11th International Conference of the Immersive Learning Research Network
Selected Academic papers for the iLRN Proceedings

DOI: https://doi.org/10.56198/54rfwf69

Demonstrating Replay for Highly Scalable and Cost-Effective User Research of Virtual Reality Learning Games

David J. Gagnon[1], Kevin Ponto[2], Luke Swanson[3] and Ross Tredinnick[4]

[1] Field Day Lab, University of Wisconsin - Madison, Madison, USA
[2] Wisconsin Institute for Discovery, University of Wisconsin - Madison, Madison, USA
djgagnon@wisc.edu
kbponto@wisc.edu
lwswanson2@wisc.edu
rdtredinnick@wisc.edu

Abstract. Observing users interacting with a learning game, often referred to as playtesting, is a critical component of usability testing. Unfortunately, this practice is expensive, requiring users and researchers to be in the same place at the same time as the participants. With Virtual Reality, this difficulty is amplified due to the experience being hidden from researcher view by default, and the extra complexity of setting up casting to an external monitor, especially with groups.
In this paper we develop and utilize a replay-based approach to usability testing that relieves these concerns. The approach uses a low-bandwidth stream of telemetry signals that are generated by the original play session. These signals are then reconstructed into a full representation of the original experience at a different time or place and leveraged to identify usability issues. Once the issues have been discovered, automated processes are developed to computationally identify their presence and severity in arbitrarily large public audiences. This work contributes a demonstrated use of replay for Virtual Reality usability research, and a novel use of replay combined with educational data mining to develop an automated process for studying large audiences at low cost.

Keywords: Replay, VR, Virtual Reality, Educational Data Mining, Usability, Game Data.

1 Introduction

As an immersive media, head mounted Virtual Reality (VR) provides unique challenges for usability researchers. The user's experience is rendered to displays only they can see and to speakers right above their ears. While they may be immersed into an interactive world, a researcher only sees someone wearing a headset and moving around oblivious to their physical surroundings. Solutions such as "casting" are invaluable but not always reliable or even feasible given that they have particular requirements on the WiFi network, such as requiring multicast capacities that are often disabled on enterprise infrastructures. Even when casting functions, it still requires the researchers to be present at the same time and place as the participant, leading to significant expense. Instead, we develop a system for VR experience replay that may be used across a variety of projects to drastically decrease the complexity and expense of usability testing.

As a demonstration of value, the replay approach was used to study the usability of *Waddle: A Penguins Tale*. While the project has employed usability focused playtests with small representative audiences several times during development, a new version was recently released containing significant changes to the design that had not been tested rigorously. To perform testing on this new version of the game, and in pursuit of a repeatable model for usability test for future projects, we adopt the following high-level guiding questions:

1. Does VR replay have the capacity to reveal previously unidentified usability concerns using data from anonymous audiences?
2. Can we use automated methods to quantify the severity of these concerns and estimate how often they occur?

The first question explores the value of replay as a new tool for our design and testing efforts. While more in-person testing could certainly be scheduled and performed, for projects that already have some amount of audience, are we able to collect usability data at a lower cost that provides insights we previously did not have?

If so, this would position replay for use in coordination with or after small scale playtesting and could drastically lower the cost and complexity of collecting detailed usage data. The second question explores the intersection of replay with educational data mining to computationally identify and count the phenomena, reducing a complex qualitative analysis into a scalable, quantitative endeavor. If successful, this capacity allows us to benefit from the rich understanding of qualitative methods for arbitrarily large audience sizes at no additional cost.

2 Prior Work

Game designers have a long history of using replays to identify usability issues and interface inefficiencies. Similar to play testing, replays can reveal where players experience confusion, fail to understand game mechanics, or encounter design flaws that inhibit gameplay [1]. Replays are also used within the broader context of educational media design as a way to assess players' thinking and affect using qualitative methods [2].

Various projects in the last few years have developed technical demonstrations of capturing VR experiences and replaying them either in VR or on desktop computers. By rendering the replay to a video file first, researchers developed automated assessments of VR surgical simulations [3]. Other projects leverage telemetry data, digitally capturing the positions and rotations of the user's hands and head-mounted display, such as the *MAGES* project [4] and *STAG* [5]. A few projects also utilize VR for immersive visualization and review of the replay data. This allows the researcher to be placed in the same virtual environment as the original player and visualize what objects were within their field of view [6]. Similarly, the movement of a player over time can be abstracted into a line within the 3d space, with nodes appearing along that line for other actions of interest such as stopping, or performing an action of interest [7]. Finally, a system called *ReRun* [8] builds on these approaches and extends them to multi-participant systems. This system also allows for the replay user to move the virtual camera to any position to best observe the players' actions.

Unfortunately, direct replays of VR experiences can be difficult for others to watch due to the rapid and jerky camera motion that is natural for the original player, but unexpected for the replay user. To combat this problem, researchers have developed a summary approach that communicates the important features of the experience that smooth the camera movements to make them more natural to an external viewer [9].

3 Method

3.1 Waddle: A Penguins Tale

Waddle: A Penguins Tale is an educational VR experience designed for use in informal contexts for ages 6+. The game was developed in consultation with penguin researchers working at the McMurdo research station in Antarctica, at Cape Royds, and refined through iterative testing at science festivals and library events over a period of 2 years. Educationally, the game is designed to teach players about the lifecycle of Andéle penguins by having them enact key events such as nest building, a mating dance, and protecting their offspring from predators (see Figure 1). From a research standpoint, the game explores the design of direct embodiment of an animal in VR and its effect on empathy [10] and learning [11]. It was developed for the low-cost Meta Quest 2 and Meta Quest 3 head mounted displays using the Unity game engine.

Fig. 1.: A screenshot from the mating dance in *Waddle.*

As a Penguin simulator, *Waddle* has a few novel user interaction patterns that proved challenging during development. Specifically, locomotion (i.e. moving in the 3d space) in *Waddle,* is not done using the VR conventions of selecting a point with the analog stick and "warping" to that location. Instead, locomotion consists of moving one's head and shoulders side-to-side in a rocking motion attempting to replicate the way a penguin "waddles." When this motion is completed, the player is warped forward in the direction they were facing by one step. Similarly, interacting with objects in the virtual environment does not utilize the conventions of using the user's hands to grip or point. Instead, the hands are replaced by the ends of the virtual penguin's flippers. As with real penguins, the flippers are used primarily as weapons (*Waddle* does not contain a swimming mechanic). To pick up objects, the beak is used. An intersection between the beak and an object captures the object, and then a second intersection drops the object. Waddle communicates these interface designs to the user via modeling the behaviors by other virtual penguins in the space as well as by using a collection of animated signs as seen in Fig. 2.

Fig. 2.: An animated instruction sign within *Waddle.*

3.2 Technical Approach

This project leverages Open Game Data (OGD) [12], a set of technologies that define an open source, modular pipeline for sending, capturing, storing, processing and distributing game interaction data. For this project, an existing instance of the Open Game Data core infrastructure was used as well as the client library for logging events from the Unity game engine (opengamedata-unity). Open Game Data's game telemetry is event based, meaning that data is sent from the game to the logging database in response to player or game actions such as using a tool, progressing to a new level, or receiving feedback from the game. These events are sent at the moment they occur, in contrast to sending a large number of events in a package at the end of a challenge or some other significant moment. Each event has data members not only to describe the event that occurred, but also to describe key parts of the game's state, contextualizing the action. Open Game Data also provides standard data members to describe the time an event occurred, versioning information, and metadata about the game and player. Of particular importance to replay, time must be recorded with a high degree of precision and accuracy.

To record replay-specific data, we defined three custom events to record information about the user's viewport (the position and rotation of their head), their left hand and their right hand. To simplify network communications for these high frequency events, we break OGD conventions and package an entire second worth of data into a single event, encoding multiple positions and rotations as an array of values that took place over the prior second. These events were recorded at 20 frames per second. Unfortunately, the opengamedata-client package did not initially support sending such a large amount of data in a single event. This shortcoming was remedied, and support for large event payloads was added and is now part of the publicly available version of the package.

We also added a parameter to a "session_start" event common in OGD games to log the random seed used by the game. This ensures that upon replay, we will be able to reconstruct any random-but-deterministic behavior as it was seen by the original player.

Additional telemetry events are recorded from *Waddle* that describe player actions, game feedback and progression through the game's challenges. Including the replay events, the game sends a total of 36 different telemetry events. 15 describe specific player actions (e.g. picking up a rock, waddle). 10 events describe feedback or actions taken by the game (e.g. visual feedback of filling in the nest). Finally, 8 events describe progression (e.g. completing the nest). All these events are sent, as they occur, to the logging infrastructure for storage and processing. Automated exports are made available for further analysis at the Open Game Data Website.

To facilitate replay, a new package was developed called opengamedata-replay, which was added to the *Waddle* Unity project. The objects for gaze (i.e. the camera rig object), right hand and left hand are selected in a simple inspector panel, as well as the telemetry data file to parse. The raw head- and hand-tracking data is processed by the replay package into an intermediate binary format and is fed into the VR input system, mimicking the signals from a live player. Frame by frame, the replayed input interacts with the game system, creating the same experience seen by the original user. The 2D display output is recorded into an mp4 formatted media file, rendering the experience from the player's perspective. Utility scripts are provided to run batches of this process for many sessions at once.

3.3 Identifying Usability Issues

In the first iteration of the project, a small sample of anonymous data from the March 2024 dataset was used to generate video replays of a small number of anonymous gameplay sessions. While very little is known about the context of play, we assume the majority was from Meta Quest users who found Waddle within the platform's app store and were not part of any formal program.

The resulting videos were generated and reviewed by the research team. Three usability concerns were identified. These issues all occurred within the first few minutes of play, during the nest building activity. The first concern was that players were attempting to use their hands (flippers in the game) to pick up rocks to build a nest, instead of their beak, an interaction not supported by the game nor actual penguin behavior. The second issue was that players would pick up a rock correctly but attempt to deliver it to a location other than the nest they were building. Finally, a third issue we discovered with replay was that players were successfully picking up a rock in their beak, then attempting to pick up more rocks before delivering them to their nest. This process demonstrated an affirmative answer to RQ1; that is, replay *does* have the capacity to reveal previously unidentified usability concerns using data from anonymous audiences. The process allowed us to identify three concerns that were previously unknown.

3.4 Developing Automated Detectors

Following these discoveries, we turned attention to the next question, can we use automated methods to quantify the severity of these concerns and estimate how often they occur? For this question we conducted a round of feature engineering and telemetry design that would provide the necessary telemetry signals for automated detection of these three patterns. While it was possible to use only the events that were currently provided by the game to define each of the usability issues, the team decided it would be much faster and of longer-term value to refine the events that were sent from the game. Unlike related work to qualitatively code and train machine learning models to detect complex interaction patterns [13], these interaction patterns were compatible with simple algorithmic analysis. Specifically, we added new events to record a "rock pickup" as well as a "Beak intersection with a rock." We also added a game state variable to track if a player currently had a rock. Following the development of these new events in the game and deployment, we created feature extractors in the opengamedata-core processing package to calculate the number of times a player picked up a rock and walked to an incorrect nest, the number of times their flippers intersected with a rock, and the number of times their beak intersected with a rock when they already had a rock.

3.5 Audience and Process for Analysis

Following the initial sample that led to the discovery of the three usability issues of interest, and the iteration of the game telemetry logging introduced in response, a second dataset was collected from anonymous users of the newest version of *Waddle,* version 10. Between April 1, 2024 and May 15, 2024, 527 sessions were recorded. After removing sessions created from older versions of the game, sessions where the players did not move from the starting position, and sessions less than 60 seconds in duration, 202 sessions remained. 95 (47%) of these sessions completed the nest building activity. As with the original sample, little is known about the external context of these sessions.

4 Results

Instances of each of the usability concerns were detected in the second dataset. While many sessions did not exhibit an instance of a particular usability issue, some exhibited a few, and others exhibited many issues of the same type. Visualizing sessions where the usability concern was present, we plot the number of occurrences of the concern per session as a histogram (see Fig. 3).

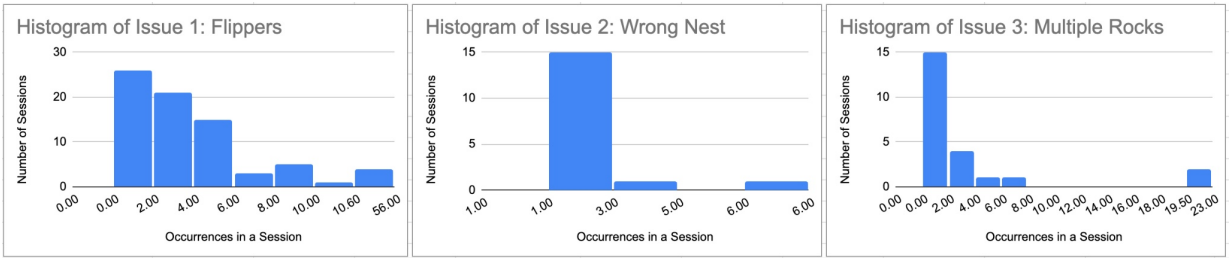

Fig. 3.: Histograms of the number of occurrences of each usability issue.

To understand the severity of each usability issue, we look for instances where the concern occurred, but was never overcome. That is, we would expect players to learn the game system through some amount of trial and error. If a player repeatedly attempts to complete a simple task incorrectly, however, and is never able to determine how to proceed, this becomes a usability concern. For Issue 1: *Using Flippers* (to Pick up a rock), we consider the issue severe if the player is never able to pick up a rock. For Issue 2: *Wrong Nest*, the issue is severe if they never deliver a rock to the correct nest. Finally, for Issue 3: *Double Rock*, the issue is severe if the player never drops a rock. The occurrences of each usability issue are quantified in Table 1.

Table 1. Usability Issue Occurrences.

	Number of Sessions	Severe Instances
Total Sessions	527	
Sessions after Filtering	202	
Issue 1: Using Flippers	75 (37%)	17 (8.4%)
Issue 2: Wrong Nest	17 (8.4%)	4 (1.9%)
Issue 3: Double Rock	23 (11%)	5 (2.5%)

5 Discussion

These results describe a comparative importance of each usability issue discovered by the replay method. While the majority of sessions did not experience any of these particular issues, we see that simply counting the number of sessions that experienced the issue is not a helpful signal on its own. Especially given the novel interactions in Waddle, it is perfectly acceptable for a player to attempt an action and not immediately succeed. However, for a smaller number of sessions, the issue became a significant factor in the experience. In these cases, the user repeatedly performed actions that demonstrate a misconception and were not able to overcome it. For these severe cases an in-game intervention should be developed, whether as additional instruction or as a real time response to player actions that trigger a usability issue detector. This analysis demonstrates that for a significant number of players, the animated signs and modeling of behaviors by other penguins was not successful at conveying the novel interaction patterns.

In regard to the research questions, we were able to demonstrate that replay not only facilitated the identification of previously unknown usability issues, but that automated methods could be used to count their relative severity. In this project, only the most rudimentary approaches were used, namely the addition of a new event to the learning product and analysis code to count and compare the presence of these events. Given prior work by the authors to compare outcomes from different versions of a game [15], to classify player groups and strategies [16], develop models to predict future behavior [17], and detect more complex behavior such as “struggle” [18], these analyses are only the beginning of what could be done with replay if more sophisticated data mining approaches were used.

6 Conclusion and Future Directions

In this project we developed technology that affords the replay of VR experiences at a later time and place from the original play experience from game telemetry data. This approach allowed us to identify usability concerns in the game design that were not previously observed through other playtesting efforts. Combined with basic data mining, these usability concerns were automatically identified across large numbers of play sessions, then counted and assessed for their severity. Moving forward, review of VR replay should be conducted with a larger number of sessions for *Waddle* until all the usability concerns have been identified using this method.

There are also several improvements to the replay system itself that follow from this project. Most pressing, the system should provide mechanisms for replaying the various hand gestures available in modern VR systems. This includes the relative position of each digit in each hand, as well as the aggregate gesture, such as pointing or grabbing. Another future direction includes the ability to view the replays directly within the game engine with moveable cameras, and ability to move forward and backward in time without rendering to a video file. An obvious expansion of this work would be to apply it beyond VR experiences and with Mixed Reality systems. Finally, an annotation system, like those utilized by Zoombinis researchers [2], should be developed to allow for a more formalized method of qualitatively studying VR game replays and the training of automated detectors. These replays could leverage prior contributions such as viewport smoothing [9] and projection mapping of the users' focus [14].

Acknowledgments

We would like to thank Jim Madsen and Diego Roman for their assistance in using this work to reach a wider audience, as well as Monae Verbeke of Institute for Learning Innovation for work to evaluate the project. We would also thank Sarah Gagnon, Eric Lang, Mary Benetti, Autumn Beauchesne, Cyril Peck and Jim Mathews for their efforts to develop the VR experience itself. Finally, we thank Renee Li for her work developing gameplay features. The developed materials are based upon work supported by the National Science Foundation under Grant No. 2116046

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Harpstead, E., MacLellan, C. J., Aleven, V., Myers, B. A.: Replay Analysis in Open-Ended Educational Games. In: Loh, C. S., Sheng, Y., Ifenthaler, D. (eds.) Serious Games Analytics: Methodologies for Performance Measurement, Assessment, and Improvement, pp. 381–399. Springer International Publishing, Cham (2015). doi: 10.1007/978-3-319-05834-4_17
2. Rowe, E., et al.: Assessing implicit computational thinking in Zoombinis puzzle gameplay. Comput. Hum. Behav. 120, 106707 (2021). doi: 10.1016/j.chb.2021.106707
3. Zia, A., et al.: Automated video-based assessment of surgical skills for training and evaluation in medical schools. Int. J. Comput. Assist. Radiol. Surg. 11(9), 1623–1636 (2016). doi: 10.1007/s11548-016-1468-2
4. Kamarianakis, M., Chrysovergis, I., Kentros, M., Papagiannakis, G.: Recording and replaying psychomotor user actions in VR. In: ACM SIGGRAPH 2022 Posters, pp. 1–2. ACM, Vancouver BC Canada (2022). doi: 10.1145/3532719.3543253
5. Basu, A.: STAG: A Tool for realtime Replay and Analysis of Spatial Trajectory and Gaze Information captured in Immersive Environments. In: 2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), pp. 43–45. IEEE, Christchurch, New Zealand (2022). doi: 10.1109/VRW55335.2022.00016
6. Ehtemami, A., Park, S. B., Bernadin, S., Lescop, L., Chin, A.: Overview of Visualizing Historical Architectural Knowledge through Virtual Reality. In: SoutheastCon 2021, pp. 1–6. IEEE, Atlanta, GA, USA (2021). doi: 10.1109/SoutheastCon45413.2021.9401850
7. Kloiber, S., et al.: Immersive analysis of user motion in VR applications. Vis. Comput. 36(10–12), 1937–1949 (2020). doi: 10.1007/s00371-020-01942-1
8. Goedicke, D., Haraldsson, H., Klein, N., Zhou, L., Parush, A., Ju, W.: ReRun: Enabling Multi-Perspective Analysis of Driving Interaction in VR. In: Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction, pp. 889–890. ACM, Stockholm, Sweden (2023). doi: 10.1145/3568294.3580211
9. Ponto, K., Kohlmann, J., Gleicher, M.: Effective Replays and Summarization of Virtual Experiences. IEEE Trans. Vis. Comput. Graph. 18(4), 607–616 (2012). doi: 10.1109/TVCG.2012.41
10. Gagnon, D. J., Ponto, K., Verbeke, M., Nathan, M., Kopp, K., Tredinnick, R.: Waddle: Developing Empathy for Adélie Penguins By Direct Embodiment in Virtual Reality. In: Haahr, M., Rojas-Salazar, A., Göbel, S. (eds.) Serious Games, LNCS, vol. 14309, pp. 227–233. Springer Nature Switzerland, Cham (2023). doi: 10.1007/978-3-031-44751-8_17

11. Ponto, K., Tredinnick, R., Verbeke, M., Kopp, K., Swanson, L., Gagnon, D.: Waddle: using virtual penguin embodiment as a vehicle for empathy and informal learning. In: 29th ACM Symposium on Virtual Reality Software and Technology, pp. 1–2. ACM, Christchurch, New Zealand (2023). doi: 10.1145/3611659.3617211
12. Gagnon, D. J., Swanson, L.: Open Game Data: A Technical Infrastructure for Open Science with Educational Games. In: Haahr, M., Rojas-Salazar, A., Göbel, S. (eds.) Serious Games, LNCS, vol. 14309, pp. 3–19. Springer Nature Switzerland, Cham (2023). doi: 10.1007/978-3-031-44751-8_1
13. Liu, X., et al.: Struggling to Detect Struggle in Students Playing a Science Exploration Game. In: Companion Proceedings of the Annual Symposium on Computer-Human Interaction in Play, pp. 83–88. ACM, Stratford ON, Canada (2023). doi: 10.1145/3573382.3616080
14. Lopez, T., Dumas, O., Danieau, F., Leroy, B., Mollet, N., Vial, J.-F.: A playback tool for reviewing VR experiences. In: Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology, pp. 1–2. ACM, Gothenburg, Sweden (2017). doi: 10.1145/3139131.3141776
15. Gagnon, D. J., et al.: Exploring players' experience of humor and snark in a grade 3-6 history practices game. In: GLS 13.0 Conference Proceedings, ETC Press, Irving, CA (2022).
16. Swanson, L., et al.: Leveraging Cluster Analysis to Understand Educational Game Player Styles and Support Design. In: GLS 13.0 Conference Proceedings, ETC Press, Irving, CA (2022).
17. Liu, X., Slater, S., Swanson, L., Metcalf, S. J., Gagnon, D. J.: Identifying When and Why Students Choose to Quit Jobs in a Science Exploration Game. Presented at: 10th International Joint Conference on Serious Games, New York University, New York City, US, 2024. [Online].
18. Liu, X., et al.: Struggling to Detect Struggle in Students Playing a Science Exploration Game. In: Companion Proceedings of the Annual Symposium on Computer-Human Interaction in Play, pp. 83–88. ACM, Stratford ON, Canada (2023). doi: 10.1145/3573382.3616080