LEARNING ANALYTICS AND PSYCHOPHYSIOLOGY: UNDERSTANDING THE LEARNING PROCESS IN A STEM GAME

A Dissertation presented to the Faculty of the Graduate School at the University of Missouri-Columbia

---

In Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy

---

by JOSEPH GRIFFIN
Dr. James Laffey and Dr. Joi Moore, Dissertation Co-Supervisors
MAY 2023

The undersigned, appointed by the dean of the Graduate School, have examined the dissertation entitled LEARNING ANALYTICS AND PSYCHOPHYSIOLOGY: UNDERSTANDING THE LEARNING PROCESS IN A STEM GAME presented by Joseph Griffin, a candidate for the degree of Doctor of Philosophy of Information Science and Learning technologies, and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. James Laffey

____

Dr. Joi Moore

____

Dr. Rose Marra

____

Dr. Sean Goggins

____

Acknowledgements

I extend my deepest gratitude to my co-chairs, James Laffey and Joi Moore, for their invaluable guidance and unwavering support. Their expertise and constructive criticism have shaped the direction of my research and strengthened its quality. Special thanks to Rose Marra, my supervisor, for her exceptional patience and encouragement.

I am grateful to Sean Goggins for his valuable insights and contributions to my research. I also want to express my heartfelt appreciation to Christi Bergin for her invaluable mentorship and guidance throughout my research efforts.

To my wife and children, thank you for your patience, motivation, and unwavering support. To my family and friends, your encouragement and belief in me have meant the world. I am truly grateful for the love, understanding, and encouragement that have sustained me throughout this journey.

I am honored to have worked with exceptional individuals and to have been surrounded by a supportive network. My heartfelt gratitude goes out to each and every one of you for your contributions and unwavering belief in my abilities. Thank you for being an integral part of my doctoral journey.

Table of Contents

Figures

Tables

LEARNING ANALYTICS AND PSYCHOPHYSIOLOGY: UNDERSTANDING THE LEARNING PROCESS IN A STEM GAME

JOE GRIFFIN

Dr. James Laffey and Dr. Joi Moore Dissertation Co-Supervisors

Abstract

This study focuses on the exploration of player experience in educational games and its potential impact on predicting learning outcomes. Specifically, the research aims to investigate the connection psychophysiology data, obtained through a summative study involving nine participants, and the results of a learning analytics model derived from a larger field test. The study incorporates eye tracking and electrodermal activity data to gain insights into the predictive power of this data.

Through the analysis of player experience data, the study sheds light on the factors that contribute to effective educational game design. By examining the eye tracking and EDA data, the researchers explored the participants' engagement levels, attention patterns, and emotional arousal during gameplay. These findings revealed a connection between spikes of visual attention and EDA during interactions with character faces as well as in game cinematics.

In conclusion, the outcomes of this study provide valuable insights for future educational game designers. By understanding the relationship between user experience indicators and learning analytics, designers can tailor game elements to enhance engagement, attention, and emotional arousal, ultimately leading to improved learning outcomes. The integration of eye tracking and EDA data in user experience studies adds a new dimension to the evaluation and design of educational games. The findings pave the way for future research in the field and highlight the importance of considering user experience as a crucial factor in educational game design and development.

Chapter 1: Introduction

Overview

We are currently in the Golden Age of Brain Research (Kaku, 2014), as we see substantial advances in understanding the biological processes of learning. Being able to measure functions of the brain at increasingly finer grained detail, makes reducing the noise to signal ratio caused by the context surrounding a learning experience more critical for accurate interpretation than ever. Serious Games, due to their inherent nature of being deep learning environments (Gee, 2003) while also being extremely controlled digital environments, are a rich context for insightful discoveries during this era of learning research. The current problem is that assessment of learning in games comes primarily from two top-down places: curricular designers and game designers, in the forms of pre and posttest assessments and embedded assessments respectively. This study will be a preliminary examination, triangulating current pre-post and embedded assessment standards with more player-centric psychophysiology measures. The findings from this study will contribute to the fields of learning analytics and psychophysiology by building a better understanding of the interconnections between the two and the systems necessary for data collection and analysis

To get a better understanding of when and where learning happens in various data rich digital environments, Learning Analytics (LA) have been developed (Kinnebrew, Loretz, & Biswas, 2013), (Harley, Trevors, Azevedo, & others, 2013), (Winne & Baker, 2013). These computational methods can be used to identify which patterns of virtual behaviors are associated with high or low performance outcomes (Xing, Guo, Petakovic, & Goggins, 2015). Although these insights can shed light on students' progress in otherwise foreign environments to most teachers; they are still limited to the behaviors trackable as logs, and possibly by the inputs chosen for analysis. These behavior patterns cannot tell us how this learning happens for an individual or why those specific game features may elicit a specific result for some members of the population, but not others.

However, by combining Learning Analytics with in-depth player experience analysis including Psychophysiology data, we can start to get a more holistic view of the students' progression. The moment-to-moment analysis of the arousal state of players begins to help us shed light on exactly what features within the game curricula had what specific impact for which players (Chanel, Rebetez, Bétrancourt, & Pun, 2011), and how all those factors impact learning overall (Morgan, 1951).

In this exploratory project I brought in nine users for three 2-hour sessions each (27 sessions total) to play Mission HydroSci (MHS), a STEM game currently being developed to teach water science and scientific argumentation. Using data collected from those sessions, I experimentally explored the arousal state of players in moments of student learning as well as exploring player behaviors during moments of high and low arousal. Understanding the relationship between arousal states, game moments, and assessment performance can help both designers and educators enact their instruction better.

Serious Games

Serious Games refers to any game, video, analog, or otherwise, created for a purpose other than entertainment (Roepke et al., 2015). For example, it can refer to a game designed to deliver a behavioral intervention (Schmidt, Laffey, Stichter, Goggins, & Schmidt, 2008), a physical rehabilitation regimen (Huo et al., 2015), or to elicit learning gains from players (Laffey, 2016). By harnessing the features naturally found within games such as: narrative, rewards, fully animated 3D contexts, just-in-time feedback, etc. and using them to teach a curriculum; serious games have the potential to be the most personalized field trip, accessible right at home or in any classroom.

Serious Games have been a rising field in education since Gee's seminal work in 2003 (Gee, 2003). By laying out 31 Learning Principles (Gee, 2005) Gee showed how every game is inherently a learning environment. Games spanning all areas of subject matter began to be developed. As a result, students who struggled to keep up with their classmates' progress through traditional educational methods have been seen as top achievers in the class (Fried, 2005). Calls for serious games research have indicated the question has begun to shift from whether games can be effective teaching tools to how games can be best made to enhance teachers' abilities to instruct and evaluate.

STEM Game Based Learning

A popular curricular focus area for serious games is STEM content (Mayo, 2009), (D'Angelo et al., 2014), (Freeman et al., 2014). One factor contributing to this is the call for America to rise in the international Science rankings requiring greater STEM education and career interest (DeJarnette, 2012). Another contributing factor is that STEM learning may be easier to automatically assess than other fields such as creative writing for example (Kapp, 2012) enabling greater potential for effective just in time feedback. The combination of these and other factors leads to a great societal need for Serious Games in Stem that also seems achievable with today's game technology.

In terms of game-based STEM learning, Quest Atlantis was a pioneer in the field, first launching in 2002. This was the environment where Barab first developed his transformational play framework (Barab, Thomas, Dodge, Carteaux, & Tuzun, 2005). Two years later, River City was developed in 2004 by Dede's team at Harvard. This team later also developed EcoMUVE, a Multi-User Virtual Environment for Learning Scientific Inquiry and 21st Century Skills (Metcalf, Clarke, & Dede, 2009). More recently Mission Biotech was developed by Sadler (Lamb, Annetta, Vallett, & Sadler, 2014). All these and many more since (Hussein et al, 2019) projects have been able to show gains in STEM scores as a result of playing serious games.

Mission HydroSci

In this study, players will engage with Mission HydroSci, a game aimed at teaching 6-8 graders water systems and scientific argumentation. The game consists of 6 units and takes on average 6-8 hours to complete. During these units' players will control a character and use it to solve various problems and accomplish different tasks, such as tracing a pollutant up a stream or distilling seawater into drinkable water. After completing all the units, players will have successfully turned a crashed landing on a distant planet into a sustainable community.

Embedded Assessments and Learning Analytics

Embedded assessments are classroom activities that provide data for specific learning outcomes. In a game context, these would be in-game tasks specifically designed to assess a certain learning objective (Ketelhut, Nelson, Schifter, & Kim, 2013). In Mission HydroSci we place embedded assessments in every Unit to get a measure of whether students understand the concepts we are presenting. Here are a few examples of how we use embedded assessments within MHS Units. In the first unit our learning objectives are not scientific in nature, but instead focus on learning the controls and interface for the game. We have several tasks which require the students to perform the basic functions of the game: movement, conversation, jumping, etc. (See Figure 1 below).

Figure 1: A screenshot from Unit 1 of the Teacher Guide showing the jumping tutorial.

The second unit teaches players about topography. In that unit, the player has become separated from the group and must use a topographic map and topographic clues to locate the rest of their team. We made this embedded assessment a side quest continuing throughout the game, where players search for pieces of their crashed spaceship based on topographic clues they receive allowing us to track their performance over time. Unit 3 features a more traditional summative embedded assessment where students are tasked to identify garden sites based on how a beneficial nutrient would disperse through a river system. Each of these are all expertly designed top-down embedded assessments. Our design team consists of game designers, instructional designers, and subject matter experts all working collaboratively to design the most engaging and educational gameplay possible.

Embedded Assessments make powerful variables for Learning Analytics. The Learning Analytics process starts by collecting logs of user interactions in a learning environment. In an online class these logs might track discussion posts made, or class resources accessed. In a serious game, these logs could be responses to multiple choice questions, or accessing their map. Logs like these and players' behaviors during embedded assessments are then all used as input for machine learning algorithms capable of handling high volume, high dimensionality data used to predict learning outcomes (e.g., post test scores). If this prediction is accurate enough; a model can be constructed to classify new students as they are engaging in the learning experience based on their anticipated assessment outcomes. This classification allows for early interventions or dynamic feedback to be delivered in the hopes of increasing students' opportunities and subsequent likelihood to learn.

Player Experience and Psychophysiology

On the Mission HydroSci team it is not enough that students are required to play our game. We want players to want to play it, and for that reason Player Experience is equally as important as curricular fidelity. To assure the game is the right balance of fun and challenge to as many middle school students as possible we run extensive Playability testing. Playability testing is a form of usability testing tailored for game play and player experience (Nacke et al., 2009). We bring players into a lab setting often to run through small portions of the game. While they are playing, we observe to see whether they complete the game as designed, if anything confuses them, and how much they enjoy the experience. Afterwards we have a short debrief where we ask them to tell us about their experience. We use this feedback, not just to make the game easier to play or to improve the embedded assessments, but also to maximize the player's engagement and interest.

One quantitative method for gauging various cognitive aspects of a user's experience is to measure psychophysiological responses. Eye tracking can be used to trace to what and how long users give their visual attention. Heart Rate can be used to have an indication of users' stress. Skin Conductance has been shown to accurately measure users' excitement, and Facial EMG can reveal users subliminal affect states. All of these reactions are measured on the millisecond level and may even occur unknowingly to the user. By triangulating multiple measures, it is not only possible to detect more complex cognitive states such as flow, but also to detect what in the media caused the reaction.

Transformational Play

While embedding quality assessments in rich game play experience may be enough to make a serious game; there is still more needed to achieve Transformational Play. Transformational Play requires players to take on a new role and develop new skills within a game-based learning environment that then transfer to real world knowledge and skills after game play (Barab, Gresalfi, & Ingram-Goble, 2010). The intent of Mission HydroSci is to take players of various science and gaming experience levels and make them ALL feel and behave like masters of natural water systems. They use their science knowledge and skills to be the hero, resolve conflicts, and ultimately save the planet. The goal is that this experience will subtly build mastery and self-efficacy, which in turn will lead players to higher masteries of science in the real world.

Statement of the Problem

The core problem is that assessment in games of whether learning is occurring or not currently only comes from two places. First, it comes from a traditional top-down expert in the form of curricular pre/post tests. Second, it comes from a game-based learning designer in the form of gameplay behavior logs. Embedded assessment scores could also be generated from either or both perspectives. Games are expertly designed systems full of lots of playful mechanics and aesthetics; so those two approaches both make sense. However, games are also active Human-Computer Interaction systems, meaning both the game and the player are processing interactions while engaged in play. When both approaches are triangulated with a third player centric approach; then the expert designers, the playful game, and the player themselves would all be included in the learning assessment. Unfortunately, there are not currently studies triangulating these three approaches into a single assessment of learning.

Purpose of the Study

The purpose of this study is to improve the impact of games on learning by advancing our understanding of (1) how Psychophysiological data can be used to enhance unsupervised learning analytics ability to predict player performance and (2) how unsupervised learning analytics can be used to predict player's arousal states.

Significance of the Study

There are many intellectual merits and broader impacts contributing to the significance of this study. This first exploratory step into the player's psychophysiological experience of a serious game will give researchers a deeper understanding of how players learn in a game-based learning environment. Is all failure or success as defined by experts or embedded assessment scores equal, or does the player's internal reaction to that experience impact its transfer? Does the player's arousal impact retention? While it may not be possible to answer these questions at this time, this study will be a step toward building on learning analytics methods, describing characteristics of player behavior, and finding effective ways to connect Psychophysiology with Learning Analytics to better predict Assessments performance outcomes.

In addition to these intellectual merits the project will advance important broader impacts. Being able to recognize the players' arousal states through gameplay behaviors would allow for adaptive gameplay opportunities based on arousal states. Games could up the tempo if players are losing interest, or slow things down if players were starting to get frustrated. Finally, as an addition to current Playability testing, this research will help our field design better games for learning by expanding on current measures to enhance game designers' understanding of player engagement.

Statement of the Research Questions

What are the Virtual Behaviors which distinguish high and low performance outcomes in MHS?

To explore this question, I used archived field test data from a 2019 testing of MHS that included over 632 students' gameplay and pre/posttest records. The primary data used were the player behavior logs collected while playing Unit 3 of MHS as well as the pretest scores associated with each player. All these data were used as the dependent variables in a clustering algorithm, which grouped players into clusters of similar behavior. Finally, the post scores were used as the independent variable in order to predict players post test score outcomes. When this was done, I had clusters of students with similar features, each related to specific post assessment outcomes. This model can now be used on future students as they are playing through the game to dynamically classify them into a cluster and predict their post assessment outcomes. This information can either be used during gameplay to dynamically react to the player, or it could be uploaded to a server in the form of a teacher notification; so, they can follow up on the student in person.

What are the specific PP patterns that indicate embodied motivational processes such as flow and attention, and what game features elicit those patterns?

For this question, I had 9 participants play through the same version of MHS that was used for the field test mentioned above. While playing, participants were measured using a Tobii Eye Tracker and an E4 Bracelet collecting heart rate, temperature, and galvanic skin response (GSR). While the participants were playing, a researcher observed the player, their gameplay, and the incoming data streams. Particular attention was paid to players' emotional reactions, and what they were doing at that time during the game (e.g., participant 7's GSR spiked after looking at the non-playable characters face during the 2nd task). After participants finished playing, we had debriefing interviews where I had a chance to follow up with them about any reactions I noticed to see if the event stuck out to them as well. When all of my participants were done, I compiled a data set of GSR spikes per player triangulated with what they were visually attending to and what in-game events were occurring when they reacted. While I feel that visual stimulus is highly relevant to GSR spikes; I do not want to discount other possible stimuli (e.g., sound effects or something external to the game like figuring out a solution).

How and to what extent can the cognitive-affective state of players be inferred from virtual behaviors?

For this triangulation question I grouped the case-study players based on GSR spikes during similar game events or while visually attending to the same stimulus. I then compared and contrasted behaviors and learning outcomes of players who had and did not have GSR spikes at specific points of gameplay to explore any patterns present. There were a few interactions which elicited a similar reaction among most participants. These exploratory findings need more investigation but may be an indication that the interaction elicits that reaction for a large portion of the general population. If so, those types of interactions may then be better utilized by designers to harness users' engagement and attention during learning experiences.

How and to what extent can high or low performance be inferred from the cognitive-affective state of players?

I answered the final triangulation question by looking at which posttest outcome clusters my 9 participants are classified into based on RQ1 and exploring if there were any patterns between those outcome clusters and GSR spikes based on RQ3. Then I could examine any similarities or differences in Psychophysiology reactions among participants in successful vs unsuccessful clusters in terms of the pre/post outcomes. If there was a lot of separation such as successful students being either bored or in flow much more frequently, while unsuccessful students were more often the opposite; then that could be a good indication that Psychophysiology could be a useful distinction in performance. While this sample is far too small to predict learning outcomes based on Psychophysiological response, exploring approaches to answering this question may be an important first step toward that goal.

Design of the Study

This study is an exploratory psychophysiology case study, within the context of a large-scale learning analytics study. Data collection took place in two separate phases. The first source is archival data from the MHS field test, which had over 600 students play through MHS during February - April 2019. These students took a pre and posttest, a demographic survey, and provided logs automatically collected through gameplay over the course of 8 days. The purpose of examining the field test data was to build classification models for examining individual cases, The second set of data was collected during summative player experience testing with 9 participants that took place during Fall 2018. Using the model generated from the log data collected in the Field Test these players were dynamically classified based on their gameplay logs. While they played eye tracking software and E4 bracelets, allowed us to collect psychophysiology data to measure their attention and reactions. The purpose of this exploratory study is to build methods and insights into how the Psychophysiology data might enhance our understanding of the player experiences and outcomes when integrated into the current MHS learning analytics systems.

Study Delimitations

A delimitation in this study is the choice not to account for the moderating factors of science interest and game experience. While both the factors of subject matter interest and experience with the medium have been shown to have significant impact on players' experiences with serious games (Bergey, Ketelhut, Liang, Natarajan, & Karakus, 2015), it was beyond the scope of this study to recruit adequately to account for those factors. In the future, a pre-survey could allow for participant screening to balance out a proper 2-by-2 experimental design (while remaining gender balanced).

A second delimitation is the choice to analyze only Unit 3 from the summative player experience testing. An important consideration in making sense of this study is that the percent completion for MHS was significantly lower than the level of completion for the comparison curriculum during the field test. The percent completion threshold required that 80% of a teacher's class reach the 4th unit of MHS (approximately half-way through the game). Only 2 of the 13 teachers met the threshold for MHS while all the comparison classes had students reach a comparable threshold for completion. This was one of the main considerations for limiting the study to Unit 3 findings only.

Key Terms

Serious Games Terms

Aesthetics: The look and feel of the game created by the visual art, sound, and narrative.

Dynamics: The outcome the play experience has on the player.

Game Moments: Specific moments of game play related to context (player action, narrative, feedback, etc.).

Mechanics: The code enabling the game systems and player actions.

Player Experience: The totality of the experience a player has including usability, playability, and outcomes.

Serious Games: A game created for a purpose other than solely entertainment such as education or wellness.

Learning Analytics Terms

Clustering: Grouping points of data within a set based on their relatedness. In general, this can either be done by repeatedly bisecting the data or by creating kernels and repeatedly adding points to them.

Embedded Assessment: A task or activity that is designed to assess a specific learning outcome, which has been placed within a larger context such as a class or game.

Learning Analytics: The process of analyzing logs collected in learning contexts with the goal of gaining greater understanding of if and how learning is occurring.

Pre/Post Assessment Outcomes: Students scores on the external pre and posttests which may be viewed in terms of gains.

Principal Component Analysis: A process for reducing the dimensionality of a data set by transforming the data such that the variable with the greatest variance lies on the first axis, the second greatest variance on the second axis, and so on. In this way the variables with the most explanatory power can be quickly identified.

Supervised Method: A category of AI training which uses pre-classified data for the model to learn to differentiate between after which raw data can be classified, vs an unsupervised approach which would have unclassified data being differentiated after which classes can be defined and applied to future data.

Psychophysiology Terms

Arousal State: A cognitive-affective state detectable through Psychophysiology technology such as the E4 bracelet used in this study. This could be a positive state such as engagement or a negative state such as frustration.

Eye Tracking: The process of tracing an infrared reflection of a user's pupils to triangulate the gaze from each eye to determine where a user is focusing their visual attention over time.

Flow: The cognitive-affective state produced when the current challenge someone is facing is perfectly balanced with their skills for overcoming that challenge, resulting in intense focus, progress, and feeling of accomplishment.

Psychophysiology: The study of examining how cognitive states manifest responses biologically.

Skin Conductance (or Galvanic Skin Response or Electro dermal Activity): A measure of how conductive the skin is due to external or internal stimulus.

Chapter 2: Literature Review

Overview

The purpose of this Literature Review is to provide a background of the research methodologies that I used in this study. The review includes three fields, the first being Learning Analytics and the second Psychophysiology, and ultimately attempts to find intersection between the two within the context of the third field, Serious Games. The analytics section primarily tracks two methods, Principal Component Analysis and Clustering, in their use in online courses, commercial games, and finally serious games. In this way the methodology is demonstrated in a purely academic setting and a purely entertainment setting before being examined in an environment used for both. Similarly, the Psychophysiology section looks at two primary methods, Eye Tracking and Galvanic Skin Response Analyses. Those methods are also detailed in non-game media as well as commercial entertainment games before being examined in academic educational game settings.

For this Literature Review, I limited all my search results to things published in 2015 or more recently. I began my search with broad terms like “Psychophysiology/Learning Analytics in Games”. This led me to a lot of articles and a large variety of methods including Clustering, Bayesian Trees, Electromyography (EMG), Galvanic Skin Response (GSR), etc. It also led me to two researchers I focused my attention on early, Lennart Nacke and Guillaume Chanel. These two researchers’ interest both centered on using Psychophysiological data to measure player experience during gameplay.

Through these sources and early methodological testing, I narrowed my search terms down very specifically to “PCA, Clustering, GSR, or Eye Tracking, to which I then added either “in serious games” or “in games ‘-serious’”. These terms produced fewer results, but usually a few relevant articles per search including the articles I reference below.

Why Research Serious Games

Games in general are inherent learning systems as described by Gee (2005) where players learn new narratives, environments, actions, etc. Serious games have the additional constraint of teaching a specific curriculum (Squire, 2003). These learning objectives need to be delivered in a way that allows the concepts to transfer to the real world in a meaningful way (Barab, Gresalfi, & Ingram-Goble, 2010). The highly controlled nature of the environments and interactions makes serious games a prime field to test various theories and work towards deeper insights into learning. Game based learning in a single player game is an independent student activity in the sense that unless the teacher’s role is explicitly designed into the experience (Stichter, Laffey, Galyen, & Herzog, 2014), they are not directly present, which standardizes curricular delivery by removing all human elements except the learner. In addition, games afford for explorative and constructive learning while also being able to log every fine detail of the environment, context, and player’s actions allowing for a deep level of analysis and insight.

Serious Games are often sought after for a desired increase in learner engagement as compared to traditional instructional methods (Hookham, Nesbitt,

& Lambkin, 2016), however this engagement is usually measured via post-play survey or in game pop-up (Shernoff, Hamari, & Rowe, 2014). Serious Games can also be designed to elicit specific gains for specific users such as increasing attention for players with ADHD (Roh & Lee, 2014) and other Learning Disabilities (García-Redondo, García, Areces, Núñez, & Rodríguez, 2019) as measured with specific validated pre/post-tests.

Learning Analytics

Learning Analytics in Online Classes

In a one-on-one tutoring scenario, a teacher is able to be intimately involved in every moment of the student's learning process, and that affords them access to subtle cues to gauge engagement and understanding. In a traditional face-to-face classroom with many students to attend to at once, teachers extend these observation practices to allow them to monitor groups. However, in an online learning environment there are fewer but more complex cues. As online learning rose in popularity; so, did the need for Learning Analytics research, seeking to identify and make sense of these sparse clues (Lias & Elias, 2011). The need for Learning Analytics was further exacerbated by the push to increase online class sizes, even to the extreme case of MOOCs (Daniel, 2012). Although these classes can deliver education on a large scale (Cress, 2014), often at a free or reduced price; the learning (instruction, assessment, and feedback) is by necessity very standardized and automated, as it is not possible for a person to monitor and provide feedback to that many students in depth. This led to a well-studied decline of student engagement causing early departures (Khalil & Ebner, 2014). At the same time, the technology capable of performing big data analysis also became more prevalent (Han Hu, Yonggang Wen, Tat-Seng Chua, & Xuelong Li, 2014). The outcomes generated by predictive Learning Analytics began helping teachers by producing a short list of which students needed extra instruction or feedback. This way they could focus their efforts where most needed instead of the impossible task of engaging with the entire class individually (Dyckhoff, Zielke, Bültmann, Chatti, & Schroeder, 2012). Enter machine learning methods focused on predicting final course evaluation scores. With enough training data, teachers could identify the patterns of behaviors of students that were most likely to fail using methods like Principal Component Analysis. Empowered with this new insight, teachers could focus their limited time on the students who needed help the most.

Principal Component Analysis (PCA) in Online Classes

Hegde's work (2016) is a good example of data reduction. Hegde's study, titled *Dimensionality reduction technique for developing undergraduate student dropout model using principal component analysis through R package,* tried to predict the course dropout rates of 150 students in two cohorts over two semesters. The data gathered for this study came from a 51-item survey. The authors were able to use PCA to reduce the necessary questions to 25 or 19 depending on the desired accuracy. PCA is necessary to understand core defining features of large datasets with complex varieties of data types.

Clustering in Online Classes

A large complex data set, even after being reduced, can still require analysis to construct meaning. Ezen-Can's study (2015), titled *Unsupervised modeling for understanding MOOC discussion forums: a learning analytics approach,* shows a popular clustering method for grouping large subsets of data into meaningful outcomes. Ezen-Can and colleague's goal was to empirically investigate the extent to which unsupervised models can provide insights into the flow of conversations among learners on a discussion forum. They studied an 8-week MOOC for teachers with 155 students. Through Discussion Board Mining they logged 550 posts from 57 distinct discussions which were then manually coded. The authors used K-medoids clustering with a greedy seed selection approach and Bayesian Information Criterion to arrive at seven clusters: agreements, opinions, declarations, disagreements, appreciations, questions, evaluations. The Ezan-Can paper is an example of how clustering can identify patterns in data but highlights the need for each cluster to be evaluated by a human afterwards to extract meaning.

Learning Analytics in Commercial Games

While analytics have been gathered privately within the games industry for quite some time; they have only recently begun to be explored academically. Although they often employ the same machine learning methods as learning analytics, it is for a different purpose. For example, players finding a successful learning strategy and sharing it with others in their community would be encouraged in an online class, and to a degree may be encouraged in a commercial game community as well. However, if enough players in a commercial game realize that this strategy wins more often than other strategies, then the gameplay can become asymmetric, usually prompting the designers to make a change to that strategy. The ideal behavior of all students demonstrating mastery successfully in an online learning environment becomes a design issue needing to be corrected in a commercial game to keep the gameplay diversely balanced.

Principal Component Analysis in Commercial Games

The authors (Drachen, Thurau, Sifa, & Bauckhage, 2014) of *A Comparison of Methods for Player Clustering via Behavioral Telemetry* studied a huge dataset from the popular online game World of Warcraft. The dataset contained gameplay logs collected over a 5-year period from 70,014 players. Drachen and colleagues were trying to examine how different clustering algorithms produced different results on the same dataset. For each player they created a vector for the highest level that player achieved by the end of the day. This created growth patterns over time for each player. The authors compared clusters generated using Archetypal Analysis, Non-negative Matrix Factorization, K-means Clustering, C-means Clustering, and Principal Component Analysis. While their results had different meanings situated within World of Warcraft Gameplay, there were essentially two main takeaways. First, different algorithms will produce different clusters, and it is important to try multiple approaches to see which works best in each context. Second, the number of clusters can also affect how clear or meaningful the results can be. While there are some algorithms which can choose a cluster number for you; the authors propose it is best left as a parameter open for testing and tweaking until the best model fit is reached for the specific problem being solved.

Clustering in Commercial Games

Manero and colleagues (Manero, Torrente, Freire, & Fernández-Manjón, 2016) attempted to validate *An instrument to build a gamer clustering framework according to gaming preferences and habits.* They sought to make a standard instrument that was efficient, but still as explanatory as possible. To do this they created a 10-item survey and administered it in eight schools across Madrid to 754 Spanish secondary students: 54% male, median age 14. The questions centered around how much and what kinds of video games they play. Using Principal Component Analysis, they were able to reduce their original 10 item survey to a simple 2 item survey. Following this dimensionality reduction the authors used K-Means Clustering to create 4 clusters of players: Everything, FPS/Sports, Casual, and Non-gamer. While not originally part of their study, the authors also found a high gender separation where males primarily fell in either the Everything or the FPS/Sports clusters and females were primarily in the Casual or Non-Gamer clusters.

Learning Analytics in Serious Games

Enter serious games, developing out of a separate school of thought these designers tried to marry the engagement of video games with the education potential of learning technology. No longer was early departure a concern (Squire, 2005) as these modules were designed to be fun, and to keep the learner persisting until the end. Scalability (Barab, Thomas, Dodge, Carteaux, & Tuzun, 2005) was no longer an issue either as games are usually designed to be played autonomously, requiring no additional teacher or students. However, because the interaction happens within the virtual environment, this creates the same absence of subtle intimacy the teacher can have naturally with students in a traditional classroom (Kebritchi, 2010). Complicating things is the fact that games are a fundamentally different structured medium than traditional education tends to use. For example, most classes online or traditional, structure the instruction so that the learning happens first followed by an assessment in which students might fail. This is repeated and then usually happens on a larger summative assessment in the form of a final exam or project at the end of the course. Video Games are quite to the contrary. They encourage failure multiple times to learn about the system and environment (Fudenberg & Levine, 1998). The embedded assessments are often designed to be repeated many times even within future tasks. For Serious Games it is not enough for the analytics system to simply report to a teacher that a student is predicted to fail a game task.

Teachers of serious games, if there is one at all, have not designed the curriculum: the instruction, the embedded assessment, nor the feedback, and in the current state of the media have very little ability to customize the in-game curricula (Baek, 2008). Because of this they likely will not be well informed of what the player is doing or has done within the game at any given time. This lack of deep knowledge of the game play necessitates that the first purpose of Learning Analytics should be to accurately relay what the player is doing to the teacher. In addition to this, games offer a substantial amount of interactivity and feedback which online course management systems simply do not match. Requiring the teacher to attend to, interpret, and respond to every player behavior counters the scalability of serious games. In this environment it becomes more efficient to identify a behavior signature that a student is falling behind on a certain task or is struggling more than you would like, and have the game deliver the necessary feedback to the learner just-in-time (Gee, 2005) rather than force the student to wait for the next teacher intervention. The second purpose of Learning Analytics then becomes to use those “found” behavior signatures to inform design for what specific feedback each student requires and how it can be embedded within game play.

Principal Component Analysis in Serious Games

Van der Graaf and colleagues (Van der Graaf, Segers, & Verhoeven, 2016) took the dimensionality reduction concepts seen above into educational games with *Discovering the laws of physics with a serious game in kindergarten.* They asked, “To what extent do children show comparable exploration and efficiency scores on the game plays, and how does individual variation in exploration and efficiency relate to executive control, reasoning ability, and vocabulary?” (p171). They examined 75 children around 5 and half years old, 41 of which were girls while they played 3 short games: The Hippo App, Flanker Fish, and Hearts and Flowers. The authors used Principal Component Analysis to confirm the three tasks measured different aspects of executive control. The researchers then conducted Mediation Analysis using ANOVA to find mediating factors. They found that exploration and attentional control were mediated by the level of each player's vocabulary competency, and efficiency and attentional control were mediated by non-verbal reasoning.

Clustering in Serious Games

One of the most rigorous articles I have found on using learning analytics in a serious game is *Investigating Epistemic Stances in Game Play with Data Mining (*Martinez-Garza & Clark, 2017). The researchers had two Goals with this study. First, they wanted to validate the Two-Stance (2SM) Framework for Game Based Learning, and then to explore automatically collecting log files of gameplay for evidence of learning. The 2SM framework is an application of the more general two-systems theory in human cognition. It posits that during educational games students enter 2 phases, or "stances": one being a learning phase where students reflect on concepts and engage in sense making; and the other being a playing phase where students test ideas and optimize their performance over multiple trials. To guide this work, they had two Research Questions. "Did students playing The Fuzzy Chronicles show evidence of dichotomous fast/slow modes of solution?", and "Is there a connection between conceptual understanding and student performance in conceptually laden challenges?" (p9). The authors ran two studies both with pre/post-tests. The first study consisted of 86 ninth graders while the second study had 123 seventh graders. Each study took 5 days of class time for the kids to play 32 levels of The Fuzzy Chronicles, a serious game teaching Newtonian kinematics. Over the course of this study 16,239 logs were collected, each being a JSON object representing an individual student's attempt to solve a certain puzzle. From these logs, 23 variables were derived to extract meaning from each log. The authors performed extensive analysis on these data. First, they did data normalization and integrity checks to assure the data were clean and accurate. Then, using Principal Component Analysis they did variable selection and dimensionality reduction. Next, they performed Clustering of Gameplay Data and Sequence Mining. Finally, they conducted Contextual Feature Mapping to extract the meaning from their results. As intended, Principal Component Analysis reduced the variables drastically. They ended up with 6 clusters: Aborters, Tinkerers, Long Aborters, Repeat Failures, Winners, and Planners. According to the 2SM framework, they were able to differentiate between low and high prior knowledge based on differences in gameplay. For example, students with higher prior knowledge were better able to plan out solutions and then execute them versus the lower prior knowledge group's more guess and check gameplay style. The 2SM appears to be a useful model to distinguish brute force trial and error gameplay from pre-planned thoughtful problem solving.

Learning Analytics Conclusion

Learning Analytics is a powerful tool capable of revealing hidden insights within large datasets. However, the utility of learning analytics is highly dependent on the proper data being collected and the proper methods being used for each specific problem being solved. Data must be pertinent to prevent misfitting and results must be actionable to make an impact. Often these data models are complex and difficult to interpret, but methods like clustering and PCA can reduce that complexity. When combined with human analysis, and human readable definitions, they can help teachers be more present and aware of student learning in more varied and novel environments, such as serious games.

Psychophysiology

Psychophysiology in Non-Game Media

Psychophysiology is a field linking the psychological (mind) and physiological (body) camps, for example the relation between a fight or flight response and an increase in heart rate (Morgan, 1951). Because these data are a collection of nervous systems responses, it is extremely fine grained (Gale & Edwards, 1983). The data exist on the millisecond level, and reactions are traced second by second (Martínez, Jhala, & Yannakakis, 2009).

Psychophysiology studies originated outside of games in media where the consumer is usually a passive listener and/or viewer. In this way participants' eyes can be tracked as their only active response indicating attention. Then data such as skin conductance and heart rate can be used to determine the audience's stimulus level at any moment. Finally, that stimulus can be valenced to determine the user's affect through more rigorous methods like facial electromyography, but also through more usable methods such as facial recognition. These three cognitive factors (attention, stimulus level and affect) are used to construct an emotional image of how the audience is experiencing the media throughout its duration.

Psychophysiology researchers have found a home in media reception studies (Ravaja, 2004). Understanding how a particular message is going to be received or how it will emotionally impact an audience is critical, whether in television, movies, or video games (Ganglbauer, Schrammel, Deutsch, & Tscheligi, 2009). Skin conductance has long been validated as a general measure of arousal, while facial EMG specifically above and below the eye has been validated with positive or negative affect (Nakasone, Prendinger, & Ishizuka, 2005). Eye-tracking allows researchers to pinpoint these physical reactions to what precise visual information participants are attending to at any given moment (Alkan & Cagiltay, 2007), (Goggins, Schmidt, Guajardo, & Moore, 2011). By combining these three points of data, we can identify elements on screen and determine whether the audience has a reaction to it, and if so, did they generally enjoy or dislike it.

Eye Tracking in Non-Game Media

Eye Tracking Data at its core are a series of XY coordinates mapped onto a 2-D surface. These raw coordinates are not very informative, and are often quickly converted into Gazes, points of prolonged attention, and Saccades, points of rapid eye movement. Jianu and Alam (2017) took that processing a step further to identify what objects are being gazed at using *A Data Model and Task Space for Data of Interest (DOI) Eye-Tracking Analyses.* This paper detailed 3 studies advancing this model. The first study asked subjects to visually track objects on screen and was primarily for testing the validity of the technology. The second study had 6 participants all in an architecture class using a web app. Finally, the third study had 16 construction workers training in a 3D simulated construction site. Each study used the same Eye Tracking methods. The construction site, however, did have a secondary rendering. While the participants were viewing a normal looking environment, a second screen featured every object of interest rendered in a unique color (e.g., solid green crane, solid red hole, etc.). Using this specially colored render the authors were able to identify what object participants were looking at simply by logging the RGB value of their Gaze point. This allowed them to talk about gazes not in terms of area on screen, but in terms of contextualized resources and hazards. This Data of Interest can relay in real time details about user tasks as their being completed, where previously these gaze points would need to be post-processed to extract meaning.

Most Eye Tracking studies assume the Eye-Mind Hypothesis, which is, where people look correlates with what people think. This was confirmed in the Jianu and Alam study when they found experts tended to use more peripheral vision, creating different Data of Interest patterns than novices. However, it is also important to note that not every gaze indicates intense concentration. Users can stare at an arbitrary point while contemplating without taking in visual information.

Skin Conductance in Non-Game Media

Like Eye-Tracking studies, Psychophysiology data usually require post processing to extract meaning. However, Green and colleagues (2014) attempted to automatically process the data. They asked, “Can the classical trough-to-peak amplitude of SCR be automated in a fashion closely matching manual scoring?” They studied 3 archived datasets of healthy adults from Duke University. The authors compared the results of 2 manual raters vs. 3 alternative automated methods. The authors detailed how they compared common methods of Amplitude, Response Latency, Rise Time, and Half-Recovery Time. In general, there was high concordance between manual and automated ratings, and when algorithms would mis-classify they would err on the side of over or under classification consistently. Because manual coding is highly time consuming it would usually prohibit large data set analysis; however automated algorithms could run big data problems without an issue.

Psychophysiology in Commercial Games

Applying PP methodologies to games introduces one major unique challenge, namely asynchronous consumption of media. In a passive media study, it does not matter what the participant does, the commercial is always 2 minutes and 30 seconds. However, in a game, player progress is entirely mitigated by player performance, allowing for variances in times when similar interesting events will occur.

With the dynamic nature of games and inherent user interaction, psychophysiology data can take on a new meaning, even being used as a controller in a few cases (Nijholt & Tan, 2007). Flow Theory could present an interesting lens for Psychophysiology studies in games. Flow is simply a user's personal skill level compared with the level of challenge they are currently experiencing on a given task, which is hypothesized to produce complete absorption in a task (Nakamura & Csikszentmihalyi, 2014). The scaffolded nature of some video games, which aims to maintain engagement through a slow and steady increase in difficulty, makes them an ideal experimental context for flow.

Researchers have recently begun to look at the ability of psychophysiology data to measure flow within the video game research community (Cowley, Charles, Black, & Hickey, 2008), (Chen, 2007). Of particular interest, Chanel et al. (Chanel, Rebetez, Bétrancourt, & Pun, 2011) used flow theory and psychophysiology data to control the difficulty of Tetris, a game which relies on scaffolded difficulty to maintain engagement as mentioned above. They showed that lack of arousal data could be linked to the boredom and apathy states elicited from low challenge tasks described by flow theory. From there the researchers were able to use affect measures to determine whether the player was negatively aroused indicating frustration or stress or positively aroused indicating flow or engagement.

Eye Tracking in Commercial Games

Bagley, Lee, and Rankin (2015) detailed their work in *The Development of an Eye-Tracking Program to Examine Working Memory During Gameplay*. They were attempting to gain insight into the causes of differences in individual working memory performance. They examined 6 participants in a lab setting with a particular focus on eye tracking validity and accuracy. Each participant followed the same process of Closed Eye Rest, Open Eye Rest, and Eye Tracking Calibration, followed by playing Mahjong. During this process participants' screens were recorded for Eye Tracking, and 4 nodes were collecting Electroencephalogram (EEG) Data. The authors first filtered the coordinates from the Eye Tribe Gaze API for off screen gazes, blinks, and any other disengagement. They then timestamped the data to synchronize it with the EEG data. The eye tracking results indicated eye tracking can be effectively captured up to every 1/45 of a second. Although EEG was not used, the general process and comparison of PP with Eye Tracking for cognitive purposes is a model for what was done for this dissertation study.

Skin Conductance in Commercial Games

In *Flow and Immersion in First-Person Shooters: Measuring the player's gameplay experience,* Nacke and Lindley, (2008) were trying to find "correlations between subjectively reported gameplay experiences and objectively measured player responses within the gameplay as measured by these psychophysiological measures in order to provide cross-validated descriptions of the emotional experience of players during gameplay." (p83) They studied 25 male college students 19-38, all of which were gamers, in a lab setting while they played three **Half-Life 2** mods over a 2-hour session. During play the authors collected a wide variety of data including Electroencephalography, Electrocardiography, Electromyography, Galvanic Skin Response, video recording, and Eye Tracking. They first filtered the data to reduce noise then ran an ANOVA. The authors found that flow states were detectable using just GSR to detect stimulated states vs boredom and then EMG to give valence to the stimulus as either frustrating or engaging. Objective measures of valence are out of scope for this study; however, some inferences can still be made based on gameplay contexts.

Psychophysiology in Serious Games

Adding a learning component to a video game does not necessarily change the Psychophysiology data collection, but it does imply a different set of research questions. Researchers can ask how the Psychophysiology signals indicate that the player is learning. Things like what players are paying attention to at any specific moment of gameplay can inform efforts to understand why they are behaving as they do. Understanding the player's emotional state can help us find new insights into how to gain the most impact out of a curriculum or garner the most retention of curricular content.

Eye Tracking in Serious Games

In this study Byun and colleagues (2014) tried using eye tracking for Serious Game Analytics. They asked, "Can Eye Tracking be used for assessment in role-playing serious games?" Three experts and 3 novices played a Military Style Search and Rescue game. Each session consisted of Instruction, Practice, Eye Tracking Calibration, and finally the Test. Like the Data of Interest (DOI) described earlier, the authors here primarily tracked Areas of Interest (boxes on the screen, AOI) as their unit of analysis for AOI Sequencing, Binning, Event Statistics, and Line Graphing. They found that experts show different eye tracking patterns than novices, and that fixations per second decreased in general when participants reported being immersed.

Skin Conductance in Serious Games

The most similar study to mine that I found was *Assessing Knowledge Retention of an Immersive Serious Game vs. a Traditional Education Method in Aviation Safety (*Chittaro and Buttussi, 2015). In this study Chittaro and Buttussi were trying to determine how gameplay vs more traditional learning methods affected a player's fear response via EDA and self-reported measures, and how both in turn affected retention. Forty-eight University students (26M and 22F) were split into two cohorts: the first would play a flight training game, and the other would learn from flashcards. The authors conducted baseline fear response measures for both groups before starting the intervention. They surveyed users for their Flight Experience, Game Experience, Fear Sensitivity, and gave each user a Pre, Post, and Retention Test. Results of the survey, sensor data, and test scores were analyzed using ANOVA. Results showed that while both interventions had similar pre/post gains, the game generated higher stimulus and fear response as well as much better retention scores. They concluded stimulus from an immersive experience may provide a deeper ingrained memory causing greater retention.

Summary

Each of these 12 articles helped shape my study in a very specific way. I used the exact methods of eye tracking (Byun et al, 2014), skin conductance (Chittaro & Buttussi, 2015), PCA (Van der Graaf, Segers, & Verhoeven, 2016), and clustering (Martinez-Garza & Clark, 2017). Also, I utilized the lessons learned, such as exploring multiple clustering options to determine the best approach for the problem and remembering that the generated clusters require a human interpretation step to extract meaning. I gained similar valuable insights from the psychophysiology literature review. While it is beyond the scope of this study, the automation process for both Eye Tracking and Skin Conductance would be necessary for carrying out this research agenda and collecting Psychophysiology data from 1,000+ students in any reasonable amount of time. I did however include the processes identified for triangulating eye tracking and skin conductance (Nacke and Lindley, 2008), and for detecting arousal during game play (Green et al, 2014). Although I did not explicitly control for gender or experience; it is always helpful to keep in mind that those factors have been shown to have a strong influence.

Mission HydroSci

Overview

In this section, I describe the research and learning objectives of Mission HydroSci. I detailed our lab's production process. Finally, I give an overview of the MHS environment we produced, and our current distribution model.

Figure 2: The cover image of the teacher guide depicting the volcanic cavern of unit 5.

Background on Mission HydroSci

Mission HydroSci is a research project designed to teach Earth Science, specifically water science, and Scientific Argumentation in a game-based learning environment. All its learning objectives align with the Next Generation Science Standards (Bybee, 2014) which indicate that learning outcomes require application of knowledge not simply recall. The MHS water science curriculum starts with topography; so, students understand how the terrain's shape determines its impact on the surface water that flows through it. The curriculum goes on to cover underground and atmospheric water, as well as touch on human impacts in the last mission. In parallel with this water learning curriculum, it teaches players scientific argumentation. MHS introduces players to what scientific claims, evidence, and reasoning are (Osborne, Henderson, MacPherson, & Szu, 2013), and how to use them together to form arguments.

The research questions for the project address the impact a game-based experience can have on science learning, but also include how best to design, implement, and distribute this curriculum.

Mechanics, Dynamics, Aesthetics Model

The MHS production process is modeled off the Mechanics-Dynamics-Aesthetics framework for designing games. While many games are driven by either mechanics (what actions the player will be able to perform) or aesthetics (what the story, art style, or environment will be) mentality, the MDA model advocates for a more holistic approach. Instead of either, it focuses first on the dynamics (what impact the experience has on the player), and then incorporates the best mechanics and aesthetics possible for that player experience (Hunicke, LeBlanc, & Zubek, 2004). In order to accomplish that as a serious games lab, we start with the learning objectives; those are the outcomes we want to impart on our players. Then three full time staff leads (art, design, and development) were each responsible for the creativity their work brings to the game. We then review that work as a team, and once each team has done the best work they are capable of; we present that to users for feedback. Through iteration, player feedback, and diverse expert review, MHS provides a rich player experience with fewer stones left unturned along the process.

The Mission HydroSci Environment

Mission HydroSci is a first-person narrative adventure where players use their newly discovered mastery of water systems to help set up a pioneer settlement on an alien planet. The learning gains in MHS have been experimentally validated and confirmed by the What Works Clearinghouse (Reeves, 2020). In the first unit, the player awakens on a space station orbiting an alien planet. Just as they are getting geared up and re-acquainted with their team, there is an explosion on the space station leading to an emergency evacuation. After the team crash lands on the planet, they must use knowledge about topography and surface water to choose a suitable location to jumpstart their settlement mission.

After getting settled, the other cadets on the mission begin to establish satellite bases to enhance the exploration and resources available. The first base established is Samantha's, where the player uses knowledge of surface water flow to aid her in gathering supplies, removing a pollutant, and planting initial gardens. Afterwards, the player uses knowledge of ground water to help Anderson gain access to fresh water in his desert base and fix a flooding problem he is having. In unit 5, the player visits Bill's tropical island and uses knowledge of atmospheric water when they become trapped by a volcano requiring them to distill seawater into drinkable water giving Bill a new idea for his factory. Finally, the player ends up on an alien shuttle orbiting the planet disguised as a moon. The player must prevent the ancient, deserted shuttle from taking off into deep space with a large portion of the planet's water supply in order to save the planet.

Unit 3 Summary

This study examines players' interactions and learning outcomes during Unit 3 of MHS. Unit 3 was primarily intended to teach two water science concepts, dissolved materials flow downstream and rivers flow based on watershed elevations, as well as to introduce reasoning, or warrant statements, into the scientific argumentation process (note: argumentation was not examined in this study). It takes place in a gardenesque area between two rivers where the non-playable character, Sam, has chosen to setup her base. Unfortunately, Sam is having some problems with scattered supplies, polluted water sources, and identifying downstream nutrient flow. To demonstrate mastery of these water science concepts two questions were determined valid after an Item Response Theory analysis performed in a previous study. These questions are shown in Appendix A.

The Mission HydroSci Delivery Model

MHS has been delivered at 9 schools across Missouri with over 800 students participating over the course of 8 class periods for 45 minutes each (See Figure 3 below) of regular science class time. MHS was designed to be available for online classes with a remote teacher, however for this field test every class was a traditional Face to Face class in a public school. The executable for the beta is given to the districts' technology coordinators who then push the game onto computers for the students to use. While tablet distribution models are still being explored, the game is currently only available on PC and Mac. When students are playing the game, behavior logs are collected in the background and sent to our server where we can use them to display students' progress to their teacher or use it for later analysis to improve the quality of the game.

DAY	GAME TIME	IN-GAME ACTIVITY & PROGRESS *(By the end of the allotted time, the average student will be…)*	CLASS TIME	OUT-OF-GAME ACTIVITY
Day 1	-	-	45 min	Pre-Test
Day 2	45 min	MHS: Halfway through U2	-	-
Day 3	30 min	MHS: Starting U3	15 min	Argumentation Introduction
Day 4	30 min	MHS: 75% through U3	15 min	U2 Follow up
Day 5	30 min	MHS: 50% through U4	15 min	U3 Follow up
Day 6	30 min	MHS: Ending U4, Starting U5	15 min	Feedback Survey
Day 7	30 min	MHS: 50% through U5	15 min	U4 Follow up
Day 8	30 min	MHS: 50% through U6	15 min	U5 Follow up
Day 9	30 min	MHS: Finished U6	15 min	Argumentation Reveiw
Day 10	-	-	45 Min	Post-Test

Figure 3: The pacing guide included in the Teacher's Guide outlining the 10-day play schedule.

Chapter 3: Methodology

Research Design

This study is an exploratory case study consisting of two rounds of data collection and analysis. The first round examines archival data from a MHS field test looking at what behaviors players who are successful on the pre/post assessment, exhibit during unit 3 (approximately 1 hour of gameplay). The second round was collected from a Player Experience Test examining how players subconsciously react to their successes and failures along the way during the same unit. This exploratory study, using best practices of Learning Analytics and Psychophysiology, hopes (1) to contribute to a growing knowledge base about the use of analytics and psychophysiological descriptions of learning and play in serious games and (2) to build a model for how new methods of study can be triangulated to give a deeper insight into the player experience to design better games for learning in the future.

Research Questions

RQ#	Research Question	Data Source	Sample Size	Analysis
RQ1	*What are the Virtual Behaviors which distinguish high and low performance outcomes in U3?*	Game Logs + Post Assessment Outcomes	806	Clustering Game Logs
RQ2	*What are the specific PP patterns that indicate embodied motivational processes such as flow and attention, and what game features elicit those patterns?*	Eye Tracking + Skin Conductance	9	Video Coding
RQ3	*How and to what extent can high or low performance be inferred from the cognitive-affective state of players?*	RQ 2 Results + RQ1 Results	4	Contrast PP Results of LA Clusters
RQ4	*How and to what extent can the cognitive-affective state of players be inferred from virtual behaviors?*	RQ1 Results + RQ2 Results	4	Comparing PP reactions during LA features

Table 1: A summary of the Research Questions, Data Sources and Analysis Methods for this study.

Field Test: Phase 1

Field Test Participants

The first round of data collection took place during the MHS summative field test (Reeves, 2020). MHS was played in the spring of 2019 by 13 middle school teachers' classes. These classes each consisted of 20-30 students in grades 6-8. These classes were from schools across Missouri, and the total student sample in the treatment group (N=806) included 51% male and 49% female, as well as 66% Caucasian, 11% African American, 6% Hispanic, 4% identifying as multi-racial, 3% Asian, 2% American Indian, with the remaining students selfidentifying as other. Of the total 806 students who participated in the treatment group of the field test, only 632 had all their logs collected from a complete playthrough qualifying them for the Learning Analytics.

Field Test Setting

These classes all played the game during their normal class schedule. Although the specific details of each classroom setup differed, the minimum requirements were that each student has access to their own computer to play MHS. Players needed to use the same computer to continue their progress from day to day.

Field Test Delivery

During the Field test each class was scheduled for 10 sessions each about 45-50 minutes. The first and last days consisted of pre and posttests respectively, and the other eight days were used to play the 6 units of MHS.

Field Test Dataset

These students were all given pre/posttests consisting of the “Middle school Affect towards Science and Technology” survey to gauge student interests, and a traditional curricular assessment consisting of both Earth: Water Science and Scientific Argumentation assessments. While playing the game for the 8 class sessions, gameplay logs were automatically collected and sent to a remote server. This large data set was used for all the Learning Analytics work in this study.

Participants played through the first available version of Mission HydroSci, a STEM game developed to teach water science and scientific argumentation.

Every time a student encountered one of the embedded assessments a log was collected. For example, in Unit 3, the player is required to plant 5 seeds in 5 of 11 different garden boxes. The target gardens are all downstream of a mysterious nutrient. If the player plants 3 seeds in the correct gardens (downstream of the nutrient source) and 2 seeds in the wrong gardens, their performance for this task would be normalized to 3/5 or 0.6. In between embedded assessments, multiple logs of the player's interactions will be collected: navigation, map use, time elapsed, etc.

From the more than 400,000 logs collected per player throughout the entirety of their gameplay sessions, 57 logged behaviors were determined to be the most predictive of their post-test performance. These key virtual behaviors broke down into the following broad categories: amount of area explored, time to complete a task, frequency of in-game tool usage, and argumentation performance (See Figure 4 Below). While the model uses data items collected from units 1 and 2 in addition to unit 3 to predict success, the behaviors specifically taking place within unit 3 will be further used to help to answer Research Question 3.

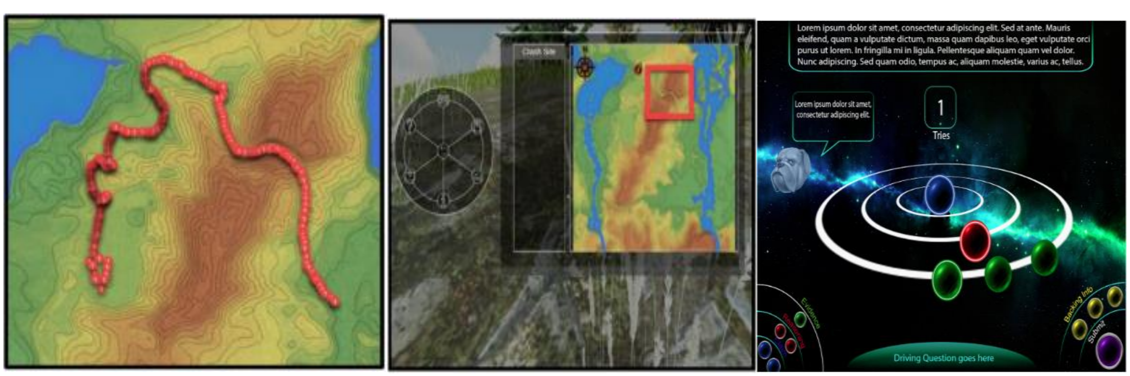


Figure 4: Left - An example navigation path reconstructed from a player's log data. Center - An example of the in-game tool menu. Right - An example argumentation scenario from within MHS.

Player Experience (PX) Testing: Phase 2

Player Experience Participants

For the second round of data collection, I recruited 10 (50% male and 50% female, 100% Caucasian) middle school students (age 11-12), who had not played the game before, to play all 6 units of MHS. These participants were recruited by word of mouth and received a $50 gift card for their participation in each session of the study. One was unable to participate due to interference with their glasses and the Tobii Eye Tracker. Each of the 9 remaining participants played MHS over the course of 3 two-hour sessions until they completed all 6 units used in the Field Test. They each took the same pre/posttests (See their scores in Appendix B), and while the same logs should have been collected as the field test group; 5 of the participants' data had missing elements causing it to be unusable for the Learning Analytics models, but still usable for the Psychophysiology questions. This left 4 usable participants including 75% females and 25% males for research question #4 about LA-PP integration.

Player Experience Setting

For this study, I set up a testing station in the study room in a research facility of the university. The room included a laptop, a Tobii Eye Tracker, and an E4 bracelet. During testing students experienced the MHS virtual world described above in Chapter 2.

Player Experience Delivery

My protocol (See Appendix C) for this Player Experience Testing (PX) mirrored the traditional MHS User Experience testing with the addition of the eye tracker and E4 bracelet. I greeted each participant and gave them a small introduction to what they would be doing that day. After answering any questions the participant had, the players completed the same MHS pretest used in the Field Test. Once completed, I sat them down at the computer and equipped them with the E4 bracelet. Before playing, the Tobii Eye Tracker needed to be calibrated, which took approximately 30 seconds. Once setup was complete, I started MHS, and sat at a nearby desk only intervening if the participant had a question. For the duration of the 2 hours, I watched the Eye Tracking and E4 data feeds taking note of my initial observations of what the player was attending to in-game while any large PP reactions were detected. Halfway through the session, the Eye Tracker battery needed to be swapped out and then re-calibrated for the second half. After the session was complete, I stepped in and stopped the participant's game session, eye tracking session, and E4 session in that order. I then conducted a debrief interview (See Appendix D) with the participant where I asked them to retell me what they did, and what they thought of their experience. After the final gameplay session, participants completed the MHS posttest (which is identical to the pretest) before being given a gift card for compensation and leaving.

Player Experience Dataset

In addition to the same pre/posttest and log data collected in the Field Test, throughout the player-experience play, an E4 bracelet was worn by participants to collect skin conductance and heart rate data, and a Tobii system was used to trace users' eye movements. Every data source was synchronized with the eye tracking video stream. This, for example, allowed me to distinguish how participants who were successful on the pre/post assessment reacted to first seeing the final crate vs participants who were less successful.

Methods of Analysis

After all the PX data had been collected for the 9 participants, I compiled the learning analytics from unit 3 of the Field Test to determine which virtual behaviors were associated with high and low performance outcomes (Xing, Guo, Petakovic, & Goggins, 2015). I classified my 4 PX participants who had complete gameplay logs into their respective categories according to the model (See Table 4 on page 56 in chapter 4).

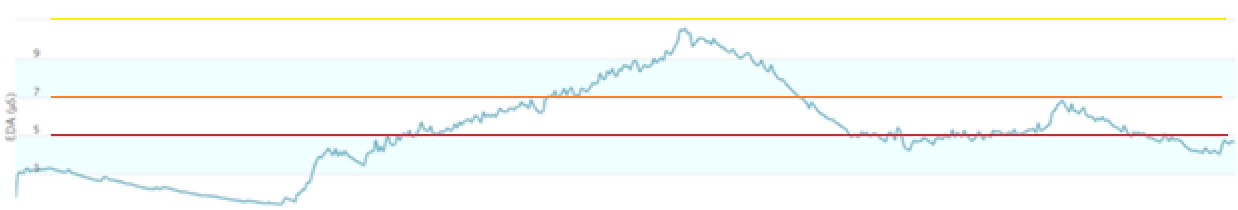


Figure 5: An example image of raw E4 EDA data output showing activity across milliseconds of time

I then examined the PP data of all 9 PX participants' unit 3 sessions in depth to create a use case for each player. The data plot shown in Figure 5 exists over a span of 10 seconds. The line plot indicates a player's electrodermal activity, or skin conductance, at any given moment. These data are shown to correlate with how “aroused” a person is in a non-valanced manner (e.g., this person could be experiencing frustration or engagement but is not experiencing apathy). The orange line indicates the chosen threshold. If this threshold is set too low (red line), too many false positives will be registered; if it is set too high (yellow line) too many significant reactions will be missed. I established a threshold of 7 (See Figure 5 Above) to define when a player’s skin conductance was at a high enough level to qualify as a spike, and then compiled a list of what the player was looking at and what was occurring to cause those reactions for each player (See Appendix G for complete list). Once the list of engagement triggers for each player was compiled, I compared the lists to determine if the players have differences in what engages them.

Finally, I examined if the differences in engagement triggers aligned with any of the determining factors for the players’ pre/posttest outcome predictions. This allowed me a first look into whether certain virtual behaviors may be associated with affective responses, as well as whether moments, which learning analytics determine are indicative of improved pre/posttest performance, elicit any affective reaction.

RQ1 Calculating Performance on the Field Test Assessment

Player performance was determined from the 632 students’ scores on the external pre/post assessment given before and after playing MHS (Reeves, Romine, Laffey, Sadler, & Goggins, 2020). Because unit 3 only had 2 assessment questions on the pre/post assessment, students could only receive a score of 0,1, or 2 (See Table 4 on page 57 of chapter 4). The average score was a 1; so, students who scored a 0 were below average and were categorized as "Low". Otherwise, students were categorized as "High" if they scored a 1 or a 2. Examining how the model links the pre/post assessment scores with players' game logs, shows which in-game behaviors correlated with greater pre/post assessment gains.

Feature Selection

While I initially planned on including every variable and running Principal Component Analysis to reduce the feature set complexity to only the most predictive features; this ultimately did not turn out to be the best way to construct a model. The features used were chosen to optimize both the model's predictive power, and its usefulness. Each feature set (e.g., navigation logs, embedded assessment outcomes, tool interaction, quest completion times, etc.) were initially tested among many others for their expected relevance. Then each feature set was tested individually to validate its relevance in determining high vs low performing students. The results of this research were chosen as features for the final model.

Predicting Field Test Outcomes

Although this is a supervised method, the data that are put into the system entirely determine the outcomes of the algorithm. Table 2 shows an example of 10 state update logs which were sent 20 times per second per player. In addition to player state logs, MHS had multiple unique "triggered" event logs that would only be sent when a certain action was taken by the player such as completing a task, responding to dialogue, or using the in-game menu. During the field test 1.2 million logs were collected in total. While I planned to enter all the player response data including player biographical info, interest results, game experience, and game feedback, these data ultimately did not improve the accuracy of the prediction outcomes in Unit 3.

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

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

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

Table 2: A printout of 10 individual player state logs each containing a: log ID, classroom ID, build version type, install ID, player Name, player ID, timestamp, platform, session ID, teacher ID, log type, unit #, build version #, Player XYZ position, Player XYZ orientation, current Quest, current Task, and Scene Name. A more legible version is included in Appendix E.

In the end, players were classified based on the logs according to their predicted post-test assessment. With this information I was able to determine what Unit 3 gameplay behaviors predicted success on the post assessment for my groups of PX students. While Random Forest became the exact algorithm used for this analysis, many were tested according to the above process until the best model was found. The Random Forrest Algorithm works by generating multiple decision trees for classification, and then ultimately selecting the class chosen by the most trees. The 57 variables included in the model carry different weights of importance in the classification with the U3 pretest score being the most important variable. The top 20 variables and their relative weights are included in Table 3 below, and all 57 are listed in Appendix F. The algorithm's peak accuracy was 91%, but due to the variance created by initial conditions the actual expected accuracy converged to 84%.

only 20 most important variables shown (out of 57)

	Overall
U3PrePerformanceL	100.00
Reasoning.2	51.50
U3.Evidence.B	36.67
seedPerformanceL	33.38
U3...Crate.throw..fail.	30.81
Crash.Diagnostics.Menu.Node.freq	30.34
movementNumber	27.46
JasperCritiqueScore	27.30
U2argumentLevel	26.12
upstreamArgScore	25.96
StateUpdateNumber	25.71
argumentLevel	25.53
U3...Toss.Sensor.Polluted.Same.Area	24.64
Quest.Menu.Node	24.27
jumpNumber	21.70
plantScore	21.34
CREIScore	21.33
Map.Menu.Node.freq	20.64
U3.Claim.II	19.95
Crash.Diagnostics.Menu.Node	19.65

Table 3: A printout of the top 20 most important weighted variables in the model. If there is a conflict in sorting a player between two variables; the algorithm will classify the player according to the more heavily weighted variable.

RQ2 Determining Emotional Response during Player Experience Testing

Player reactions were collected through two primary means, a Tobii Eye-Tracker and an E4 bracelet. By using these two methods I was able to measure key elements of attention and engagement. These two measures combined gave me an accurate representation of when players got excited and what stimulated that reaction.

Using Eye Tracking to Determine Player Attention during Player Experience Testing

The Tobii Eye-Tracker traces the reflection of users' pupils to determine at what XY location on the screen they are looking. From the raw XY data, Tobii automatically compiles fixation and saccade data. While the saccades are essentially unseen sequences of time by the player as they move their eyes from one focus to another, the fixations are exactly what a player is looking at, during moments of focus. Tobii also compiles the fixations into gazes and scans. Gazes are periods of sustained fixation such as looking at a character or environment, while scans are much faster visual processing such as reading. After the sessions I analyzed the data to determine exactly what, and how much, players paid attention to in the game visually. Contrarily, if a player never looked at something in the game, I could be sure that they were at least visually unaware of it.

Using the E4 Bracelet to Calculate Player Engagement during Player Experience Testing

The E4 wristband is one of the least intrusive ways to collect Psychophysiology data from players particularly in comparison to facial EMG electrodes. The device slips onto any wrist in minutes and measures heart rate, temperature, and skin conductance. Spikes in skin conductance indicating emotional arousal allowed me to determine whether players were engaged or bored at any given time during play. By setting the threshold for skin conductance at 7, I was able to compile the list of interactions below of every moment (See Appendix G) of gameplay in U3 for which my PX participants were stimulated. By comparing across all participants, I was able to determine the similarities and differences in emotional reactions to gameplay among players also detailed in chapter 4 below.

Triangulating Game Behaviors, Player Experience, and Assessment Performance

Although Learning Analytics and Psychophysiology studies have been done independently on learning games in the past, the gap this study begins to fill is putting the two together. Is the player's emotional reaction a good indicator of whether they will repeat a failing game strategy or change it? Can player emotions be extracted or predicted from in game behaviors? These are the questions that this study begins to explore. With much larger numbers of participants generating PP data, it would have been possible to add the emotional responses directly into the machine learning algorithm detailed above to determine the most distinguishing behaviors associated with each reaction. However, while this PX case study has far too few participants for those results to be generalizable at all, the current study will evaluate methods and provide insights needed to make a larger sample size study most productive.

RQ3 Predicting Pre/Post Assessment Performance based on Player Experience

By examining the outcomes of the posttest outcomes for each of the players in the PX testing; I was able to begin to look at how different emotional reactions may be related to the patterns of learning outcomes for each group. For example, being stimulated by certain avatars may cause a rise in attention to the curricular content delivered by that avatar. Or reactions to embedded assessment items could indicate whether the player was confident, guessing, or surprised by their outcome which may help to gauge player's mastery level of a particular learning objective. Noting similarities and differences in the PP results of the PX participants as well as their post assessment outcomes, is a good first step toward finding an indication that the players' emotional reaction impacts their performance. If it is possible to correlate PP and assessment outcomes; it may be possible to develop a hypothesis that could be tested with larger sample sizes to build predictive models. With greater numbers it would be possible to empirically validate whether certain PP reactions were indeed predictive of posttest performance outcomes.

RQ4 Predicting Player Experience based on in Game Behaviors

Finally, by examining the previously generated model's features to determine which behaviors defined each group we can begin to determine if any gameplay behaviors may indicate certain emotional reactions from players. For example, if successfully planting seeds in each garden tends to produce an arousal that may indicate the player is feeling confident in their seed planting. Conversely if a player misses all the gardens and has an arousal that may indicate they are frustrated. By finding whether certain key defining gameplay behaviors for the model's clusters are associated with engagement-related events, we can begin to get an indication of how a player's emotional experience can be determined by their gameplay and how that may be impacting their eventual learning performance. An interesting example pattern would be player's getting aroused after placing the first seed in an incorrect garden, then subsequently planting the seeds in the correct gardens. A pattern like this may indicate that the player was initially confused by the task, but after reflecting was able to improve their performance. Depending on the complexity, and of course the validity, of these emotional gameplay signatures, it may be possible in the future to check specific behavioral heuristics during gameplay to identify emotional reactions without any Psychophysiology measures.

Chapter 4: Results

What are the Virtual Behaviors which predict high performance outcomes in MHS?

RQ 1 utilizes insights gained through learning analytics to begin to answer this question. The answer to this question is a Random Forrest Model containing 57 various gameplay behaviors which can all be found in Table 4. Although these predictive virtual

U3 Pre-test Score	dungeon Explored Area	U2 argument Level	U3 Same Area Sensors
U3 Post-test Score	explored Area	U2 bigger Arg Score	U3 Clean Sensors
U3 Pre-test Performance	dialogue Ave Speed	U2 Jasper Critique Score	U3 Downstream Clean
U3 Post-test Performance	hover Node Freq	U2 Backing Info Menu Node frequency	U3 Garden Score
average Speed	U1 argument Level	U2 Chat Log Menu Node freq	U3 Reasoning 1
trigger Number	U1 tutorial Arg Score	U2 Crash Diagnostics Menu Node freq	U3 Reasoning 2
movement Number	U1 Backing Info Menu Node	U2 Help Menu Node freq	U3 Reasoning 3
mission Complete Number	U1 Chat Log Menu Node	U2 Map Menu Node freq	U3 Reasoning 4
State Update Number	U1 Crash Diagnostics Menu Node	U2 Quest Menu Node freq	U3 Reasoning 5
dialogue Number	U1 Help Menu Node	U3 Crate fails	U3 Claim I
arf Related Number	U1 Map Menu Node	U3 Crate successes	U3 Claim II
hot Key Number	U1 Quest Menu Node	U3 Crate Score	U3 Evidence A
toggle Number	U2 find Team Ave Score	U3 Polluted Sensors	U3 Evidence B
jump Number	U2 CREI Score	U3 Downstream Polluted	U3 upstream Arg Score
arg Number			

Table 4: This table shows the 57 variables in the LA model that were determined to be predictive of students Unit 3 posttest outcomes.

behaviors were modeled from the 806 Field Test participants, the results shown in the Appendix are the 57 specific log values for each of the 4 PX participants.

	JG UX 2	JG UX 3	JG UX 4	JG UX 6	Accuracy
Model Prediction	H	H	H	H	75%
Actual Ouctomes	H	L	H	H	N/A

Table 5: This shows the model prediction outcomes for my PX participants whose gameplay logs could be included in the Learning Analytics modeling.

Table 5 shows how participants were classified according to high and low performance on the posttest and the model's accuracy percentage for just these 4 participants. High performance is defined as achieving high scores (1 or 2 out of 2) on the post assessment as described in the Data Collection Section above. Low performance would then be relatively low scores (0 out of 2) on the same post assessment. While performance can also be viewed as the gains players made from their pre-test score to their post test score, using gain scores essentially doubles the classification complexity from 2 groups (Low vs High) to 4 groups (LL, LH, HL, HH) and was out of scope for this exploratory study. The benefit of having a large, diverse population of participants from the 2019 field test created a wide distribution of high and low performance across the entire group. In this case, there ended up being 449 High performers and 183 Low performers in the field test, and 3 High performers and 1 Low performer in my PX test (shown here in Table 6).

Participants\Metrics	U3 Pre-test Score	U3 Post-test Score	U3 Pre-test Performance	U3 Post-test Performance
JG UX 2	2.00	2.00	H	H
JG UX 3	1.00	0.00	H	L
JG UX 4	0.00	1.00	L	H
JG UX 6	0.00	1.00	L	H

Table 6: This shows the posttest performance outcomes for my PX participants whose gameplay logs could be included in the Learning Analytics modeling and their respective coding.

Once the performance data distribution was established and classified, various Clustering methods such as Random Forrest and Tree Bagging were used to identify patterns in groups of players at either end of the performance spectrum. For example, depending on the context of a certain task, a shorter time to completion could be either an indicator of successful performance (e.g., finding the team in Unit 2 quickly by using the map) or an indicator of poor performance (e.g., failing the garden task quickly by choosing the closest, but wrong gardens). The results of this research question are a set of virtual behaviors which differentiate between high and low posttest performance outcomes. The complete list of features chosen are listed in Appendix G.

Participants\Metrics	U3 Pre-test Score	U3 Reasoning 2	U3 Garden Score	U3 Crate fails	movement Number	dialogue Number
JG UX 2	2.00	2.00	0.17	2.00	0.49	0.06
JG UX 3	1.00	3.80	0.13	1.00	0.20	0.30
JG UX 4	0.00	0.00	0.19	0.00	0.65	0.05
JG UX 6	0.00	2.00	0.17	0.00	0.19	0.10

Table 7: Six of the top ten most predictive features out of the fifty-seven features selected in total to construct the model predicting posttest performance for Unit 3 with their respective values from 4 PX participants as examples.

Table 7 shows some interesting LA findings. In total 57 features (See Appendix F for complete list note the logs in the bottom column are all specific to Unit 3 gameplay) were used to predict players' U3 posttest performance with an accuracy of 84%. This Table contains just 6 of these features, which are particularly understandable and actionable. These 6 items were also all among the top 10 weighted variables in the model, meaning they were among the 10 most predictive variables. Pre-Test performance was one of the top weighted variables in multiple models tested. Students who scored high on the test before playing were almost always more likely to score high on the same test after playing than those who scored low on the pretest. “U3 Garden Score” was a calculated value indicating performance on the third task, which happens to be the summative embedded assessment item for Unit 3. In the case of my PX participants, JG UX 3 had the lowest score on this item and was also the only Low performer in the post test. “U3 Crate Fails” on the other hand is the first task of the unit, and JG UX 2 actually had the worst score, this is a good example of why some items such as pre-test scores carry more weight than others. “Dialogue Number” and “Movement Number” are not unit 3 specific values but rather a count of the total dialogue and movement the player has experienced. However, the second task in this unit involves quite a bit of movement in order to trace a pollutant back to its source. A low value here may indicate that players performed well on this task even though this pattern does not seem to be present in my PX participants data. In summary these 57 behaviors when taken together can accurately predict a player’s outcome 84% of the time, and the 6 behaviors discussed here are among the most predictive of all the variables.

What are the specific Psychophysiology (PP) patterns that indicate embodied motivational processes such as attention and engagement, and what game features elicit those patterns?

RQ 2 is primarily a Psychophysiology question. Examining users’ eye tracking data during EDA spikes measured with the E4 bracelet as described in the

Analysis section above resulted in a set 28 of in-game PP events (Listed in Table 8, See Appendix G for the values and descriptions, note event totals tally all PP reactions) when the players

Sam Morning	Arg Final	Crate 4	Cube Tutorial
Same Base 1	Battery Cutscene	Sam "Ta Da!"	Dungeon Cinematic
Crate Cutscene	Sam Garden Intro	1st Sensor	First Pump
Toppo Video	Garden Pump	Halfway	Dungeon Complete
Crate 1	Super Tree	3 clean in a row	First Garden
Crate 2	Alien Ruins	Holo- Toppo	Second Garden
Crate 3	Key Finish	Arg 1	Last Garden

Table 8: This shows the 28 in game events that generated EDA spikes in the 9 PX participants.

entered the affective arousal state, a.k.a. their EDA rose above the threshold set at 7. There were many interesting insights identifiable from even this small sample of nine participants (see Table 9). One example of a feature which consistently seemed to elicit engagement was Face-to-face interaction with the non-playable character (NPC)

Partipants\Moments	Player Totals	Same Base 1	Sam "Ta Da!"	Holo- Toppo	Battery Cutscene	Super Tree
JG UX 1	7	0	1	0	0	1
JG UX 2*	9	1	0	1	1	0
JG UX 3*	10	1	1	1	1	0
JG UX 4*	13	1	1	1	1	1
JG UX 5	8	1	1	0	1	1
JG UX 6*	5	1	0	0	0	1
JG UX 7	7	0	0	1	1	0
JG UX 8	11	1	1	1	1	1
JG UX 9	6	1	1	0	1	0
Event Totals		7	6	5	7	5

Table 9: PP data from 9 PX participants showing the total number of PP events each experienced as well as the top 5 in game events generating PP reactions. 1 indicates the player had a PP reaction at that moment and 0 indicates they did not.

avatars, which in the case of Unit 3 was Sam and Holo-Topo. Both times players interacted with Sam in this way over half of the players had PP events (n=7, and n=6). There was also one interaction with a Hologram NPC, Holo-Topo, which resulted in 5 players having PP events. After that cutscenes, small spans of non-playable gameplay in which there is usually more detailed and important animation being shown, seemed to generate engagement (n=7 and n=5). There were many other events: dialogue statements, quest updates, item interaction, and task completion, which produced PP events in less than half of the participants (n=1-4) that have been detailed in Appendix G. The final interesting result is noting the lack of PP events in the final third of the unit. There is interaction with Sam towards the beginning, and interesting cutscenes in the middle, but the last quest of the unit lacks either of these resulting in no major PP events. In summary, it appears that NPC interaction, and cutscene viewing are two of the most stimulating features of Unit 3 in Mission HydroSci.

How and to what extent can high or low performance be inferred from the cognitive-affective state of players?

Partipants\Moments	Peromance Group	Crate 1	Crate 3	Arg Final	Dungeon Cinematic	Same Base 1
JG UX 2*	High	0	0	0	0	1
JG UX 3*	Low	1	1	1	1	1
JG UX 4*	High	0	0	0	0	1
JG UX 6*	High	0	0	0	0	1
Performance Totals		Low	Low	Low	Low	All

Table 10: PP data from 4 PX participants showing how the PP events relate to PX participants' Post Test performance. This portion shows 4 events that were only experienced by the Low Posttest performer and 1 event experienced by all PX participants.

RQ 3 is the first triangulation question. It begins with the results from RQ 2 and applies the analysis of RQ 1. The results for this analysis are detailed in the bottom row of Appendix G. “High” or “Low” indicate that a PP event was experienced (i.e., in game event caused EDA stimulation ABOVE the threshold level) by ONLY high or low post test performers respectively. This is the most important result for this research question, even though unfortunately there were no “High” events found. Other than “Dungeon Cinematic” the other 3 "Low” values are all on embedded assessment items, possibly indicating the player was confused or frustrated, rather than engaged, with the feedback and results they received (See Table 10 above). Next in order of importance would be “2H” and “1/1”. These values are inverses of each other indicating the low performer

Partipants\Moments	Sam Morning	Sam "Ta Da!"	Super Tree	Alien Ruins	Key Finish	First Garden
JG UX 2*	0	0	0	0	1	0
JG UX 3*	0	1	0	1	0	0
JG UX 4*	1	1	1	1	1	1
JG UX 6*	1	0	1	0	0	1
Performance Totals	2H	1/1	2H	1/1	2H	2H

Table 11: PP data from 4 PX participants showing how the PP events relate to PX participants' Post Test performance. This portion shows 4 events that were only experienced by 2 High Posttest performers and 2 events experienced by 1 High and 1 Low PX participants.

reacted the same as one of the three high performers (See Table 11 above). With a much larger data set (such as the 632 in the FT) this sort of overlap would be expected between the groups of performers. However, with this small sample it is not possible to determine if this pattern is generalizable. Next least important is the “2/1” value. This indicated that 2 high performers experienced the same reaction as the low performer, if this pattern continued at scale, it would not be very informative. Finally, least indicative of a relationship, and therefore least important are the ”All” and the “None” values. While it is great for engagement if an in-game event reliably excites all players who experience it, those data are not helpful for distinguishing between high and low future posttest performance. “None” indicates that while some of my PX participants experienced a PP event, none of the participants with log data experienced one. If this pattern continued at scale, it may be too weak a signal to be useful for model construction.

One limitation of this study is the lack of valence measurements to accompany the E4 arousal data. These data could indicate that a student was frustrated rather than excited, perhaps at receiving poor feedback on various tasks. There were 2 events which resulted in 1 high performer and 1 low performer reacting. With more participants this could end up being an event in which most high performers do not react indicating low performance. All the other events had 2 or more of the high performing players aligned with the single low performing player, which correlates with the overall majority of players rather than a specific subgroup of performers. Unfortunately, due to the low number of participants in the PX study (n = 9) and the even lower number whose logs were not corrupted (n = 4); these results are not as predictive as are the results from RQ1; however, they are good indications of areas for interesting future research. While these results do not have enough participants to test whether PP events can be predictive of posttest performance, this method of data representation and analysis would make sense for testing that hypothesis at scale.

How and to what extent can the cognitive-affective state of players be inferred from virtual behaviors?

RQ 4 is the second triangulation question and the inverse of RQ 3. This question starts with the results from RQ 1 and then applies the analysis in RQ 2. The driving curiosity behind this question is if we have successfully identified indicators that lead to specific pre/post assessment outcomes through Machine Learning Analysis of logs and player behaviors, and those patterns correlate with specific PP event reactions; then it may be possible to detect those player reactions through gameplay behaviors. These results are detailed in the bottom highlighted section of Appendix F and a selection related to the embedded assessment tasks in U3 are in Table 12 below. During the crate return task, the first task of the unit, there were 2 events which only the low performer reacted to. Upon checking the results, the low performer only successfully returned 2 crates (the fourth crate is not present in the logs). Because these crate successes and failures are tallies it is not possible to determine whether the low performer reacted negatively to

Participants\Metrics	U3 Crate successes	U3 Polluted Sensors	U3 Same Area Sesnors	U3 Garden Score
JG UX 2	1.00	7.00	3.00	0.17
JG UX 3	2.00	50.00	20.00	0.13
JG UX 4	3.00	11.00	14.00	0.19
JG UX 6	3.00	4.00	7.00	0.17

Table 12: This table shows the LA results from the embedded assessment tasks in U3. JG UX 3 ended up being the Low Performer and their results show them on the outskirts of performance even among these 4. Looking at the PP events during these embedded assessments also indicates that JG UX 3 was reacting to some of these events perhaps feeling confusion or frustration for performing poorly.

their 2 failures, positively to their 2 successes or a combination of both. However, it does seem like variance in the performance of this task may be related to PP reactions during the task and eventual assessment performance. The only PP result from the pollution task (second task of the unit) was that 2 high performers reacted to the “halfway to the pollution” feedback dialogue. Upon inspecting the feature data from the log those 2 players also spent the least amount of time downstream in polluted portions of the river. This result may indicate that the dialogue statement positively reinforced students who were already succeeding. While the garden task did have one event (the first garden planting) where 2 high performing users reacted, this did not correlate with any pattern in the log feature data. This could be due to the data being condensed into a single score for this embedded assessment task and may indicate some value in adding the first garden planting’s results into the performance prediction model.

Chapter 5: Conclusion

Overview of Findings

Prior to this study LA have been used to gather interaction data and predict student performance outcomes, and PP has been used to get an in depth look at how media consumer's subliminally experience media, but rarely have the two been used in conjunction. This study is intended to be an initial exploration of the impact PP data could have on the LA prediction process and results during game-based learning. The first two RQs ask straightforward LA and PP questions regarding what the predictors of posttest performance in MHS are and what are the in-game events which cause players to become stimulated. While these questions have both been asked and answered many times in other contexts it was important to establish these findings within the MHS context. The final questions examine two possible ways the data could be used in conjunction to enhance the outcomes they traditionally support. One possibility being to better predict learning outcomes by incorporating learners' subliminal reactions into the prediction process, and another being using LA methods to predict PP events without requiring the advanced physical sensors. To this end a variety of methods have been employed during a summative Player Experience test of MHS including clustering, eye tracking, and PP data collection. The initial findings are lists of LA and PP variables that were shown to be important, and the latter findings are explorations into how the data could be used together.

Discussion of RQ1 Findings

What are the Virtual Behaviors which predict high performance outcomes in MHS?

This study found 57 behaviors (See Appendix F) that predict a high performance on the Unit 3 questions of the MHS posttest. While these specific findings are novel and grounded in the context of MHS, they align with previous research on serious games and learning analytics.

One specific example of the relationship between in-game behaviors and players' learning processes was found in the study by Martinez-Garza and Clark (2017) mentioned earlier, which emphasized the impact of prior knowledge on players problem solving vs guessing behaviors. Our findings build upon this growing body of research by providing further evidence of how guessing during gameplay on tasks such as the Unit crate collection, pollution testing, and garden planting tasks can negatively impact the learning outcomes. These findings can be utilized in 3 interesting ways. Most directly this classification can be used on a student-by-student basis to alert teachers of potential player learning issues before they finish MHS and complete the posttest, allowing them to intervene and potentially course correct that student. With additional programming these findings could be directly incorporated into the game so that the gameplay dynamically course corrects players automatically as they play. For future researchers, while these results are specific to this MHS context there may be a pattern among the patterns of various serious games and this study provides one more data set to help determine that. Perhaps different genres (e.g., 3D adventure vs 2D puzzle) tend to have different predictor behaviors or perhaps games designed for different ages. The more studies done like this, the closer we are to finding that out.

Discussion of RQ2 Findings

What are the specific Psychophysiology (PP) patterns that indicate embodied motivational processes such as attention and engagement, and what game features elicit those patterns?

The specific pattern to identify these embodied processes was measuring a GSR spike above 7 using the E4 wristband during a Gaze period detected by the Tobii Eye Tracker. 28 features (See Appendix G) were discovered to elicit a reaction in at least one of my participants. Again, these individual findings were specific to the MHS context, but overall, these psychophysiological findings support the same findings as Nacke and Lindley (2008), that psychological states, such as positive emotions and arousal, can be accurately derived from physiological responses during gameplay; providing insight into the cognitive and affective processes involved in game experiences. The key takeaways here were identifying two key contexts that repeatedly caused spikes for my participants. The first context was NPC interaction. When a human face came on screen most participants were immediately engaged and attending to it. This could be because NPC interactions are somewhat infrequent in MHS (only 3 in the hour-long unit 3 gameplay), however as these NPCs are often vehicles for conveying curricular content; this finding is very interesting for serious game designers. The second engaging context across users were in-game cutscenes. These are more expected to be exciting as they often show more complex animations and interactions, like explosions. With knowledge like this serious game designers are better equipped to keep their students engaged and deliver content in ways that players will pay attention to.

Discussion of RQ3 Findings

How and to what extent can high or low performance be inferred from the cognitive-affective state of players?

By incorporating these PP data as inputs into a traditional LA data pipeline and analysis process it is possible to determine if performance can be inferred from a player's cognitive-affective state. While this study was small and exploratory, 4 states were found to be correlated exclusively with low performance indicating this may be a signal of a struggling student. In addition, 5 states were found to be correlated with a majority of high performers indicating that with a greater population these may also be powerful predictors. While we cannot judge the generalizability of these specific findings they do offer new directions for serious games research that have long been supported by the literature. As Chittaro and Buttussi, (2015) confirmed the relationship between intense emotional reactions and memory recall in a learning game, findings like this, if validated at scale, would provide a rich new insight into the broader connection between emotional reactions and learning in serious games. Deeper insight into not just what behaviors predict learning, but what learners' reactions to those behaviors mean will help begin to narrow down precisely when and how learning occurs; so ultimately better learning experiences can be designed.

Discussion of RQ4 Findings

How and to what extent can the cognitive-affective state of players be inferred from virtual behaviors?

If PP reactions are known to have an impact on learning; it, then becomes important to identify PP states as efficiently as possible. By targeting these PP states as outcomes of our traditional LA data pipeline and analysis process it may be possible to infer the cognitive-affective state of a player from virtual behaviors. The results of this study showed all 3 of the main tasks and embedded assessments were both included in the LA model for predicting posttest performance, as well as the list of items that triggered PP events. These exploratory data indicate that players were reacting to their in-game task performance, a key learning predictor. Specifically, players who have PP reactions to successful performance may perceive that they are making valuable progress and may be motivated to continue learning. On the other hand, players who have PP reactions to unsuccessful performance may feel frustrated and understand that they are not learning effectively. This is the most novel finding of this study. Mapping out what in game behaviors, such as successes or failures are likely to trigger certain PP states could allow for a purely behavioral model that first predicts a player's emotional state and then makes a more confident prediction on performance. Validating that players are reacting to their in-game performance could be an initial indication they are subliminally aware of whether learning is taking place. This finding would be valuable to learning researchers trying to identify the process as well as serious game designers testing whether in-game feedback is effective.

Future Directions

Help for Learning Analysts

Choosing an appropriate Learning Analytics method involved a rigorous process of exploring a variety of models and assessing their accuracy in predicting posttest results based on player performance. Tree Bagging was another method considered, however it's prediction accuracy on the training data was only 82%. Ultimately the Random Tree Forest model demonstrated the best model fit for this data with a prediction accuracy of 84%. Its ability to handle nonlinear relationships and interactions among variables, combined with its high accuracy and robustness, made it the ideal choice for predicting player performance in my study. The careful selection of this model helped ensure that my results were reliable and relevant, and while the same method might not work for all serious games, the same method testing process should work for any project.

Logging "everything" is the ultimate desire for any big data specialist but is not exactly practical development-wise or analysis-wise. These results from the PP methods described above can help to identify important types of events which may not currently be captured. Important curricular items like embedded assessments should be examined in detail to determine if and where players may be having PP reactions, such as the results of the first garden planting task. Additionally non-curricular items like NPC interactions and cutscene viewing time could be counted to determine in general how engaging of an experience the player is having. By having a greater sense of how game features impact a player's experience, researchers would be able to better understand how that experience relates to previous and subsequent behavior. With better measurement tools, greater numbers of participants, and more automated data processing it would be possible to incorporate PP results directly into the LA performance prediction model providing much richer insights into the Learning Process.

Help for Psychophysiology Researchers

Games are playful emotional environments. This media provides a different context for rich insights into the user's learning experience, focused on engaging with media rather than passively consuming it. Chittaro and Buttussi (2015) found that emotional intensity can increase memory retention. My exploratory study also found a correlation between participants PP states, in game behaviors, and learning outcomes. Giving participants a more active, emotionally charged, role could help uncover new methods and insights for PP analysis. With more technology focused on non-linear dynamic eye-tracking, and more PP analysis automation, much larger studies can be run with greater insights being discovered.

Help for Serious Game Designers

Tailoring game development to a specific audience requires intense formative assessment with large amounts of user feedback to assure that the final product appeals to them. The kind of PX testing described above gives an extremely deep perspective of play-testers experience which can in turn help serious game designers create more engaging and effective learning experiences. For example, in an early pilot test version of the game a small water drip animation would play when the players emerged from underwater that consistently produced PP results like the face-to-face NPC interaction found above. Designers could utilize these known stimulating game features to increase players engagement and attention either at key moments of curricular content or during slow portions of the unit such as the final third of unit 3 examined above.

Help for Learning Researchers in General

This is the Golden Age of Brain Research and to drive discovery we need to be constantly finding new ways to dive deeper into learning minds before, during, and after the actual moments where learning happens to uncover the contexts and processes that enable it to occur. By exploring how to map out the learning process for a large number of users in entirely defined contexts, we can begin to find the patterns and anomalies. Those patterns help define new theories, while the anomalies help drive experimental testing to determine the limits of those theories. The technology and time are ripe for these results; however, studies like this will require rich, controlled learning environments to provide highly detailed data to fuel the research. This will only come with more Serious Game development research like MHS as well as more supporting technology and interest in measuring the Player Experience within those games.

Limitations

A limitation of this study was the use of only 9 participants. Due to the limited knowledge and lack of prior large-scale studies on these Psychophysiology phenomena, it is not appropriate to ask teachers and students to use their class time setting up and using the potentially burdensome artifacts required to collect the Psychophysiology data of this study on a significant number of students. Due to this limitation, the PP portions of this study are being approached as exploratory rather than as experimental. In the future it may be possible to have participants in each class wear an E4 bracelet or another Psychophysiology collection device during play. By distributing the devices among all participating classes in a large field test, such as the one used for the analytics portion of this study, it may be possible to achieve a significant number of participants.

The second major limitation of this study stems from our lack of access to a usable and scalable facial EMG system. The long-term purpose for which this study is an initial step, is to investigate the benefits of incorporating PP data into the LA process. To that end the process for acquiring the PP data must be scalable to generate the large amounts of data necessary for the analytics to produce accurate usable results. While not having EMG is a limitation of my dissertation study, I hope to include it in future studies; if I can find a less intrusive sensor to collect the data.

The third limitation of this study is the use of linear static eye tracking during gameplay. This resulted in the eye-tracking being analyzed separately and in a much more time-consuming, and probably less insight generating, manner. After conducting this exploratory study, I have found *GazeMapping* by iMotions which can collect facial EMG through webcams and does not require training to set up properly. Using this or another “nonlinear dynamic eye tracking solution” would enable us to automate the eye tracking analysis into our usual learning analytic pipeline.

Conclusion

Michio Kaku described these times as the golden age of brain research. His statement sets the stage for risky research and inspires researchers to make big breakthroughs. This exploratory study has been a small first step into taking these risks and making these breakthroughs. Discerning the impact that a player’s emotional experience has on their learning outcomes could pave the way for deeper discoveries and insights. Player’s take notice and respond positively to in-game characters and cutscenes, and this spike in attention and engagement may be able to be harnessed by serious game designers for delivering key instructional content. Ensuring games are designed, developed, and tested not just from a behavioral perspective but also from an emotional perspective will be an important next step towards these goals.

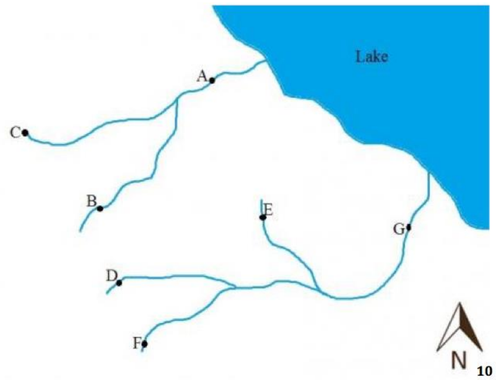
The present study aimed to enhance our understanding of the role of PP data in predicting player performance in the context of learning games. The results suggest that it may be feasible to use PP data to predict player performance, and additionally that LA may be useful in predicting player arousal states. This study contributes to the growing body of research on the importance of PP and LA in game-based learning, and by continuing to identify the types of in-game behaviors that trigger psychophysiological reactions, serious game designers will be better able to deliver key instructional content for maximizing learning outcomes.

Future research should continue to explore the use of psychophysiological data in game-based learning, with a focus on developing more accurate and efficient methods for measuring and analyzing these data. Additionally, future studies could investigate the impact of specific game features and design elements on psychophysiological responses and learning outcomes. Ultimately, a more comprehensive understanding of the role of psychophysiological experiences in game-based learning could lead to the development of more effective and engaging learning games.

Appendices

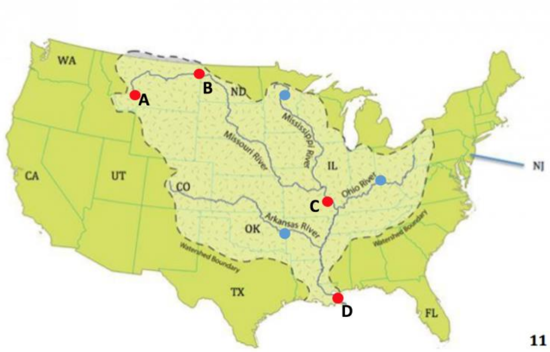
Appendix A - Unit 3 Assessment Items

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?
A. Site F B. Site E C. Site G D. Site B

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

Appendix B – Pre and Post Test Scores

Score Content	Water Science Knowledge													U3			U3								Argumentation Knowledge												CK	AK
Item #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	1	2	3	4	5	6	7	8	9	10	11	12	T	T
JG UX 1-Pre	0	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	0	1	0	0	1	1	1	1	1	1	1	1	1	0	1	0	17	10
JG UX 1-Post	1	1	1	1	1	1	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	0	0	1	1	0	1	1	1	1	1	1	1	1	0	19	10
JG UX 2-Pre	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	1	1	1	1	1	1	1	21	10
JG UX 2-Post	0	1	1	1	1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	21	11
JG UX 3-Pre	0	0	1	1	1	1	0	1	1	0	0	0	1	1	0	1	0	1	1	1	1	1	0	1	1	0	0	1	1	1	1	1	1	0	1	0	15	8
JG UX 3-Post	0	0	1	0	1	1	1	1	1	0	1	0	1	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	16	11
JG UX 4-Pre	1	1	1	1	1	1	0	1	1	0	0	0	1	0	0	1	0	1	1	1	1	1	0	0	1	0	1	1	1	0	1	1	1	0	1	0	15	8
JG UX 4-Post	1	1	1	1	1	1	0	1	1	0	1	0	1	1	0	0	0	1	1	1	0	1	1	0	1	1	1	1	1	1	0	1	1	0	1	0	16	9
JG UX 5-Pre	0	0	1	0	1	1	1	0	0	0	0	1	0	1	1	1	0	1	1	1	0	1	0	0	1	0	1	1	1	1	0	0	1	0	1	0	12	7
JG UX 5-Post	1	0	1	0	1	1	1	0	1	0	0	1	0	1	1	1	0	1	0	1	1	1	0	1	1	0	1	1	1	1	1	0	1	0	1	0	15	8
JG UX 6-Pre	0	0	1	1	1	1	1	0	1	0	0	0	1	0	0	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	0	14	10
JG UX 6-Post	0	0	1	1	1	1	1	0	1	0	0	0	0	1	1	1	0	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	0	1	0	15	9
JG UX 7-Pre	0	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	1	1	1	1	1	1	1	1	1	18	11
JG UX 7-Post	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	1	1	1	1	1	1	1	1	1	18	11
JG UX 8-Pre	1	0	0	1	1	1	0	1	1	0	0	0	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	16	11
JG UX 8-Post	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	19	11
JG UX 9-Pre	1	1	0	1	1	0	1	1	1	0	0	1	0	1	0	1	1	1	1	1	1	1	1	0	1	1	0	1	1	1	0	0	1	1	1	0	17	8
JG UX 9-Post	1	1	1	1	1	1	0	1	1	1	0	0	1	1	0	1	1	1	1	1	1	1	0	0	1	1	0	1	1	1	1	1	1	0	1	1	18	10

Appendix C - Player Experience Protocol

Pre PX Testing
1. Schedule Participant
2. Reserve the Collab and Study Room
3. Make sure “MHS Spring FT 18.exe” is installed on the computer
4. Charge the E4 Bracelet

PX Testing
1. Greet Participant
2. Have Participant’s parent sign Consent Form
3. Have Participant sign Youth Assent Form
4. Orient Participant to the activities for the day
5. Calibrate Tobii Eye Tracker and E4 Bracelet for Participant
6. Have Participant play “MHS Spring FT 18.exe”
7. Help Participant only when directly asked
8. Debrief Participant with PX Interview Questions
9. Give the Participant a gift card

Post PX Testing
1. Transfer the Eye Tracking Video and PP data stream to the external HDD
2. Return the Tobii Eye Tracker to the IE Lab
3. Return the external HDD and the E4 Bracelet to MULE Games

Appendix D - Player Experience Interview Questions

What did you all think of MHS?

Was it fun?

What was your favorite part?
What was your least favorite part?
What would you change about it?

Did you feel like you learned anything by playing MHS?

What did you learn?
Did you think MHS was a fun way to learn science, or not much better than regular science lessons?

If we wanted students to learn even more about water systems while playing MHS, do you have any suggestions for how we could improve MHS?

What did you think of the story?

Was it interesting? Did you like it?
Was anything confusing or did you have any questions about it?

What would be some good rewards for completing missions in the game?
What fun stuff should we add to the game?

What did you think of the first level on the space station before it crashed?

Do you feel like it was a good way to get started in the game?

Were you ever confused on what to do?

What did you think of the characters?

Who was your favorite?
Who was your least favorite?
What would you change about them?

What did you think of exploring the outside environments like the mountains, river valley, and desert?

Was it fun to explore the environments?

What was your favorite?
What was your least favorite?
What would you change about those parts?

Did you find any of our secrets or collectibles?

What did you find?
Would you like more of that stuff?

What did you think of exploring the Alien Ruins with the power cubes?

Was it fun to explore the Alien Ruins?

What was your favorite part?
What was your least favorite part?
What would you change about the ruins?

What was your characters purpose for going through the puzzles?

What did you find?
Would you like more of that stuff?

What did you think of the Argumentation System with the drag and drop circles?

Was it fun to make the Arguments?

What was your favorite part?
What was your least favorite part?
What would you change about the argumentation?

We have heard that some players may have just guessed while doing the arguments. Tell us how you figured them out?

Did you read each of the options before trying one?
Did you read each of the feedbacks, and were they helpful?
Was there a point when you stopped reading the feedback?

What was your character's purpose for making the arguments?

Would you like more chances to do arguments in MHS?

What did you think of the final mission to the moon?

Was it fun?
Was it scary or tense?
What would you like to see happen at the end?

Thanks for all your help today….one last question…..if you could make any change to MHS that you wanted what would it be (not counting bugs, crashes or slow performance).

Appendix E – Example Log Data From MHS

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

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

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

Appendix F – LA Model Values for PX Participants

Participants\Metrics	U3 Pre-test Score	U3 Post-test Score	U3 Pre-test Performance	U3 Post-test Performance	average Speed	trigger Number	movement Number	mission Complete Number	State Update Number	dialogue Number	arf Related Number	hot Key Number	toggle Number	jump Number	arg Number	dungeon Explored Area	explored Area	dialogue Ave Speed	hover Node Freq
JG UX 2	2.00	2.00	H	H	2.12	0.07	0.49	0.01	0.33	0.06	0.01	0.01	0.00	0.02	0.01	0.80	0.72	1.82	72.00
JG UX 3	1.00	0.00	H	L	3.67	0.30	0.20	0.10	0.21	0.30	0.20	0.01	0.00	0.10	0.10	0.72	0.60	3.50	170.00
JG UX 4	0.00	1.00	L	H	1.28	0.05	0.65	0.01	0.21	0.05	0.00	0.00	0.00	0.03	0.01	0.68	0.65	1.49	19.00
JG UX 6	0.00	1.00	L	H	3.40	0.08	0.19	0.02	0.56	0.10	0.01	0.01	0.00	0.01	0.01	0.71	0.59	2.12	42.00

Participants\Metrics	U1 argument Level	U1 tutorial Arg Score	U1 Backing Info Menu Node	U1 Chat Log Menu Node	U1 Crash Diagnostics Menu Node	U1 Help Menu Node	U1 Map Menu Node	U1 Quest Menu Node	U2 find Team Ave Score	U2 CREI Score	U2 argument Level	U2 bigger Arg Score	U2 Jasper Critique Score	U2 Backing Info Menu Node freq	U2 Chat Log Menu Node freq	U2 Crash Diagnostics Menu Node freq	U2 Help Menu Node freq	U2 Map Menu Node freq	U2 Quest Menu Node freq
JG UX 2	2.00	1.00	3.50	0.00	0.00	0.00	2.88	2.00	0.33	1.00	4.00	2.00	0.14	2.00	0.00	0.00	0.00	92.00	6.00
JG UX 3	2.00	1.00	1.00	0.00	2.50	0.00	3.50	4.00	0.00	0.10	3.00	2.00	0.00	1.00	0.00	2.00	0.00	119.00	2.00
JG UX 4	1.00	1.00	0.00	0.00	0.00	0.00	2.80	0.00	1.00	0.67	4.00	2.00	0.20	0.00	0.00	0.00	0.00	5.00	0.00
JG UX 6	5.00	1.00	2.00	2.33	1.00	1.00	3.10	0.00	1.00	0.81	4.00	2.00	0.20	1.00	3.00	4.00	1.00	68.00	0.00

Participants\Metrics	U3 Crate fails	U3 Crate successes	U3 Crate Score	U3 Polluted Sensors	U3 Downstream Polluted	U3 Same Area Sesnors	U3 Clean Sensors	U3 Downstream Clean	U3 Garden Score	U3 Reasoning 1	U3 Reasoning 2	U3 Reasoning 3	U3 Reasoning 4	U3 Reasoning 5	U3 Claim I	U3 Claim II	U3 Evidence A	U3 Evidence B	U3 upstream Arg Score
JG UX 2	2.00	1.00	0.33	7.00	106.00	3.00	12.00	3.00	0.17	2.00	2.00	2.18	2.25	1.00	1.71	2.40	2.60	2.20	1.00
JG UX 3	1.00	2.00	0.67	50.00	350.00	20.00	13.00	5.00	0.13	3.50	3.80	4.00	4.00	4.00	3.80	3.50	1.00	1.00	1.00
JG UX 4	0.00	3.00	1.00	11.00	174.00	14.00	12.00	1.00	0.19	0.00	0.00	2.67	4.00	4.00	2.50	0.00	2.00	2.67	0.00
JG UX 6	0.00	3.00	1.00	4.00	258.00	7.00	9.00	7.00	0.17	0.00	2.00	4.00	3.00	1.00	3.00	0.00	2.00	2.17	2.00

U3 Pre-test Score	Raw Assessment Score out of 2	U2 argument Level	The number of times the player attempted the argument.
U3 Post-test Score	Raw Assessment Score out of 2	U2 bigger Arg Score	Argument Success Outcome (2=completed; 1=hints; 0=failed)
U3 Pre-test Performance	Raw Score Classification (H=1 or 2 & L=0)	U2 Jasper Critique Score	A normalized value of how many attempts the player used to critique.
U3 Post-test Performance	Raw Score Classification (H=1 or 2 & L=0)	U2 Backing Info Menu Node freq	The amount of time the player viewed the backing info.
average Speed	The Average Speed it took players to complete a task.	U2 Chat Log Menu Node freq	The amount of time the player viewed the chat log menu.
trigger Number	The total number of "triggered" events players interacted with.	U2 Crash Diagnostics Menu Node freq	The amount of time the player viewed the crash diagnostics.
movement Number	A normalized value of how much the player moved throughout the unit	U2 Help Menu Node freq	The amount of time the player viewed the help menu.
mission Complete Number	Percentage of how much of the total Game the player completed	U2 Map Menu Node freq	The amount of time the player viewed the map menu.
State Update Number	A normalized value of how much the player's state change throughout the unit	U2 Quest Menu Node freq	The amount of time the player viewed the quest menu.
dialogue Number	A normalized value of how many dialogue statements the player encountered dialouge.	U3 Crate fails	The number of crates the player threw into the wrong river.
arf Related Number	A normalized value of how many times the player interacted with the arf menu.	U3 Crate successes	The number of crates the player threw into the correct river.
hot Key Number	A normalized value of how many times the player used hot keys.	U3 Crate Score	A normalized value of how successful the player was with the crates.
toggle Number	A normalized value of how many times the player toggled the escape menu.	U3 Polluted Sensors	The number of sensors the player threw in polluted water.
jump Number	A normalized value of how many times the player jumped.	U3 Downstream Polluted	The amount of polluted area the player explored.
arg Number	A normalized value of how many times the player attempted the argument.	U3 Same Area Sesnors	The number of sensors the player threw in areas they already explored.
dungeon Explored Area	A normalized value of how much the player moved inside the dungeon.	U3 Clean Sensors	The number of sensors the player threw in clean water.

explored Area	A normalized value of how much the player moved around the exterior terrain.	U3 Downstream Clean	The amount of clean area the player explored.
dialogue Ave Speed	The average amount of time dialogued statements remained on screen.	U3 Garden Score	A normalized value of how successful the player was with the gardens.
hover Node Freq	The number of times the player got off and on the hoverboard.	U3 Reasoning 1	The amount of time the player viewed the first reasoning statement.
U1 argument Level	The number of times the player attempted the argument.	U3 Reasoning 2	The amount of time the player viewed the second reasoning statement.
U1 tutorial Arg Score	Argument Success Outcome (1=completed; 0=failed)	U3 Reasoning 3	The amount of time the player viewed the third reasoning statement.
U1 Backing Info Menu Node	The amount of time the player viewed the backing info.	U3 Reasoning 4	The amount of time the player viewed the fourth reasoning statement.
U1 Chat Log Menu Node	The amount of time the player viewed the chat log menu.	U3 Reasoning 5	The amount of time the player viewed the fifth reasoning statement.
U1 Crash Diagnostics Menu Node	The amount of time the player viewed the crash diagnostics.	U3 Claim I	The amount of time the player viewed the first claim statement.
U1 Help Menu Node	The amount of time the player viewed the help menu.	U3 Claim II	The amount of time the player viewed the second claim statement.
U1 Map Menu Node	The amount of time the player viewed the map menu.	U3 Evidence A	The amount of time the player viewed the first evidence statement.
U1 Quest Menu Node	The amount of time the player viewed the quest menu.	U3 Evidence B	The amount of time the player viewed the second evidence statement.
U2 find Team Ave Score	A discretized and then normalized value of how long the player searched for the team.	U3 upstream Arg Score	Argument Success Outcome (2=completed; 1=hints; 0=failed)
U2 CREI Score	A normalized value of how many attempts the player used to classify arguments.		

Appendix G – PP Moments for PX Participants

Partipants\Moments	Peromance Group	Player Totals	Sam Morning	Same Base 1	Crate Cutscene	Toppo Video	Crate 1	Crate 2	Crate 3	Crate 4	Sam "Ta Da!"	1st Sensor	Halfway	3 clean in a row	Holo-Toppo
JG UX 1	N/A	7	0	0	1	0	0	1	0	0	1	0	0	0	0
JG UX 2*	High	9	0	1	0	1	0	0	0	1	0	0	1	0	1
JG UX 3*	Low	10	0	1	0	0	1	0	1	0	1	0	0	0	1
JG UX 4*	High	13	1	1	1	0	0	0	0	0	1	0	1	0	1
JG UX 5	N/A	8	0	1	0	0	1	0	0	0	1	0	0	0	0
JG UX 6*	High	5	1	1	0	0	0	1	0	0	0	0	0	0	0
JG UX 7	N/A	7	0	0	0	0	1	0	0	1	0	0	0	1	1
JG UX 8	N/A	11	0	1	1	0	0	0	0	0	1	1	0	0	1
JG UX 9	N/A	6	0	1	0	0	0	0	0	1	1	0	0	0	0
Event Totals			2	7	3	1	3	2	1	3	6	1	2	1	5
Performance Totals			2H	All	1H	1H	Low	1H	Low	1H	1/1	None	2H	None	2/1

Partipants\Moments	Arg 1	Arg Final	Battery Cutscene	Sam Garden Intro	Garden Pump	Super Tree	Alien Ruins	Key Finish	Cube Tutorial	Dungeon Cinematic	First Pump	Dungeon Complete	First Garden	Second Garden	Last Garden
JG UX 1	0	0	0	1	0	1	0	0	0	0	1	1	0	0	0
JG UX 2*	1	0	1	0	0	0	0	1	0	0	0	0	0	0	1
JG UX 3*	0	1	1	0	0	0	1	0	0	1	0	0	0	0	1
JG UX 4*	0	0	1	0	0	1	1	1	0	0	0	1	1	0	1
JG UX 5	0	0	1	0	0	1	1	0	0	1	1	0	0	0	0
JG UX 6*	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0
JG UX 7	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1
JG UX 8	0	0	1	1	0	1	0	0	1	1	0	0	1	0	0
JG UX 9	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0
Event Totals	1	1	7	2	1	5	3	2	1	3	2	2	3	2	4
Performance Totals	1H	Low	2/1	None	None	2H	1/1	2H	None	Low	None	1H	2H	None	2/1

Sam Morning	The dialogue statement telling the player to go see Sam	Arg Final	A dialogue statement letting the player know they finished the argument.
Same Base 1	The first time the player sees Sam in the Unit.	Battery Cutscene	A cutscene showing the polluted battery being dug up.
Crate Cutscene	The cutscene first showing where the crates are.	Sam Garden Intro	The player sees same for the second time in the unit.
Toppo Video	A short video explaining downstream pollution.	Garden Pump	A cutscene introducing the garden pumps.
Crate 1	A cut scene showing whether the players succeeded or failed when they threw crate 1.	Super Tree	A cutscene showing the super tree when the player first discovers it.
Crate 2	A cut scene showing whether the players succeeded or failed when they threw crate 2.	Alien Ruins	A cinematic cutscene showing the player the alien ruins.
Crate 3	A cut scene showing whether the players succeeded or failed when they threw crate 3	Key Finish	The player successfully completing the key puzzle to enter the dungeon.
Crate 4	A cut scene showing whether the players succeeded or failed when they threw crate 4.	Cube Tutorial	A cutscene showing the players how to use the cubes.
Sam "Ta Da!"	The cut scene showing Sam’s new base after receiving the crates.	Dungeon Cinematic	A cinematic showing an overview of the dungeon.
1st Sensor	An animation showing the results of the tutorial where a player throws the first sensor.	First Pump	The first time a player recovers a pump for completing a dungeon wing.
Halfway	A dialogue statement that shows where the player is halfway through the pollution task.	Dungeon Complete	When the player completes the dungeon and receives the final pump.
3 clean in a row	A reaction a single player had after messing up 3 in a row.	First Garden	The first time a player installs a pump at a garden.
Holo-Toppo	A hologram version of Toppo where the players talked to right before the argument.	Second Garden	The second time a player installs a pump at a garden.
Arg 1	The first time the player submitted an argument.	Last Garden	The third time a player installs a pump at a garden.

References

Alkan, S., & Cagiltay, K. (2007). Studying computer game learning experience through eye tracking. *British Journal of Educational Technology*, *38*(3), 538–542. https://doi.org/10.1111/j.1467-8535.2007.00721.x

Baek, Y. K. (2008). What Hinders Teachers in Using Computer and Video Games in the Classroom? Exploring Factors Inhibiting the Uptake of Computer and Video Games. *CyberPsychology & Behavior*, *11*(6), 665–671. https://doi.org/10.1089/cpb.2008.0127

Bagley, J., Lee, P.-L., & Rankin, Y. (2015). The Development of an Eye-Tracking Program to Examine Working Memory During Gameplay.

Barab, S., Thomas, M., Dodge, T., Carteaux, R., & Tuzun, H. (2005). Making learning fun: Quest Atlantis, a game without guns. *Educational Technology Research and Development*, *53*(1), 86–107.

Barab, S. A., Gresalfi, M., & Ingram-Goble, A. (2010). Transformational Play: Using Games to Position Person, Content, and Context. *Educational Researcher*, *39*(7), 525–536. https://doi.org/10.3102/0013189X10386593

Bergey, B. W., Ketelhut, D. J., Liang, S., Natarajan, U., & Karakus, M. (2015). Scientific Inquiry Self-Efficacy and Computer Game Self-Efficacy as Predictors and Outcomes of Middle School Boys' and Girls' Performance in a Science Assessment in a Virtual Environment. *Journal of Science Education and Technology*, *24*(5), 696–708. https://doi.org/10.1007/s10956-015-9558-4

Bybee, R. W. (2014). NGSS and the Next Generation of Science Teachers. *Journal of Science Teacher Education*, *25*(2), 211–221. https://doi.org/10.1007/s10972-014-9381-4

Byun, J. H., Loh, C. S., & Zhou, T. (2014). Assessing play-learners' performance in serious game environments by using in situ data: Using eye tracking for Serious Game Analytics. In *Annual*

Conference of the Association for Educational Communications and Technology (AECT), Jacksonville, FL.

Chanel, G., Rebetez, C., Bétrancourt, M., & Pun, T. (2011). Emotion Assessment From Physiological Signals for Adaptation of Game Difficulty. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *41*(6), 1052–1063. https://doi.org/10.1109/TSMCA.2011.2116000

Chen, J. (2007). Flow in games (and everything else). *Communications of the ACM*, *50*(4), 31–34.

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

Cowley, B., Charles, D., Black, M., & Hickey, R. (2008). Toward an understanding of flow in video games. *Computers in Entertainment*, *6*(2), 1. https://doi.org/10.1145/1371216.1371223

Cress, U. (2014). Proceedings of the European MOOC Stakeholder Summit 2014. Retrieved from https://www.researchgate.net/profile/Anja_Lorenz/publication/263543544_Open_Online_Courses_in_the_context_of_higher_education_an_evaluation_of_a_German_cMOOC/links/54941c560cf2e1b6095f97bc.pdf

D'Angelo, C., Rutstein, D., Harris, C., Bernard, R., Borokhovski, E., & Haertel, G. (2014). Simulations for STEM learning: Systematic review and meta-analysis. *Menlo Park: SRI International*. Retrieved from https://pdfs.semanticscholar.org/6bf2/15af93f56403e2fe8cb8affe798d65b1142d.pdf

Daniel, J. (2012). Making sense of MOOCs: Musings in a maze of myth, paradox and possibility. *Journal of Interactive Media in Education*, *2012*(3). Retrieved from http://jime.open.ac.uk/article/view/2012-18/sirjohn.ca/

DeJarnette, N. (2012). America's children: Providing early exposure to STEM (science, technology, engineering and math) initiatives. *Education*, *133*(1), 77–84.

Drachen, A., Thurau, C., Sifa, R., & Bauckhage, C. (2014). A comparison of methods for player clustering via behavioral telemetry. *arXiv Preprint arXiv:1407.3950*.

Dyckhoff, A. L., Zielke, D., Bültmann, M., Chatti, M. A., & Schroeder, U. (2012). Design and Implementation of a Learning Analytics Toolkit for Teachers. *Journal of Educational Technology & Society*, *15*(3), 58–76.

Ezen-Can, A., Boyer, K. E., Kellogg, S., & Booth, S. (2015). Unsupervised modeling for understanding MOOC discussion forums: a learning analytics approach (pp. 146–150). ACM Press. https://doi.org/10.1145/2723576.2723589

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, *111*(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Fried, R. L. (2005). The Game of School: Why We All Play It, How It Hurts Kids, and What It Will Take to Change It. *Education Review // Reseñas Educativas; 2005DO - 10.14507/er.v0.594*. Retrieved from http://edrev.asu.edu/edrev/index.php/ER/article/view/594

Fudenberg, D., & Levine, D. K. (1998). *The Theory of Learning in Games*. MIT Press. Retrieved from https://books.google.com/books?id=G6vTQFluxuEC

Gale, A., & Edwards, J. (1983). Psychophysiology and individual differences: Theory, research procedures, and the interpretation of data. *Australian Journal of Psychology*, *35*(3), 361–379. https://doi.org/10.1080/00049538308258749

Ganglbauer, E., Schrammel, J., Deutsch, S., & Tscheligi, M. (2009). Applying psychophysiological methods for measuring user experience: possibilities, challenges and feasibility. In *Workshop on user experience evaluation methods in product development*. Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.189.3410&rep=rep1&type=pdf

García-Redondo, P.; García, T.; Areces, D.; Núñez, J.C.; Rodríguez, C. (2019) Serious Games and Their Effect Improving Attention in Students with Learning Disabilities. *Int. J. Environ. Res. Public Health*, *16*, 2480. https://doi.org/10.3390/ijerph16142480

Gee, J. P. (2003). What video games have to teach us about learning and literacy. *Comput. Entertain.*, *1*(1), 20–20.

Gee, J. P. (2005). Good video games and good learning. In *Phi Kappa Phi Forum* (Vol. 85, p. 33). THE HONOR SOCIETY OF PHI KAPPA PHI. Retrieved from http://norcalwp.org/pdf/Gee--Learning_Principles_Articles.pdf

Gee, J. P. (2005). Learning by design: Good video games as learning machines. *E-Learning and Digital Media*, *2*(1), 5–16.

Green, S. R., Kragel, P. A., Fecteau, M. E., & LaBar, K. S. (2014). Development and validation of an unsupervised scoring system (Autonomate) for skin conductance response analysis. *International Journal of Psychophysiology*, *91*(3), 186–193. https://doi.org/10.1016/j.ijpsycho.2013.10.015

Goggins, S., Schmidt, M., Guajardo, J., & Moore, J. L. (2011). 3D Virtual Worlds: Assessing the Experience and Informing Design. *International Journal of Social and Organizational Dynamics in IT*, *1*(1), 30–48. https://doi.org/10.4018/ijsodit.2011010103

Han Hu, Yonggang Wen, Tat-Seng Chua, & Xuelong Li. (2014). Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. *IEEE Access*, *2*, 652–687. https://doi.org/10.1109/ACCESS.2014.2332453

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

Hegde, V. (2016). Dimensionality reduction technique for developing undergraduate student dropout model using principal component analysis through R package. In *Computational*

Intelligence and Computing Research (ICCIC), 2016 IEEE International Conference on (pp. 1–6). IEEE.

Hookham, G., Nesbitt, K., and Kay-Lambkin, F. (2016) Comparing usability and engagement between a serious game and a traditional online program. In Proceedings of the Australasian Computer Science Week Multiconference (ACSW '16). Association for Computing Machinery, New York, NY, USA, Article 54, 1–10. https://doi.org/10.1145/2843043.2843365

Hunicke, R., LeBlanc, M., & Zubek, R. (2004). MDA: A formal approach to game design and game research. In *Proceedings of the AAAI Workshop on Challenges in Game AI* (Vol. 4, p. 1722).

Huo, Z., Griffin, J., Babiuch, R., Gray, A., Willis, B., Marjorie, S., & Sun, S. (2015). Examining the feasibility of a Microsoft Kinect™ based game intervention for individuals with anterior cruciate ligament injury risk. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE* (pp. 7059–7062). IEEE.

Hussein, M., Ow, S., Cheong, L., Thong, M., & Ebrahim, N. (2019). Effects of Digital Game-Based Learning on Elementary Science Learning: A Systematic Review. In *IEEE Access*, vol. 7, pp. 62465-62478, 2019, doi: 10.1109/ACCESS.2019.2916324.

Jianu, R., & Alam, S. S. (2017). A Data Model and Task Space for Data of Interest (DOI) Eye-Tracking Analyses. *IEEE Transactions on Visualization and Computer Graphics*, 1–1. https://doi.org/10.1109/TVCG.2017.2665498

Kaku, M. (2014). *The Future of the Mind: The Scientific Quest to Understand, Enhance, and Empower the Mind*. Knopf Doubleday Publishing Group. Retrieved from https://books.google.com/books?id=4SiwAAAAQBAJ

Kapp, K. M. (2012). *The Gamification of Learning and Instruction: Game-based Methods and Strategies for Training and Education*. Wiley. Retrieved from https://books.google.com/books?id=M2Rb9ZtFxccC

Kebritchi, M. (2010). Factors affecting teachers' adoption of educational computer games: A case study. *British Journal of Educational Technology*, *41*(2), 256–270. https://doi.org/10.1111/j.1467-8535.2008.00921.x

Ketelhut, D., Nelson, B., Schifter, C., & Kim, Y. (2013). Improving Science Assessments by Situating Them in a Virtual Environment. *Education Sciences*, *3*(2), 172–192. https://doi.org/10.3390/educsci3020172

Khalil, H., & Ebner, M. (2014). MOOCs completion rates and possible methods to improve retention-a literature review. In *World Conference on Educational Multimedia, Hypermedia and Telecommunications* (Vol. 2014, pp. 1305–1313). Retrieved from https://www.researchgate.net/profile/Martin_Ebner2/publication/306127713_MOOCs_completion_rates_and_possible_methods_to_improve_retention-A_literature_review/links/57bb349c08aefea8f0f44ce9.pdf

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

Laffey, J. (2016). Mission HydroSci: Distance Learning through Game-Based 3D Virtual Learning Environments. In *Handbook of Research on Gaming Trends in P-12 Education* (pp. 421–441). IGI Global.

Lamb, R. L., Annetta, L., Vallett, D. B., & Sadler, T. D. (2014). Cognitive diagnostic like approaches using neural-network analysis of serious educational videogames. *Computers & Education*, *70*, 92–104. https://doi.org/10.1016/j.compedu.2013.08.008

Lias, T. E., & Elias, T. (2011). *Learning Analytics: The Definitions, the Processes, and the Potential*.

Manero, B., Torrente, J., Freire, M., & Fernández-Manjón, B. (2016). An instrument to build a gamer clustering framework according to gaming preferences and habits. *Computers in Human Behavior*, *62*, 353–363. https://doi.org/10.1016/j.chb.2016.03.085

Martínez, H. P., Jhala, A., & Yannakakis, G. N. (2009). Analyzing the impact of camera viewpoint on player psychophysiology. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on* (pp. 1–6). IEEE. Retrieved from http://ieeexplore.ieee.org/abstract/document/5349592/

Martinez-Garza, M. M., & Clark, D. B. (2017). Investigating Epistemic Stances in Game Play with Data Mining: *International Journal of Gaming and Computer-Mediated Simulations*, *9*(3), 1–40. https://doi.org/10.4018/IJGCMS.2017070101

Mayo, M. J. (2009). Video Games: A Route to Large-Scale STEM Education? *Science*, *323*(5910), 75–79. https://doi.org/10.1126/science.1166900

Metcalf, S. J., Clarke, J., & Dede, C. (2009). Virtual worlds for education: River City and EcoMUVE. In *MiT6 International Conference* (pp. 1–6).

Morgan, C. T. (1951). The psychophysiology of learning. In *Handbook of experimental psychology* (pp. 758–788). Oxford, England: Wiley.

Nacke, L., Drachen, A., Kuikkaniemi, K., Niesenhaus, J., Korhonen, H. J., Hoogen, W. M., … De Kort, Y. A. (2009). Playability and player experience research. In *Proceedings of DiGRA 2009: Breaking New Ground: Innovation in Games, Play, Practice and Theory*. DiGRA. Retrieved from http://www.diva-portal.org/smash/record.jsf?pid=diva2:835637

Nacke, L., & Lindley, C. A. (2008). Flow and immersion in first-person shooters: measuring the player's gameplay experience. In *Proceedings of the 2008 Conference on Future Play: Research, Play, Share* (pp. 81–88). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1496998

Nakamura, J., & Csikszentmihalyi, M. (2014). The Concept of Flow. In M. Csikszentmihalyi (Ed.), *Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihalyi* (pp. 239–263). Dordrecht: Springer Netherlands. Retrieved from http://dx.doi.org/10.1007/978-94-017-9088-8_16

Nakasone, A., Prendinger, H., & Ishizuka, M. (2005). Emotion recognition from electromyography and skin conductance. In *Proc. of the 5th International Workshop on Biosignal Interpretation* (pp. 219–222). Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.104.4937&rep=rep1&type=pdf

Nijholt, A., & Tan, D. (2007). Playing with your brain: brain-computer interfaces and games. In *Proceedings of the international conference on Advances in computer entertainment technology* (pp. 305–306). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1255140

Osborne, J., Henderson, B., MacPherson, A., & Szu, E. (2013). Building a Learning Progression for Argumentation. American Educational Research Association Conference.

Ravaja, N. (2004). Contributions of Psychophysiology to Media Research: Review and Recommendations. *Media Psychology*, *6*(2), 193–235. https://doi.org/10.1207/s1532785xmep0602_4

Reeves. T. C., Romine, W., Laffey, J., Sadler, T., & Goggins, S. (2020). Distance learning through game-based 3D virtual learning environments: Mission HydroSci (External Evaluation Report 1). University of Georgia. (ERIC Document Reproduction Service No. 600497).

Roepke, A. M., Jaffee, S. R., Riffle, O. M., McGonigal, J., Broome, R., & Maxwell, B. (2015). Randomized Controlled Trial of SuperBetter, a Smartphone-Based/Internet-Based Self-Help Tool to Reduce Depressive Symptoms. *Games for Health Journal*, *4*(3), 235–246. https://doi.org/10.1089/g4h.2014.0046

Roh, C., Lee, W. (2014) A Study of the Attention Measurement Variables of a Serious Game as a Treatment for ADHD. *Wireless Pers Commun* **79**, 2485–2498. https://doi.org/10.1007/s11277-014-1744-9

Ryu, M., & Ketelhut, D. J. (2014). ENGAGE. ELABORATE. EVALUATE!: Virtual Environment-Based Assessments of Science Content and Practices. *Science Scope*, *37*(5), 44–49.

Schmidt, M., Laffey, J., Stichter, J., Goggins, S., & Schmidt, C. (2008). The design of iSocial: A three-dimensional, multi-user, virtual learning environment for individuals with autism spectrum disorders to learn social skills. *International Journal of Technology, Knowledge and Society*, *4*(2), 29–38.

Shernoff, D., Hamari, J. & Rowe, E. (2014). Measuring Flow in Educational Games and Gamified Learning Environments. In J. Viteli & M. Leikomaa (Eds.), *Proceedings of EdMedia 2014--World Conference on Educational Media and Technology* (pp. 2276-2281). Tampere, Finland: Association for the Advancement of Computing in Education (AACE). Retrieved January 2, 2023 from https://www.learntechlib.org/primary/p/148041/.

Squire, K. (2003). Video games in education. *Int. J. Intell. Games & Simulation*, *2*(1), 49–62.

Squire, K. (2005). Changing the game: What happens when video games enter the classroom. *Innovate: Journal of Online Education*, *1*(6). Retrieved from http://website.education.wisc.edu/~kdsquire/tenure-files/26-innovate.pdf

Stichter, J. P., Laffey, J., Galyen, K., & Herzog, M. (2014). iSocial: Delivering the Social Competence Intervention for Adolescents (SCI-A) in a 3D Virtual Learning Environment for Youth with High Functioning Autism. *Journal of Autism and Developmental Disorders*, *44*(2), 417–430. https://doi.org/10.1007/s10803-013-1881-0

Van der Graaf, J., Segers, E., & Verhoeven, L. (2016). Discovering the laws of physics with a serious game in kindergarten. *Computers & Education*, *101*, 168–178. https://doi.org/10.1016/j.compedu.2016.06.006

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

Xing, W., Guo, R., Petakovic, E., & Goggins, S. (2015). Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory. *Learning Analytics, Educational Data Mining and Data-Driven Educational Decision Making*, *47*, 168–181. https://doi.org/10.1016/j.chb.2014.09.034

Vita

Joe Griffin was born in St. Louis, Missouri, in 1988. After finishing high school in 2007, he moved to Columbia Missouri. He studied Computer Science at the University of Missouri. He received his Bachelor of Science in 2011. After working for his Advisor James Laffey, on a project to teach social competence to children with autism in a virtual online world, he became interested in game base learning. He then continued his studies at the University of Missouri, where he received his Master of Education in Learning Technologies and Design, which he completed in 2013. He completed his Ph.D. in Information Science and Learning Technology from the University of Missouri in 2023. Since beginning to work for the University of Missouri in 2008, he has collaborated on over $18 million in grant funded game based learning projects.