JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Volume 12(1), 293–321. https://doi.org/10.18608/jla.2025.8105

Game-Based Learning Prediction Model Construction: Toward Validated Stealth Assessment Implementation

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

Abstract

Game-based learning (GBL) is increasingly recognized as an effective tool for teaching diverse skills, particularly in science education, due to its interactive, engaging, and motivational qualities, along with timely assessments and intelligent feedback. However, more empirical studies are needed to facilitate its wider application in school curricula. A significant challenge is designing and implementing valid in-game assessments crucial for measuring student progress and providing reliable references for instructors' intervention decisions. Stealth assessment, guided by the evidence-centred design (ECD) framework, offers a promising solution but requires more specific guidelines for full effectiveness. In this study, we present a granular, framework-supported pipeline to systematically implement stealth assessments in a GBL environment. This pipeline involves constructing an ECD framework, generating features, selecting appropriate models, preprocessing data, evaluating model performance, and conducting model inference on a black-box computational model. We validate the effectiveness of this pipeline by assessing the performance of these computational models and identifying distinct behavioural patterns between high and low performers. Our analysis highlights potential areas for improvement in the design of stealth assessments within digital games for learning. Furthermore, we discuss the generalizability of the proposed pipeline and outline limitations for future research to address.

Notes for Practice

- Our study details and validates a systematic approach for developing and applying stealth assessment via a granular embedded logging system. We demonstrate this approach's efficacy in game-based learning (GBL), outlining strategies for stealth assessment structuring, feature generation, computational model selection and training, performance evaluation, and inference. Our findings underscore the importance of selecting suitable frameworks for each procedure to enhance the feasibility, efficiency, and effectiveness of stealth assessments. Importantly, our defined process initiates the formation of a guideline for implementing stealth assessment in other GBL contexts.
- Concerning the feature generation process, we advise using a suitable framework and performing multi-level classification on the data based on the information reflected. This enhances model interpretability and enables analytics at various granularity levels to meet research requirements.
- The performance of the computational model suggests that combining in-game learning progress, as indicated by embedded assessment scores, with behaviours yields the most accurate predictions of learning outcomes.
- Implementing a surrogate model, commonly a white-box model, is a practical approach for interpreting black-box models. Through detailed analysis of inference results, we identify distinct behavioural patterns between high- and low-outcome students in the game.
- Drawing on insights from model inference results and the iterative design paradigm within the evidence-centred design framework, we discuss how to continuously refine our proposed pipeline for establishing stealth assessments and offer recommendations for designing and developing adaptive stealth assessments in GBL environments.

Keywords

Serious games, learning analytics, game-based learning, logging system, computational model, conceptual model and frameworks, machine learning, prediction model, educational data mining, performance measurement.

Submitted: 23/06/2023 — **Accepted:** 20/01/2024 — **Published:** 18/03/2025

[1] *Corresponding author Email: wldh6@umsystem.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: griffinjg@missouri.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*
[5] *Email: gogginss@missouri.edu Address: Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, Missouri, USA. ORCID iD: https://orcid.org/0000-0002-4331-147X*

1. Introduction

Game-based learning (GBL) is an influential tool in science education, enhancing student learning efficiency and skill mastery (M. Wang & Zheng, 2021; L. H. Wang et al., 2022). It fosters competencies vital for science learning, such as motivation, conceptual understanding, and science process skills (Laffey et al., 2017). In contrast with entertainment games, educational games focus on skill development and knowledge acquisition. Their interactive nature offers an engaging medium for students, allowing the repetitive practice of a progressive science curriculum under guidance or self-paced learning (Breuer & Bente, 2010; Maryani & Hidayat, 2019; Fadila et al., 2023). Their replay feature promotes learning from failure and strategy refinement (Zhang & Rutherford, 2022). Teachers, in turn, can enhance pedagogical designs for improved, timely support (Shohel et al., 2022). The effectiveness of GBL depends on the availability of relevant, timely information for optimal adaptation during gameplay (Sevcenko et al., 2021).

While the potential benefits of GBL are widely recognized, particularly in fostering problem-solving skills and enabling adaptive learning through timely feedback, the effectiveness of GBL is also subject to debate, with several controversial aspects that warrant careful consideration. Recent studies have identified potential downsides, such as cognitive overload, equity accessibility issues, difficulties in skill transfer to real-world applications, and challenges in measuring complex learning outcomes. For example, cognitive overload can occur if the game design is too complex or not well aligned with learning objectives, potentially reducing efficiency (Sevcenko et al., 2021; Seyderhelm & Blackmore, 2023). Additionally, concerns about equity and accessibility persist, as some students may lack access to the necessary technology or digital literacy skills, raising issues of inclusivity in digital learning environments (Haas & Tussey, 2022; Rohmani & Pambudi, 2023). Moreover, there are ongoing challenges regarding the transferability of skills acquired through GBL to real-world contexts, with evidence suggesting that skills developed in game environments do not always translate into improved academic or practical performance (Cerra et al., 2022; Nietfeld, 2020).

Furthermore, there are concerns that game elements may overshadow educational objectives, resulting in shallow learning where the focus shifts from educational content to game mechanics. Recent literature suggests that overly gamified environments may prioritize entertainment over meaningful learning, potentially diminishing deeper cognitive engagement and retention of material (Bernecker & Ninaus, 2021; Manzano-León et al., 2021). Additionally, accurately measuring complex learning outcomes—such as critical thinking, creativity, and problem-solving—remains a significant challenge in GBL. Despite the extensive data generated by these environments, there is a lack of standardized metrics and validated models for assessing these higher-order skills effectively (Zhu et al., 2023; Strukova et al., 2023). These controversies underscore the need for a nuanced approach to integrating and assessing GBL within educational curricula, ensuring that its potential benefits are realized while addressing its limitations.

Active engagement with educational games bolsters students' problem-solving skills (Rosydiana et al., 2023). These interactions produce extensive data traces, which, when designed to align with game learning objectives, can enhance students' learning approaches and offer valuable insights for teaching (Georgiadis et al., 2019). To harness these benefits, practitioners have integrated a data collection system, or logging system, and corresponding assessments directly into game design, in contrast to traditional instruction where assessments are often an afterthought (Loh et al., 2016; Zhu et al., 2023). Such integration enables educational games to act as intelligent tutoring systems, offering timely feedback that supports student learning and encourages adaptive strategies during gameplay (Gee, 2003; V. J. Shute, 2008; Hooshyar et al., 2016; Ke et al., 2019; Yu et al., 2022).

Stealth assessment, a formative method also known as "assessment for learning," is notably beneficial in GBL environments (J. P. Rowe et al., 2009; V. J. Shute, 2011b; Mislevy et al., 2003; Baker et al., 2010). It operates on three primary principles: (1) unobtrusive data collection, (2) evaluation of complex competencies like critical thinking, and (3) sequential and detailed monitoring of learning progress to offer personalized feedback. Integrating stealth assessment in these environments is a complex, resource-intensive task requiring multidisciplinary collaboration (computer science, education, psychology, statistics, etc.). A comprehensive framework for integrating game logs, feature engineering, computational model building, and model interpretation is essential to maximize the utility of stealth assessment for both students and teachers in a GBL scenario.

Interdisciplinary collaboration within educational game design requires a robust alignment of technical tasks and learning objectives, with learning as the ultimate objective. The evidence-centred design (ECD) approach serves as a foundational methodology for stealth assessment, validated by numerous studies across different educational games (V. J. Shute, 2011a; V. Shute & Ventura, 2013; V. J. Shute et al., 2016; V. Shute et al., 2017; V. J. Shute & Rahimi, 2021; Moore & Shute, 2017; Min et al., 2020; Henderson et al., 2020, 2022). ECD incorporates three main components: competency, evidence, and task models. The competency model probabilistically represents students' skills and knowledge. The evidence model demonstrates how student behaviour observations can reveal their competencies. The task model defines challenges that produce evidence for inferring students' competency levels (V. J. Shute, 2011a; V. Shute & Ventura, 2013; Ma et al., 2015).

1.1 Need for Comprehensive Stealth Assessment Design and Evaluation

Although prior research indicates the potential of stealth assessment, educational game creators face a range of approaches for developing models and identifying game-specific features, with limited comprehensive evaluations of their comparative performance (Jeon et al., 2023; Gomez et al., 2023; Georgiadis et al., 2021; Fang et al., 2023).

Our study builds upon the development of Mission HydroSci (MHS), a 3-D GBL environment. MHS employs a co-curricular design that complements teacher interaction. It is developed alongside the curriculum to teach middle school students water science knowledge and scientific argumentation. Each unit in MHS corresponds to a specific curriculum objective, shaping the game mechanics. In MHS, players become junior scientists on a mission to establish a settlement on the newly discovered planet WAT247. They can explore the game world, search for hints, and complete quests in different formats such as puzzles, item searches, and pathfinding.

Along with the game, there is an integrated logging system. The system captures events related to in-game activities and learning progress. Our logging system draws inspiration from the works of Carvalho and colleagues (2015) and Serrano-Laguna and colleagues (Serrano-Laguna et al., 2017). Carvalho and colleagues introduced the Activity Theory-based Model of Serious Games (ATMSG) to define and deconstruct explicit content for the logging system to collect. This content includes key features reflecting students' in-game progress, behaviour, and decisions, recorded in chronological sequences of trace data. Serrano-Laguna and colleagues proposed an Experience API (xAPI) standardization for structuring defined content into data statements in JSON format, saved in remote databases.

Further application and empirical evidence are needed to validate the effectiveness of combining both frameworks, ATMSG (Fokides et al., 2019; Alonso-Fernández, Freire, et al., 2021) and xAPI (Schardosim Simão et al., 2018; Heinemann et al., 2022), for logging system design and development, a gap our paper aims to address. Additionally, due to the diverse nature of the in-game activities within MHS, we faced challenges in designing appropriate data coding schemes and generating relevant features from raw logs for model construction with interpretability for education and learning purposes. In this study, we propose a potential approach for generalizing the feature generation process based on a complex learning system like MHS.

To summarize, our main goal is to develop a stealth assessment for evaluating students' learning outcomes related to water science knowledge taught in Unit 3 of MHS. To validate its effectiveness, we used gameplay log data from over 300 students and their external post-assessment scores as the learning standard for constructing and evaluating prediction models. We integrated our model prediction procedure into the ECD approach and employed the Integrated Design of Event-stream for Analysis (IDEFA) framework (Owen & Baker, 2020) to guide our feature generation process. We then preprocessed the feature set for model training, evaluated model performance to verify the effectiveness of our stealth assessment, and conducted model inference for result interpretation. Based on the inference results, we discussed the generalizability of our comprehensive pipeline for implementing a stealth assessment within MHS.

1.2 Research Questions

More specifically, to achieve our research goal, we probe and present solutions for the following research questions:

> **Research question 1 (RQ1):** Apply, adapt, and extend existing models and frameworks for MHS to discover how and to what extent they are valid for stealth assessment within MHS.
>
> **Research question 2 (RQ2):** How do the inference results from interpreting the black-box computational model inform future stealth assessment design and GBL development?

2. Literature Review

2.1 High-Level Conceptual Models Guiding the Game Design and Learning Analytics

Data-driven methods, including learning analytics and educational data mining, are crucial in education, but their validity rests on the quality of data collection aligned with research goals. The system design and implementation framework play a significant role in ensuring the appropriateness of collected data. Various validated models and frameworks have been proposed to guide this process, each emphasizing different areas.

This study builds on prior research on conceptual models for learning analytics within GBL. De Freitas and colleagues (2006) introduced a framework incorporating pedagogic considerations, mode of representation, learner specification, and context, primarily suited for pre-existing video games. However, limitations such as restricted instructor usage, lack of dimension transformation, and limited learning context present opportunities for future research. Further, Gunter and colleagues (2006) proposed the RETAIN model for new game design, blending three validated paradigms: Gagne's nine events of instruction, Keller's ARCS model of motivation, and Bloom's learning domains. Despite showing potential for embedding learning content into game design, the RETAIN model requires further illustration and case studies for validation and linking pedagogical design with specific game elements.

Arnab and colleagues (2015), addressing Gagne's model's shortcomings, developed a Learning Mechanics-Game Mechanics (LM-GM) model that effectively translates pedagogical practices into game mechanics. Validated against Alan Amory's Game Object Model (GOM) (Amory, 2007) through two user evaluations, LM-GM was superior in three aspects: (1) providing detailed descriptions of relationships between learning mechanics and game elements, (2) accessibility and applicability, and (3) enhancing user understanding of how games promote learning. However, the authors recognize the need for enhanced and precise evaluation techniques for their model components, especially those related to learning measurements.

In recent years, the game industry's swift advancement has given rise to increasingly sophisticated video games, enabling educators to use them for teaching complex skills such as critical thinking and problem-solving. This progression has created a demand for comprehensive frameworks for designing and evaluating game mechanics and pedagogical goals in GBL environments. Addressing this need, Carvalho and colleagues (2015) presented ATMSG, building upon and extending previous models. Applied in user evaluations of five educational games, ATMSG demonstrated its superior capacity for precise evaluation of game mechanics and pedagogical elements compared to the LM-GM model. This model's strengths lie in its ability to thoroughly analyze system components as the game progresses and articulate the relationships between game components and learning goals for various stakeholders. Although ATMSG is more beneficial for expert users like game designers or researchers, the authors see its potential as a blueprint for creating analysis tools in GBL environments.

Inspired by ATMSG's meticulous analytical abilities, we applied them to guide the design and development of our adaptive logging system, a game-embedded analysis tool. Specifically, they assisted in determining what game content to log for learning analytics and generating features for the predictive model.

2.2 Game Logs Application as Learning Measurements

Advanced technologies in education have popularized the use of logs for profiling application use across several domains, including social computing environments (Ayzenberg et al., 2012; Goggins et al., 2010; Park & Cho, 2010), massive open online courses (MOOCs) (Goggins et al., 2016; N. Li et al., 2015), computer-supported collaborative learning (CSCL) (Xing et al., 2014; Goggins et al., 2011; Martínez-Monés et al., 2011), and 3-D virtual learning environments (Agudo-Peregrina et al., 2014; Ma et al., 2015; Grover et al., 2017).

A robust logging system should record detailed, timestamped sequences of user-system interactions to unveil learning-related usage patterns, supporting multi-level analytics. It should also be capable of real-time representation of user learning progress when necessary (Ventura & Shute, 2013; Goggins et al., 2010; Hauge et al., 2014). Furthermore, the logging system should evolve with the learning game, ensuring the capture of vital information on how game modifications affect learning outcomes (Kim et al., 2019). Lastly, it should function unobtrusively, allowing uninterrupted user engagement with the system, resulting in more accurate learning measurements (Loh et al., 2015).

The use of game logs for learning analytics has been widely explored within GBL. These logs are pivotal for studying areas like subject-matter knowledge and skills (Nguyen et al., 2020; Emerson et al., 2020; Feng & Yamada, 2019), complex competencies (Lee et al., 2019; Niemelä et al., 2020; Cloude et al., 2020; Wen et al., 2018; Sabourin et al., 2013; Seaton et al., 2019), and performance assessment design (Gibson & Clarke-Midura, 2015; Westera et al., 2014; Loh & Sheng, 2015). When combined with external data from sources like eye-tracking or emotion detection, game logs provide enhanced, more precise learning analytics (Lee et al., 2019; Emerson et al., 2020; Cloude et al., 2020).

Building on the established use of game logs in learning analytics, recent studies have applied advanced methods like machine learning and reinforcement learning to analyze game data and achieve adaptive learning experiences in GBL environments (F. Chen et al., 2020; Cardia da Cruz et al., 2020; Rahimi et al., 2023). For example, Chen and colleagues (2020) employed support vector machines and long short-term memory networks to predict learning outcomes from students' game logs. Researchers have also used reinforcement learning to adjust game difficulty in real time based on players' performance, demonstrating potential in maintaining engagement and enhancing learning outcomes (Cardia da Cruz et al., 2020; Rahimi et al., 2023). These AI-driven techniques highlight the potential of game logs in understanding and predicting learning outcomes, designing dynamic assessments, and creating adaptive learning experiences. However, limitations such as limited exploration of behavioural features, small sample sizes, short-term evaluations, and simplified difficulty metrics hinder real-world implementation. The authors encourage future research to investigate additional game log features to enhance predictive accuracy and the effectiveness of reinforcement learning approaches in GBL environments.

Researchers have also explored various in-game features—representing behaviours, decision-making, and progress—to analyze learning patterns across different expertise levels (Díaz-Ramírez, 2020; F.-Y. Li et al., 2021; Liu et al., 2022). Examined features include tool usage, navigation patterns, task progress, and earned rewards. Recent studies have focused on real-time visualization of game logs to enhance educational outcomes. Vidakis and colleagues (2020) collected real-time data on student interactions in "ThimelEdu," capturing behaviours like navigation and decision-making and visualizing data for educators to adapt teaching strategies in real time. Similarly, Calvo-Morata and colleagues (2020) used game analytics to validate "Conectado," a game designed to raise cyberbullying awareness. They captured students' decisions and emotional responses, creating dashboards that offer insights into engagement and empathy development. Real-time monitoring enables educators to refine interventions, tailoring them to meet specific learning objectives and enhance educational impact. However, comprehensive studies combining these features are scarce, necessitating more empirical research.

However, most studies use game logs in the context of specific environments, which limits their wider applicability. With the growing prominence of GBL, it's important to develop universal standards for logging system design (Shoukry et al., 2014; Pérez-Colado et al., 2022; Lu et al., 2023). Recent research highlights that standardized data collection methods can greatly benefit future research, such as facilitating cross-study comparisons and supporting cross-platform tool development (Serrano-Laguna et al., 2014; Vidakis et al., 2020; Alonso-Fernández, Calvo-Morata, et al., 2021). xAPI, a model proposed by Serrano-Laguna and colleagues, is a high-level standard for logging system design. This model aims to ensure that the data collected is efficient and effective for measuring learning goals across various game environments (Serrano-Laguna et al., 2017). The xAPI model has inspired our design for high-level data collection processes.

Informed by the frameworks of xAPI and ATMSG, we designed the high-level data structure and determined the fine-grained content to capture. Yet, a disconnect remains between the data collected and the learning goals we aim to measure or predict (Jeon et al., 2023). A framework that guides the implementation of performance assessments, seamlessly integrated with the game progression, can bridge this gap (V. J. Shute & Rahimi, 2021; Udeozor et al., 2024). Such a framework ensures that the collected data can accurately and unobtrusively measure the targeted learning outcomes as the game unfolds. As reviewed in the next section, previous research has created, developed, and validated frameworks for implementing stealth assessments in GBL environments. However, these studies have also acknowledged the need for further empirical evaluations to validate these frameworks in diverse educational contexts (V. J. Shute et al., 2016; Min et al., 2020; Georgiadis et al., 2019; V. J. Shute & Rahimi, 2021; Udeozor et al., 2024).

2.3 Stealth Assessment in a GBL Environment

As society evolves, today's youth must master complex competencies, including 21st-century skills (Romero et al., 2015), to keep pace with the modern world. However, teaching these skills and assessing students' progress presents significant challenges. Emerging technologies offer solutions by enabling the development of embedded assessments to augment learning processes. Unlike traditional assessments, such as paper-based exams, embedded assessments offer many benefits: (1) They unobtrusively gather continuous, multifaceted learner data, providing objective, comprehensive results without disrupting learning or creating test anxiety. (2) Utilizing machine technologies, they provide real-time scores based on learner actions and progress, offering quantitative feedback to improve learning. (3) By integrating into learning systems like game-based environments, they measure learning in context and in real time, unlike traditional pre- and post-learning assessments. This immediacy accurately reflects learner progress, making these "stealth assessments" a valuable tool for educators and researchers.

Göbel and colleagues (2009; 2013), building on their early research in story-based edutainment and serious games, put forth a stealth assessment framework for story-based digital educational games (DEGs), named Narrative Game-based Learning Objects (NGLOB). They validated this framework with two computer-based games. However, its use has declined recently, possibly due to the constraints of the narrative genres and the specific needs of GBL environments where their methods were applied.

Shute and colleagues (2011b), unrestricted by the narrative game genre, conducted various studies on a stealth assessment model in GBL environments, utilizing the ECD approach (Mislevy et al., 2003). This model was validated across diverse educational games, assessing competencies such as mathematical skills, problem-solving, conscientiousness, calculus abilities, and creativity (V. Shute et al., 2017; V. J. Shute et al., 2016; Moore & Shute, 2017; Smith et al., 2019; V. J. Shute & Rahimi, 2021). The ECD model encompasses three key components: (1) **competency model (CM)**, which defines the knowledge and skills to be assessed; (2) **evidence model (EM)**, which identifies in-game behaviours or progress revealing the competencies and their statistical relationship with CM variables; and (3) **task model (TM)**, which outlines in-game situations or quests through which students demonstrate their competency progress.

These components enable practitioners to examine learning behaviour patterns and estimate competence levels in a timely manner. Shute's studies primarily focus on discerning relationships between different in-game behaviour-derived indicators and assessed competencies using only Bayesian networks (BNs). BNs effectively visualize complex relationships, including time factors, in a manner that keeps data useful and manageable (Champion & Elkan, 2017; Heine, 2021; Belland et al., 2017; Mouri et al., 2016). However, developing BNs is labour-intensive, time-consuming, and costly, to ensure accurate representation of

learning in the final structure. Furthermore, even a robust BN, based on an appropriate prior distribution, usually requires substantial data for validation. The outcomes may be too specific to the experiment, limiting their applicability in other GBL environments.

Exploring beyond BNs, Grover and colleagues (2017) applied the ECD framework to measure computational thinking (CT) in block-based programming environments like Alice, combining hypothesis-driven methods with data-driven learning analytics. Their approach integrates real-time data from student programming activities with pre-defined indicators of CT skills, such as debugging, use of conditionals, and iterative design, to enhance the accuracy of assessments and provide formative feedback to learners. This hybrid framework illustrates another application of ECD beyond traditional GBL, where it supports stealth assessment in educational programming environments.

Furthermore, Lester and colleagues examined the use of machine learning models, such as random forest, support vector machine, and recurrent neural networks, for stealth assessments within GBL environments (Akram et al., 2018; Min et al., 2020; Henderson et al., 2020; Gupta et al., 2021; Henderson et al., 2022). They identified in-game behaviours linked to targeted knowledge and skills, integrating these models into the ECD framework. This resulted in novel stealth assessment frameworks with various benefits: streamlining data preprocessing (Min et al., 2020), enabling the operation of stealth assessments in domains and educational content where prior data and labels are unavailable (Henderson et al., 2022), and infusing diverse data types (Henderson et al., 2020). However, these models' complexity hinders interpreting how individual indicators predict learning outcomes.

Addressing the limitations of BNs (Champion & Elkan, 2017; Heine, 2021; Belland et al., 2017; Mouri et al., 2016) and the challenges of other modelling approaches in identifying game-specific behaviours related to learning (Akram et al., 2018; Min et al., 2020; Henderson et al., 2020; Gupta et al., 2021; Henderson et al., 2022), Georgiadis and colleagues (2019) developed a computational prototype to conceptualize various approaches to stealth assessment. This work uses simulated data to verify their prototype and explore numerous modelling techniques to inform future stealth assessment designs within serious games. It offers a comprehensive range of potential computing approaches for stealth assessment researchers. However, this work's applicability is limited due to the absence of a specific GBL system, a human evaluation environment, and a clear connection between these models and learning outcomes.

2.4 Gaps in the Literature

Despite significant advancements in GBL analytics and the development of various conceptual models and frameworks, several gaps remain in the literature. Existing models, such as those proposed by De Freitas and colleagues (2006), Gunter and colleagues (2006), and Arnab and colleagues (2015), face limitations like restricted applicability, lack of transformation between pedagogical and game elements, limited learning contexts, and the need for enhanced evaluation techniques—particularly concerning learning measurements. Advanced analytic methods, including machine learning and reinforcement learning, have shown potential in utilizing game logs for adaptive learning experiences. However, their real-world implementation is hindered by limited exploration of behavioural features, small sample sizes, short-term evaluations, and simplified difficulty metrics.

Additionally, there is a scarcity of comprehensive empirical studies that combine various in-game features to analyze learning patterns, necessitating more research in this area. The lack of universal standards for logging system design further limits the generalizability of findings across different GBL environments. Frameworks guiding the implementation of performance assessments, such as stealth assessments integrated within game progression, require additional empirical validation in diverse educational contexts. Complex modelling approaches like BNs and advanced machine learning models, while useful, pose challenges in terms of development effort, interpretability, and applicability.

Therefore, there is a critical need to develop accessible, learning-centred frameworks that facilitate the transition from raw game logs to meaningful feature sets and computational models. To address these gaps, we present our study involving MHS, a comprehensive GBL environment designed for middle school students. By applying our proposed framework for constructing learning prediction models within MHS, we aim to enhance the accuracy of measuring and predicting targeted learning outcomes, thereby bridging the gap between data collection and educational objectives in GBL environments. This approach not only validates our framework but also contributes to the broader goal of making advanced analytics more accessible and effective in educational games.

3. Game Context and Data Collection

3.1 MHS

MHS is a comprehensive GBL curriculum conceptualized as a 3D transformational role-play platform for middle school students (see online Appendix B (https://bit.ly/3QvTeBP) for detailed information on game content). Within this interactive context, students play the role of novice scientists, exploring and managing a remote planet's water systems and topography with the overarching objective of utilizing these resources for human survival. MHS encompasses numerous assistive tools,

facilitating the learning of decision-making processes, puzzle-solving techniques, and scientific argumentation. Integral to MHS is a logging system conceived and developed with the game, following the ATMSG and xAPI frameworks.

To identify the content to be captured during gameplay, we applied the ATMSG framework to deconstruct and analyze gameplay and learning activities, ensuring that each log event aligns with the framework. The ATMSG framework's hierarchical structure—encompassing gaming, learning, and instructional activities—was instrumental in defining which actions and events should be logged in MHS. This systematic categorization of actions, tools, and goals enabled us to capture meaningful data on player engagement and learning outcomes.

- **Gaming activities:** Player interactions with the game were mapped to specific events, such as movement, trigger, and quest-/task-related events. Movement events capture details like direction (forward, left, right, back), state (start, end), and navigation tool use (e.g., hoverboard). These events monitor player exploration and spatial engagement. Trigger events log interactions with objects, recording object ID, action type (e.g., Lift, Drop, Press), and object state changes, reflecting problem-solving approaches. Quest and task events track progress, logging when players initiate or complete quests or tasks.

- **Learning activities:** Actions aligned with learning objectives were logged through events capturing scientific argumentation, decision-making, and knowledge application. Argumentation events log players' use of the argumentation engine, capturing session openings, engagement with nodes, and argument construction, providing insights into critical thinking and reasoning. Dialogue events track the start and end of dialogues and player choices, offering data on how players engage with educational narratives and assess understanding.

- **Instructional activities:** Instructional elements, including in-game and teacher-led activities, were tracked through tool events and hotkey use, monitoring players' access to support tools and the impact on their performance. Tool events, such as the argumentation engine and AI companion tools, provide insight into tool utilization, while hotkey events reveal player fluency with controls and resource navigation strategies.

The ATMSG framework effectively aligned game design with educational objectives. Each quest or challenge was designed to support learning goals, such as understanding water flow dynamics or properties of water-soluble materials. The logging system captured relevant data—such as argumentation, dialogue choices, and tool use—offering insights into players' problem-solving, decision-making, and concept application.

For the collection and storage of this content in data statements within remote servers, such as a learning record store (LRS), for real-time applications and advanced analysis, we applied the xAPI framework. This framework provides a technical standard for tracking and recording behavioural trace data within serious games. By mapping ATMSG components to xAPI statements and making necessary adjustments to align with MHS and our research objectives, we designed and developed the embedded logging system. To evaluate the effectiveness of this logging system, we conducted an empirical study using data collected during the first field test of MHS (Lu et al., 2023).

The MHS curriculum is organized into six units. Unit 1, a tutorial, introduces navigation, in-game tools, and a system for scientific argumentation. Unit 2 focuses on topography, teaching students to find in-game teammates and compare watershed sizes. Our research primarily examines Unit 3, which instructs students on water flow and the properties of water-soluble materials. We have deliberately excluded data from other units in the current study. This decision is premised on the fact that subsequent units do not influence a student's performance in Unit 3. Focusing on Unit 3 also presents a manageable scope for furthering stealth assessment research in line with our research questions. Our analysis of Unit 3 is intended to lay the groundwork for future investigations into stealth assessments, both for our subsequent work and for other researchers in the field.

3.2 Data Collection

The second field test for MHS was conducted over two weeks in the spring semester of 2019. The participating teachers, drawn from 13 middle schools across nine school districts, were required to have at least two class periods from 6th to 8th grade. These schools were situated in mid-sized cities and small rural communities, and the student sample comprised 1,110 students of varied ethnicities and genders. 806 students from 35 classes went through MHS as a game-based curriculum. Among those students, 632 of them completed pre- and post-tests for all constructs, qualifying for the analytic sample.

We implemented three distinct assessment instruments for the pre- and post-tests. These included student tests of affect toward science and technology (MAST) (Romine et al., 2017), water science content knowledge (Reeves et al., 2020), and scientific argumentation (Reeves et al., 2020). The tests were administered via Qualtrics, an online assessment platform. The primary objective of this study was to construct a stealth assessment, informed by the ECD framework, to evaluate students' learning outcomes associated with Unit 3's curriculum related to water science content knowledge. This evaluation was based on the features extracted from the students' in-game logs. These logs, saved on a remote MongoDB database, were transformed

into data frames via the R software for additional analysis and model construction. Following a thorough data cleaning process and trimming to eliminate incomplete records, the final dataset comprised 354 students with comprehensive log records and corresponding assessment scores.

4. Framework-Supported Pipeline for Valid Stealth Assessment Implementation

This section outlines our approach for feature set identification, which we used to train and evaluate our computational models. We also detail integrating the ECD framework in affirming our stealth assessment application within MHS. Feature engineering, a method for pinpointing potent learning predictors from vast volumes of student gaming data, is crucial for developing durable, interpretable prediction models (Owen & Baker, 2020; Guyon & Elisseeff, 2003; Sao Pedro et al., 2012; Fogarty, 2006). Essentially, MHS feature engineering transforms raw data into meaningful information through the fusion of expert judgment and iterative mathematical operations, adapting continually with curriculum and game design based on distinct learning goals. To optimize our final feature set, we implemented recommendations from Zheng and Casari (2018) and Butcher and Smith (2020) and used the feature engineering framework Integrated Design of Event-stream for Analysis (IDEFA) (Owen & Baker, 2020) to guide our process.

4.1 Overview of the IDEFA Framework

The IDEFA framework is a systematic process for creating and refining data features from event-stream data, with the ultimate goal of producing robust models for behaviour prediction. It consists of several key phases: data design and collection, base feature aggregation, feature engineering, and iterative analysis. While the IDEFA framework includes data design and collection as a core element, in this study, we employed a different way to design and develop the integrated logging system within MHS using the ATMSG and xAPI frameworks, as described in the previous section.

Although we did not follow IDEFA's guidance for data collection, the remaining three phases—base feature aggregation, feature engineering, and iterative analysis—were fully integrated into our workflow, as follows: (1) **Base feature aggregation:** Once raw data was collected through our customized logging system, we followed the IDEFA framework's principles to identify and aggregate key features from the event-stream data. (2) **Feature engineering:** We applied mathematical operators to transform base features into new variables that better captured player behaviours, ensuring that the data was suitable for predictive modelling. (3) **Iterative analysis:** Using IDEFA's iterative process, we tested and refined the engineered features, optimizing our models for predicting learning outcomes.

By combining the flexibility of our logging system with the structured processes from IDEFA for feature generation, we ensured that the features generated were optimized for predicting our targeted learning objectives.

4.2 Stealth Assessment Goal Set-Up

According to the IDEFA framework, the initial step in implementing a stealth assessment is to establish a clear goal or research question. This involves defining the specific learning outcomes that the stealth assessment aims to measure. In this study, we chose to use the sum of three post-assessment scores to quantify students' water science knowledge in MHS Unit 3. Below, we first justify the use of the summed scores both conceptually and statistically.

First, the assessment designer confirmed that these three items evaluate distinct but interconnected aspects of the water flow dynamics topic in Unit 3. To statistically validate their unidimensionality, we conducted both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) (Widaman & Helm, 2023). The EFA results indicated that the first factor had an eigenvalue of 2.5, explaining 83% of the total variance, while the second factor had an eigenvalue of 0.4, suggesting that a single factor captures the construct sufficiently. The factor loadings for each item were 0.73, 0.85, and 0.71, further supporting the unidimensional structure. CFA demonstrated a good model fit, with $\chi^2(2) = 4.23$ ($p = 0.12$), $RMSEA = 0.06$, and $CFI = 0.93$. The Cronbach's alpha (Tavakol & Dennick, 2011) of 0.79 also exceeded the acceptable threshold of 0.7, confirming the reliability of aggregating the scores. More details on the assessment items and analysis can be found in online Appendix E (https://bit.ly/3QvTeBP).

Given the statistical confirmation of unidimensionality, we chose to sum the three item scores for several reasons: (1) **Enhanced privacy:** Using the sum of assessment scores as the dependent variable reduces data granularity. Given that participants input their real names into our logging system, individual item scores could reveal specific strengths and weaknesses, which might be sensitive information. (2) **Comprehensive measurement:** Each item assesses a different yet interconnected aspect of water flow dynamics. By summing the scores, we create a broader measure of the student's overall understanding of these interrelated concepts. The factor analysis results further validate that the three items contribute to a unified measurement, supporting our decision to combine them. (3) **Increased statistical power:** Summing the three items increases the variability in the data, which enhances the statistical power of our analysis. A greater range of scores allows for more robust detection of significant relationships or differences in student learning. In contrast, analyzing only one item would limit the variability and reduce our ability to uncover meaningful patterns. (4) **Improved reliability:** Summing the scores across multiple

items improves reliability in two ways: by reducing item-specific bias and by balancing random errors. A single item may disproportionately reflect one aspect of the construct or vary in difficulty. Summing the items helps to mitigate this issue, ensuring a more balanced assessment of student knowledge. Additionally, error averaging ensures that random errors, such as misunderstanding a question or a momentary lapse in concentration, do not overly influence the final score. Summing scores tends to cancel out these random errors, leading to a more stable and reliable estimate of overall performance.

To determine whether students demonstrated significant learning outcome improvements after playing MHS, we conducted a paired t-test (Student, 1908) on the summation of pre- and post-assessment scores relevant to water science knowledge in MHS Unit 3. The results indicated significant score improvement ($M = 1.447$, $SD = 0.971$ for pre-assessment and $M = 2.009, SD = 0.961$ for post-assessment), with a t-value ($df = 353$) of -9.094 and a p-value of 0.0. Following this analysis, we categorized student performance into two groups: high performers, defined as students with post-test scores greater than or equal to 2, and low performers, defined as those with scores lower than 2.

4.3 Feature Engineering and Selection

4.3.1 Base Feature Aggregation

According to the IDEFA framework, the initial step in the feature generation process is base feature aggregation. This step involves identifying variables or features from the base event-stream data, or raw log data, that are significant for analysis and then aggregating values for each feature within a defined game walkthrough timeframe or window.

A robust feature set should comprehensively cover the investigated data, with each feature strongly correlated to the target variable, and should minimize overlap or high correlation between features to enhance model interpretability (Owen & Baker, 2020). Following these principles, we established three high-level feature categories: gameplay behaviours, embedded assessment scores, and external information. These categories were determined through a full MHS team consensus, reflecting various aspects of students' information. Detailed explanations of each category are as follows: (1) **Gameplay behaviours:** This category describes students' trajectories, actions, and choices in the gaming procedure. (2) **Embedded assessment scores:** This category contains in-game achievement markers representing signs of learning progress about subject-matter knowledge. (3) **External information:** This category includes features collected outside the game, such as pre- and post-assessment scores, demographic information, and sensor streams (e.g., eye-tracker, emotion detector, motion sensors).

Following the initial aggregation, we conducted an intermediate categorization, further subcategorizing each feature category. For gameplay behaviours, inspired by our previous study (Lu et al., 2023), we identified nine subcategories with high predictive power: (1) the size of the explored game area; (2) the speed of task completion; (3) tool-using status; (4) in-game item interactions; (5) argumentation construction; (6) event type shares; (7) dialogue reading behaviours; (8) game replay times; and (9) other information (e.g., which instructor guided the student).

For embedded assessment scores, we created two subcategories according to whether the score measures learning outcomes in the current unit or the previous units: (1) previous embedded assessment scores, which measure the learning outcomes of subject-matter knowledge marked by game logs collected from previous units (Units 1 and 2 in this study), and (2) current embedded assessment scores, which have the same function as described above but are collected in the current unit (Unit 3 in this study).

We divided external information into three subcategories based on our prior work demonstrating predictive validity and the aspects of students' information collected through external sources: (1) demographic information, (2) assessment scores, and (3) other external information. In this study, we primarily focused on assessment scores within the external information category, using students' pre- and post-assessment scores related to water science knowledge in Unit 3 of MHS.

4.3.2 Feature Engineering

Following the base feature aggregation process, we proceeded with feature engineering using mathematical operators. This step involved iterative feature engineering and selection procedures. Within each subcategory, we engaged in several brainstorming sessions with team members to create new features based on those identified in the previous step. We utilized mathematical functions (e.g., summation, ratio, multiplication, descriptive statistics) and data transformation techniques (e.g., discretization, one-hot encoding, and scaling methods).

This iterative process involved extensive collaboration among team members, during which we brainstormed potential new features based on the identified base features. Each newly engineered feature was evaluated for its relevance and contribution to understanding students' learning behaviours. The resulting features provided deeper insights into students' decision-making processes, engagement levels, and learning trajectories within the game.

4.3.3 Iterative Analysis

We then conducted a systematic iterative feature selection procedure by examining each feature's variance and its correlation with the targeted learning outcome, as defined in Section 4.2, Stealth Assessment Goal Set-Up. Given that our dependent variable is categorical, with "high-performance" and "low-performance" categories, and our independent variables include both

numeric and categorical features, we employed the Mann–Whitney U test (Mann & Whitney, 1947) for numeric features and Fisher's exact test (Fisher, 1935) for categorical features to determine significant correlations with the targeted learning outcome. Features included in the final feature set met several criteria: they exhibited sufficient variance (interquartile range greater than 0.1), had a correlation value exceeding a predetermined validity threshold (p-value for the statistical test less than 0.1 along with corresponding effect size (Glass, 1966; Cramér, 1999) value larger than 0.1), and provided adequate interpretability for generating insights after model fitting. The final feature set includes 57 features.

Online Appendix A (https://bit.ly/3QvTeBP) provides detailed descriptions of each feature within the final feature set. Additionally, online Appendix C (https://bit.ly/3QvTeBP) provides more comprehensive descriptions of each feature subcategory and corresponding exploratory analyses.

4.4 Model Selection

Our final feature set, developed through the aforementioned processes, exhibits the following characteristics: (1) **High-dimensional and diverse:** The set comprises a total of 57 features encompassing a wide variety of types, including numeric features (e.g., object interaction frequency, tool usage frequency, choice node hovering frequency) and categorical features (e.g., teacher IDs, explored area size, argumentation performance). (2) **Non-linear relationships:** Many relationships between our features and learning outcomes are likely non-linear. For instance, tool usage frequencies or event-type shares may have complex, non-linear effects on post-test learning outcomes. (3) **Class imbalance:** The dataset exhibits class imbalance (e.g., a greater number of high-performing students than low-performing students), which needs to be addressed to avoid biased predictions. (4) **Noise and outliers:** The feature set includes noisy data and potential outliers, such as rare in-game item interactions or atypical explored area sizes that are not representative of most students. (5) **Sparse and low-frequency features:** Some features, such as dialogue-triggering frequencies or specific object interaction behaviours (e.g., sensor usage frequency, crate delivery success), may be sparse or occur infrequently. (6) **Complex feature interactions:** The feature set likely contains complex interactions between features, such as how dialogue reading behaviours and argumentation construction speed combine to predict learning outcomes.

Based on these characteristics, we selected the following algorithms for predicting the desired learning outcomes: Bayesian generalized linear model (BGL) (Albert, 1988), distance weighted discrimination with polynomial kernel (DWD) (Marron et al., 2007), random forest (RF) (Breiman, 2001), support vector machines with class weights (SVM) (Huang & Du, 2005), and model-averaged neural network (NN) (Abrahart & See, 2000). Although these algorithms employ different approaches, they are all capable of effectively handling feature sets with the characteristics mentioned above. Each algorithm not only shares capabilities with the others but also offers unique advantages, which justify their inclusion in this study. Utilizing multiple algorithms can generate more validated prediction results, and the findings from each can provide supplementary support for the others. Key unique advantages of each selected algorithm that contribute to the construction of our prediction model are briefly described as follows.

The **DWD** algorithm establishes decision boundaries that avoid overemphasizing outliers or disproportionately favouring the majority class, enhancing its ability to predict outcomes for underrepresented low-performing students. Additionally, DWD's polynomial kernel effectively captures complex interactions between features derived from student actions (e.g., tool usage, task completion speed, argumentation construction), enabling a robust distinction between high and low performers. By employing class weights, **SVM** addresses class imbalance by assigning greater importance to the minority class (low performers), ensuring that its influence is not overshadowed. This approach refines the decision boundary and improves prediction accuracy for underrepresented groups. Furthermore, SVM's nonparametric nature makes it resilient to violations of distribution assumptions and suitable for smaller sample sizes, which is common in educational datasets. The model-averaging technique in **NNs** reduces the risk of overfitting, particularly in a feature set combining categorical and continuous data. By integrating predictions from multiple NNs, the model enhances generalization across diverse learning patterns. This approach is particularly effective when certain features, such as gameplay strategies or pre-test performance, risk skewing predictions. The averaging mechanism ensures balanced predictions and applicability to various student profiles.

The Bayesian framework of **BGL** is particularly effective at managing uncertainty inherent in features such as interaction frequencies or completion times. It also allows the integration of prior knowledge through the use of informative priors. For example, if prior research or domain expertise suggests that features representing pre-knowledge and performances from earlier game tasks (e.g., pre-assessment scores, embedded scores from previous units) are related to the target learning outcome, this information can be incorporated into the model. A key strength of BGL is its ability to generate interpretable probabilistic predictions, providing an understanding not only of the influence of various features but also of the model's confidence in its predictions. This interpretability is critical in educational settings, where actionable insights based on model predictions are essential. **RFs** are particularly well suited for handling datasets with mixed feature types, minimizing the need for extensive pre-processing. By constructing multiple decision trees based on random subsets of features, RF reduces overfitting and manages high-dimensional data effectively. Its ability to handle class imbalance by averaging results across many trees improves representation of minority classes, and its robustness to missing data ensures consistent performance even with incomplete

game logs.

To generate the necessary features from raw logs, conduct the analytical process, and construct the machine learning classifiers described above, we utilized R and several R packages. Notably, we employed the "Caret" package, developed by Max Kuhn (Kuhn, 2019), which provides a comprehensive suite of functions for creating predictive models.

4.5 ECD Framework Mapping

According to the descriptions regarding each component of the ECD framework in Section 2, Literature Review, we mapped our study to the framework shown in Table 1.

Table 1. Defining specific content for each component with ECD framework mapping to our study.

ECD Component	Specific Content Mapped to Our Study
Competency model	We used the summation of three post-test assessment scores measuring students' water science knowledge in MHS Unit 3 and categorized the summation score into high and low performers based on whether the score is smaller than 2 or not.
Task model	For this component, we included 12 quests until the end of MHS Unit 3. Each quest's detailed description can be found in online Appendix B (https://bit.ly/3QvTeBP).
Evidence model	In this model, we involved a final feature set containing a total of 57 features as the learning evidence and five machine learning models to help us solve our learning outcome prediction problem.

4.6 Model Training Preprocessing

As described in Section 4.2, Stealth Assessment Goal Set-Up, our target or dependent variable consists of two classes: high performers and low performers. Due to the criteria used to categorize these classes, the variable exhibits an imbalanced class distribution, with 251 high and 103 low performers. This imbalance could significantly hinder model performance (Guo et al., 2008; Elrahman & Abraham, 2013; Buda et al., 2018). To address this issue, we employed subsampling techniques (Kaur et al., 2019), such as down-sampling, up-sampling, synthetic minority over-sampling technique (SMOTE), and random over-sampling examples (ROSE). Each method was applied to the training dataset and evaluated using the testing dataset, with only the most effective methods being reported. Specifically, SMOTE was adopted for the BGL, SVM, and NN models, while up-sampling was selected for the RF and DWD models. These techniques were chosen for their ability to enhance the proportion of the minority class, thereby improving the model's predictive accuracy for the low performers. Detailed descriptions of the subsampling applications are provided in online Appendix D (https://bit.ly/3QvTeBP).

4.7 Model Inference

In an ideal scenario, stealth assessment would reliably indicate students' learning outcomes throughout game phases, enabling instructors to intervene promptly. However, challenges and gaps exist that prevent stealth assessments from functioning as expected. A promising initial step toward bridging these gaps is identifying potential indicators that reflect students' behaviour patterns at different learning outcome levels. Model inference is a valuable tool for identifying these indicators, as it helps elucidate the impact of each feature on learning outcomes. However, many machine learning models are "black-box" models, prioritizing prediction accuracy over interpretability due to numerous parameters and nonlinear transformations, which tend to obscure the relationship between the engineered features and the intended learning outcomes within the game design (Min et al., 2020). Fortunately, several methods help interpret black-box models, such as feature importance rates (Breiman, 2001), partial dependence plots (PDPs) (Molnar et al., 2020), Shapley additive explanations (SHAPs) (Lundberg & Lee, 2017), and local interpretable model-agnostic explanations (LIMEs) (Ribeiro et al., 2016).

After investigation, we decided to use BGL as a surrogate model to interpret the RF model, which achieved the highest test accuracy rate based on our dataset. The surrogate model approach involves training a simpler, interpretable model (such as a linear model or decision tree) to approximate the predictions of a more complex, black-box model, thereby providing insights into the black-box model's decision-making process (Guidotti et al., 2018; Molnar et al., 2020). Specifically, we first used the trained RF model to generate predictions on the training dataset. Then, we constructed a new dataset where the features remained the same as the original dataset, but the target variable was replaced with the predictions from the RF model. The final step was to train a BGL model on this new surrogate dataset to learn the mapping between the features and the RF predictions.

In addition to BGL's demonstrated effectiveness in predicting learning outcomes with our feature set (as shown in Table 2 and corresponding descriptions), several key reasons informed our choice to interpret the black-box model using BGL: (1) **Probabilistic interpretations:** BGL provides a probabilistic framework, offering insights into the uncertainty and variability of predictions, which is valuable for understanding confidence intervals and the reliability of model outputs. Methods like feature importance and PDPs typically offer point estimates without quantifying uncertainty. (2) **Prior information incorporation:**

BGL allows the inclusion of prior information about parameters, guiding the model to plausible solutions, especially when data is scarce or noisy. Other methods, like SHAP and LIME, do not directly incorporate prior knowledge, potentially resulting in less robust interpretations. (3) **Regularization and multicollinearity management:** BGL incorporates regularization through priors, helping manage multicollinearity and overfitting. Feature importance and PDPs do not address multicollinearity, which can skew interpretations. (4) **Simplicity and communication:** The linear structure of BGL makes it easy to interpret and communicate results to non-technical stakeholders. Coefficients have a direct and intuitive interpretation, unlike SHAP values and LIME explanations, which can be complex. Feature importance scores are easy to understand but lack detail on feature interactions. (5) **Inference and predictive distribution:** BGL provides not just point estimates but entire predictive distributions, crucial for understanding the full range of potential outcomes and their probabilities. Other methods typically focus on individual predictions or feature contributions without offering a comprehensive view of predictive distributions.

5. Results

Having constructed the framework-supported pipeline to predict MHS learning outcomes, we evaluated its efficacy by reviewing the performance of the machine learning models.

5.1 Model Training and Results

To assess the predictive ability of student learning outcomes, we divided our 57-feature dataset into three groups: embedded assessment scores, in-game behaviours, and full features. The first group contains 10 features denoting in-game learning progress, the second contains 45 features representing individual in-game actions, and the third combines all previous features with additional information gathered outside the game, such as the summation of pre-test scores and instructor IDs.

Each model algorithm was trained using the three feature groups. Prior to training, features underwent Yeo–Johnson transformation (Yeo & Johnson, 2000), followed by centring and scaling. We partitioned the data, allocating 80% of the samples for training and 20% for testing. The training process involved 30-fold cross-validation, repeated 30 times, with hyperparameters tuned via greedy search across 30 randomly formed combinations. Model evaluation was conducted using accuracy, sensitivity, and specificity metrics derived from the confusion matrix to assess the models' ability to predict high and low performers' learning outcomes. Table 2 provides detailed information on model performance.

Table 2. Model prediction performance. The bolded numbers represent the highest values within each of the performance measurement metrics—accuracy, sensitivity, and specificity—within each of the three different feature sets.

	Embedded Assessment Scores			In-Game Behaviours			All Features		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
BGL	0.64	0.86	0.1	0.69	0.84	**0.3**	0.84	0.88	**0.75**
DWD	0.71	**1**	0	0.7	**0.94**	0.1	0.81	0.84	**0.75**
RF	0.61	0.78	0.2	0.71	0.9	0.25	**0.86**	0.96	0.6
SVM	**0.76**	0.9	**0.4**	**0.73**	0.9	**0.3**	0.83	**1.0**	0.4
NN	**0.76**	0.9	**0.4**	0.63	0.78	0.25	0.81	0.86	0.7

As indicated in Table 2, our first finding is that all five models, when utilizing the dataset with all features, achieve accuracy metrics exceeding the 80% benchmark, which is considered a good prediction accuracy rate within educational contexts (Bird et al., 2021).

For each feature set, the results reveal that the SVM and NN algorithms offer the best test accuracy and specificity rates, while the DWD algorithm provides the best sensitivity rate for the feature set of embedded assessment scores. When using the feature set of in-game behaviours, SVM produces the best test accuracy and specificity rates, DWD provides the best sensitivity rate, and BGL matches the specificity rate of SVM. With the full feature set, RF leads in test accuracy, both BGL and DWD achieve the highest specificity rate, and SVM offers the highest sensitivity rate.

Table 2 compares the three different feature sets and shows that the accuracy rate using all features is significantly higher than that using the embedded assessment scores and in-game behaviours feature sets. Additionally, the specificity rate using all features is significantly improved compared to the embedded assessment scores and in-game behaviours feature sets. To validate these differences, we conducted the Kruskal–Wallis (KW) test (Kruskal & Wallis, 1952) along with Dunn's (D) test (Dunn, 1964) for post hoc pairwise comparisons.

Regarding the accuracy rate, the KW test yielded a significant difference among feature sets ($\chi^2(2) = 9.51$, p-value = 0.01), and the D test indicated that the accuracy using all features was significantly higher than that using embedded assessment scores (p-value = 0.018) and in-game behaviours (p-value = 0.008). For the specificity rate, the KW test also revealed significant differences ($\chi^2(2) = 8.77$, p-value = 0.01), and the D test showed that specificity using all features was significantly higher than that using embedded assessment scores (p-value = 0.014) and in-game behaviours (p-value = 0.017). The differences in

sensitivity rates across the three feature sets were not statistically significant, as confirmed by the significance tests mentioned above.

From the analysis, it is evident that the feature set with all features provides better prediction outcomes. However, specificity rates are generally lower than accuracy and sensitivity rates, indicating a weaker model capacity for identifying low performers. Overall, RF yields the highest accuracy rate with the full feature set. Notably, BGL, with all features, offers the most balanced model performance, with all measures surpassing the threshold of 75%. The effectiveness of lower-complexity models like BGL, as noted by Zheng and Casari (2018), underscores the effectiveness of our feature engineering and model selection processes.

In summary, the results validate the stealth assessment framework for MHS and underscore the value of combining multiple evidence sources (e.g., in-game behaviours and embedded assessment scores) to improve predictive accuracy. The approach integrates the **IDEFA** framework for feature generation, machine learning models tailored to the feature set, and the **ECD** framework to ensure the relevance of students' in-game activities for predicting learning outcomes. These findings suggest the potential for generalizing this framework to other digital learning environments. However, challenges remain, such as lower specificity in detecting low performers and insufficient representation of subtle learning behaviours. Further iterative refinements, guided by the **iterative design paradigm** within the ECD framework, are needed to enhance model accuracy and generalizability, as discussed in Section 6.1.

5.2 Model Inference Results

Using BGL for model inference provides insight into the construction and effect of each feature on the target variable. This analysis yields each feature's estimated coefficient and highest-density interval (HDI), revealing the nature and significance of each feature's influence on learning outcomes. Detailed inference results are in Table 3. We summarized key findings as follows:

- **Map exploration:** Larger map exploration sizes are associated with lower probabilities of high-level learning outcomes, especially in areas with major quests. This finding suggests that excessive exploration may lead to distractions or inefficient use of time, detracting from focused learning efforts.
- **Task completion time:** Prolonged task completion times correlate positively with high-level learning outcomes. This implies that students who spend more time on tasks tend to engage more deeply and comprehensively, leading to better learning outcomes.
- **Tool usage:** Frequent use of topographic maps, quest descriptions, background information, and game help, coupled with slower average chat log use, is associated with high-level learning outcomes. These tools facilitate students' understanding and problem-solving within the game. In contrast, frequent use of chat logs and side-quest-related information correlates with lower learning outcomes, suggesting that these tools may distract students from their major learning objectives.
- **Interaction with in-game items:** Correct placement of pollution sensors enhances learning outcomes, while redundant sensor usage or excessive crate delivery negatively impacts learning outcomes. This suggests that interactions with items that assist in completing major quests lead to better learning outcomes, whereas redundant interactions with unrelated items result in worse outcomes.
- **Argumentation performance:** Higher argumentation performance in Unit 3 and frequent hovering over choice nodes in the argumentation system suggest high-level learning outcomes. Detailed interactions within choice nodes, text reminders, and illustrations, as well as spending time to carefully read and digest information, help improve learning outcomes.

Table 3. The model inference outcome, produced by BGL, provides several insights. Each feature falls into a category listed in the "Category" column. The "Feature Name" column lists the features, and the "Coefficient" column presents the median value of each estimated coefficient. In the fourth column, the HDI functions similarly to a confidence interval and serves as a primary index for determining a feature's significance. We deem a feature significant if its HDI doesn't encompass zero. Significant features are marked in two colours: orange indicates a positive correlation with students' high-level learning outcomes, while blue denotes a negative correlation.

Category	Feature Name	Coefficient	HDI	Significance
Map exploration	mainMapSize	−0.6	[−1.7, −0.09]	Yes
	dungeonMapSize	−0.4	[−0.9, 0.04]	No
Task completion	taskAveSpeed	0.6	[0.14, 0.98]	Yes
Tool using	mapFreq	0.3	[−0.4, 1.04]	No
	chatFreq	−0.9	[−2.03, 0.18]	No
	crashFreq	−0.6	[−1.36, 0.2]	No
	questFreq	0.6	[0.12, 1.07]	Yes
	backingFreq	0.4	[−0.32, 1.19]	No
	helpFreq	0.5	[0.11, 0.97]	Yes
	mapSpeed	−0.2	[−0.61, 0.15]	No
	chatSpeed	1	[−0.07, 2.14]	No
	crashSpeed	−0.2	[−0.98, 0.53]	No
	questSpeed	−0.1	[−0.56, 0.35]	No
	backingSpeed	−0.1	[−0.87, 0.6]	No
	helpSpeed	-0.5	[−0.95, −0.14]	Yes
Item interaction	pollutedSensor	1.3	[0.33, 2.22]	Yes
	downStreamSensor	0.5	[−1.54, 2.53]	No
	sameAreaSensor	−2.4	[−3.69, −1.08]	Yes
	findAreaSensor	−0.7	[−1.64, 0.36]	No
	cleanSensor	0.3	[−0.26, 0.81]	No
	failCrate	−0.7	[−1.35, 0]	No
	successCrate	−2	[−3.3, −0.75]	Yes
Argumentation	U3argumentLevel	0.1	[−0.16, 0.4]	No
	U2argumentLevel	0	[−0.24, 0.22]	No
	nodeHoverFreq	1	[0.5, 1.58]	Yes
	claimIISpeed	0.3	[0.02, 0.66]	Yes
	evidenceBSpeed	0.5	[0.18, 0.85]	Yes
	reason3Speed	0.1	[−0.24, 0.44]	No
	reason4Speed	−0.6	[−0.95, −0.19]	Yes
	claimISpeed	−0.3	[−0.65, 0.06]	No
	reason5Speed	0.1	[−0.21, 0.44]	No
	evidenceASpeed	−0.1	[−0.4, 0.28]	No
	reason2Speed	−0.1	[−0.52, 0.21]	No
	reason1Speed	−0.5	[−0.85, −0.04]	Yes
Event shares	itemTrigger	-0.7	[−1.61, 0.25]	No
	movement	−1.3	[−3.16, 0.4]	No
	missionProgress	0.2	[−0.51, 0.87]	No
	dialogue	−0.7	[−2.19, 0.82]	No
	toolUsing	−0.7	[−1.98, 0.44]	No
	hotkey	1.3	[0.19, 2.53]	Yes
	argument	−1	[−1.51, −0.42]	Yes
	jump	−0.4	[−0.77, −0.02]	Yes
	toggleBoard	0	[−0.33, 0.48]	No
Dialogue	dialogueAvgSpeed	0	[−0.38, 0.35]	No
Replay time	trial	1.1	[0.24, 2.14]	Yes
	tutorial	0	[−0.29, 0.33]	No
	biggerWatershed	0	[−0.44, 0.44]	No

continued on next page...

. . . continued from previous page

Category	Feature Name	Coefficient	HDI	Significance
Embedded score	upstreamArgument	-0.1	$[-0.53, 0.26]$	No
	CREi	1.2	$[0.69, 1.8]$	Yes
	JasperCritique	0.6	$[0.1, 1.16]$	Yes
	findTeam	1.3	$[0.65, 1.9]$	Yes
	gardenPlant	-0.4	$[-0.99, 0.1]$	No
	crateDelivery	0.4	$[-1.21, 1.96]$	No
External information	preTestLowLevel	-2.4	$[-3.18, -1.7]$	Yes
	teacherID	0	$[0, 0.01]$	No

- **Game events:** Certain game events, such as hotkey usage and quest-/task-related actions, positively correlate with learning outcomes. Conversely, other event shares negatively impact outcomes, except for toggle board usage. This implies that students concentrating on completing major quests and knowing how to apply appropriate in-game tools perform better in corresponding learning outcomes.
- **Dialogue reading speed:** Average dialogue reading speed appears to have little impact on learning outcomes.
- **Replay frequency:** Our findings indicate that students who replayed the game more frequently exhibited higher learning outcomes, suggesting that repeated exposure or practice can reinforce mastery of the content.
- **Embedded assessment scores:** High embedded assessment scores in the CREi system, which tasks students with identifying the correct argumentation component (claim, reason, evidence) by directing a soccer ball appropriately, critical dialogue with Jasper (an in-game non-player character), team finding, and crate delivery tasks indicate high-level learning outcomes. In contrast, scores in argumentation in Unit 3 and garden planting slightly decreased learning outcomes. Other scores have little impact on learning outcomes.
- **External sources:** A student's pre-test performance positively correlates with their post-test performance, implying that pre-existing knowledge plays a vital role in learning progression. However, instructor guidance seems to have little effect on students' learning outcomes in this study, suggesting that the game's design and individual student engagement are more critical factors.

Each category, except for the dialogue feature category, has features significantly correlating with learning outcomes. Potential reasons why the feature representing students' dialogue reading behaviours shows a significant contribution individually but decreases in significance when combined with other features include the following: (1) **Feature interaction:** The feature's effect might depend on the presence of other features, with interaction effects either enhancing or diminishing its importance when combined. (2) **Overfitting issue:** The feature might overfit the target variable when considered alone but fail to generalize well when other features are included. (3) **Redundant information:** The feature might provide information already captured by a combination of other features, rendering its individual contribution insignificant in the combined model.

6. Discussion

Stealth assessment in serious games enables the unobtrusive collection of data to monitor learning progress and evaluate complex competencies. This study introduces a granular, framework-based pipeline for implementing stealth assessments in a GBL environment, structured around the ECD framework, feature generation, computational model selection, and performance evaluation. A key novelty is the development and validation of this approach to effectively predict learning outcomes by capturing in-game behaviour patterns. Additionally, we integrate two established frameworks—ATMSG and xAPI—for logging system design, addressing gaps in the empirical application of these frameworks for data logging in educational games. Our approach also incorporates the IDEFA framework for feature generation, advancing current methodologies that require further empirical validation.

This study makes a significant contribution by incorporating a wide range of in-game behaviours into the dataset for model construction, addressing the gap in empirical research on combining in-game features to analyze learning patterns. While there is room for further expansion, our approach includes a broader feature set than many previous studies, advancing the application of learning analytics in GBL. Furthermore, we introduce a surrogate white-box model to interpret the black-box computational models used in stealth assessments. This blend of interpretability and predictive power offers a novel solution for identifying distinct behavioural patterns between high- and low-performing students, a method not extensively validated in educational games. Finally, this research addresses key gaps by providing generalizable guidelines for implementing stealth assessments across diverse educational contexts, contributing to the broader body of research.

6.1 Further Thoughts Regarding RQ1 Results: Potential Refinements Following ECD's Iterative Design Paradigm

As discussed in Section 4, the ECD framework was instrumental in structuring our stealth assessment pipeline for MHS, providing a systematic approach for translating in-game behaviours into insights about student learning. One of ECD's core strengths is its iterative design paradigm (L. Wang et al., 2015; Ke & Shute, 2015; V. J. Shute et al., 2016; Grover et al., 2017), which we will adopt in future iterations to continually refine the pipeline. This iterative process will enable us to revisit and enhance each stage—feature aggregation, feature engineering, model selection, and task design—based on empirical findings. By doing so, we expect to improve the pipeline's predictive accuracy, model interpretability, and generalizability across different educational contexts.

6.1.1 Iterative Adjustments in Feature Generation

Initial analysis and expert input revealed several areas for improvement in our feature set, particularly in managing intercorrelated features, addressing prediction specificity for low-performing students, and enhancing generalizability across game contexts.

First, during feature aggregation, we opted to delete student records with over 30% missing data. However, this approach may have discarded critical information, especially for low-performing students, whose gameplay logs often had more missing data due to early disengagement. Future refinements should explore more sophisticated imputation techniques to retain these samples and better capture learning challenges faced by lower performers (Keerin, 2021).

Second, while our feature selection process relied on manual curation based on expert judgment, we may have overlooked potentially valuable data. For instance, we only included interactions with key pedagogical objects, but non-pedagogical interactions could also provide insight into learning behaviours. Future iterations could involve automatic feature selection methods, applying advanced techniques to evaluate all potential features systematically. This approach would save time, retain relevant information, and improve the pipeline's generalizability (Jalota & Agrawal, 2021).

Additionally, certain features, such as the size of the explored game area and dialogue events, require further refinement. For exploration behaviours, future adjustments could capture both breadth and depth, such as repeated visits to the same areas, to better predict performance across different student groups. Dialogue event features could be further differentiated by categorizing dialogues related to gameplay progression versus those related to instructional content, ensuring that subtle learning behaviours are better represented in the model.

In terms of feature engineering, future refinements could address correlated features within subcategories, which may introduce noise and hinder model performance. Techniques like independent component analysis, principal component analysis, or autoencoders could be employed to reduce redundancy and improve the model's predictive power (Ray et al., 2021). Moreover, combining features from different subcategories through feature extraction methods could help capture inter-categorical relationships, providing richer insights and enhancing model performance.

6.1.2 Computational Model Refinement

Although the current study demonstrated promising results, several areas could benefit from further iteration to enhance model performance and improve generalizability across different game contexts.

One key area for refinement is the generalizability of the model selection process to other GBL environments. While models were carefully selected based on the final feature set's characteristics, their performance may be limited to the specific context of this study. Additionally, the model selection process did not involve a comprehensive search for optimal algorithms aligned with the feature set. Moreover, the current study relied heavily on RF without leveraging the combined strengths of multiple models. To address these issues, future iterations could explore ensemble learning (EL) techniques. EL allows for the flexible combination of different base models, with the ability to add or remove models depending on performance. Various voting schemes, such as hard voting (using the best-performing model's result) and soft voting (weighted combination of model outputs), provide flexibility in determining the final result. Implementing EL could enhance model robustness and improve generalizability across various game contexts within MHS and other GBL environments (Siddique et al., 2021).

Another area for refinement is model performance, particularly in improving specificity for identifying low-performing students. While RF achieved high overall accuracy, it struggled to capture the nuanced behaviours of lower performers. Future work could involve testing simpler models, such as BGL, through extensive hyperparameter tuning. BGL's probabilistic framework offers greater interpretability and may provide insights into low-performing students' learning behaviours, which complex models might overlook. This refinement would prioritize not only accuracy but also the ability to generate actionable insights for instructors to support struggling students.

Finally, addressing dataset imbalance remains a critical area for future iterations. Although the current study employed several resampling techniques, future studies could incorporate more advanced methods such as the adaptive synthetic sampling approach (ADASYN), borderline-SMOTE, and cluster-SMOTE (Wongvorachan et al., 2023). Paired with refined feature selection methods like recursive feature elimination with cross-validation (RFECV), the Boruta algorithm, and mutual

information-based feature selection, these approaches would help ensure that models capture learning patterns across the full spectrum of student performance, particularly in underrepresented groups (Dhal & Azad, 2022).

6.1.3 Task Model Adjustment

In addition to feature and model refinements, the task model, which involves defining in-game tasks to generate evidence for learning prediction models, can also benefit from the iterative design paradigm of ECD.

One key area for refinement is the involvement of in-game activities that elicit clearer evidence of student learning behaviours, particularly in distinguishing different learning levels. For example, tasks such as "arguing which watershed is bigger" and "convincing Bill the pollutant is nearby" (see Table 16 in Appendix B (https://bit.ly/3QvTeBP)) already integrate argumentation construction, a critical learning objective. However, future iterations could introduce more detailed in-game activities that encourage varied levels of reasoning and decision-making. For instance, argumentation tasks could be broken down into multiple attempts (e.g., an attempt is from starting to construct an argument to submitting a final response). By recording which specific components were selected during argument construction—especially in cases where students submit incorrect arguments—we could identify which aspects of the argumentation process pose challenges for students. This more granular breakdown would help us understand where students struggle, providing stronger evidence of their learning progress and identifying the specific skills they need to improve.

Additionally, task refinement could involve segmenting broader tasks (e.g., transporting supplies or tracing a pollutant source) into smaller, more specific actions to enable more detailed tracking of student progress. For example, breaking the task of tracing a pollutant into distinct phases—such as identifying clean versus polluted river sections or applying sensors at different water points—could offer deeper insights into how students apply their knowledge of water flow dynamics. These adjustments would generate clearer, more actionable evidence that can be used to enhance predictive models.

6.1.4 More Granular Competency Model

In future iterations of our stealth assessment framework, we plan to adopt a more granular competency model that evaluates learning outcomes at the item level, while ensuring appropriate protection of student privacy. Although this study used the summation of three assessment items to measure overall knowledge of water flow dynamics, we recognize the limitations in capturing item-specific insights. Future studies will focus on item-level analyses to gain a deeper understanding of student competencies across different aspects of the topic. This approach will help identify specific strengths and weaknesses, allowing for more targeted instructional interventions. By refining the competency model to reflect distinct knowledge areas, we aim to improve the interpretability of assessment results, providing educators with more precise and actionable insights.

Ultimately, future iterations will aim to refine the task design (task model in ECD), the evidence generation process (evidence model in ECD, including features and computational models), and the competency measurement (competency model in ECD). By aligning in-game activities more closely with learning objectives and collecting more precise evidence, we can improve the accuracy of predictions related to student performance. Specifically, adopting a more granular competency model will allow us to conduct item-level analyses, providing deeper insights into individual student competencies and enhancing the precision of our learning assessments. This iterative refinement will ensure that tasks in MHS not only engage students but also yield meaningful insights into their learning trajectories. The continuous optimization of these models, under the ECD's iterative design paradigm, will further enhance the robustness and generalizability of our stealth assessment framework across various educational contexts.

6.2 Implications for RQ2's Results: Insights from Interpreting Black-Box Models

In addressing **RQ2**, we discuss potential methods for interpreting black-box models and explain why we chose the surrogate model method to infer black-box models, specifically using BLG to interpret the RF model. Our inference method offers a viable solution for understanding how individual features impact the targeted learning outcome while achieving highly accurate predictions from black-box models. By consolidating inference results, we identified key features that signal students' learning outcomes for MHS, completing the stealth assessment tool. A thorough examination of the inference results led us to conclude that nearly every feature category has features significantly impacting students' learning outcomes, except the dialogue category. Specifically, the findings suggest several deductions, as described in the following paragraphs.

Previous research has demonstrated that patterns of student engagement with in-game tools correlate with their level of expertise (Kang et al., 2017). Our analysis reveals that high-performing students exhibit greater efficiency and purpose when using in-game tools. Specifically, they engage with these tools primarily to verify preconceived ideas rather than to generate new ones. Furthermore, high performers display a superior ability to maintain sustained attention compared to their low-performing peers when searching for appropriate tools to complete quests. Additionally, high performers frequently and quickly use tools that do not contain lengthy texts, indicating a stronger motivation or enhanced capability to gather information from multiple sources for quest completion.

Insights into students' problem-solving strategies and knowledge absorption can be gleaned from their interactions with in-game items (V. J. Shute et al., 2016). We observed that high-performing students interacted with relevant items less frequently

but with a clear purpose. In contrast, low-performing students interacted with items more frequently and without a clear purpose. This suggests that high performers are more likely to engage with items purposefully or after careful consideration than their low-performing peers. This pattern is evident in how these groups interacted with crates that needed to be delivered to one of two rivers. High performers adjusted their actions based on feedback from previous delivery outcomes, whereas low performers tended to ignore feedback and did not alter their actions.

Event share features illuminate how students distribute their game time, indicating that students concentrating on completing major quests and knowing how to apply appropriate in-game tools perform better in corresponding learning outcomes. It may also suggest that high performers exhibit a keen understanding of game mechanics and maintain focus on their objectives. In contrast, low performers are more easily distracted and struggle to filter relevant information from their surroundings.

Regarding the argumentation system, we observed that high performers tended to read information at different choice nodes more carefully and compared them more frequently, indicating a deeper understanding of the curriculum. In contrast, low performers were more efficient at filtering incorrect nodes and identifying correct ones. Notably, argumentation performances in both Unit 2 and Unit 3 showed no significant correlation with our targeted learning outcomes, contrary to our expectations. This discrepancy may be attributed to inadequate feature generation, as we only included logs showing the frequency of correct and incorrect answer submissions. We did not include logs representing student interactions with assistant tools or behaviour sequences within the argumentation system. Additionally, inappropriate game mechanics design within the argumentation system might have contributed to this issue. First, the system allows unlimited attempts to find the correct answer, which can diminish the incentive for deeper understanding and engagement. Second, it lacks explicit mechanisms to help students connect the evidence displayed in the system with the evidence they gathered during gameplay, thereby weakening the overall coherence of the argumentation process. Furthermore, the content embedded within the constructed argumentation has limited association with the content knowledge that MHS Unit 3 aims to teach.

Embedded assessment scores showed mixed correlations with targeted learning outcomes. While some scores demonstrated a positive association, one score exhibited a negative correlation, contrary to our expectations. This discrepancy may be due to unclear quest instructions or immature game graphics and mechanics. Additionally, some embedded scores showed minimal correlation with learning outcomes, indicating the need for further feature engineering to find better measurements for these scores. Notably, students who replayed the game multiple times outperformed their peers, implying that a mechanism allowing students to replay similar tasks as practice is crucial for educational game design and development. This finding also suggests that incorporating a stealth assessment to track the frequency of repeated game content or mechanics is important for accurately measuring in-game performance and targeted learning outcomes. As for the dialogue category, we discuss potential reasons for the diminishing significance of this feature category that warrant further analysis and may emphasize the complex interplay between various gameplay elements and their collective impact on learning outcomes.

Our study underscores the multifaceted nature of learning in educational games and highlights the critical role of specific gameplay features in enhancing or impeding high-level learning outcomes. These insights have potential implications for the design of stealth assessments, pedagogical activities, and educational games, which will be elaborated on in the following section.

6.3 Implications for Stealth Assessment Design

Carvalho and colleagues (2015) suggested that their framework is particularly suitable for "expert usage," implying that stakeholders refining their games with ATMSG analytics benefit more from this framework. Conversely, those using the games without being willing to modify them may be better served by simpler frameworks handling less granular log data due to cognitive load limitations. Our study, however, indicates that research groups should integrate a granular logging system capable of collecting comprehensive, feature-rich data, especially during the early stages of research, regardless of their specific usage contexts. Similar opinions are echoed by Rowe and colleagues (2017) and Ke and colleagues (2019).

Even experienced experts often struggle to identify necessary game logs for constructing effective stealth assessments, requiring numerous rounds of adjustment and testing with granular logging systems to ensure that no key information is omitted (F. Chen et al., 2020). Despite extensive brainstorming and discussion in our study, we found our feature selections insufficient to fully exploit stealth assessment potential, which includes real-time formative feedback during gameplay (Min et al., 2020), adaptive learning experiences (V. J. Shute & Rahimi, 2021), and user-friendly dashboards for instructors to monitor students' gameplay in real time (F. Chen et al., 2020). Therefore, we recommend that future studies construct stealth assessments within complex GBL environments to ensure that embedded logging systems can collect extensive, detailed data reflecting students' actions and contextual information. Ideally, this system should evolve alongside the game (Ke et al., 2019).

Based on our study, we recommend that future research consider the following game logs to assess students' learning outcomes unobtrusively:

- **Tool utilization tracking:** Logs on how students use in-game tools can distinguish low and high performers. Detailed sequences of tool-usage actions provide insights into decision-making processes and problem-solving strategies (Kang

et al., 2017). Efficiency metrics, such as time taken to locate and utilize tools, indicate understanding of game mechanics, suggesting a correlation between high efficiency and greater expertise. Differentiating between tools used for verifying ideas versus generating new insights helps assess cognitive strategies and learning styles.

- **Instruction interaction tracking:** Logs recording how students interact with instructions, including time spent reading and the sequence of steps followed, can measure students' learning outcomes. These actions reflect how students absorb new information and their learning processes during gameplay.
- **Game world exploration tracking:** Detailed logs of exploration activities, including areas visited, time spent, and interactions, help identify patterns of distraction or focus loss. This data is crucial for understanding navigation and attention allocation within the game environment (Loh et al., 2016).
- **In-game item interaction tracking:** Capturing detailed information on interactions with in-game items, including time spent, frequency, and sequence, helps identify students who engage meaningfully with content versus those who struggle (Yang et al., 2021).
- **Decision-making points tracking:** Identifying key decision points and the sequences of choices made at these points, including time spent reading, provides insights into how decision-making skills correlate with learning outcomes, which is also supported by the study of Snow and colleagues (2015).
- **Feedback responsiveness measurement:** Logs tracking how students respond to in-game feedback, including changes in behaviour after receiving feedback, such as adjusting strategies or correcting mistakes, can differentiate high performers who use feedback effectively from low performers who may ignore it.
- **Comprehensive event share logging:** Detailed logs capturing how students distribute their time across various game events and activities, including the duration and frequency of interactions with different game elements, help identify students' learning processes and moments when they deviate from their objectives. This provides insights into their attention management.
- **Replay mechanics and practice:** Logs of replay activities capturing data on tasks replayed, frequency of replays, and outcomes of each attempt help measure students' learning progress during gameplay. This can identify valuable tasks for practice and how repeated practice affects learning outcomes.
- **Performance metrics through different game stages:** Learning objective–specific performance metrics, such as embedded scores, track students' learning outcomes in different game stages. Our findings indicate that combining in-game activities with embedded assessment scores or performance metrics leads to the highest accuracy. Therefore, we suggest that stealth assessments involve a comprehensive standard rubric to guide the creation of these performance metrics according to educational game design.

6.3.1 Key Suggestions to Designers and Practitioners Regarding Stealth Assessment Design and Development Process

Our process for constructing the stealth assessment emphasizes the importance of iterative design and testing in refining both the stealth assessment and the educational game. By integrating continuous improvement cycles, data-driven adjustments, and key user involvement, we aim to create a robust and effective learning environment. This environment should align game content with learning objectives, maintaining engagement and motivation without causing excessive distractions. We offer the following suggestions for future researchers conducting iterative design and testing for refining stealth assessments:

- **Continuous improvement:** Practitioners should frequently conduct pilot tests of new logging and assessment features with a small group of students. These pilot tests facilitate gathering preliminary data and necessary adjustments before wider implementation. User feedback and observed data patterns from these tests guide the iterative design and testing processes, continually refining logging systems and stealth assessments. Regular updates and improvements ensure that assessment tools remain effective and relevant (L. Wang et al., 2015; V. J. Shute et al., 2016). The iterative process also helps align game mechanics with curriculum goals, confirming that students learn from the game. Designing in-game tasks and elements that directly support learning objectives, such as real-world scenarios, helps students apply their knowledge to solve problems (Ke & Shute, 2015).
- **Data-driven adjustments:** Regular reviews of logged data allow for data-driven adjustments to stealth assessments, ensuring the collection of sufficient and relevant events with corresponding context information. Analyzing the correlation between learning-objective performance metrics or embedded assessment scores and targeted learning outcomes enables the refinement of stealth assessments (Ke & Shute, 2015). Focusing on assessments with strong positive correlations and reworking or replacing those with weak or negative correlations can improve the reliability and effectiveness of the stealth assessment.

- **Key user involvement:** Involving both students and educators in the design and testing process ensures that the system meets instructional needs and provides valuable insights into the student learning process (Hicks, 2021). Educator feedback is essential for constructing stealth assessments that align with the game's educational content. This feedback helps identify necessary indicators or features for educators' use, enhancing the overall effectiveness of the stealth assessment tool. Matching educators' perceptions with the goals of stealth assessments provides instructors with useful references for determining intervention moments, thereby enhancing the educational value of the GBL environment.

6.4 Future Research: Considering Implications for Affective Metrics

Sections 6.1, 6.2, and 6.3 outlined potential future studies centred on iterative refinements to features, models, and the design of both the stealth assessment and the game itself. In this section, we propose new avenues for research, focusing specifically on the development of affective metrics to further enhance the predictive accuracy of our learning outcome model. Based on our study, although our prediction model surpasses an acceptable accuracy threshold, there is noticeable room for improvement. While examining our results, we identified some interesting patterns not directly confirmed by the current features. We propose that constructing affective metrics based on features representing in-game activities could enhance our learning outcome prediction model, providing enhanced interpretability and appropriately measuring those patterns. These affective metrics can measure attention, engagement, motivation, and cognitive load levels. Additionally, these metrics can shape critical elements within instructor dashboards, aiding teachers in making appropriate intervention decisions. When combined, these affective metrics generate richer information for measuring complex patterns.

Based on our findings, we summarize the following implications for generating affective metrics from features representing in-game activities:

- **Attention and focus measurement:** By analyzing the duration of uninterrupted interactions with specific tools or tasks, we can track students' ability to sustain attention over extended periods. This measurement helps identify individuals who excel at maintaining concentration and those who might benefit from additional support. Logging instances when students divert from primary tasks to engage with irrelevant game elements provides critical data on distraction events. These instances are valuable for identifying game aspects that may cause students to lose focus. Analyzing patterns of focus and distraction, such as deviations from set tasks or prolonged periods of inactivity, can indicate levels of attention that significantly affect students' learning outcomes (Lin et al., 2019).

- **Engagement and motivation metrics:** Tracking indicators such as the frequency and duration of interactions with specific gameplay elements, optional tasks, and replay activities allows us to assess overall engagement and motivation levels, which often correlate with targeted learning outcomes. Studies by Chen and colleagues (2019) and Dabbous and colleagues (2022) identified that participants willing to engage in optional or extra activities are likelier to have higher engagement and learning outcomes. Additionally, analyzing how often students voluntarily use optional tools or participate in extra–game world exploration and conversations with non-playable characters offers insights into their motivation levels (David Des Armier Jr. & Skrabut, 2016). Investigating the correlation between these metrics and targeted learning outcomes helps determine whether highly engaging elements support learning objectives or merely entertain. Monitoring instances where high engagement does not lead to high learning outcomes allows for refining game features to align more closely with educational goals.

- **Cognitive load measurement:** Cognitive load could be measured by (1) calculating the time duration students use to complete tasks and comparing this duration with the total gameplay duration; (2) measuring the time spent reading instructions relative to the total gameplay time to estimate the cognitive load of absorbing new information; (3) monitoring the frequency and duration of interactions with in-game items to derive a metric of cognitive effort related to puzzle-solving, where high frequency and duration of interactions with certain items may indicate a high cognitive load when solving corresponding problems (Sevcenko et al., 2021); and (4) tracking students' speed, frequency, and accuracy with tool usage in various contexts to reflect how well they handle the cognitive demands of tool selection and application.

  Suppose all students exhibit high levels of cognitive load in certain activities. In that case, these game sessions may require further investigation to determine if they are too difficult, which may negatively affect learning outcomes (Chang et al., 2017). Conversely, if only a few students show high cognitive load levels, instructors could consider providing additional support to those students.

Last but not least, our prediction model results indicate a significant discrepancy in the model's ability to detect low-performing students compared to high-performing ones. We hypothesize that this discrepancy may be due to the lack of specific features representing students' off-task behaviours. The absence of such features could pose a key obstacle to implementing serious games with embedded stealth assessment in classrooms (Sabourin et al., 2013; Carpenter et al., 2020). Off-task behaviours, known contributors to ineffective learning (J. P. Rowe et al., 2009; Baker et al., 2004; Beserra et al., 2019;

Baker et al., 2010), are complex to measure accurately due to their contextual nature (Carpenter et al., 2020). Although we identified potential indicators of such behaviours in MHS through model inference—including extraneous map exploration, non-goal-oriented interactions with items, and hasty task completion—further analysis is required to confirm their correlation with off-task behaviours. Once confirmed, these indicators could be incorporated into a visualization dashboard for instructors, enabling more effective monitoring of student gameplay performance.

Since the game logs in MHS contain rich behavioural information designed to measure learning, we demonstrated a pipeline for creating a stealth assessment system grounded in student outcomes. It remains challenging for researchers to identify the underlying reasons for these behaviours based solely on logging data, and this remains a limitation of stealth assessment. This gap can be closed using a comprehensive framework that extends beyond the game and analysis of its logs, as introduced by Grover and colleagues (2017). Grover's approach encourages incorporating multi-source data, such as video recordings, interviews, surveys, biometric measurements, and focus groups. This approach assists in interpreting the behaviours reflected in the logs. Understanding the reasons behind these behaviours will help generate additional feature metrics, such as those mentioned above, and improve the accuracy and effectiveness of our stealth assessment.

7. Acknowledgements

The authors gratefully acknowledge Dr. Sean P. Goggins for his invaluable guidance and insightful feedback throughout the research and writing process. We extend our sincere thanks to Dr. Eric Wulff for his substantial contributions to developing and refining the rubrics for embedded assessment, as well as for providing organized external assessment instruments and related analytical results. His expertise significantly enhanced the rigour and depth of this work. We are deeply appreciative of the team at Adroit Studios Gaming Lab at the University of Missouri for their support, resources, and facilities, as well as their expert feedback regarding the MHS project. Their collaboration has been instrumental in the success of this research. Special thanks go to the editorial team at the *Journal of Learning Analytics*, including Dr. Hassan Khosravi (production editor) and Julia Cochrane (copyeditor) for their meticulous editing and proofreading efforts, which greatly improved the clarity and quality of this paper. Finally, we are grateful to the anonymous reviewers for their constructive feedback and thoughtful suggestions, which have substantially strengthened the final manuscript.

8. Declaration of Conflict of Interest

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

9. Funding

This work is supported by the US Department of Education's Institute of Education Sciences (R305A150364) and the i3 program (U411C140081). The views expressed are those of the project team and do not reflect the funders.

References

Abrahart, R. J., & See, L. (2000). Comparing neural network and autoregressive moving average techniques for the provision of continuous river flow forecasts in two contrasting catchments. *Hydrological Processes*, *14*(11-12), 2157–2172. https://doi.org/10.1002/1099-1085(20000815/30)14:11/12%3C2157::AID-HYP57%3E3.0.CO;2-S

Agudo-Peregrina, Á. F., Iglesias-Pradas, S., Conde-González, M. Á., & Hernández-García, Á. (2014). Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning. *Computers in Human Behavior*, *31*, 542–550. https://doi.org/10.1016/j.chb.2013.05.031

Akram, B., Min, W., Wiebe, E., Mott, B., Boyer, K. E., & Lester, J. (2018). Improving stealth assessment in game-based learning with LSTM-based analytics. In K. E. Boyer & M. Yudelson (Eds.), *Proceedings of the 11th International Conference on Educational Data Mining* (EDM 2018), 15–18 July 2018, Buffalo, New York, USA (pp. 208–218). EDM. https://educationaldatamining.org/files/conferences/EDM2018/EDM2018_Preface_TOC_Proceedings.pdf

Albert, J. H. (1988). Computational methods using a Bayesian hierarchical generalized linear model. *Journal of the American Statistical Association*, *83*(404), 1037–1044. https://doi.org/10.1080/01621459.1988.10478698

Alonso-Fernández, C., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., & Manjón, B. F. (2021). Data science meets standardized game learning analytics. In T. Klinger, C. Kollmitzer, & A. Pester (Eds.), *Proceedings of the 2021 IEEE Global Engineering Education Conference* (EDUCON 2021), 21–23 April 2021, Vienna, Austria (pp. 1546–1552). IEEE. https://doi.org/10.1109/EDUCON46332.2021.9454134

Alonso-Fernández, C., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2021). Improving evidence-based assessment of players using serious games. *Telematics and Informatics*, *60*, 101583. https://doi.org/10.1016/j.tele.2021.101583

Amory, A. (2007). Game object model version II: A theoretical framework for educational game development. *Educational Technology Research and Development*, *55*(1), 51–77. https://doi.org/10.1007/s11423-006-9001-x

Arnab, S., Lim, T., Carvalho, M. B., Bellotti, F., de Freitas, S., Louchart, S., Suttie, N., Berta, R., & De Gloria, A. (2015). Mapping learning and game mechanics for serious games analysis. *British Journal of Educational Technology*, *46*(2), 391–411. https://doi.org/10.1111/bjet.12113

Ayzenberg, Y., Hernandez Rivera, J., & Picard, R. (2012). FEEL: Frequent EDA and event logging—a mobile social interaction stress monitoring system. In *CHI EA 2012: CHI '12 Extended Abstracts on Human Factors in Computing Systems*, 5–10 May 2012, Austin, Texas, USA (pp. 2357–2362). ACM. https://doi.org/10.1145/2212776.2223802

Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-task behavior in the cognitive tutor classroom: When students "game the system." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI 2004), 24–29 April 2004, Vienna, Austria (pp. 383–390). ACM. https://doi.org/10.1145/985692.985741

Baker, R. S., D'Mello, S. K., Rodrigo, M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, *68*(4), 223–241. https://doi.org/10.1016/j.ijhcs.2009.12.003

Belland, B. R., Walker, A. E., & Kim, N. J. (2017). A Bayesian network meta-analysis to synthesize the influence of contexts of scaffolding use on cognitive outcomes in STEM education. *Review of Educational Research*, *87*(6), 1042–1081. https://doi.org/10.3102/0034654317723009

Bernecker, K., & Ninaus, M. (2021). No pain, no gain? Investigating motivational mechanisms of game elements in cognitive tasks. *Computers in Human Behavior*, *114*, 106542. https://doi.org/10.1016/j.chb.2020.106542

Beserra, V., Nussbaum, M., & Oteo, M. (2019). On-task and off-task behavior in the classroom: A study on mathematics learning with educational video games. *Journal of Educational Computing Research*, *56*(8), 1361–1383. https://doi.org/10.1177/0735633117744346

Bird, K. A., Castleman, B. L., Mabel, Z., & Song, Y. (2021). Bringing transparency to predictive analytics: A systematic comparison of predictive modeling methods in higher education. *AERA Open*, *7*, 23328584211037630. https://doi.org/10.1177/23328584211037630

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Breuer, J. S., & Bente, G. (2010). Why so serious? On the relation of serious games and learning. *Eludamos: Journal for Computer Game Culture*, *4*(1), 7–24. https://doi.org/10.7557/23.6111

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, *106*, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011

Butcher, B., & Smith, B. J. (2020). Feature engineering and selection: A practical approach for predictive models: By Max Kuhn and Kjell Johnson. Boca Raton, FL: Chapman & Hall/CRC Press, 2019, xv + 297 pp., $79.95(H), ISBN: 978-1-13-807922-9 [Review]. *The American Statistician*, *74*(3), 308–309. https://doi.org/10.1080/00031305.2020.1790217

Calvo-Morata, A., Rotaru, D. C., Alonso-Fernández, C., Freire-Morán, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2020). Validation of a cyberbullying serious game using game analytics. *IEEE Transactions on Learning Technologies*, *13*(1), 186–197. https://doi.org/10.1109/TLT.2018.2879354

Cardia da Cruz, L., Sierra-Franco, C. A., Silva-Calpa, G. F. M., & Barbosa Raposo, A. (2020). A self-adaptive serious game for eye-hand coordination training. In X. Fang (Ed.), *HCI in games. HCII 2020. Lecture notes in computer science* (pp. 385–397, Vol. 12211). Springer International Publishing. https://doi.org/10.1007/978-3-030-50164-8_28

Carpenter, D., Emerson, A., Mott, B. W., Saleh, A., Glazewski, K. D., Hmelo-Silver, C. E., & Lester, J. C. (2020). Detecting off-task behavior from student dialogue in game-based collaborative learning. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Proceedings of the 21st International Conference on Artificial Intelligence in Education, Part I* (AIED 2020), 6–10 July 2020, Ifrane, Morocco (pp. 55–66). ACM. https://doi.org/10.1007/978-3-030-52237-7_5

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

Cerra, P. P., Álvarez, H. F., Parra, B. B., & Cordera, P. I. (2022). Effects of using game-based learning to improve the academic performance and motivation in engineering studies. *Journal of Educational Computing Research*, *60*(7), 1663–1687. https://doi.org/10.1177/07356331221074022

Champion, C., & Elkan, C. (2017). Visualizing the consequences of evidence in Bayesian networks. *arXiv preprint arXiv:1707.00791*. https://arxiv.org/abs/1707.00791

Chang, C.-C., Liang, C., Chou, P.-N., & Lin, G.-Y. (2017). Is game-based learning better in flow experience and various types of cognitive load than non-game-based learning? Perspective from multimedia and media richness. *Computers in Human Behavior*, *71*, 218–227. https://doi.org/10.1016/j.chb.2017.01.031

Chen, C.-H., Law, V., & Huang, K. (2019). The roles of engagement and competition on learner's performance and motivation in game-based science learning. *Educational Technology Research and Development*, *67*(4), 1003–1024. https://doi.org/10.1007/s11423-019-09670-7

Chen, F., Cui, Y., & Chu, M. W. (2020). Utilizing game analytics to inform and validate digital game-based assessment with evidence-centered game design: A case study. *International Journal of Artificial Intelligence in Education*, *30*(4), 481–503. https://doi.org/10.1007/s40593-020-00202-6

Cloude, E. B., Dever, D. A., Wiedbusch, M. D., & Azevedo, R. (2020). Quantifying scientific thinking using multichannel data with Crystal Island: Implications for individualized game-learning analytics. *Frontiers in Education*, *5*, 21. https://doi.org/10.3389/feduc.2020.572546

Cramér, H. (1999). *Mathematical methods of statistics*. Princeton University Press. https://press.princeton.edu/books/paperback/9780691005478/mathematical-methods-of-statistics

Dabbous, M., Kawtharani, A., Fahs, I., Hallal, Z., Shouman, D., Akel, M., Rahal, M., & Sakr, F. (2022). The role of game-based learning in experiential education: Tool validation, motivation assessment, and outcomes evaluation among a sample of pharmacy students. *Education Sciences*, *12*(7), 434. https://doi.org/10.3390/educsci12070434

David Des Armier Jr., C. E. S., & Skrabut, S. (2016). Using game elements to increase student engagement in course assignments. *College Teaching*, *64*(2), 64–72. https://doi.org/10.1080/87567555.2015.1094439

De Freitas, S., & Oliver, M. (2006). How can exploratory learning with games and simulations within the curriculum be most effectively evaluated? *Computers & Education*, *46*(3), 249–264. https://doi.org/10.1016/j.compedu.2005.11.007

Dhal, P., & Azad, C. (2022). A comprehensive survey on feature selection in the various fields of machine learning. *Applied Intelligence*, *52*(4), 4543–4581. https://doi.org/10.1007/s10489-021-02550-9

Díaz-Ramírez, J. (2020). Gamification in engineering education—An empirical assessment on learning and game performance. *Heliyon*, *6*(9), e04972. https://doi.org/10.1016/j.heliyon.2020.e04972

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, *6*(3), 241–252. https://doi.org/10.1080/00401706.1964.10490181

Elrahman, S. M. A., & Abraham, A. (2013). A review of class imbalance problem. *Journal of Network and Innovative Computing*, *1*, 332–340. https://cspub-jnic.org/index.php/jnic/article/view/42

Emerson, A., Cloude, E. B., Azevedo, R., & Lester, J. (2020). Multimodal learning analytics for game-based learning. *British Journal of Educational Technology*, *51*(5), 1505–1526. https://doi.org/10.1111/bjet.12992

Fadila, N. N., Saidah, K., & Wendha, D. D. N. (2023). Development of interactive multimedia based on educational games of plant parts and their functions to improve student learning outcomes. *Jurnal Pijar Mipa*, *18*(5), 666–669. https://doi.org/10.29303/jpm.v18i5.5472

Fang, Y., Li, T., Huynh, L., Christhilf, K., Roscoe, R. D., & McNamara, D. S. (2023). Stealth literacy assessments via educational games. *Computers*, *12*(7), 130. https://doi.org/10.3390/computers12070130

Feng, X., & Yamada, M. (2019). Effects of game-based learning on informal historical learning: A learning analytics approach. In M. Chang, H.-J. So, L.-H. Wong, J.-L. Shih, & F.-Y. Yu (Eds.), *Proceedings of the 27th International Conference on Computers in Education* (ICCE 2019), 2–6 December 2019, Kenting, Taiwan (pp. 505–514). APSCE. https://doi.org/10.58459/icce.2019.496

Fisher, R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society*, *98*(1), 39–82. https://www.jstor.org/stable/2342435

Fogarty, J. A. (2006). *Constructing and evaluating sensor-based statistical models of human interruptibility* [Doctoral dissertation, Carnegie Mellon University] [AAI3241594]. http://reports-archive.adm.cs.cmu.edu/anon/usr/ftp/hcii/CMU-HCII-06-100.pdf

Fokides, E., Atsikpasi, P., Kaimara, P., & Deliyannis, I. (2019). Factors influencing the subjective learning effectiveness of serious games. *Journal of Information Technology Education: Research*, *18*, 437–466. https://doi.org/10.28945/4441

Gee, J. P. (2003). What video games have to teach us about learning and literacy. *Computers in Entertainment (CIE)*, *1*(1), 20. https://doi.org/10.1145/950566.950595

Georgiadis, K., van Lankveld, G., Bahreini, K., & Westera, W. (2019). Learning analytics should analyse the learning: Proposing a generic stealth assessment tool. In *Proceedings of the 2019 IEEE Conference on Games* (CoG), 20–23 August 2019, London, UK (pp. 1–8). IEEE. https://doi.org/10.1109/CIG.2019.8847960

Georgiadis, K., van Lankveld, G., Bahreini, K., & Westera, W. (2021). On the robustness of stealth assessment. *IEEE Transactions on Games*, *13*(2), 180–192. https://doi.org/10.1109/TG.2020.3020015

Gibson, D., & Clarke-Midura, J. (2015). Some psychometric and design implications of game-based learning analytics. In P. Isaías, J. M. Spector, D. Ifenthaler, & D. G. Sampson (Eds.), *E-learning systems, environments and approaches: Theory and implementation* (pp. 247–261). Springer International Publishing. https://doi.org/10.1007/978-3-319-05825-2_17

Glass, G. V. (1966). Note on rank biserial correlation. *Educational and Psychological Measurement*, *26*(3), 623–631. https://doi.org/10.1177/001316446602600307

Göbel, S., de Carvalho Rodrigues, A., Mehm, F., & Steinmetz, R. (2009). Narrative game-based learning objects for story-based digital educational games. *Narrative*, *14*, 16. https://www.kom.tu-darmstadt.de/papers/GdMS09_533.pdf

Göbel, S., & Mehm, F. (2013). Personalized, adaptive digital educational games using narrative game-based learning objects. In K. Bredl & W. Bösche (Eds.), *Serious games and virtual worlds in education, professional development, and healthcare* (pp. 74–84). IGI Global. https://doi.org/10.4018/978-1-4666-3673-6.ch005

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 *Proceedings of the 2010 IEEE Second International Conference on Social Computing* (SocialCom 2010), 20–22 August 2010, Minneapolis, Minnesota, USA (pp. 500–507). IEEE. https://doi.org/10.1109/SocialCom.2010.79

Goggins, S. P., Galyen, K., Petakovic, E., & Laffey, J. M. (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

Goggins, S. P., Laffey, J., & Amelung, C. (2011). Context aware CSCL: Moving toward contextualized analysis. In H. Spada, G. Stahl, N. Miyake, & N. Law (Eds.), *Connecting Computer-Supported Collaborative Learning to Policy and Practice: CSCL2011 Conference Proceedings. Volume II—Short Papers & Posters*, 4–8 July 2011, Hong Kong, China. International Society of the Learning Sciences. https://doi.org/10.22318/cscl2011.591

Gomez, M. J., Ruipérez-Valiente, J. A., & García Clemente, F. J. (2023). A framework to support interoperable game-based assessments as a service (GBAaaS): Design, development, and use cases. *Software: Practice and Experience*, *53*(11), 2222–2240. https://doi.org/10.1002/spe.3254

Grover, S., Basu, S., Bienkowski, M., Eagle, M., Diana, N., & Stamper, J. (2017). A framework for using hypothesis-driven approaches to support data-driven learning analytics in measuring computational thinking in block-based programming environments. *ACM Transactions on Computing Education*, *17*(3). https://doi.org/10.1145/3105910

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5). https://doi.org/10.1145/3236009

Gunter, G., Kenny, R. F., & Vick, E. H. (2006). A case for a formal design paradigm for serious games. *The Journal of the International Digital Media and Arts Association*, *3*(1), 93–105. https://www.academia.edu/29030853/A_Case_for_a_Formal_Design_Paradigm_for_Serious_Games_iDMAa_and_IMS_conference

Guo, X., Yin, Y., Dong, C., Yang, G., & Zhou, G. (2008). On the class imbalance problem. In M. Guo, L. Zhao, & L. Wang (Eds.), *Proceedings of the 2008 Fourth International Conference on Natural Computation* (ICNC), 18–20 October 2008, Jinan, China (pp. 192–201, Vol. 4). IEEE. https://doi.org/10.1109/ICNC.2008.871

Gupta, A., Carpenter, D., Min, W., Rowe, J. P., Azevedo, R., & Lester, J. C. (2021). Multimodal multi-task stealth assessment for reflection-enriched game-based learning. In *Multimodal Artificial Intelligence in Education at the 22nd International Conference on Artificial Intelligence in Education* (MAIED@ AIED), 14 June 2021, Utrecht, Netherlands (pp. 93–102). CEUR. https://ceur-ws.org/Vol-2902/paper9.pdf

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection (L. P. Kaelbling, Ed.). *Journal of Machine Learning Research*, *3*, 1157–1182. https://dl.acm.org/doi/10.5555/944919.944968

Haas, L., & Tussey, J. (2022). Equity and engagement through digital storytelling and game-based learning. In *Research anthology on developments in gamification and game-based learning* (pp. 1451–1472). IGI Global. https://doi.org/10.4018/978-1-7998-5770-9.ch013

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 *Proceedings of the 2014 IEEE 14th International Conference on Advanced Learning Technologies* (ICALT 2014), 7–10 July 2014, Athens, Greece (pp. 230–232). IEEE. https://doi.org/10.1109/ICALT.2014.73

Heine, C. (2021). Towards modeling visualization processes as dynamic Bayesian networks. *IEEE Transactions on Visualization and Computer Graphics*, *27*(2), 1000–1010. https://doi.org/10.1109/TVCG.2020.3030395

Heinemann, B., Ehlenz, M., Görzen, S., & Schroeder, U. (2022). xAPI made easy: A learning analytics infrastructure for interdisciplinary projects. *International Journal of Online & Biomedical Engineering*, *18*(14). https://doi.org/10.3991/ijoe.v18i14.35079

Henderson, N., Acosta, H., Min, W., Mott, B., Lord, T., Reichsman, F., Dorsey, C., Wiebe, E., & Lester, J. (2022). Enhancing stealth assessment in game-based learning environments with generative zero-shot learning. In A. Mitrovic & N. Bosch

(Eds.), *Proceedings of the 15th International Conference on Educational Data Mining* (EDM 2022), 24–27 July 2022, Durham, UK (pp. 171–182). International Educational Data Mining Society. https://educationaldatamining.org/edm2022/proceedings/2022.EDM-long-papers.15/index.html

Henderson, N., Kumaran, V., Min, W., Mott, B., Wu, Z., Boulden, D., Lord, T., Reichsman, F., Dorsey, C., Wiebe, E., & Lester, J. (2020). Enhancing student competency models for game-based learning with a hybrid stealth assessment framework. In A. N. Rafferty, J. Whitehill, C. Romero, & V. Cavalli-Sforza (Eds.), *Proceedings of the 13th International Conference on Educational Data Mining* (EDM 2020), 10–13 July 2020, online (pp. 92–103). International Educational Data Mining Society. https://educationaldatamining.org/files/conferences/EDM2020/papers/paper_158.pdf

Hicks, D. (2021). *Stealth assessment: Teacher's perceptions of how digital-based educational games influence teaching and learning in the middle school mathematics classroom* [Doctoral dissertation, Piedmont College]. https://www.proquest.com/openview/a10b0dc17805485238724aa3376201b3/1

Hooshyar, D., Ahmad, R. B., Yousefi, M., Fathi, M., Horng, S.-J., & Lim, H. (2016). Applying an online game-based formative assessment in a flowchart-based intelligent tutoring system for improving problem-solving skills. *Computers & Education*, *94*, 18–36. https://doi.org/10.1016/j.compedu.2015.10.013

Huang, Y.-M., & Du, S.-X. (2005). Weighted support vector machine for classification with uneven training class sizes. In *Proceedings of the 2005 International Conference on Machine Learning and Cybernetics* (ICMLC 2005), 18–21 August 2005, Guangzhou, China (pp. 4365–4369, Vol. 7). IEEE. https://doi.org/10.1109/ICMLC.2005.1527706

Jalota, C., & Agrawal, R. (2021). Feature selection algorithms and student academic performance: A study. In D. Gupta, A. Khanna, S. Bhattacharyya, A. E. Hassanien, S. Anand, & A. Jaiswal (Eds.), *International conference on innovative computing and communications. Advances in intelligent systems and computing* (pp. 317–328, Vol. 1165). Springer. https://doi.org/10.1007/978-981-15-5113-0_23

Jeon, H., He, H., Wang, A., & Spooner, S. (2023). Modeling student performance in game-based learning environments. *arXiv preprint arXiv:2309.13429*. https://arxiv.org/abs/2309.13429

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

Kaur, H., Pannu, H. S., & Malhi, A. K. (2019). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys (CSUR)*, *52*(4). https://doi.org/10.1145/3343440

Ke, F., Parajuli, B., & Smith, D. (2019). Assessing game-based mathematics learning in action. In D. Ifenthaler & Y. J. Kim (Eds.), *Game-based assessment revisited* (pp. 213–227). Springer International Publishing. https://doi.org/10.1007/978-3-030-15569-8_11

Ke, F., & Shute, V. (2015). Design of game-based stealth assessment and learning support. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious games analytics: Methodologies for performance measurement, assessment, and improvement* (pp. 301–318). Springer International Publishing. https://doi.org/10.1007/978-3-319-05834-4_13

Keerin, P. (2021). A comparative study of missing value imputation methods for education data. In *Proceedings of the 29th International Conference on Computers in Education* (ICCE 2021), 22–26 November 2021, online. Asia-Pacific Society for Computers in Education. https://library.apsce.net/index.php/ICCE/article/view/4233

Kim, Y. J., Ruipérez-Valiente, J. A., Philip, T., Louisa, R., & Klopfer, E. (2019). Towards a process to integrate learning analytics and evidence-centered design for game-based assessment. In *Companion Proceedings of the Ninth International Conference on Learning Analytics and Knowledge* (LAK 2019), 4–8 March 2019, Tempe, Arizona, USA (pp. 204–205). SoLAR. https://www.solaresearch.org/wp-content/uploads/2019/08/LAK19_Companion_Proceedings.pdf

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*(260), 583–621. https://doi.org/10.1080/01621459.1952.10483441

Kuhn, M. (2019). The caret package. https://topepo.github.io/caret/available-models.html

Laffey, J. M., Griffin, J., Sigoloff, J., Lander, S., Sadler, T., Goggins, S., Kim, S. M., Wulff, E., & Womack, A. J. (2017). Mission HydroSci: A progress report on a transformational role playing game for science learning. In *Proceedings of the 12th International Conference on the Foundations of Digital Games* (FDG 2017), 14–17 August 2017, Hyannis, Massachusetts, USA. ACM. https://doi.org/10.1145/3102071.3106354

Lee, J. Y., Donkers, J., Jarodzka, H., & van Merriënboer, J. J. (2019). How prior knowledge affects problem-solving performance in a medical simulation game: Using game-logs and eye-tracking. *Computers in Human Behavior*, *99*, 268–277. https://doi.org/10.1016/j.chb.2019.05.035

Li, F.-Y., Hwang, G.-J., Chen, P.-Y., & Lin, Y.-J. (2021). Effects of a concept mapping-based two-tier test strategy on students' digital game-based learning performances and behavioral patterns. *Computers & Education*, *173*, 104293. https://doi.org/10.1016/j.compedu.2021.104293

Li, N., Kidziński, Ł., Jermann, P., & Dillenbourg, P. (2015). MOOC video interaction patterns: What do they tell us? In G. Conole, T. Klobučar, C. Rensing, J. Konert, & E. Lavoué (Eds.), *Design for teaching and learning in a networked*

world. EC-TEL 2015. Lecture notes in computer science (pp. 197–210, Vol. 9307). Springer International Publishing. https://doi.org/10.1007/978-3-319-24258-3_15

Lin, Y.-C., Hsieh, Y.-H., Hou, H.-T., & Wang, S.-M. (2019). Exploring students' learning and gaming performance as well as attention through a drill-based gaming experience for environmental education. *Journal of Computers in Education*, *6*(3), 315–334. https://doi.org/10.1007/s40692-019-00130-y

Liu, M., Cai, Y., Han, S., & Shao, P. (2022). Understanding student navigation patterns in game-based learning. *Journal of Learning Analytics*, *9*(3), 50–74. https://doi.org/10.18608/jla.2022.7637

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. (2015). Measuring the (dis-) similarity between expert and novice behaviors as serious games analytics. *Education and Information Technologies*, *20*(1), 5–19. https://doi.org/10.1007/s10639-013-9263-y

Loh, C. S., Sheng, Y., & Ifenthaler, D. (2015). Serious games analytics: Theoretical framework. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious games analytics: Methodologies for performance measurement, assessment, and improvement* (pp. 3–29). Springer International Publishing. https://doi.org/10.1007/978-3-319-05834-4_1

Lu, W., Griffin, J., Sadler, T. D., Laffey, J., & Goggins, S. P. (2023). Serious game analytics by design: Feature generation and selection using game telemetry and game metrics—toward predictive model construction. *Journal of Learning Analytics*, *10*(1), 168–188. https://doi.org/10.18608/jla.2023.7681

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*. https://arxiv.org/abs/1705.07874

Ma, J., Han, X., Yang, J., & Cheng, J. (2015). Examining the necessary condition for engagement in an online learning environment based on learning analytics approach: The role of the instructor. *The Internet and Higher Education*, *24*, 26–34. https://doi.org/10.1016/j.iheduc.2014.09.005

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, *18*(1), 50–60. https://www.jstor.org/stable/2236101

Manzano-León, A., Camacho-Lazarraga, P., Guerrero, M. A., Guerrero-Puerta, L., Aguilar-Parra, J. M., Trigueros, R., & Alias, A. (2021). Between level up and game over: A systematic literature review of gamification in education. *Sustainability*, *13*(4), 2247. https://doi.org/10.3390/su13042247

Marron, J. S., Todd, M. J., & Ahn, J. (2007). Distance-weighted discrimination. *Journal of the American Statistical Association*, *102*(480), 1267–1271. https://doi.org/10.1198/016214507000001120

Martínez-Monés, A., Harrer, A., & Dimitriadis, Y. (2011). An interaction-aware design process for the integration of interaction analysis into mainstream CSCL practices. In S. Puntambekar, G. Erkens, & C. Hmelo-Silver (Eds.), *Analyzing interactions in CSCL: Methods, approaches and issues* (pp. 269–291). Springer US. https://doi.org/10.1007/978-1-4419-7710-6_13

Maryani, I., & Hidayat, N. (2019). Interactive game: A step to reduce science learning difficulties of elementary school students. In R. N. D. Irmawati, D. Sulisworo, E. Arroyo, I. Tokoro, R. C. I. Prahmana, & A. Azhari (Eds.), *Proceedings of the First International Conference on Progressive Civil Society* (ICONPROCS 2019), 19 February 2019, Yogyakarta, Indonesia (pp. 100–103). Atlantis Press. https://doi.org/10.2991/iconprocs-19.2019.20

Min, W., Frankosky, M. H., Mott, B. W., Rowe, J. P., Smith, A., Wiebe, E., Boyer, K. E., & Lester, J. C. (2020). DeepStealth: Game-based learning stealth assessment with deep neural networks. *IEEE Transactions on Learning Technologies*, *13*(2), 312–325. https://doi.org/10.1109/TLT.2019.2922356

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, *2003*(1), i–29. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable machine learning—a brief history, state-of-the-art and challenges. In I. Koprinska, M. Kamp, A. Appice, C. Loglisci, L. Antonie, A. Zimmermann, R. Guidotti, Ö. Özgöbek, R. P. Ribeiro, R. Gavaldà, J. Gama, L. Adilova, Y. Krishnamurthy, P. M. Ferreira, D. Malerba, I. Medeiros, M. Ceci, G. Manco, E. Masciari, … J. A. Gulla (Eds.), *ECML PKDD 2020 workshops. ECML PKDD 2020. Communications in computer and information science* (pp. 417–431, Vol. 1323). Springer International Publishing. https://doi.org/10.1007/978-3-030-65965-3_28

Moore, G. R., & Shute, V. J. (2017). Improving learning through stealth assessment of conscientiousness. In A. Marcus-Quinn & T. Hourigan (Eds.), *Handbook on digital learning for K–12 schools* (pp. 355–368). Springer International Publishing. https://doi.org/10.1007/978-3-319-33808-8_21

Mouri, K., Okubo, F., Shimada, A., & Ogata, H. (2016, July). Bayesian network for predicting students' final grade using e-book logs in university education. In J. M. Spector, C.-C. Tsai, D. G. Sampson, Kinshuk, R. Huang, N.-S. Chen, &

P. Resta (Eds.), *Proceedings of the 2016 IEEE 16th International Conference on Advanced Learning Technologies* (ICALT 2016), 25–28 July 2018, Austin, Texas, USA (pp. 85–89). IEEE. https://doi.org/10.1109/ICALT.2016.27

Nguyen, H. A., Hou, X., Stamper, J., & McLaren, B. M. (2020). Moving beyond test scores: Analyzing the effectiveness of a digital learning game through learning analytics. In A. N. Rafferty, J. Whitehill, C. Romero, & V. Cavalli-Sforza (Eds.), *Proceedings of the 13th International Conference on Educational Data Mining* (EDM 2020), 10–13 July 2020, online. International Educational Data Mining Society. https://educationaldatamining.org/files/conferences/EDM2020/papers/paper_182.pdf

Niemelä, M., Kärkkäinen, T., Äyrämö, S., Ronimus, M., Richardson, U., & Lyytinen, H. (2020). Game learning analytics for understanding reading skills in transparent writing system. *British Journal of Educational Technology*, *51*(6), 2376–2390. https://doi.org/10.1111/bjet.12916

Nietfeld, J. L. (2020). Predicting transfer from a game-based learning environment. *Computers & Education*, *146*, 103780. https://doi.org/10.1016/j.compedu.2019.103780

Owen, V. E., & Baker, R. S. (2020). Fueling prediction of player decisions: Foundations of feature engineering for optimized behavior modeling in serious games. *Technology, Knowledge and Learning*, *25*(2), 225–250. https://doi.org/10.1007/s10758-018-9393-9

Park, H.-S., & Cho, S.-B. (2010). Building mobile social network with semantic relation using Bayesian NeTwork-based Life-log Mining. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing* (SocialCom 2010), 20–22 August 2010, Minneapolis, Minnesota, USA (pp. 401–406). IEEE. https://doi.org/10.1109/SocialCom.2010.64

Pérez-Colado, I. J., Pérez-Colado, V. M., Martinez-Ortiz, I., Freire, M., & Fernandez-Manjón, B. (2022). Using e-learning standards to improve serious game deployment and evaluation. In I. Kallel, H. M. Kammoun, & L. Hsairi (Eds.), *Proceedings of the 2022 IEEE Global Engineering Education Conference* (EDUCON 2022), 28–31 March 2022, Tunis, Tunisia (pp. 2077–2083). IEEE. https://doi.org/10.1109/EDUCON52537.2022.9766573

Rahimi, M., Moradi, H., Vahabie, A.-h., & Kebriaei, H. (2023). Continuous reinforcement learning-based dynamic difficulty adjustment in a visual working memory game. *arXiv preprint arXiv:2308.12726*. https://arxiv.org/abs/2308.12726

Ray, P., Reddy, S. S., & Banerjee, T. (2021). Various dimension reduction techniques for high dimensional data analysis: A review. *Artificial Intelligence Review*, *54*(5), 3473–3515. https://doi.org/10.1007/s10462-020-09928-0

Reeves, T., Romine, W., Laffey, J., Sadler, T., & Goggins, S. (2020). Distance learning through game-based 3D virtual learning environments: Mission Hydro Science. Evaluation report for Mission HydroSci. https://files.eric.ed.gov/fulltext/ED605283.pdf

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD 2016), 13–17 August 2016, San Francisco, California, USA (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778

Rohmani, R., & Pambudi, N. (2023). A critical review of educational games as a tool for strengthening digital literacy. *International Journal of Multidisciplinary: Applied Business and Education Research*, *4*(5), 1483–1493. https://doi.org/10.11594/ijmaber.04.05.10

Romero, M., Usart, M., & Ott, M. (2015). Can serious games contribute to developing and sustaining 21st century skills? *Games and Culture*, *10*(2), 148–177. https://doi.org/10.1177/1555412014548919

Romine, W. L., Sadler, T. D., & Wulff, E. P. (2017). Conceptualizing student affect for science and technology at the middle school level: Development and implementation of a measure of affect in science and technology (MAST). *Journal of Science Education and Technology*, *26*(5), 534–545. https://doi.org/10.1007/s10956-017-9697-x

Rosydiana, E. A., Sudjimat, D. A., & Utama, C. (2023). The effect of digital learning media using scratch game based learning on student problem solving skills. *Jurnal Penelitian Pendidikan IPA*, *9*(11), 10010–10015. https://doi.org/10.29303/jppipa.v9i11.4876

Rowe, E., Asbell-Clarke, J., Baker, R. S., Eagle, M., Hicks, A. G., Barnes, T. M., Brown, R. A., & Edwards, T. (2017). Assessing implicit science learning in digital games. *Computers in Human Behavior*, *76*, 617–630. https://doi.org/10.1016/j.chb.2017.03.043

Rowe, J. P., McQuiggan, S. W., Robison, J. L., & Lester, J. C. (2009). Off-task behavior in narrative-centered learning environments. In V. Dimitrova, R. Mizoguchi, B. du Boulay, & A. Graesser (Eds.), *Frontiers in artificial intelligence and applications. Ebook Volume 200: Artificial intelligence in education* (pp. 99–106). IOS Press. https://doi.org/10.3233/978-1-60750-028-5-99

Sabourin, J. L., Shores, L. R., Mott, B. W., & Lester, J. C. (2013). Understanding and predicting student self-regulated learning strategies in game-based learning environments. *International Journal of Artificial Intelligence in Education*, *23*(1), 94–114. https://doi.org/10.1007/s40593-013-0004-6

Sao Pedro, M. A., Baker, R. S. J. d., & Gobert, J. D. (2012). Improving construct validity yields better models of systematic inquiry even with less information. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, J. Masthoff, B. Mobasher, M. C. Desmarais, & R. Nkambou (Eds.), *User modeling, adaptation, and personalization. UMAP 2012. Lecture notes in computer science* (pp. 249–260, Vol. 7379). Springer. https://doi.org/10.1007/978-3-642-31454-4_21

Schardosim Simão, J. P., Mellos Carlos, L., Saliah-Hassane, H., da Silva, J. B., & da Mota Alves, J. B. (2018). Model for recording learning experience data from remote laboratories using xAPI. In *Proceedings of the 2018 XIII Latin American Conference on Learning Technologies* (LACLO 2018), 1–5 October 2018, São Paulo, Brazil (pp. 450–457). IEEE. https://doi.org/10.1109/LACLO.2018.00081

Seaton, J. X., Chang, M., & Graf, S. (2019). Integrating a learning analytics dashboard in an online educational game. In A. Tlili & M. Chang (Eds.), *Data analytics approaches in educational games and gamification systems* (pp. 127–138). Springer Singapore. https://doi.org/10.1007/978-981-32-9335-9_7

Serrano-Laguna, Á., Martínez-Ortiz, I., Haag, J., Regan, D., Johnson, A., & Fernández-Manjón, 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

Serrano-Laguna, Á., Torrente, J., Moreno-Ger, P., & Fernández-Manjón, B. (2014). Application of learning analytics in educational videogames. *Entertainment Computing*, *5*(4), 313–322. https://doi.org/10.1016/j.entcom.2014.02.003

Sevcenko, N., Ninaus, M., Wortha, F., Moeller, K., & Gerjets, P. (2021). Measuring cognitive load using in-game metrics of a serious simulation game. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.572437

Seyderhelm, A. J., & Blackmore, K. L. (2023). How hard is it really? Assessing game-task difficulty through real-time measures of performance and cognitive load. *Simulation & Gaming*, *54*(3), 294–321. https://doi.org/10.1177/10468781231169910

Shohel, M. M. C., Ashrafuzzaman, M., Naomee, I., Tanni, S. A., & Azim, F. (2022). Game-based teaching and learning in higher education: Challenges and prospects. In C.-A. Lane (Ed.), *Handbook of research on acquiring 21st century literacy skills through game-based learning* (pp. 78–106). IGI Global. https://doi.org/10.4018/978-1-7998-7271-9.ch005

Shoukry, L., Göbel, S., & Steinmetz, R. (2014). Learning analytics and serious games: Trends and considerations. In *Proceedings of the 2014 ACM International Workshop on Serious Games* (SeriousGames 2014), 7 November 2014, Orlando, Florida, USA (pp. 21–26). ACM. https://doi.org/10.1145/2656719.2656729

Shute, V., Ke, F., & Wang, L. (2017). Assessment and adaptation in games. In P. Wouters & H. van Oostendorp (Eds.), *Instructional techniques to facilitate learning and motivation of serious games* (pp. 59–78). Springer International Publishing. https://doi.org/10.1007/978-3-319-39298-1_4

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

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Shute, V. J. (2011a). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503–523). Information Age Publishing.

Shute, V. J. (2011b). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503–524, Vol. 55). IAP Information Age Publishing. https://psycnet.apa.org/record/2011-11269-020

Shute, V. J., & Rahimi, S. (2021). Stealth assessment of creativity in a physics video game. *Computers in Human Behavior*, *116*, 106647. https://doi.org/10.1016/j.chb.2020.106647

Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior*, *63*, 106–117. https://doi.org/10.1016/j.chb.2016.05.047

Siddique, A., Jan, A., Majeed, F., Qahmash, A. I., Quadri, N. N., & Wahab, M. O. A. (2021). Predicting academic performance using an efficient model based on fusion of classifiers. *Applied Sciences*, *11*(24), 11845. https://doi.org/10.3390/app112411845

Smith, G., Shute, V., & Muenzenberger, A. (2019). Designing and validating a stealth assessment for calculus competencies. *Journal of Applied Testing Technology*, *20*(S1), 52–59. https://www.jattjournal.net/index.php/atp/article/view/142702

Snow, E. L., Allen, L. K., & McNamara, D. S. (2015). The dynamical analysis of log data within educational games. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious games analytics: Methodologies for performance measurement, assessment, and improvement* (pp. 81–100). Springer International Publishing. https://doi.org/10.1007/978-3-319-05834-4_4

Strukova, S., Ruipérez-Valiente, J. A., & Mármol, F. G. (2023). Adapting knowledge inference algorithms to measure geometry competencies through a puzzle game. *ACM Transactions on Knowledge Discovery from Data*, *18*(1). https://doi.org/10.1145/3614436

Student. (1908). The probable error of a mean. *Biometrika*, *6*(1), 1–25. https://doi.org/10.2307/2331554

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Udeozor, C., Abegão, F. R., & Glassey, J. (2024). Measuring learning in digital games: Applying a game-based assessment framework. *British Journal of Educational Technology*, *55*(3), 957–991. https://doi.org/10.1111/bjet.13407

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

Vidakis, N., Barianos, A. K., Trampas, A. M., Papadakis, S., Kalogiannakis, M., & Vassilakis, K. (2020). In-game raw data collection and visualization in the context of the "ThimelEdu" educational game. In H. C. Lane, S. Zvacek, & J. Uhomoibhi (Eds.), *Computer supported education. CSEDU 2019. Communications in computer and information science* (pp. 629–646, Vol. 1220). Springer International Publishing. https://doi.org/10.1007/978-3-030-58459-7_30

Wang, L. H., Chen, B., Hwang, G. J., Guan, J. Q., & Wang, Y. Q. (2022). Effects of digital game-based STEM education on students' learning achievement: A meta-analysis. *International Journal of STEM Education*, *9*(1), 26. https://doi.org/10.1186/s40594-022-00344-0

Wang, L., Shute, V., & Moore, G. R. (2015). Lessons learned and best practices of stealth assessment. *International Journal of Gaming and Computer-Mediated Simulations (IJGCMS)*, *7*(4), 66–87. https://doi.org/10.4018/IJGCMS.2015100104

Wang, M., & Zheng, X. (2021). Using game-based learning to support learning science: A study with middle school students. *The Asia-Pacific Education Researcher*, *30*(2), 167–176. https://doi.org/10.1007/s40299-020-00523-z

Wen, C.-T., Chang, C.-J., Chang, M.-H., Fan Chiang, S.-H., Liu, C.-C., Hwang, F.-K., & Tsai, C.-C. (2018). The learning analytics of model-based learning facilitated by a problem-solving simulation game. *Instructional Science*, *46*(6), 847–867. https://doi.org/10.1007/s11251-018-9461-5

Westera, W., Nadolski, R., & Hummel, H. (2014). Serious gaming analytics: What students' log files tell us about gaming and learning. *International Journal of Serious Games*, *1*(2), 35–50. https://journal.seriousgamessociety.org/index.php/IJSG/article/view/9/pdf_56

Widaman, K. F., & Helm, J. L. (2023). Exploratory factor analysis and confirmatory factor analysis. In H. Cooper, M. N. Coutanche, L. M. McMullen, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology: Data analysis and research publication* (2nd edition, pp. 379–410). American Psychological Association. https://doi.org/10.1037/0000320-017

Wongvorachan, T., He, S., & Bulut, O. (2023). A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. *Information*, *14*(1), 54. https://doi.org/10.3390/info14010054

Xing, W., Wadholm, B., & Goggins, S. (2014). Learning analytics in CSCL with a focus on assessment: An exploratory study of activity theory-informed cluster analysis. In *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge* (LAK 2014), 24–28 March 2014, Indianapolis, Indiana, USA (pp. 59–67). ACM. https://doi.org/10.1145/2567574.2567587

Yang, D., Zargar, E., Adams, A. M., Day, S. L., & Connor, C. M. (2021). Using interactive e-book user log variables to track reading processes and predict digital learning outcomes. *Assessment for Effective Intervention*, *46*(4), 292–303. https://doi.org/10.1177/1534508420941935

Yeo, I.-K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*(4), 954–959. https://doi.org/10.1093/biomet/87.4.954

Yu, J., Ma, W., Moon, J., & Denham, A. (2022). Developing a stealth assessment system using a continuous conjunctive model. *Journal of Learning Analytics*, *9*(3), 11–31. https://doi.org/10.18608/jla.2022.7639

Zhang, Q., & Rutherford, T. (2022). Grade 5 students' elective replay after experiencing failures in learning fractions in an educational game: When does replay after failures benefit learning? In *Proceedings of the 12th International Conference on Learning Analytics and Knowledge* (LAK 2022), 21–25 March 2022, online (pp. 98–106). ACM. https://doi.org/10.1145/3506860.3506873

Zheng, A., & Casari, A. (2018). *Feature engineering for machine learning: Principles and techniques for data scientists* (1st edition). O'Reilly Media. https://www.oreilly.com/library/view/feature-engineering-for/9781491953235/

Zhu, S., Guo, Q., & Yang, H. H. (2023). Beyond the traditional: A systematic review of digital game-based assessment for students' knowledge, skills, and affections. *Sustainability*, *15*(5), 4693. https://doi.org/10.3390/su15054693

1. Appendix A

A.1 Gameplay Metrics Construction

A.1.1 Size of explored game area

Feature Name In Dataset	Brief Description
Explored Area	A categorical variable, including 4 levels, describes the size of the explored area of unit 3's main game map. The higher the level, the larger the map size the student explored during gameplay.
DungeonExplored Area	A categorical variable, including 4 levels, describes the size of the explored area of unit 3's dungeon map. The higher the level the larger the map size the student explored during the game procedure.

Table 4. Brief descriptions regarding the features under the category of the size of the explored game area.

A.1.2 Speed of task completion

Feature Name In Dataset	Brief Description
AverageSpeed	A categorical variable, including 4 levels, describes the size of the explored area of unit 3's main game map. The higher the level, the larger the map size the student explored during gameplay.

Table 5. Brief descriptions regarding the features under the category of the speed of the task completion.

A.1.3 Tool using status

Feature Name in Dataset	Brief Description
Map.menu.node.freq	An integer variable that measures how many times a student opens the in-game tool, map, to seek for a path to fulfill tasks or solve puzzles.
Chat.log.menu.node.freq	An integer variable measures how often a student opens the chat log tool to review conversations that happened with in-game characters for important information extraction.
Crash.diagnostics.menu.node.freq	An integer variable that measures how many times a student opens the in-game tool, crash diagnostics, to figure out what issues happened on the spaceship mainly for fixing it and performing the side quests.
Quest.menu.node.freq	An integer variable that measures how many times a student opens the in-game tool, quest menu, to review requests for fulfilling the current quest.
Backing.info.menu.node.freq	An integer variable that measures how many times a student opens the in-game tool, backing information, to check the storyline and background information.
Help.menu.node.freq	An integer variable that measures how many times a student opens the in-game tool, help menu, to find out solutions for problems related to game operation.
Map.menu.node	A numeric variable that measures the student's average speed of checking the tool, map, to seek for a path to fulfill tasks or solve puzzles.
Chat.log.menu.node	A numeric variable that measures the student's average speed of checking the tool, chat log, to review conversations that happened with in-game characters for important information extraction.

Crash.diagnostics.menu.node	A numeric variable that measures the student's average speed of checking the tool, crash diagnostics, to figure out what issues happened on the spaceship mainly for fixing it and performing the side quests.
Quest.menu.node	A numeric variable that measures the student's average speed of checking the tool, quest menu, to review requests for fulfilling the current quest.
Backing.info.menu.node	A numeric variable that measures the student's average speed of checking the tool, backing information, to check the storyline and background information.
Help.menu.node	A numeric variable which measures the student's average speed of checking the tool, help menu, to find out solutions for problems related to game operation.

Table 6. Brief descriptions regarding the features under the category of the tool using status.

A.1.4 In-game items interactions

Feature Name in Dataset	Brief Description
U3...TOSS.SENSOR..POLLUTED	This data was collected during the quest which asks students for throwing out sensors and find out the pollutant river source based on the sensors' signals. The variable records the frequency of throwing out the sensor with the signal showing that this river area has a pollutant. This is an integer variable.
U3...TOSS.SENSOR.DOWNSTREAM	This data was collected during the quest which asks students for throwing out sensors and find out the pollutant river source based on the sensors' signals. The variable records the frequency of throwing out the sensor with the signal showing that the student throws sensors into the river's down area which is not in the search scope. This is an integer variable.
U3...TOSS.SENSOR.POLLUTED.SAME.AREA	This data was collected during the quest which asks students for throwing out sensors and find out the pollutant river source based on the sensors' signals. The variable records the frequency of throwing out the sensor with the signal showing that the student threw sensors into the river's area which has already been checked by previously thrown sensors. This is an integer variable.
U3...TOSS.SENSOR..SUCCESS	This data was collected during the quest which asks students for throwing out sensors and find out the pollutant river source based on the sensors' signals. The variable records the frequency of throwing out the sensor with the signal showing that the student throws sensors into the river's area which is not polluted. This is an integer variable.
U3...TOSS.SENSOR.DOWNSTREAM.CLEAN	This data was collected during the quest which asks students for throwing out sensors and find out the pollutant river source based on the sensors' signals. The variable records the frequency of throwing out the sensor with the signal showing that the student throws sensors into the river's down area which is clean and not in the searching scope. This is an integer variable.
U3...CRATE.THROW..FAIL	This data was collected during the quest which asks students for delivering crates into the correct river based on the river flow. The variable records the frequency of how many crates the player delivers to the wrong river. This is an integer variable.

U3...CRATE.THROW..SUCCESS	This data was collected during the quest which asks students for delivering crates into the correct river based on the river flow. The variable records the frequency of how many crates the player delivers to the correct river. This is an integer variable.

Table 7. Brief descriptions regarding the features under the category of in-game items interactions.

A.1.5 Argumentation-related gaming behaviors

Feature Name in Dataset	Brief Description
HOVERNODEFREQ	The variable records how many times the student hovered on a node, which will trigger out a popup text box showing the detail information associated with the node. This is an integer variable.
U3.CLAIM.II	The variable records the speed the student used to read a specific node, which is called "U3.Claim.ll" at this case. This is a numeric variable with the interval scaled from 1 to 4.
U3.EVIDENCE.B	The variable records the speed the student used to read a specific node, which is called "U3.Evidence.B" at this case. This is a numeric variable with the interval scaled from 1 to 4.
REASONING.3	The variable records the speed the student used to read a specific node, which is called "Reasoning.3" at this case. This is a numeric variable with the interval scaled from 1 to 4.
REASONING.4	The variable records the speed the student used to read a specific node, which is called "Reasoning.4" at this case. This is a numeric variable with the interval scaled from 1 to 4.
U3.CLAIM.I	The variable records the speed the student used to read a specific node, which is called "U3.Claim.l" at this case. This is a numeric variable with the interval scaled from 1 to 4.
REASONING.5	The variable records the speed the student used to read a specific node, which is called "Reasoning.5" at this case. This is a numeric variable with the interval scaled from 1 to 4.
U3.EVIDENCE.A	The variable records the speed the student used to read a specific node, which is called "U3.Evidence.A" at this case. This is a numeric variable with the interval scaled from 1 to 4.
REASONING.2	The variable records the speed the student used to read a specific node, which is called "Reasoning.2" at this case. This is a numeric variable with the interval scaled from 1 to 4.
REASONING.1	The variable records the speed the student used to read a specific node, which is called "Reasoning.1" at this case. This is a numeric variable with the interval scaled from 1 to 4.

Table 8. Brief descriptions regarding the features under the category of argumentation-related gaming behaviors.

A.1.6 Gaming logging event shares

Feature Name In Dataset	Brief Description
TriggerNumber	The logging system will record a triggering event when the student interacts with in-game items, such as boxes, river boarders, buttons, and so on. The variable is calculated by summing up all triggering events divided by the summation of the total number of events recorded during the whole game procedure.
MovementNumber	The logging system will record a movement event when the student presses the keyboard button of A, W, S, and D and move around in the game environment. The variable is calculated by summing up all movement events divided by the summation of the total number of events recorded during the whole game procedure.

MissionCompleteNumber	The logging system will record a mission-complete event when the student receives, makes progress to, or perform a quest or a task. The variable is calculated by summing up all mission-complete events divided by the summation by the total number of events recorded during the whole game procedure.
StateUpdateNumber	The logging system will record a triggering event when the student switch between different game scenes or does something that makes the game system update some data. The variable is calculated by summing up all state-update events divided by the summation by the total number of events recorded during the whole game procedure.
DialogueNumber	The logging system will record a dialogue event when the student triggers a dialogue box out and makes operations, such as making choices to different dialogue branches, pressing the button moving to the next dialogue or pressing the button moving to the previous dialogue for reviewing. The variable is calculated by summing up all dialogue events divided by the summation of the total number of events recorded during the whole game procedure.
ArfRelatedNumber	The logging system will record an ARF-related event when the student interacts with ARF(AI) panel to use in-game tools for seeking hints related to solutions. The variable is calculated by summing up all ARF-related events divided by the summation of the total number of events recorded during the whole game procedure.
HotkeyNumber	The logging system will record a Hotkey event when the student press hotkeys for checking in-game tools, such as mini map, quest reminder, dialogue records, and so on. The variable is calculated by summing up all hotkey events divided by the summation of the total number of events recorded during the whole game procedure.
ToggleNumber	The logging system will record a toggle event when the student uses a toggle board, a flyable skateboard to navigate in the game environment. The variable is calculated by summing up all toggle events divided by the summation of the total number of events recorded during the whole game procedure.
JumpNumber	The logging system will record a jump event when the student jumps in the game environment. The variable is calculated by summing up all jump events divided by the summation of the total number of events recorded during the whole game procedure.
ArgNumber	The logging system will record an argument event when the student makes progress or interactions in a 2D game scene for argumentation construction. The variable is calculated by summing up all argumentation events divided by the summation of the total number of events recorded during the whole game procedure.

Table 9. Brief descriptions regarding the features under the category of log event type shares.

A.1.7 Gaming performance assessment

Feature Name in Dataset	Brief Description
SeedPerformance	This data is collected when the student is asked to plant seeds in the correct locations based on water flow. The student will receive high performance when he or she plants less than two seeds into the wrong location, otherwise, he or she will be ranked as low performance. This is a categorical variable.

ArgumentLevel	This data is collected after students construct a complete argumentation during playing Unit 3 and submitting the results for feedback. There are 5 levels under this categorical variable. The higher the level, the better performance students receive in argumentation construction sessions. E.g. students who submit correct answers without any failed trial will reach the 5th level, and students who submit just one failed answer without any correct submission will receive the 1st level.
U2ArgumentLevel	This data is collected after students construct a complete argumentation during playing Unit 2 and submitting the results for feedback. There are 5 levels under this categorical variable. The higher the level, the better performance students receive in argumentation construction sessions. E.g. students who submit the correct answer without any failed trial will reach the 5th level, and students who submit just one failed answer without any correct submission will receive the 1st level.

Table 10. Brief descriptions regarding the features under the category of gaming performance assessment.

A.1.8 Dialogue Reading Statement

Feature Name in Dataset	Brief Description
DialogueSpeed	This variable is numeric. It saves the average speed students used to read dialogues. It is scaled into the interval from 1 to 4. The higher the value the slower students read dialogues.

Table 11. Brief descriptions regarding the features under the category of dialogue reading statement.

A.1.9 Other Information

Feature Name in Dataset	Brief Description
TeacherId	This variable reflects which teacher leads or guides the student to play the game.
Trial	It represents how many times the student replays the game or repeats the same quests or tasks.

Table 12. Brief descriptions regarding the features under the category of other information.

A.2 Embedded Assessment Score Description

A.2.1 Embedded assessment score related to previous units

Feature Name in Dataset	Brief Description	Formula to Calculate the Score
TutorialArgScore	This EA score is calculated after the student finding out a proper claim during the argument construction tutorial quest that happened in Unit 1	1 point for correct submission on 1st attempt; 0 points for anything else.
BiggerArgScore	This EA score is calculated after the student forms a proper argumentation that clarifies which watershed is bigger than the other during the argumentation quest that happened during Unit 2.	2 points for finding correct answers within 3 attempts; 1 point for finding out correct answers within 4 attempts; 0 points for no answer finding for more than 4 attempts.
UpStreamArgScore	This EA score is calculated after the student forms a proper argumentation that clarifies where the pollutant source is during the argument construction happening in Unit 3.	2 points for finding correct answers within 3 attempts; 1 point for finding out correct answers within 6 attempts; 0 points for no answer finding for more than 6 attempts.

CREIScore	This EA score is calculated when the student enters into an environment where an avatar asks them to figure out a complete argumentation structure. This quest happens in Unit 2.	1 point for each correct choice, -0.33 points for each incorrect choice.
JasperCritiqueScore	This EA score is calculated when the student triggers out a dialogue box, chatting with an avatar named Jasper, and needs to make a choice to decide if Jasper's critique is correct or not. This quest happens in Unit 2.	1 point for selecting "you forgot evidence; 0 points for either "Jasper you are right; or "Jasper you forgot the claim."
FindTeamAveScore	This EA score is calculated when the student needs to use an in-game mini-map to figure out the location based on topologic characteristics. This quest happens in Unit 2.	0.5 points for opening the map; 1 point for finding the team in 3 minutes or less; 0 points for anything else.

Table 13. Brief descriptions and calculating formula regarding the features under the category of embedded assessment score related to previous units.

A.2.2 Embedded assessment score related to current unit

Feature Name in Dataset	Brief Description	Formula to Calculate the Score
PlantScore	This EA score is calculated when the student needs to figure out where to plant seeds based on pumps' locations along the pollutant river. This quest happens in Unit 3.	1 point for Selecting a correct pump location; -1/2 points for selecting an incorrect pump location.
CrateDeliveryScore	This EA score is calculated when the student needs to choose the correct river for delivering crates to an avatar Sam based on the river flow. This quest happens in Unit 3.	1 point for correct crate placement.

Table 14. Brief descriptions and calculating formula regarding the features under the category of embedded assessment score related to current units.

A.3 External information

A.3.1 Pretest and post-test outcomes

Feature Name in Dataset	Brief Description
U3PrePerformance	Pretest score related to Unit 3's curriculum knowledge
U3PostPerformance	Post-test score related to Unit 3's curriculum knowledge

Table 15. Brief descriptions regarding the features under the category of external information.

2. Appendix B

As described in the game context and data collection section, Mission HydroSci (MHS) is a 3D game-based learning environment designed and developed associated with comprehensive and sophisticated curriculum integration. Each key quest of MHS can be seen as an efficient marker of learning achievement.

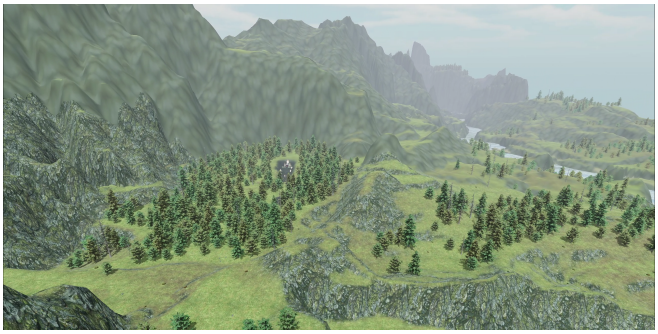
Table 16 displays the game unit, what quests are included in a specific unit, and a detailed description corresponding to one quest. As you can see, there is one quest in Unit 1, five in Unit 2, and five in Unit 3, eleven in total. Notably, we mentioned previously that there are 12 quests involved in this study and defined in the task model of the Evidence-Centered Design (ECD) approach, which is because the quest of collecting samples from eastern and western waterfalls is actually two successive quests.

Figure 1 shows several example screenshots displaying how the game world or surroundings look like when students were engaging in completing some key quests mentioned above.

Game Unit	Quest Name	Detailed Description
Unit 1	Tutorial unit	In this unit, students will talk to each key non-playable character, learn open and get familiar of each in-game tool through menus or hotkeys, know ways to navigate in the game world, and know about the interface of the argumentation system to understand how to construct a complete scientific argument
Unit 2	Find the team	After crash-landing on a new planet, the main character (controlled by the player) must locate the rest of the crew. To accomplish this, they must interpret the topographic map and carefully observe their surroundings. However, there is no wayfinding assistance provided during this quest.
	Collect samples from eastern and western waterfalls	Based on the conversations with NPCs, players need to find the positions of the eastern and western waterfalls. By investigating the samples of two waterfalls, players need to collect appropriate evidence describing the characteristics of each waterfall. In this way, they could deduce the conditions of each waterfall's watershed and prepare later scientific argumentation or debate with NPCs.
	Argue which watershed is bigger	Dr. Toppo (one of the NPCs) will invite players to the argumentation system to construct a complete argument that makes sense with collected evidence from the waterfalls. The argumentation system mimics the solar system, where the claim works as the sun, and reason and evidence work as planets around the sun. The planets represent evidence position in the further interstellar orbit than the planets representing reasons. Players need to choose the correct claim, reason and evidence from available choices displayed in the left corner of the system.
	Jasper's proposal	Through a conversation with Jasper (Another NPC), you will debate with him to determine if his proposal about the new place is logical with the information the player collected from today's environment.
	CREI system	To fix the system of the AI ARF, players will enter into a system called CREI to practice the definitions of three components of scientific argumentation. Players will see a screen showing different sentences, and they need to judge which component the sentence represents by throwing balls in the direction showing the correct component.
Unit 3	Sam's supplies	Players will meet Samantha (NPC) at her garden base as she is just starting. To help her build up the garden, players need to transport supplies to Sam's Garden base through the river. Players must deliver 4 crates to the river stream to finish the quest. There are two river streams where players must investigate their water flows to decide which is the correct stream to transport. After players deliver each crate to a certain river stream, a dialogue will show the feedback on whether the stream is correct.

Game Unit	Quest Name	Detailed Description
	Collect pumps from the alien ruins	After finding the pollutant source, Sam told us she found a huge tree near an intersection of the river branches and doubted that some river branches were also polluted by the battery core. To ensure her thought, we need to enter into an alien ruin to collect pumps that allow us to plant Sam's plant seeds into the mini gardens along the river to test which branch was polluted. Players must apply what they learned regarding water flows to unlock those pumps and solve puzzles within the alien ruin. The general format of the puzzle is to find and carry a cube from the surroundings, put it into the water channel, and guide it to the destination by managing the water flow direction through a controlling panel.
	Trace the source of pollutant	After receiving the supplies, Sam found the river is polluted. She provides players with sensors which will light red when the river spot is polluted and green when it's clean. Players need to take advantage of the sensors and investigate the characteristics of the river, such as water flow direction, whether in a river branch or its surrounding environment, to find the source of the pollutant, which is a crashed battery core.
	Plant seeds	After getting the pumps, players can plant Sam's seeds into the garden along the river to trace how the dissolved pollutant materials spread along the river flow. Players need to observe the river conditions to judge which mini garden to plant to accurately trace the dissolved pollutant materials' flow direction. Each time players plant the seed will trigger a dialogue showing Sam's feedback regarding whether the mini-garden is polluted.
	Convince Bill the pollutant is nearby	After finding the position of the battery core, Bill (NPC) will invite us to enter the argumentation system to construct a complete scientific argument to convince him where the battery core is. The players must choose the correct reasoning to connect the pre-decided evidence and claim logically within the system.

Table 16. Detailed description for each game quest involved in the task model of the Evidence-Centered Design (ECD) approach.

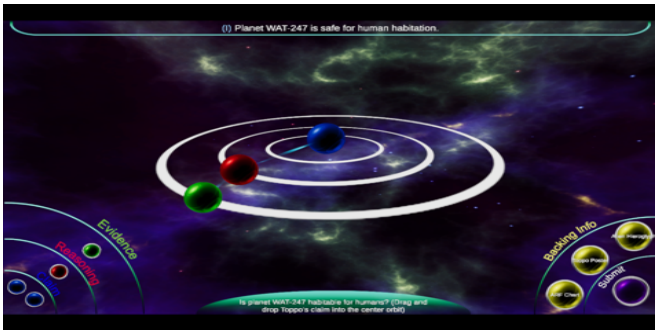


(a) Students just arrived at a new planet, trying to find other team members.

(b) Students are picking up the crate for later delivery.

(c) Students just delivered a crate and Sam gave feedback regarding the delivery.

(d) Students just found the source of the pollution based on the sensors reflection



(e) Students are constructing an argument within the argumentation system

(f) A giant tree representing the source of dissolvable materials, after this scene, students need to find how the dissolvable materials spread along the river by installing pumps.

(g) Students installed the pumps in one spot along the river. Sam gave feedback according to the result.

Figure 1. example screenshots about part of quests involved in the task model of Evidence-Centered Design (ECD).

3. APPENDIX C: Identifying Optimal Features: Methods and Materials

This appendix offers detailed descriptions of each of the subcategories within each feature category. We also present extensive visualizations and descriptions that delineate the outcomes of our exploratory analyses and descriptive statistics pertaining to the engineered features. This supplementary information further expands upon the procedures outlined in Section 4.2.

C.1 Detailed descriptions of feature subcategories

Based on the high-level feature categories described within the main text of this article, we continue to subcategorize each of them. For gameplay behaviors, we have the following feature subcategories:

1. **Size of explored game area**: it describes how large a student went around or explored in different game maps.
2. **Speed of task completion**: It represents how fast a student completes a game task.
3. **Tool using status**: It depicts in-game tools using status.
4. **In-game item interactions**: It depicts how students interact with in-game items.
5. **Argumentation construction**: It includes how students perform when constructing a complete and reasonable scientific argumentation.
6. **Event type shares**: There are ten logging events describing different in-game behaviors students spent their time on. It describes how students distribute or allocate their game time to some extent.
7. **Dialogue reading behaviors**: it depicts how frequently and fast a student reads dialogues to receive helpful information or promote game progress.
8. **Game replay times**: It represents how often students repeatedly replayed the same content.

For embedded assessment scores, we have the following subcategories:

1. **Previous embedded assessment scores**: It contains scores measuring the learning outcomes of subject-matter knowledge marked by game logs collected from previous units.
2. **Current embedded assessment scores**: It includes scores having the same function as described above but collected in the current unit.

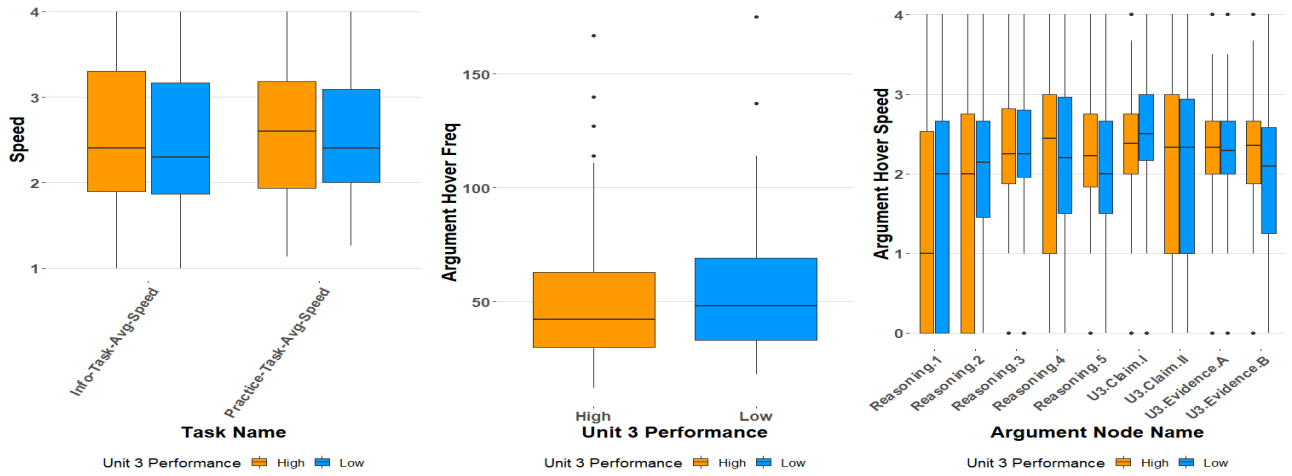
For external information, the feature subcategories are shown in the following:

1. **Demographic information**: It includes students' general information, such as gender, age, and social-economic status. Considering content limitations, we did not include students' demographic information in this study.
2. **Assessment scores**: It contains assessment scores reflecting students' expert levels of subject-matter knowledge, which we want them to learn from the game. External instruments collect all scores under this subcategory.
3. **Other external information**: This subcategory contains all other information related to participating students except the aforementioned information. It could be information about students' personalities, learning styles, or gamer types. It also could be some streams of information collected from high-frequency sensors, such as eye trackers, motion sensors, or motion detectors. Given that the primary goal of this article is to investigate the potential of game logs to predict students' learning outcomes, we do not include any feature under this category.

C.2 Exploratory Analysis and Descriptive Statistics Regarding Generated Features

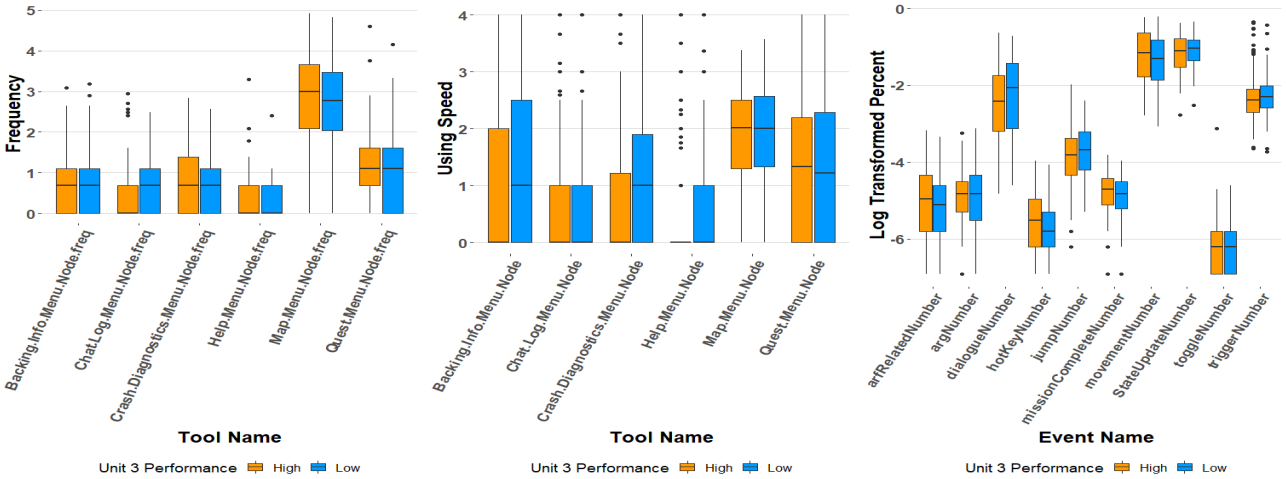
The relationship of each feature to the target learning outcome directly or indirectly affects the prediction model's performance, influencing our feature engineering and the predictive power of each feature. We conducted an exploratory analysis to comprehend how each feature might predict the target learning outcome. Given that the data types of the selected features can be categorized into numeric and categorical, we selected two statistical methods appropriate to these types: the Mann-Whitney U test (Mann & Whitney, 1947) for numeric features and Fisher's Exact test (Fisher, 1935) for categorical features. These nonparametric tests are designed to handle features that do not conform to statistical assumptions about their probability distributions, and they produce robust and accurate results even when the classification groups are small and unbalanced. Additionally, to further validate the results of the above nonparametric tests and help audiences understand the practical significance of the results, we also calculated the Effect Size for both nonparametric tests. Specifically, Rank-Biserial Correlation (Glass, 1966) for the Mann-Whitney U test, while Cramer's V (Cramér, 1999) for Fisher's Exact test.

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License (CC BY 4.0)



(a) Task average completing speeds **(b)** Arg-nodes hovering frequency **(c)** Arg-nodes hovering speeds

Figure 2. The distributions of completion speeds, argumentation hovering frequency, and argumentation hovering speeds.



(a) Tool using frequency **(b)** Tool using speeds **(c)** Event type percentages

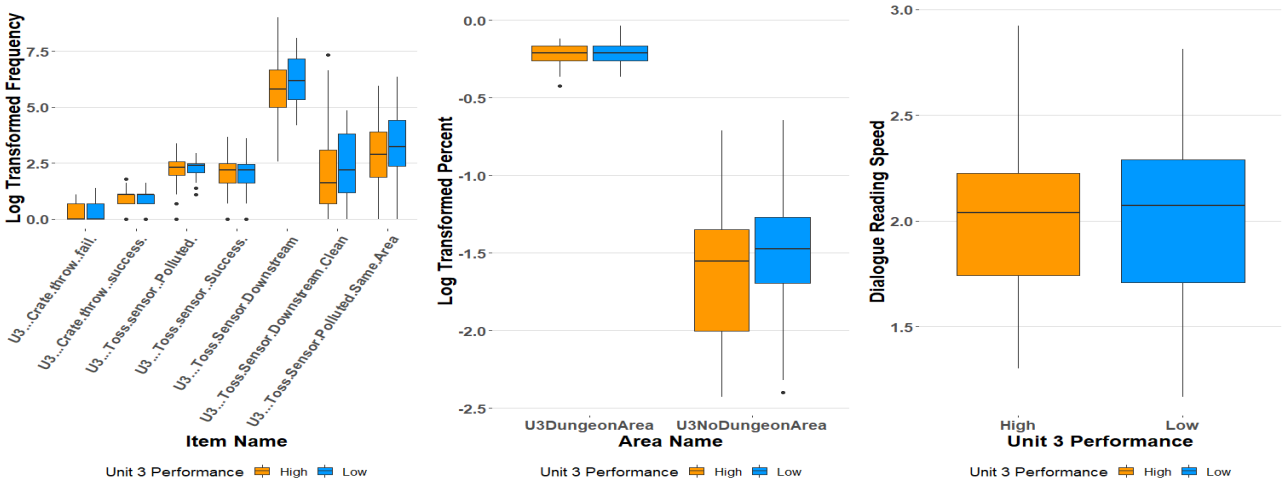
Figure 3. The distributions of tool use frequency, tool use speed, and event type percentages.

In this section, we examine the validity of both numeric and categorical variables against our categorical dependent variable, Unit 3 Post-Test Performance. The following subsections are labeled ”Categorical vs. Numeric” and ”Categorical vs. Categorical,” respectively. This exploratory data analysis aids us in identifying and iterating on the features with the greatest predictive potential for subsequent model construction.

Categorical V.S. Numerical Figures 2, 3, 4, and 5 display the distributions of relevant numeric features, color-coded by Unit 3’s post-content assessment scores. The high-score group is represented in orange, and the low-score group in blue. Table 17 enumerates numeric features that demonstrate statistically significant differences between the two performance groups.

Dependent Variable	Independent Variable	Mann-Whitney U Value	Effect Size (Rank-biserial correlation)
U3 Post Performance	Node-evidence-b-hover-speed	12247*	0.16
	Node-claim-I-speed	9266.	-0.13
	Node-reasoning-1-speed	9352.5.	-0.12
	Tool-chat-log-freq	9341.	-0.12
	Tool-map-freq	9419.5.	-0.11
	Tool-crash-diagnostics-speed	9317.	-0.12
	Tool-backing-info-speed	9417.	-0.11
	Event-dialogue	9229.5*	-0.13
	Event-Arf-tool	11758.5.	0.11
	Event-hotkey	12256**	0.16
	Item-freq-down-stream-sensor	9002.5*	-0.15
	Item-freq-polluted-sensor	8915.5*	-0.16
	Item-freq-crate-fail	9075.5*	-0.14
	Embedded-score-upstream-arg	12025*	0.14
	Embedded-score-CREI	12162.5**	0.16
	Embedded-score-find-team	11879.5*	0.12
	Embedded-score-Jasper	11799*	0.11
	Area-size-non-dungeon-area	8969.5*	-0.15

Table 17. Mann-Whitney U Testing results related to independent numeric variables. The mark ".", "*", "**", and "***" following the testing result value represents the significance degree reflected by p-value. "." indicates the p-value is less than 0.1 but larger or equal to 0.05; "*" indicates the p-value is less than 0.05 but larger or equal to 0.01; "**" indicates the p-value is less than 0.01 but larger or equal to 0.001; "***" indicates the p-value is less than 0.001. The Effect Size (Rank-biserial correlation ranges from -1 to 1 where: 1 indicates a perfect positive association; -1 indicates a perfect negative association; 0 indicates no association. An absolute effect size value of 0.1 or greater suggests a potentially meaningful association between the dependent and independent variables warranting further analysis.



(a) Item interaction frequencies **(b)** Area exploration sizes **(c)** Dialogue reading speeds

Figure 4. The distributions of item interaction speed, area exploration, and dialogue reading speed.

Integrating our analysis from Table 17 and Figures 2, 3, 4, and 5, we can deduce the following:

1. High-performing students, those with high post-content assessment scores, tend to read nodes (such as Reasoning.1 and Claim. I) significantly faster than their peers, but they spend considerably more time on the Evidence B node.

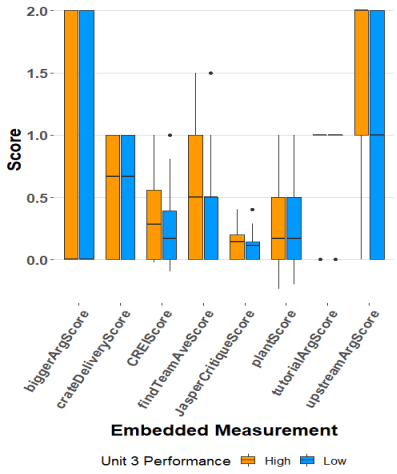
2. These high-performing students use the map tool significantly more frequently than their lower-performing peers. However, they infrequently open the chat log, which is more commonly used by the low-performing group. They also spend less time using tools that display background game information and the spaceship's crash diagnostics information.

3. High-performing students significantly outpace the low-performing group in utilizing the game's "hot keys", such as in-game tools and references to Arf, the in-game AI. These tools provide helpful information for solving puzzles and completing tasks. Conversely, low-performing students generate more dialogue events during their gameplay.

4. High-performing students are more accurate in the game tasks, delivering fewer boxes to the wrong river and throwing fewer sensors into polluted areas than low-performing students.

5. High-performing students achieve significantly higher scores on embedded assessments, particularly in tasks related to completing Unit 3's argument construction, understanding the complete argument structure during the CREI system, selecting the correct answer when defending their opinion to Jasper, and efficiently finding the team location using the MHS map tool.

6. Interestingly, high-performing students cover less area in the main game scene of Unit 3 than their low-performing counterparts.



(a) Embedded assessment scores

Figure 5. The distributions of embedded assessment scores.

Although these observations lack statistical significance, the patterns revealed in Figures 2, 3, 4, and 5 provide additional potential insights for distinguishing high- and low-performing students:

1. High-performing students generally spend more time completing tasks, regardless of whether these tasks involve acquiring new knowledge or applying and practicing previously learned information.

2. High-performing students hover less frequently over argument text nodes than their low-performing counterparts.

3. They spend more time reading argument nodes associated with Reasoning 4 and 5, which represent the two most potentially correct answers within the argumentation system.

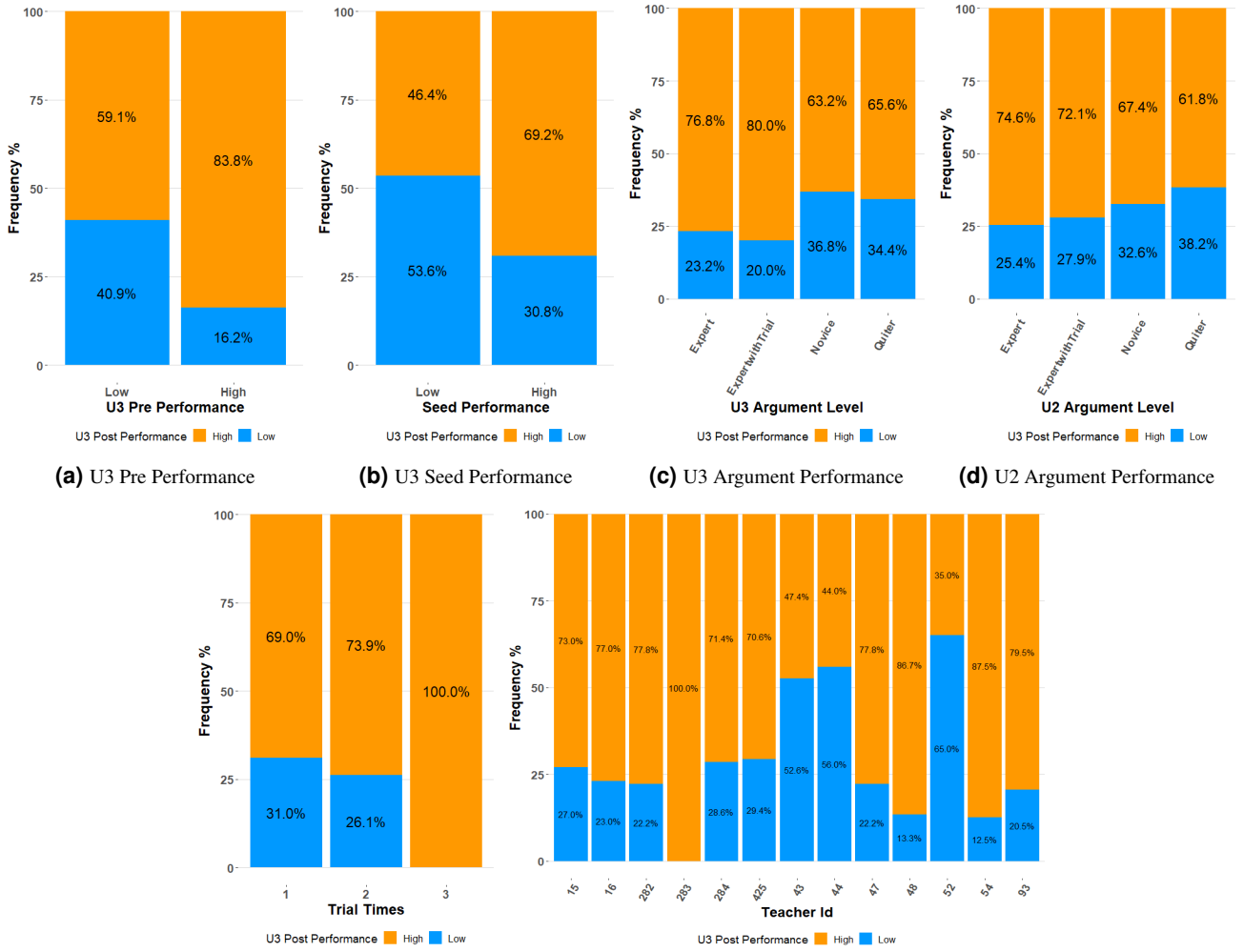
4. High-performing students tend to scan the "Help" tool more swiftly, dedicating more time to the tool that provides detailed quest information.

5. They generate a larger share of log events related to mission completion and movement within game areas, but a smaller share related to interactions with in-game items.

6. High-performing students typically achieve higher, or in some cases equal, embedded assessment scores compared to low-performing students.

7. Lastly, high-performing students demonstrate quicker reading speeds, as indicated by the reduced time spent on reading dialogues.

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License (CC BY 4.0)



(a) U3 Pre Performance **(b)** U3 Seed Performance **(c)** U3 Argument Performance **(d)** U2 Argument Performance

(e) Game replay times **(f)** Teacher Ids

Figure 6. The distributions of categorical features

Categorical V.S. Categorical Figure 6 illustrates the distributions of the relevant categorical features. From this figure, we observe that high-performing students typically achieve higher pre-test scores for the content of Unit 3 and exhibit superior seed-planting performance, implying that they planted three or more seeds in locations where growth is likely. While these high-performing students may commit occasional errors when planting the first seed, they also demonstrate better performance in argument construction and exhibit a higher frequency of game replay. Notably, Subfigure (f) of Figure 6 suggests that the teacher employing MHS impacts student performance.

Table 18 further shows that the features of Unit 3's pre-content assessment scores, seed performance, argumentation performance, and the classroom teacher exhibit significant differences between the two performance groups.

Dependent Variable	Independent Variable	Fisher's-Exact Value	Effect Size (Cramér's V)
U3 Post Performance	U3 pre-performance	21.75***	0.26
	seed performance	17.267***	0.18
	U3 argument performance	7.6.	0.15
	Teacher Id	33.241***	0.32

Table 18. Fisher's Exact Testing results related to independent categorical variables. The mark ".", "*", "**", and "***" following the testing result value represents the significance degree reflected by p-value. "." Indicates the p-value is less than 0.1 but larger or equal to 0.05; "*" indicates the p-value is less than 0.05 but larger or equal to 0.01; "**" indicates the p-value is less than 0.01 but larger or equal to 0.001; "***" indicates the p-value is less than 0.001. The Effect Size ranges from 0 to 1 where: 1 indicates a perfect association; 0 indicates no association. An effect size value of 0.1 or greater suggests a potentially meaningful association between the dependent and independent variables warranting further analysis.

4. Appendix D: Computational Model Selection and Training Preprocessing

D.1 Model Training Preprocessing

As delineated in Section 4.1, our target variable, the post-test score, is transformed into a binary variable. A score of 1 represents high performance, and 0 represents low performance, which posits the modeling task as a classification problem. Notably, the target variable exhibits class imbalance: 251 students are categorized as high performers, while 103 students are deemed low performers. In classification problems, such imbalance can adversely affect model performance, particularly when the problem is complex (Guo et al., 2008; Elrahman & Abraham, 2013; Buda et al., 2018). A powerful technique for addressing this issue during data preprocessing is subsampling (Kaur et al., 2019). We selected widely used and extensively validated subsampling algorithms, such as down-sampling, up-sampling, SMOTE, and ROSE. Besides their proven track record, these techniques preserve data integrity while filtering out noise. Moreover, up-sampling, SMOTE, and ROSE effectively ensure that individuals, who belong to smaller demographic groups and are thus not easily discernible, are still identifiable in our results reporting.

1. **Down-sampling**: randomly subset all the classes in the training set so that their class frequencies match the least prevalent class, which is the low-level class in our case.

2. **Up-sampling**: randomly sample (with replacement) the minority class, the low-level class, to be the same size as the majority class, the high-level class.

3. **SMOTE**: an algorithm to over-sample minority class, low-level class, by creating synthetic examples, follows the formula of $(x_{new}, y_{new}) = (x_{select}, y_{select}) + rand(0,1) * (diffx_{neighbor}, diffy_{neighbor})$, in which the *rand(0,1)* is a random number between 0 and 1.

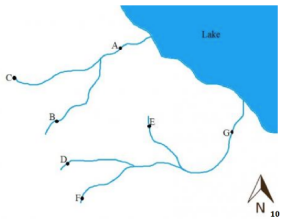
4. **ROSE**: an algorithm generates new artificial data from the classes using a smoothed bootstrap approach. The artificial data is generated based on neighbors of the centered observation with a set of unimodal, symmetric probability distributions. It combines techniques of oversampling the minority class, the low-level class, and the undersampling majority class, the high-level class.

Each of the four subsampling techniques is applied individually to the training dataset. Subsequently, model performance is evaluated using the testing dataset. However, for conciseness and clarity, we only report results pertaining to the subsampling technique that yielded the highest model performance.

5. Appendix E: Detailed Description of the Assessment Instrument and Corresponding Factor Analysis Process

E.1 Assessment Instrument: Water Flow Dynamics of Water Science Content Knowledge

For the following items, refer to the image below which represents an overhead view of a watershed



14. Suppose there is pollution at Site D, which location would you expect would become polluted?

- Site F
- Site E
- Site G
- Site B

Figure 7. First item of the assessment instrument.

15. Based on the watershed map, which location would you predict to have the lowest elevation?

- Site D
- Site F
- Site E
- Site G

Figure 8. Second item of the assessment instrument.

For the following items, refer to the image below, which highlights the Mississippi River watershed, which is outlined by a dashed line.

17. The image above highlights the Mississippi river watershed. If the red dots indicate polluted water and the blue dots indicate clean water, where is the most likely source of the pollution?

- A
- B
- C
- D

Figure 9. Third item of the assessment instrument.

E.2 Factor Analysis Process and Results

Exploratory Factor Analysis (EFA) is a data-driven statistical technique used when we do not have a preconceived notion of how many factors (latent variables) underlie the observed data. The goal of EFA is to discover the underlying structure of relationships between the variables (items) and to group these items into factors that represent underlying latent constructs. Confirmatory Factor Analysis (CFA) is a theory-driven statistical technique used when we have a predefined hypothesis about the factor structure of the data. The purpose of CFA is to confirm whether the data fits the hypothesized factor model. CFA requires us to specify the number of factors and which items are expected to load on each factor before running the analysis.

EFA and CFA are complementary techniques that can work together to validate the factor structure of a set of variables. EFA helps explore the underlying structure of the data and identify potential factors, which will be applied first. CFA will be applied after EFA to verify the identified factor structure with model fit indices and loadings.

E.2.1 Exploratory Factor Analysis

The first thing we did for EFA was to determine how many components should be generated based on the data for the factor analysis, which involved drawing a Scree plot. Figure 10 is the scree plot we drew based on the assessment data.

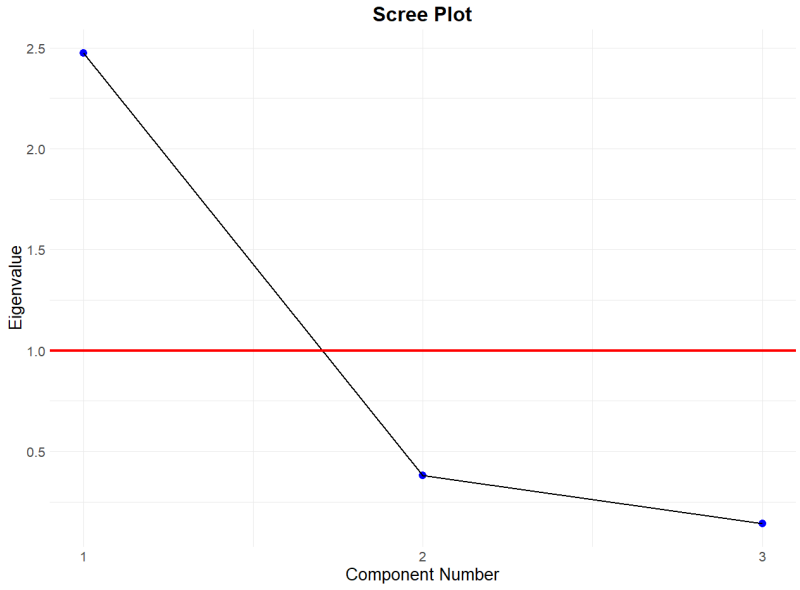


Figure 10. Scree plot for component number determination

The figure shows that one component is the most appropriate for the factor analysis based on the data since the first factor's eigenvalue is 2.1, much larger than the threshold value of 1.0. The other two factors have eigenvalues much less than the threshold value, which further supports the decision that one component is most appropriate for the EFA for our study.

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License (CC BY 4.0)

With the determination of the component number, then we conducted EFA with our assessment data. The result is shown within Table 19.

	MR1	H2	U2	Com
Item 1	0.73	0.53	0.47	1
Item 2	0.85	0.72	0.28	1
Item 3	0.71	0.50	0.50	1

Table 19. Exploratory Factor Analysis

MR1 represents the factor loadings, which indicates how strongly each item correlates with the underlying factor. In our case, there is one extracted factor (MR1), which represents the latent construct, representing water flow dynamics. Item 1 has a strong correlation with the factor with a factor loading of 0.73. Factor loading of 0.85, which is very high, indicating that Item 2 strongly reflects the latent construct. Item 3 has a factor loading of 0.71, also indicating a strong correlation with the factor. In summary, all three items have factor loadings greater than 0.7, showing that they all have strong relationships with the latent factor. Factor loadings above 0.7 typically indicate that the items are highly representative of the underlying construct.

H2 indicates commonalities, representing the proportion of variance in each item explained by the factor. It is calculated by squaring the factor loadings. For Item 1, the commonality is 0.53, meaning that the factor explains 53% of the variance in Item 1. Similarly, the factor explains 72% and 50% of the variance in Items 2 and 3, respectively. Generally speaking, the communities indicate that the factor explains a substantial portion of the variance in all three items. Item 2 has the highest commonality, suggesting that it is the most strongly related to the factor, while Item 3 is slightly less strongly related.

U2 reflects uniqueness, representing the proportion of variance in each item that is not explained by the factor. It reflects that error variance or the item-specific variance. For Item 1, uniqueness of 0.47 means that the factor does not explain 47% of the variance in Item 1 and is unique to the item. For Item 2, 28% of the variance in Item 2 is not explained by the factor. Item 3 has 50% of variance is not explained by the factor. Summarily, Items 1 and 3 have relatively higher uniqueness values, suggesting that a moderate portion of their variance is unique and not explained by the factor. In contrast, Item 2 has a relatively low uniqueness, meaning it is strongly explained by the factor.

Com represents complexity, which indicates how many factors each item loads on. A complexity value of 1 means that the item loads on a single factor. All three items have a complexity value of 1, meaning that they load exclusively on a single factor. This supports the conclusion that the items are unidimensional and measure a single latent construct.

E.2.2 Confirmatory Factor Analysis

Since the primary purpose of this study is to explore the latent structure of the assessment items, EFA was initially employed to identify the underlying factors. To validate the structure identified through EFA, CFA was conducted as a complementary, post-hoc analysis. Given that the factor loadings and structure were already explored during the EFA stage, the focus of CFA is on verifying the model fit. Therefore, in this section, we present the model fit indices from CFA to confirm that the hypothesized factor structure aligns with the observed data. These indices provide a succinct and clear indication of the model's validity, without redundant reporting of factor loadings already examined in EFA. Table 20 presents the results of the model fit indices.

Fit Index	Value
Chi-square	4.23 (p = 0.12)
RMSEA	0.06
CFI	0.93
TLI	0.92

Table 20. Model fit indices for confirmatory factor analysis

- **Chi-square**: The chi-square value of 4.23 with a p-value of 0.12 indicates a non-significant result ($p > 0.05$), which is considered desirable in CFA. A non-significant chi-square suggests that there is no significant difference between the model's predicted values and the observed data, meaning the model fits well.
- **RMSEA (Root Mean Square Error of Approximation)**: An RMSEA value of 0.06 indicates reasonable model fit, as values below 0.05 are generally considered to reflect a close fit to the data. Values between 0.05 and 0.08 indicate reasonable fit, while values above 0.10 suggest poor fit.
- **CFI (Comparative Fit Index)**: A CFI value of 0.93 reflects a good model fit. CFI values range between 0 and 1, with values above 0.95 indicating excellent fit, 0.90 to 0.95 indicating a good or acceptable fit, 0.85 to 0.90 representing marginal fit, and values below 0.85 indicating poor fit.

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License (CC BY 4.0)

- **TLI (Tucker-Lewis Index)**: The TLI value of 0.92 further supports the model's reasonable fit. Similar to the CFI, a TLI above 0.95 is indicative of an excellent fit, while values between 0.90 and 0.95 suggest a good fit. Values below 0.90 may signal a need for model improvement, particularly in terms of parsimony.

The CFA fit indices presented in Table 20 suggest that the hypothesized model fits the data very well. These results provide strong evidence that the factor structure identified through EFA is well-supported by the data when validated through CFA, making this model a robust representation of the underlying construct.

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License (CC BY 4.0)