A Framework for XReality Serious Games

Nicole Kosoris
Advanced Human Integration Branch
Georgia Tech Research Institute
Atlanta, GA
nicole.kosoris@gtri.gatech.edu

Brian Liu
Electronic Systems Laboratory
Georgia Tech Research Institute
Atlanta, GA
brian.liu@gtri.gatech.edu

Brad Fain
Center for Advanced Communications Policy
Georgia Institute of Technology
Atlanta, GA
brad.fain@cacp.gatech.edu

***Abstract*—Virtual and Augmented Reality capabilities have advanced remarkably in the last decade. Serious Games in VR and AR have become not only possible, but their development has become rapid and inexpensive. Commercial devices and development environments are now sufficient to enable data-driven application development for research focused applications. Despite these outstanding innovations, Virtual and Augmented Reality present unique design challenges in the research space. Development of five different data-driven research applications in VR and AR over the last several years has illustrated that data collection in such applications has considerations unique to XReality. Working from a unified framework – that is, beginning with an analysis plan, developing with an analyst as part of the team, and developing with an understanding that virtual data collection has additional requirements that don't apply when one can rely on the constancy of real-world physics – is critical to the advancement of these technologies as valid and valuable research tools. This paper presents such a framework for developing data-driven Serious Games for research purposes.**

***Index Terms*—Human Performance, Study Design, Data Collection and Analysis, System Design and Analysis, Virtual Reality, Augmented Reality**

I. Background

In the late 2010's, Virtual Reality crossed a critical threshold wherein "... the field of 3D UI is becoming more mature." [1] Despite decades of overblown claims for VR systems, the late 10's systems delivered the ability to move around in physical reality while interacting with minimally believable virtual environments, and models therein began to be recognizable analogues for their physical counterparts. The number of VR systems sold skyrocketed during early 2020 such that the Oculus Quest – unique for its consumer-gaming price point and lack of need for a separate premier gaming computer – and the Valve Index were backordered for more than three months.

Emotional responses can be evoked in VR in a uniquely isolated environment – we can know exactly what the user is exposed to. Similarly, they can be tracked, somewhat, with direct correlation to stimuli using Galvanic Skin Response, Gaze Tracking, EEG, or other physiological signals [2]. Although we cannot always tell which emotion is evoked, we can tell if we're mimicking a real-world situation well enough to evoke changes in stress and model human response [3].

Research has been done in many systems that are more accurately called "simulations", and technology for simulation has been used in aviation, in particular, for decades. Although we will draw heavily from aviation simulation literature here, the primary difference between a simulation and a serious game is that a serious game is not a direct analogue for a physical event [4]. It can certainly be argued that some of the software we will discuss merely contains "game elements" or somewhat trivial "Extrinsic Gamification" [5], however, rather than being serious games. This being "The use of game thinking and elements in non-game contexts." [6].

Data-driven architecture – that is, designing with data collection in mind – in serious games has been discussed by [7] and [8]. Working according to the guidelines for data driven games proposed by [9] gives you a good starting point for ensuring that you give data due consideration, but, particularly when extended into 3 dimensions, it can also provide an overconfidence based on intuition that causes critical issues.

The problem is that, although Data is important, once it is removed from context it can lack characteristics necessary to its interpretation. Data by itself isn't really information, [10]. Iteratively pilot testing incremental builds is the traditional solution for finding these kinds of errors and resolving them, but although iterative development and frequent small pilot projects of a system are important, "a comprehensive formative evaluation must include analysis. Much like the differences between working with two-dimensional and three-dimensional mathematics, the change in complexity when moving to Virtual Reality complicates analysis, where human intuition suggests that it would, by contrast, more accurately reflect the physical world and require less accommodation for its change to virtual context.

II. Related Work

Over the last several years, we have created seven total VR Serious Games applications. The first three we will discuss here are variants on a system called iDETECT, based on a system for testing cognitive impairment called DETECT [11]. The systems of iDETECT began as a multimodal test whose most novel sub-test treated physical movement as an input. In its next transformation, iDETECT was re-envisioned in Virtual Reality to be supported with additional game elements as seen in Figure 1. In a final variant, iDETECT was transformed into a juvenile-targeted Serious Game called BrainBuddy.

The iDETECT system attempted to use physical movement as a system input. In this platform, head motion was sensed

978-1-6654-0652-9/22/$31.00 ©2022 IEEE

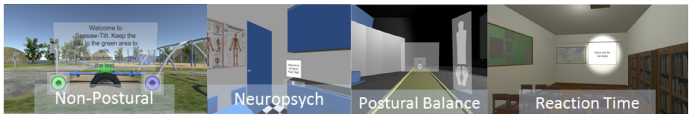

Fig. 1. Virtual Reality iDETECT System, 3rd Generation

Fig. 2. ARTEMIS HUD Overlay

and translated into control of an on-screen element. One of the primary motivations for moving to Virtual Reality was to increase user immersion and to move towards a gamified system, in order to gain a more sensitive test of neuropathways for detecting concussions. The effectiveness of in-game assessment has been studied and meta-studied [12] [13], and engagement or motivation were found to be key in every case where it was given consideration. Therefore, although "[t]he most general definition [of serious games] is any game that has a purpose beyond simply entertaining the player [14]", the more the researcher can engage and immerse the user, the more they can improve the sensitivity of their assessments.

Although we successfully captured physical movement's roll, pitch, and yaw using a custom sensor suite, analysis ended up running far beyond the expected end of the project. Problems included that our calculated physics did not precisely match real-world physics, and that determining "good" head movement was nontrivial. Where a physical recording could be assessed for "wobbliness" of head-movement by a reasonably trained technician, months of machine learning work was unable to extract good movement from bad without largely overlapping the variance between individuals.

iDETECT was later expanded into a commercial VR platform, the Samsung Gear VR. We additionally used the Unity game engine. Unfortunately, we made assumptions about the "real world" nature of the physics available to us on a commercial platform. Rather than simplifying analysis, we now needed an extra step to determine the accuracy of the physics employed by the platform itself. The physics were significantly closer to reality, but this presented an additional lesson in analytical assumptions.

The other systems that developed our methods discussed here feed into a XR Optimization and Interaction Lab (XOIL). Successful educational games tailor themselves to the user, in particular providing feedback and challenges appropriate to the user's demonstrated in-game responses [15]. Our first project series focused specifically on HUD optimization, one of which flew a quad-copter with brain control, one which worked on HUD optimization for First Responder teaming, and a third which extended the HUD optimization work into human-machine teaming. These projects taught us a lot about the need for participants to engage with the target systems at a basic level before attempting the study work. That is, humans needed to work with the BCI at a desktop to move a cube, and to work with a VR heads-up display to understand the headset and controllers, before being introduced to a mission objective such as a quad-copter or a remote controlled vehicle.

Augmented Reality Testing of Equipment in Multiple Immersive Simulations (ARTEMIS) is a testing platform to enable objective comparisons on the efficacy of new methods and devices in scenarios important to first responders. Designed from the ground up in collaboration with the police and rescue organizations, ARTEMIS is data-driven and focused on meeting the needs of first responders. ARTEMIS is a comprehensive system capable of creating, modifying, and testing new scenarios, technologies, and capabilities. Not only a simulation system, ARTEMIS is a modular collection and analysis platform that can be expanded or incorporated into other systems – from low-tech surveys to mobile sensor data and beyond.

ARTEMIS includes a scenario management system, the ability to model and integrate novel future technologies, and an after action reporting tool that can integrate Virtual Reality recordings, survey information, and real-time tagged administrator feedback. ARTEMIS natively captures objective data and allows the tagging and integration of subjective data to measure efficiency, effectiveness, and user experience, driving viability experiments with repeatable outcomes. An example HUD is shown in Figure 2.

In order for ARTEMIS to provide accurate feedback on the viability of new technologies and innovations, scenarios must be repeatable with differing tools and technologies without overly biasing results by "learning" the specifics of the scenario presented. The optimization of which elements should be static and which could be changeable within a scenario required analysis throughout system development. Developed for NIST, whose primary interest is in creating standards, ARTEMIS integrated an analyst in the early phases of design, who has caught many of our assumptions about what the data would tell us. The overall purpose of the platform was to generate data for comparison, given any set of XR or real-world scenarios and associated modifications to be tested.

III. XR Data Analysis

A. Piloting

There is an assumption that piloting alone can find – and enable the fix of – issues with data collection. In an ordinary system, this is reasonable; in some cases this has worked in XR as well, such as with the "fall" condition. In other cases, the XR system makes a series of interdependent assumptions that a pilot test may be inadequate to catch – or, even if piloted and caught, the very interdependence may make it impossible to resolve those concerns without starting over. At a minimum, piloting must include analysis, even when the results of such piloting are not scientifically valid due to things like population size or recruitment type. That is, one might unofficially recruit co-located researchers to ensure the equipment works as planned. Such data is unpublishable, but can and should be analyzed as a part of pilot testing.

An example of a case where piloting resolved the concern was the "fall" condition of the balance test. Though it was initially anticipated that "fall" data might inform the impairment decision, it was eventually determined by the clinical team that it was far more likely to be caused by a user's unfamiliarity with the system, or with some kind of glitch in the IMU data stream. Thus, invisible "bumpers" were added to the ends of the beam, and the ball was not permitted to fall off.

An example of a case where piloting didn't resolve the problem was the very first balance test. We made assumptions about the test that weren't reflected as expected. There's an intuitive relationship between head movement and the data obtained from an accelerometer. It simply made sense that if we knew where a person's head was at all times, we would be able to tell if they were "wobbling". It's an easy thing to see in the physical world, and so of course it should be easy to detect with sensors. There were two problems with our assumptions. The first is that the data itself cannot be analyzed without transformation; it has to first be converted into a continuous line signal, which then needs to be analyzed via signal-processing methods using an assumption that the continuous line has been faithfully recreated from the samples. The second problem was that the test inadequately reflected reactions in the physical world. That is, the "physics" employed in the test did not match real-world physics, and so our analysis, failing to take this into account, was flawed. Although we could tell a "wobbly" person from a non-wobbly person, in a larger group this difference was utterly lost in the data that told us who could quickly encode a change in physics.

Completely different physics of movement are employed in many human sports. For example, in Ice Skating you are continually placing your body in a position that a brain trained on walking will tell you is "falling". Riding a bike requires similarly novel physical assumptions. Yet once learned, these assumptions typically carry forward well; most people find they can ride a bicycle even after years without an attempt. Even aside from the physical skills required, it takes people different amounts of time to encode the change in physics that these kinds of activities require, and long term exposure to these kind of situations physically changes the brain [16]. The requisite neuroplasticity changes with age, but is also known to differ widely among individuals [17].

Interacting with VR already requires some of this change. That portion can be resolved by acquiring a base skill in VR systems. Developers of VR serious games need to be extraordinarily careful, however, and aware of whether their particular application will require a change in physics. Although it is best avoided where possible, if it is necessary the developer must be aware. They might need to provide for training of varying lengths in the requisite change, or at a minimum provide an analysis plan that supports abstracting the encoding of the novel physics separately from the aspects to be tested.

B. *Base Skill Acquisition*

Further discussion is valuable on Base Skill in VR as a whole. One problem in VR Serious Games is that learning affects tend to be marked, simply because users are learning to interact with a virtual environment. The problem is that this learning effect will totally hide the effects of different tests; that is, the user's learning curve for working in VR is more substantial than the statistical difference between test conditions. This has been seen in Serious Games before, as "...serious games developed by researchers without formal training and insight in game design, risk [missing] scientific effects [18]." Such a problem is compounded in XReality.

This has also been seen before in other simulation systems, such as in flight simulators. Papers such as Koglbauer2016 quantify time spent on the simulator, but assume the reader understands that [19] simulator-specific training is a necessary first step. This implies that if a dependent skill has not been acquired, the validity of our behavioral predictions will be impossible to isolate from the noise of skill acquisition in a particular simulator. Similar to in those scenarios, the solution is to first acclimatize users to the VR system, lest valuable information be lost in the user learning to move and operate within the virtual environment.

Many consumer programs in VR are suitable for such a task, it is simply the need to have users undertake one associated with the needed movements and interactions in their system that must be considered. There is an assumption that, because users are moving in real space, they will find the interaction natural. We have found that users take time to accustom themselves to the interface. Not simply the controller, but also to learn to trust that the virtual "walls" will reliably prevent them from walking into physical walls.

Users also often have a perception of "looking silly" that requires engagement with an immersive game to overcome. Again, this must be well-matched to the activities that will be required of the game under study.

The "win" condition issue was not caught for this reason. The initial analysis showed a difference between our pilot controls and concussed patients. We then did a long-term study with high school student football players. We saw very few concussions. More problematic, however, was that players got so much better at each of the subsequent tests that comparing baseline scores to concussion scores directly was complex and problematic.

We moved to use a machine learning algorithm from that point, but found an additional complication: the methods capable of analyzing our data required the sets of data to be the same size, and we had allowed participants to "win" rounds by successfully keeping the ball within the center target for 3 seconds. This made data recorded from some rounds substantially shorter than data recorded from longer rounds –

and, once the participants moved to a higher level of capability with the application, the challenge of achieving this was no longer sufficient to allow comparison of the participant's data to show differences.

IV. Framework

There are three components of perspective that are helpful in developing XR Serious Games. Firstly, develop to the analysis: that is, start with not only what game mechanics intuitively seem to support your investigation, but with the particular analytical methods you plan to use to analyze it. Some analysis methods require data that may be non-intuitive. It is necessary to you're your game mechanics in the light of the kinds of data that need to be collected, to prevent invalidating your test with a poorly designed mechanic that didn't appear to be harmful.

Second, develop with an analyst. As data becomes increasingly complex, the methods for its analysis increase in concert. Some methods have very specific requirements, and Signal Processing and Machine Learning are vast and complex fields that many developers have inadequate knowledge of. Working with an analyst from the beginning will keep you informed, and will help develop to the analysis.

Finally, develop for the adaptation. This is a broad, and in some ways impossible directive. It requires that we discover the perspective of not only the intended users, but of the information consumers who will engage with the results. It also requires that we attempt the monumental task of isolating individuality and interpretation of specific VR mechanics from changes that more broadly impact user perception. Nevertheless, it is imperative to attempt it. Where standard HCI research best practices demand that we collect only the data we're certain to need to answer a particular question, and de-identify data such that an individual cannot be recognized, this could be counterproductive and may even be impossible in a VR setting.

A. Develop to the Analysis

The first "Virtual Reality Serious Game" we developed stretched the definitions of either set of terms. iDETECT was an expansion of a previous work, DETECT. DETECT was a psychological battery used to detect cognitive issues in an aging population, and was based on a tablet. The purpose of iDETECT was to discover whether that battery, as well as a series of related tests, could be used to discover symptoms of concussion. We created custom hardware to allow device input, and a heads-up display so that individuals taking the assessment wouldn't be distracted by their environment. The "games" were simple, two-dimensional tasks largely generated by the neurologists on the team.

The systems recorded certain data, such as position, orientation, and acceleration in 3 dimensions of the user's head with an ArduIMU sensor added to a customized VUZIX 2-D display run from an Android tablet, all integrated into 3D printed goggles, custom cabling, and a ruggedized 3D printed case. As it was obvious to the Subject Matter Experts when a person was "wobbly" based on their head movements, it seemed intuitive that tracking head movement would directly indicate how "wobbly" that person had been while interacting with the virtual system.

Some months later we moved to the analysis phase. We found that the physics engine being used for the ball and beam application wasn't using real world physics, and that this was a major issue for analysis. What was on the screen at the moment when the participant was interacting was actually critical data to determining the appropriateness of actions they were undertaking. This produced lesson one when designing serious games, which is even more critical when applied to VR: design not just from the data out, but from the analysis out. Knowing what data you think you want is essentially useless; knowing what information you need is the best way to drive finding the data you'll need.

This may have been a transparent step in other application development, but it is less obvious in VR and AR what the connection will be between your participant and the data you'll need. Every piece of every interaction could, theoretically, be tracked in VR, but it's impractical if not impossible to save that volume of data in real time. You need to choose which data, and in order to do that, you need to know as much as possible about the information you plan to generate. Otherwise, you risk losing something important, like the position the participant was tilting their head at during the precise moment when they saw a certain movement or object on the screen.

B. Develop with the Analyst

When we ported the iDETECT system to Unity, we thought we had solved most of our data problems. We were using a physics engine developed by a top-tier company, and surely that would be enough to give us the data that intuition told us would show how "wobbly" someone was. It appeared we were correct in pilot experiments, but it wasn't until the long-term study that we learned about other effects that might mask our differentiating factors. The response of the items on the screen that the participant was manipulating were being encoded into the participants' perception; when they didn't align with real-world physics, participants encoded this new physics with different levels of capability. Thus, despite being visually "more" or "less" wobbly in response, the data was incapable of indicating which participants were adapting better, which were having trouble encoding, which were still learning the controls, and other such measures.

Having an analyst on the team from the beginning might not have resolved all of that concern, but the analyst could have told us that the way we were collecting data would limit our future analysis options. Although the data collected was of the quality we needed, we had made a design choice that was specific to the analysis methods we intended to employ.

One of the best insights we found by developing in a team where the analyst was integrated from the beginning was the utility of seeing what is on the participant's screen in real time, and the ability to flag or mark data for review. Every system developed has wanted some kind of administrator oversight, the ability to view or play back exactly what the participant saw, and the ability of an administrator to control, interrupt, or in some cases interact with the system.

Often times, this has been a “nice to have” discovered in the final phases of testing, and difficult or impossible to add at that point. Working with an analyst from the beginning of ARTEMIS caused the co-development of the analyst’s interface, enabling the addition or modification of many kinds of interference that had not been anticipated. The insight thus proved invaluable, as the resultant system proved far more capable of handling unexpected events and interactions than would have been otherwise possible. This is, in certain ways, a thing that seems to obvious (and the reader argues that the analyst simply isn’t available). Yet the critical point here is that you must not simply consider what the data represents, but what you will do with the data once you have it.

That is, there is a natural assumption that one can interpret all the complexities of head movement if only one has an adequate set of sensors. This may be true, but the actual interpretation may be far beyond the scope of ordinary analysis, may involve a set of experts, and may result in data that is far too noisy. You may be able to watch a video of a person moving their head and say ”their movement is too wobbly”, or watch it alongside the view of the person within the HUD and say their reactions are delayed, but your human mind is intuiting a wealth of concepts that you may not find so intuitively available in sensor output form.

C. Develop for the Adaptation

In most data collection methods, the goal, in order to protect user privacy, is to collect only the data you absolutely need. You simply can’t do that when you’re developing a VR serious game and will need to “recreate” some kind of physical movement within the space. Privacy has to be protected in other ways, such as by using participant identifiers that have been dissociated from their identities.

Since the virtual world is not, in fact, the physical world, we cannot make the same cause and effect assumptions. Sometimes, anomalous readings are the fault of a display glitch, or a sensor error, or a calculation flaw. Since we are not directly observing, we cannot always know what has occurred. It can be necessary to recreate not only the actions of the user, but the exact construction of the VR environment as seen from the user’s perspective at that moment in time. Such reconstruction requires a wealth of data.

At present, some of the greatest protection is the very complexity of the system. Analyzing, for example, a gaze pattern as the signature of an individual is not simplistic. Although a user can be reliably identified by the pattern of their keypresses on a keyboard [20], you may not be able to isolate that kind of data “signature” from VR simulation analysis work. The differences between datasets are often too relationally dependent – that is, all of the factors contributing to the individual’s change in capability are more individually different than the differences between the test conditions.

In ARTEMIS, we stored head and hand position and orientation at 60 Hz during scenario execution. Gaze was also tracked, but rather than directly tracking eye positions, we tracked the interaction of the ”gaze” raycast with specific objects in the scene. Although this stores precise information about gaze, it somewhat limits the ability to identify the person by their eye movement data. It is possible, and even probable, that even the movements stored will be identifying information in the future, but it is also relatively likely that exact position data will be needed to create analyses of movement in the scenario.

D. Serious Games Data

Having defined the perspective of the framework, we need to examine the data itself. There are three kinds of data that are generated by VR environments on which we need to be able to perform analysis. The first concrete step in establishing an analysis plan is to determine what kind of data you are looking at, and what kind of information you are looking to generate from it.

Firstly, some data from VR scenarios is Simple Discrete Data (SDD). SDD data points can stand entirely on their own, or with the kinds of framing provided by a unit of measure. This data is easy to analyze by hand, as well as being easy to understand and visualize. We know precisely what to do with SDD. It requires no novel methods or collection. Examples of this include time to complete, button presses, or other kinds of data that could be collected in any interface and are neither unique to VR nor interdependent on simulated physics. Example analyses include Chi Square test of Association and Simple Linear Regression.

The second type of data we can collect is Combined Discrete Data (CDD). CDD is characterized by a dependence on the simulation, simulated context, or a sensor. CDD requires an association to ground truth or a collection of data to be valuable. The analysis itself can be complex, but is fully explicable, and results of CDD analyses are easy to understand and visualize.

The main “gotcha” in CDD is the dependence and interdependence. Developing from the start with an analysis plan and/or developing with an analysis perspective is typically enough to manage these data. CDD can be analyzed by traditional means, leveraging established methods for interpreting complex data, but such analysis is significantly simplified by computer. Example analyses for CDD include mXn Within or Between-subjects ANOVA.

The final, and most difficult type of data we receive from a XR system is Continuous Signal Data (CSD). Not to be confused with data points taken from a continuous range, CSD represents data samples taken at a frequency that is intended to roughly approximate or reconstruct real time data. CSD is difficult to interpret, and often information produced from CSD analysis is still difficult to display.

CSD analysis requires computer-based methods. It always requires multiple steps to transform CSD into information. For example, given a motion capture data set: first signal processing is required, then analysis of signal characteristics, and only then can further analysis, often machine-learning based, produce information. The results, when they can classify one kind of data against another, are often separate for reasons that are difficult to understand or display.

It is often easier to interpret CSD with a human watching after-action real-world video of participants alongside screen capture, tagging points of interest, than it is to intuit these items from sensor data alone.

V. Conclusions

These seven projects – iDETECT, iDETECT VR, Brain Buddy, the XOIL projects (QuadCopter BCI, HUD optimization, and HUDs for Human-Machine Teaming) and ARTEMIS –have defined our research capabilities in VR and AR. The lessons learned from the first five systems were brought forward into the data framework developed for ARTEMIS. In complex, highly inter-dependent XR serious games systems it is not enough to pursue data-driven design practices, one must consider the analysis as central to the practice of good development. Piloting an XR system is not enough, and can come far too late to make the necessary changes.

Considering your application from the perspective of analysis will drive you to create a system that gives you information from the beginning, rather than simply data. Working in collaboration with an analyst will allow you to see when that information is potentially problematic.

The interdependent nature of XR data and its simulation environment can cause a data-driven architecture to produce data points that are, although valid, useless without the context in which they were gathered. Piloting is not enough to find this in every case. Having your users learn to use the XR system, to find its interactions natural, can help to un-bury your differentiating information. It is still worth collecting all of the data that would allow the VR scenario's complete recreation, despite the intuition to not collect data you don't need to avoid the risks such data may pose to confidentiality.

VI. Future Work

Future work on this framework will focus on two areas: identifying and limiting the scope of VR "learning curve" impacts, and on simulation interdependent data. In the first case, it would be valuable to establish categories of VR systems in which to learn the interaction types, in order to correctly prepare participants for interacting with those systems. Additionally, this could identify markers to establish when participants have passed a general threshold of XR system familiarity, as these designs and their expected behaviors become more common. Perhaps some day they will become as common as mouse usage, and researchers will be able to assume that familiarity with computer systems is adequate to count the participant prepared.

The other major area of further research is in expanding the methods for recognizing, characterizing, analyzing, and displaying data that has simulation interdependence (CDD and CSD data). The more explicit the guidelines for discovering and handling these, the easier it will be for any development team to collect usable data. The more usable the data, the more valuable the insights gained will be. Valuable insight into user behavior is the goal of all Serious Games research. No environment is more suited to advancing knowledge of human behavior and interaction, but our data analysis and methods will be critical to uncovering those insights in the coming years.

References

[1] D. A. Bowman, E. Kruijff, J. J. LaViola, and I. Poupyrev, *3D User Interfaces: Theory and Practice*. USA: Addison Wesley Longman Publishing Co., Inc., 2004.

[2] A. Schall, "The future of ux research: Uncovering the true emotions of our users," *User Experience Magazine*, vol. 15, no. 2, 2015.

[3] R. Taib and S. Berkovsky, "Modeling humans via physiological and behavioural signals," *Interactions*, vol. XXVII, May-June 2020.

[4] V. Narayanasamy, K. W. Wong, C. C. Fung, and S. Rai, "Distinguishing games and simulation games from simulators," *Comput. Entertain*, vol. 4, April-June 2006.

[5] K. Kapp, *The gamification of learning and instruction: Game-based methods and strategies for training and education. San Francisco, CA: Pfeiffer*. 01 2012.

[6] A. Marczewski, "Game thinking. even ninja monkeys like to play: Gamification, game thinking and motivational design," *CreateSpace Independent Publishing*, vol. 1.

[7] Y. Kim and J. A. Ruipérez-Valiente, "Data-driven game design: The case of difficulty in educational games," pp. 449–454, 09 2020.

[8] L. Luo, H. Yin, W. Cai, M. Lees, N. Othman, and S. Zhou, "Towards a data-driven approach to scenario generation for serious games," *Computer Animation and Virtual Worlds*, vol. 25, pp. 395–404, 05 2014.

[9] S. Cooper, "A framework for scientific discovery through video games," *Association for Computing Machinery and Morgan Claypool,*. New York, NY USA.

[10] S. Bansal, "What is the difference between data and information?," Dec 2021.

[11] J. M. Barker, D. W. Wright, F. C. Goldstein, J. Ockerman, J. J. Ratcliff, and M. C. Laplaca, "The detect system: portable, reduced-length neuropsychological testing for mild traumatic brain injury via a novel immersive environment," *Journal of Medical Engineering and Technology*, vol. 31, pp. 161–169, May-June 2007.

[12] P. Sweetser and P. Wyeth, "Gameflow: a model for evaluating player enjoyment in games," *Comput. Entertain.*, vol. 3, July 2005.

[13] M. Ventura and V. Shute, "The validity of a game-based assessment of persistence," *Computers in Human Behavior*, vol. 29, pp. 2568–2572, 11 2013.

[14] S. Cooper, *A Framework for Scientific Discovery Through Video Games*. New York, NY, USA: Association for Computing Machinery and Morgan Claypool.

[15] C. Linehan, B. Kirman, S. Lawson, and G. Chan, "Practical, appropriate, empirically-validated guidelines for designing educational games," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1979–1988.

[16] Y. Nigmatullina, P. J. Hellyer, P. Nachev, D. J. Sharp, and B. M. Seemungal, "The neuroanatomical correlates of training-related perceptuo-reflex uncoupling in dancers," *Cerebral Cortex*, pp. 554–562.

[17] A. Koralek, X. Jin, J. Long, R. Costa, and J. Carmena, "Corticostriatal plasticity is necessary for learning intentional neuroprosthetic skills," *Nature*, pp. 331–335.

[18] V. V. Abeele, J. Wouters, A. G. Pol Ghesquière, and L. Geurts, "Game-based assessment of psycho-acoustic thresholds: Not all games are equal!," *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, pp. 331–341, 2015.

[19] I. Koglbauer, R. Michael, and R. Braunstingl, "Positive effects of combined aircraft and simulator training on the acquisition of visual flight skills.," *Brain Behavior and Evolution*, vol. 20, pp. 309–318, 01 2016.

[20] K. Longi, J. Leinonen, H. Nygren, J. Salmi, A. Klami, and A. Vihavainen, "Identification of programmers from typing patterns," *Proceedings of the 15th Koli Calling Conference on Computing Education Research*, pp. 60–67, 2015.