. https://doi.org/10.1016/j.compedu.2015.03.023

Chaves, R. O., von Wangenheim, C. G., Furtado, J. C. C., Oliveira, S. R. B., Santos, A., & Favero, E. L. (2015). Experimental evaluation of a serious game for teaching software process modeling. *IEEE Transactions on Education*, *58*(4), 289–296. https://doi.org/10.1109/TE.2015.2411573

Chittaro, L., & Buttussi, F. (2015). Assessing knowledge retention of an immersive serious game vs. a traditional education method in aviation safety. *IEEE Transactions on Visualization and Computer Graphics*, *21*(4), 529–538. https://doi.org/10.1109/TVCG.2015.2391853

Clarke-Midura, J., Rosenheck, L., & Groff, J. (2015). Designing games for learning and assessment: The radix endeavor. In *Proceedings of the 14th International Conference on Interaction Design and Children* (IDC 2015), 21–24 June 2015, Boston, MA, USA (pp. 343–346). ACM. https://doi.org/10.1145/2771839.2771912

Cutumisu, M., Blair, K. P., Chin, D. B., & Schwartz, D. L. (2015). Posterlet: A game-based assessment of children's choices to seek feedback and to revise. *Journal of Learning Analytics*, *2*(1), 49–71. https://doi.org/10.18608/jla.2015.21.4

Fernández, C. A. (2016). *Gaming learning analytics for serious games* [eprint], Universidad Complutense Madrid. https://eprints.ucm.es/38711

Fridenson-Hayo, S., Berggren, S., Lassalle, A., Tal, S., Pigat, D., Meir-Goren, N., O'Reilly, H., Ben-Zur, S., Bölte, S., Baron-Cohen, S., & Golan, O. (2017). "Emotiplay": A serious game for learning about emotions in children with autism: Results of a cross-cultural evaluation. *European Child & Adolescent Psychiatry*, *26*(8), 979–992. https://doi.org/10.1007/s00787-017-0968-0

Fulmore, Y. (2015). Video games and the customization of learning: Interactive narratives as a promising design framework for crafting inclusive educational environments. In *Proceedings of the Third Conference on GenderIT* (GenderIT 2015), 24 April 2015, Philadelphia, PA, USA (pp. 32–39). ACM. https://doi.org/10.1145/2807565.2807710

Gauthier, A., & Jenkinson, J. (2018). Designing productively negative experiences with serious game mechanics: Qualitative analysis of game-play and game design in a randomized trial. *Computers & Education*, *127*, 66–89. https://doi.org/10.1016/j.compedu.2018.08.017

Goggins, S. P., Gallagher, M., Laffey, J., & Amelung, C. (2010). Social intelligence in completely online groups—Toward social prosthetics from log data analysis and transformation. In *IEEE Second International Conference on Social Computing* (SocialCom 2010), 20–22 August 2010, Minneapolis, MN, USA (pp. 500–507). IEEE. https://doi.org/10.1109/SocialCom.2010.79

Goggins, S. P., Galyen, K., & Laffey, J. (2010). Network analysis of trace data for the support of group work: Activity patterns in a completely online course. In *Proceedings of the 16th ACM International Conference on Supporting Group Work* (GROUP 2010), 7–10 November 2010, Sanibel Island, FL, USA (pp. 107–116). https://doi.org/10.1145/1880071.1880089

Goggins, S., Galyen, K., Petakovic, E., & Laffey, J. (2016). Connecting performance to social structure and pedagogy as a pathway to scaling learning analytics in MOOCs: An exploratory study. *Journal of Computer Assisted Learning*, *32*(3), 244–266. https://doi.org/10.1111/jcal.12129

Graves, I., McDonald, N., & Goggins, S. P. (2016). Sifting signal from noise: A new perspective on the meaning of tweets about the "big game." *New Media & Society*, *18*(2), 293–312. https://doi.org/10.1177/1461444814541783

Hauge, J. B., Berta, R., Fiucci, G., Manjón, B. F., Padrón-Nápoles, C., Westra, W., & Nadolski, R. (2014). Implications of learning analytics for serious game design. In *IEEE 14th International Conference on Advanced Learning Technologies* (ICALT 2014), 7–9 July 2014, Athens, Greece (pp. 230–232). IEEE. https://doi.org/10.1109/ICALT.2014.73

Kang, J., Liu, M., & Qu, W. (2017). Using gameplay data to examine learning behavior patterns in a serious game. *Computers in Human Behavior*, *72*, 757–770. https://doi.org/10.1016/j.chb.2016.09.062

Kapur, M. (2008). Productive failure. *Cognition and Instruction*, *26*(3), 379–425. http://www.jstor.org/stable/27739887

Kiili, K., Moeller, K., & Ninaus, M. (2018). Evaluating the effectiveness of a game-based rational number training—In-game metrics as learning indicators. *Computers & Education*, *120*, 13–28. https://doi.org/10.1016/j.compedu.2018.01.012

Laffey, J., Griffin, J., Sigoloff, J., Sadler, T., Goggins, S., Womack, A., Wulff, E., & Lander, S. (n.d.). *Mission HydroSci: Meeting learning standards through gameplay* [Unpublished manuscript]. College of Education & Human Development, University of Missouri. https://files.eric.ed.gov/fulltext/ED600153.pdf

Liu, M., Lee, J., Kang, J., & Liu, S. (2016). What we can learn from the data: A multiple-case study examining behavior patterns by students with different characteristics in using a serious game. *Technology, Knowledge and Learning*, *21*(1), 33–57. https://doi.org/10.1007/s10758-015-9263-7

Loh, C. S., & Li, I.-H. (2015). *Predicting the competency improvement for serious games analytics: Action-sequences, game grids, PLS DA and JMP* [Paper presentation]. Discovery Summit 2015, San Diego, CA, USA. https://doi.org/10.13140/RG.2.1.3997.4889

Loh, C. S., Li, I.-H., & Sheng, Y. (2016). Comparison of similarity measures to differentiate players' actions and decision-making profiles in serious games analytics. *Computers in Human Behavior*, *64*, 562–574. https://doi.org/10.1016/j.chb.2016.07.024

Loh, C. S., & Sheng, Y. (2014). Maximum similarity index (MSI): A metric to differentiate the performance of novices vs. multiple-experts in serious games. *Computers in Human Behavior*, *39*, 322–330. https://doi.org/10.1016/j.chb.2014.07.022

Loh, C. S., Sheng, Y., & Li, I.-H. (2015). Predicting expert–novice performance as serious games analytics with objective-oriented and navigational action sequences. *Computers in Human Behavior*, *49*, 147–155. https://doi.org/10.1016/j.chb.2015.02.053

Martin, C. K., Nacu, D., & Pinkard, N. (2016). Revealing opportunities for 21st century learning: An approach to interpreting user trace log data. *Journal of Learning Analytics*, *3*(2), 37–87. https://doi.org/10.18608/jla.2016.32.4

Palmquist, S. D., & Crowley, K. (2007). Studying dinosaur learning on an island of expertise. In R. Goldman, R. Pea, B. Barron, & S. J. Derry (Eds.), *Video research in the learning sciences* (pp. 271–286). Routledge. https://doi.org/10.4324/9780203877258

Renninger, K. A. (2012). Interest and its development. In R. M. Ryan (Ed.), *The Oxford Handbook of Human Motivation* (pp. 167–187). Oxford University Press.

Renninger, K. (2010). Working with and cultivating the development of interest, self-efficacy, and self-regulation. In D. D. Preiss & R. J. Sternberg (Eds.), *Innovations in educational psychology: Perspectives on learning, teaching and human development* (pp. 107–138). Springer. https://psycnet.apa.org/record/2010-06004-005

Renninger, K., Cai, M., Lewis, M., Adams, M., & Ernst, K. (2011). Motivation and learning in an online, unmoderated, mathematics workshop for teachers. *Educational Technology Research and Development*, *59*(2), 229–247. https://doi.org/10.1007/s11423-011-9195-4

Serrano-Laguna, A., Martínez-Ortiz, I., Haag, J., Regan, D., Johnson, A., & Fernandez-Manjon, B. (2017). Applying standards to systematize learning analytics in serious games. *Computer Standards & Interfaces*, *50*, 116–123. https://doi.org/10.1016/j.csi.2016.09.014

Shute, V., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games*. The MIT Press. https://doi.org/10.7551/mitpress/9589.001.0001

Towne, T. J., Boot, W. R., & Ericsson, K. A. (2016). Understanding the structure of skill through a detailed analysis of individuals' performance on the Space Fortress game. *Acta Psychologica*, *169*, 27–37. https://doi.org/10.1016/j.actpsy.2016.05.006

van Leeuwen, A. (2015). Learning analytics to support teachers during synchronous CSCL: Balancing between overview and overload. *Journal of Learning Analytics*, *2*(2), 138–162. https://doi.org/10.18608/jla.2015.22.11

Ventura, M., & Shute, V. (2013). The validity of a game-based assessment of persistence. *Computers in Human Behavior*, *29*(6), 2568–2572. https://doi.org/10.1016/j.chb.2013.06.033

Appendix

Detailed Description Related to Unit 3 Key Quests and Selection of Learning Performance Standard in This Study

In this section, we will discuss how we distinguish expert students and novice students. As we described in the research context section, Mission HydroSci (MHS) is a serious game designed and developed with comprehensive and sophisticated curriculum integration in mind. Each key quest of MHS can be seen as an efficient marker of learning achievement. This judgment builds on the previous journal article, which indicated that high overall performance in a game, which combines game mechanics directly with learning objectives, reflected players' knowledge of curriculum subject matter rather than chance. The authors proved the statement by examining correlation indexes and regression outcomes based on post-test outcome as well as in-game performance measurements (Kiili et al., 2018). Instead of identifying experts a priori, we identify our expert category of players for MHS Unit 3 by breaking down the key quests in the unit and choosing "expertise criteria" that are durable and also emergent in the field study. The next paragraph explains how we arrived at "successful seed planting" as a key differentiator of expertise in this unit.

There are four key quests in this unit. Each quest teaches students one aspect of knowledge about water flow. The first quest is to deliver crates of supplies to the river that can carry the crate to Sam's base (Figure .11a and .11b). There are two rivers, and students need to check the map tool to identify where Sam's base is and which river passes nearest to the base. The second quest is to toss sensors into the river to help identify the source of pollution (Figure .11c). Completing this quest demonstrates that students know how soluble materials flow with a river. The third quest asks students to choose a correct combination of nodes, categorized in groups of evidence, reason, and claim, to construct a scientific argument. During the procedure, students can review up-to-now water flow knowledge and what a good argumentat looks like in the game (illustrated in Figure .11d). This quest is where students demonstrate the ability to construct a scientific argument about water flow.

The fourth key quest is planting seeds in Sam's garden (Figure .11e and .11f). Done correctly, garden pods are spread along the river following a strange, huge tree, which is affected by a toxic water-soluble material. Students need to place the garden pods where they will be influenced by the pollution. Seeds planted in correct pods will grow rapidly into huge plants similar to the strange, huge tree. We make this quest our evaluation criterion for expertise because (1) it is where students summarize their water knowledge by integrating knowledge of water flow direction and water-soluble materials; (2) a successful result for this quest means players have completed previous key quests without abandoning the game; and (3) as the final key quest of Unit 3, which is directly integrated with curriculum content, seed planting measures players' overall game performance in Unit 3, which may reveal signs of players' learning performance, since it has been proved to have high correlation with players' learning performance (Kiili et al., 2018).

Figure .11. Example Screen Shots of Key Quests in Unit 3*

*Figures (a) and (b) focus on crate placement, and (c) focuses on pollution in the game. Figure (d) is a screen shot of our argumentation system where students integrate evidence, claims, and reasoning to construct scientific arguments. Figures (e) and (f) focus on the seed-planting quest.