Projective Replay Analysis: A Reflective Approach for Aligning an Educational Game to its Goals

Erik Harpstead
Ph.D. Thesis Proposal
May 16th, 2016
Human-Computer Interaction Institute, Carnegie Mellon University

Committee:
Vincent Aleven (Chair) (HCII, CMU)
Jodi Forlizzi (HCII, CMU)
Jessica Hammer (HCII & ETC, CMU)
Sharon Carver (Psychology, CMU)
Jesse Schell (ETC, CMU)

ABSTRACT

Educational games have become an established paradigm of instructional practice, however, there is still much to be learned about how to design games so that they can be the most beneficial to learners. An important consideration when designing an educational game is whether there is good alignment between its content goals and the instructional behaviors it makes in order to reinforce those goals. What is needed is a better way to define and evaluate this alignment in order to guide the educational game design process. This thesis explores ways to operationalize this concept of alignment and demonstrates an analysis technique that helps educational game designers measure the alignment of both current educational game designs as well as prototypes of future iterations.

In my work thus far, I have explored the use of replay analysis, which analyzes player experience in terms of in-game replay files rather than traditional analytics data, as a means of capturing gameplay experience for the evaluation of alignment between an educational game's feedback and its stated goals. The majority of this work has been performed in the context of *RumbleBlocks*, an educational game that teaches basic structural stability and balance concepts to young children. This work has highlighted that *RumbleBlocks* likely possesses a misalignment in how it teachers the concept of designing for a low center of mass to students. It has also lead to suggestions of design iterations for future implementations of the game. This work has shown that replay analysis can be used to evaluate the alignment of an educational game and suggests future directions.

In the proposed work, I plan to demonstrate an extension of replay analysis that I call Projective Replay Analysis, which uses recorded student replay data in new versions of the game in order to evaluate whether alignment has improved. To do this, I plan to implement two forms of projective replay: Literal replay, which replays past player actions through a new game version exactly as they were originally recorded; and Flexible, which uses prior player actions as training

data for AI player models, which then play through a new game version as if they were players. Finally, to assess the validity of this method of game evaluation, I will perform a close-the-loop study with a new population of human play testers to validate whether the conclusions reached through virtual methods correspond to those reached in a normal playtesting situation.

This work will make contributions to the fields of human-computer interaction, by exploring the benefits of limitations of different replay paradigms for the evaluation of interactive systems; learning sciences, by establishing a novel operationalization of alignment for instructional moves; and educational game design, by providing a model for using Projective Replay Analysis to guide the iterative development of an educational game.

Table of Contents

INTRODUCTION

I have gotten used to starting my publications by saying that there is growing interest in the use of games for education. However, at this point in my career, I believe it more appropriate to say that games are an established mode of instructional practice. James Gee's seminal book [20], that arguably jump started the modern field, was published over a decade ago, when I was just graduating from middle school. I went to elementary school just blocks away from the central headquarters of the Minnesota Educational Computing Consortium (MECC), the creators of *Oregon Trail*. Games have been a central component of the educational system for as long as I have been in that system. Yet in the learning sciences, we seem to be conflicted about whether or not they can be beneficial for learning. Every few months we see yet another literature review claiming that evidence is mixed as to whether or not games can produce measurable learning gains [65]. I believe this is the wrong question to ask. At best it is already solved, requiring only an existence proof, of which there are several [9,17,22], and at worst it is fundamentally unanswerable [14,33]. At this point, the far more useful question to ask is how do we make educational games that are beneficial for learning.

I am not the first person to ask this question. There are several frameworks on educational game design that describe the qualities of good educational games [2,5,6]. However, the vast majority of these frameworks provide few recommendations for the actual design *process*. I like to describe this issue by way of an analogy. To me, the promise of educational games is like the promise of the new world across the sea. As a field, we have already had a few explorers make their way there and they have brought back learning theories that give us a map of which parts of the continent would be most fruitful to settle. In effect, we know what we want to create; all we have to do is design a voyage that sails a straight line to the land of effective educational games. However, the process of crossing that gulf between your current reality and your goal is never that easy, in sailing or in design, and it can be fraught with unforeseen perils. The winds of stakeholder opinion could change, technology can break down, and all manner of storms can come to toss a design off course. When these problems strike, it does no good to have a well-detailed map that tells you the promised land of effective instruction is west if you have lost the sense of which direction you are heading. So, rather than develop yet another definition of a good educational game, I instead want to focus on how better to steer toward the desired outcome of a good game.

What is missing from many of these prior conceptions of educational game design is a notion of design as a process. To address this gap in the literature, I propose the Projective Replay Analysis paradigm, an approach that uses novel game analytics techniques to assist designers in aligning their game to its instructional goals. This approach enables designers to carry on a reflective conversation with computational models of their target demographic.

Expected Contributions

Overall, I expect my work will make contributions to the fields of human-computer interaction, learning science, and educational game design.

In human-computer interaction, replay paradigms have been an established means for the evaluation of interactive software. In exploring Projective Replay Analysis, I am probing the boundaries of how the nature of particular design changes affect the validity of replay based methodologies. The results of my planned studies will provide insight into the benefits and limitations of a literal replay paradigm and hopefully present encouraging support for AI enabled flexible replay as a valid method for evaluating future design iterations in the absence of new player data.

In the learning sciences, my focus on measurements of the alignment of instructional moves to goals stands to benefit the literature on alignment research, which is typically more focused on the alignment of assessments to goals. In particular my definition of alignment in terms of a solution space is a novel way to approach the alignment of instructional moves. Further, my advocacy of alignment within a common reference frame will hopefully provide clarity to prior results in the field.

In educational game design, I expect to demonstrate Projective Replay Analysis as a method that can inform the practice of educational game design in the future. In particular, establishing an approach that adheres to the tenants of reflective practice and maintains a connection to student data stands to benefit many educational game design efforts. Finally, the reference implementation of Projective Replay will stand as a model for how other researchers and practitioners can apply my approach in their own work.

ALIGNMENT

The fundamental question that this thesis looks to answer is: “Is my educational game good?” and relatedly "If I were to make this change could I tell if it was better?” While these questions may seem almost uselessly broad, they have power in how they can be asked of any educational game designed for any subject. These questions embody the kinds of move-testing hypotheses that designers make at the heart of the iterative design process [58]. Of course, a certain amount of operationalization is required to make these questions useful. In particular, what does it mean for an educational game to be good? The goodness of an educational game could be defined in a number of ways but an important aspect that I focus on in this thesis is whether the game meets the educational goals that were set for it. I refer to this relationship between a particular design and its goals as alignment.

The desirability of alignment is clear as a matter of design process, just as it is intuitively desirable to stay on course when sailing. A complication arises, however, in how best to translate this general desirability into some sort of measure that can be used as a yardstick to inform a design process. While it is not unreasonable to assume that good products will result from good practice, the educational game design process can be complicated by issues like expert blind spot [31], the complexity of ill-defined domains [40] and the wickedness of design problems in general [11]. What is needed is an operationalization that can take alignment from a general idea of *process* toward a *property* of a particular design that can be used to inform whether a current design is better than another.

Alignment is originally a term from instructional design literature. Good definitions, however, are hard to come by. In most discussions, alignment is used as a term to frame other issues such as content/construct validity in assessment [36], or in comparing the connection of state assessments to their respective standards [43]. One of the more concise definitions that does exist comes from Cohen who defines Instructional Alignment as “... the extent to which stimulus conditions match among three instructional components: intended outcomes, instructional processes, and instructional assessment” [16]. What this essentially says is that the goals, instruction, and assessment of a particular educational setting are all about the same thing. This is important because if the constructs underlying goals, instruction, or assessment differ it results in learners and instructors talking past each other. It would be unfair to assess someone on calculus when you taught them algebra, and if you really wanted them to be learning chemistry then that assessment will not tell you anything useful anyway.

Cases of misalignment are rarely so glaringly obvious. For example, a recent study of the commercially successful algebra game *DragonBox* found that students tested before and after playing the game showed no improvement on a paper and pencil algebra test [39]. While both the game and assessment claim to be about algebra, a lack of test improvement would suggest that there is some kind of misalignment between the instructional behavior of the game and its stated goals. Alternatively, it could be argued that there is some deficiency in how the particular test embodied the domain of algebra. Both the game and the assessment possess the face validity to claim to be about algebra, but given the empirical result, they cannot both be right. These kinds of ambiguities highlight one of the issues of Cohen’s definition of alignment: if the only condition of alignment is that all the elements of goals, instruction, and assessment be the same, then any one of them can be suspect. Since every element is equally capable of failure, any critique can be dismissed by saying whichever element highlights an error is simply wrong.

I refer to this conundrum as instructional alignment’s n-body problem. In physics, the n-body problem refers to trying to predict the individual motions of n bodies, accounting for all of their individual gravitational interactions on each other. The problem is easy for two bodies but gets significantly harder, if not impossible, as more bodies are added to the system. Similarly, alignment is almost always discussed as a relationship between two things (e.g., state assessments and their related standards [43]) but there are many such binary relationships in a larger educational system. For example, one could discuss how well aligned a given unit of instruction is to its educational goals, or how well aligned the learning goals of an instructional product are to what society believes to be a desirable set of skills. Just as in physics based n-body problems, measuring alignment requires the assumption of a common reference frame. In traditional educational practice this common frame is normally provided by state standards and standardized assessments, but those constructs are not relevant to all contexts nor available for all domains, particularly when such domains are ill-defined [40].

To prevent this kind of crosstalk in practice, the instructional design literature has provided several methods for helping to ensure the alignment of instruction. One of the more influential such processes is Backwards Design [64]. In Backwards Design, instructional goals form the common reference frame against which all other aspects of instruction are considered. In this model, instructional goals manifest as a collection of specific understandings that an instructor wants their learners to acquire. For each of those goals the instructor must decide on a collection of measures, ideally multiple for each goal, that they will use to get a sense of whether a learner possess a desired level of understanding in each goal. Finally, they design instructional activities that are likely to lead learners to progress on the measures of understanding.

While Backwards Design has been influential as a model for instructional design, it does have some problems in being adapted as a model for educational game design. Firstly, Backwards Design is primarily directed toward practicing teachers in traditional classroom contexts. This setting allows for instructional design in an environment that affords rich multifaceted assessment and the ability to pivot instruction readily according to the instructor's judgment of learner behavior. Essentially, Backwards Design works because it is not concerned with shaping the form of instructional behaviors or environments but of the understandings of learners directly. Designers of educational games and technologies have to take a different approach because they are a step removed from the learner. While ultimately their goal is to craft a learner's understanding, educational game designers have to accomplish that by crafting the instructional behaviors of an environment that works in their absence.

Further complicating the educational game design task is the second order nature of game design [56]. Second order design refers to the idea that while a game designer has direct input on the rules and mechanics of a game system, the player experience is only indirectly created by these mechanics and systems. In interacting with a game, players become part authors of their own experience through their choices and actions within the framework provided by the designer. This dynamic nature of game play [28] can lead players to discover nuances and implications within a game's system that the designers themselves never anticipated [35]. When a game is being designed for an educational purpose, it becomes important to be able to anticipate the kinds of experiential variations that players are likely to explore, since the game's reactions to player variation will have to satisfy that educational purpose in the absence of the designer's direct input.

What educational game designers need is a formulation of alignment that allows them to consider how well the instructional behavior of their game aligns to their instructional goals. Typically, this is approximated by using some kind of assessment external to the game that can act as a reference frame between instruction and goals. Quite a lot of work has gone into methods for aligning *assessments* to particular *goal constructs* [45]; however, these methods do not necessarily provide strong affordances for making the leap between *instructional behavior* and *goals*. The typical paradigm of assessment driven instructional design is to define some form of pre-posttest that is administered before and after an instructional intervention. If an improvement on this test is observed then the usual conclusion is that the instruction was good because it did its job. While this is a gold standard scientific design, I would argue it is not *informative* but rather *confirmatory*. If the result is negative or inconclusive then little can be said about what aspect of the instructional behavior within a system broke down. This is a problem for iterative design because through the course of iteration a designer is likely to deal with far more bad prototypes than good ones [21].

In order to account for the complexities of games' second order nature we need an informative operationalization of alignment that can do more than confirm whether or not a game is working. Such an operationalization needs a way of relating a game's instructional moves (i.e., feedback) to an assessment of player understanding within the context of the game itself. Drilling analysis down to a finer grain size of instructional moves allows for a more nuanced evaluation of alignment issues than is afforded by the monolithic treatment of a pre-posttest external to the game. To this end, I propose to define alignment with a paradigm of internal assessment linked to instructional feedback moves that the game makes in response to player actions. The essential logic behind this paradigm is that an educational game is about what it incentivizes, and a game should be about the same thing you want to be teaching.

Before I close the loop on my operationalization of alignment there are still a couple definitions to cover. What do I mean by feedback moves and how do I define a player action? To answer these questions I want to make what might seem like an unintuitive jump to the idea of game spaces.

All games take place in some kind of space. One way of thinking about this space is as the Magic Circle [56] that forms a boundary around the game and gives its components meaning (e.g., outside the Magic Circle, *Monopoly* money is just paper). However, game spaces have other useful properties beyond their ability to give endogenous value to game elements. Another useful lens for talking about game space is in terms of *functional* game space [57]. The functional space of a game can be thought of as the space in which the game *really* takes place. For example, while the game of *Monopoly* is printed as a two-dimensional board, in terms of functional space, the game is really a single one-dimensional loop of properties. Further, each property is actually a zero-dimensional space, as the particular placement of a player piece within the bounds of a property is meaningless.

Using a lens of functional space allows us to define a player action as anything that meaningfully changes the state of the functional space of a game. Extending this concept further, a solution to a game challenge or level is a collection of player actions that lead to a goal state. This allows us to define the solution space to a challenge or level as the set of all pathways through, or configurations of, functional game space that lead to goal states. I make the assumption that in creating their solution to an in-game challenge a player is expressing their understanding of the knowledge and skills (i.e., knowledge components) required to solve that challenge [30]. Under this assumption there would arise a qualifiable difference between solutions created by someone who understands a concept and someone who does not. Markers of this difference can be formalized as a separation function of the solution space in terms of target domain principles.

Feedback can be defined in similar terms as the collection of changes to the functional game space that the game communicates to players in response to their actions. While this collection of changes is potentially very large it can be useful to think about it in terms of different channels [57] that are functionally distinct. For example, successfully solving an in-game puzzle might result in a bright green check mark, animated fireworks, and a “woo-hoo” sound; all of these responses are multiple channels of conveying the same message; namely, that the player succeeded. Similar to how players with different understandings will generate different portions of a solution space, feedback can act as a separation function over the solution space, where a certain collection of solutions will be give one type of feedback while the rest will be given another.

Using this operationalization, solutions within a solution space can arrive at one of four designations, best thought of as the 2x2 matrix shown in Figure 1. Two quadrants in this matrix are desirable and, if solutions consistently land in either one of these quadrants, this indicates that the game is well aligned. Solutions that are highly principled would ideally be given some form of positive feedback, which would imply that the game is reinforcing target concepts to the player. Similarly, solutions that are unprincipled should be given negative feedback, which would mean that the game is discouraging deviations from target concepts, allowing a player to learn from their mistakes. Solutions would ideally *not* fall into the other two quadrants, where principled solutions are discouraged or unprincipled solutions are reinforced. In these cases, the game is sending contradictory feedback to students, at best confusing them and at worst fostering misconceptions.

		Domain Judgment	
		Unprincipled	Principled
Game Feedback	Positive	**Bad**	**Good**
	Negative	**Good**	**Bad**

Figure 1. A matrix showing the possible alignment interpretations of student solutions based on the agreement between domain judgment and game feedback.

Coming full circle, I define the alignment of an educational game as the level of agreement between the game’s separation function of its solution space (i.e., feedback), and a domain-principle-based separation function of the solution space.

Looking at alignment as agreement between how feedback separates a solution space and how assessment separates the same space, it would seems like the obvious design solution is to base feedback directly on an in-game assessment; however, this is not always possible. For example, the target domain might be ill-defined [40] meaning that there is no single strong domain theory that could be used to create feedback rules, or any such rules would be subject to debate by experts. Alternatively, the instructional goals might be heuristic in nature (e.g., a wider base leads to a more stable structure); in which case any decision point used to define “good enough” will be inherently arbitrary. Further, the layering of game mechanics onto a pedagogical activity entails a certain level of contrivance that invites the possibility of misalignment by accident.

In all of these cases, playtesting and iteration are essential to see if the instructional behaviors of a game are aligning with designers’ expectations. Since the instructional behaviors of a game encompass the results of many different design decisions, we need a definition of alignment that is capable of functioning at a fine grain size. I believe my definition of alignment as an agreement between separations of a solution space serves this purpose.

GAME ANALYTICS AND PROJECTIVE REPLAY ANALYSIS

In order to use my notion of alignment to guide the design of effective educational games it is necessary to have a way of measuring the alignment of the current state of a design. Beyond measurement of a current design, it would also be desirable to have some way of estimating the potential that future design iterations hold with regard to alignment. I propose that Projective Replay Analysis is an approach that can facilitate measuring how aligned a current game design is to its goals and enable the ability to forecast the potential future alignment of a next design iteration. In this section, I describe the structure of the Projective Replay Analysis approach as well as the related game analytics literature that informs its design.

Game Analytics

The measurement of player experience has been a longstanding interest of both game user research and serious games research [37,59]. Many practitioners and researchers have explored different ways of measuring player experience [46] including self-report and subjective surveys [10,50], biometric and physical response [44,47], and data mining and analytics approaches [18,63]. Among these approaches, analytics is the most promising for guiding alignment analysis in terms of solution space.

Game analytics research is generally concerned with the application of game log data (also commonly referred to as telemetry) to answer questions about game players and game design . There are many approaches to game analytics, and several ways to describe the relationships between different methods (see [59] for several different taxonomies). In moving toward a design informative process using analytics it is necessary to understand where analytics sits within the iterative design process. In his influential account, Donald Schön describes the design process as a reflective conversation with a situation [58]. This conversation iterates through stages of reframing a design situation, moving to improve the situation, and then evaluating whether the move worked. Commonly, analytic approaches typically occupy the evaluation stage of this loop in providing a picture of a current design in terms of some framing. Existing approaches also support reframing to varying extents in their ability to ask questions beyond their original intent. What is generally missing, however, is the ability to explore move testing. Rather than employing analytics as a tool throughout the conversation with a situation, a designer must step away from their context and ask for it to be analysed before being able to form new design hypotheses to test.

Using Schön's notion of a reflective conversation as a guide I categorize different game analytics techniques into one of three groups based on two distinctions. The first distinction I make is whether an approach is generally measure-then-record or record-then-measure.

Measure-then-record approaches to game analytics work by performing the actual measurement of a desirable feature within the game and then recording only the result of that measurement. This is generally the realm of very large-scale metric-based [18] approaches to understanding player experience where key performance indicators that designers want to record are known well in advance [32]. The benefit of these approaches is that they generally require minimal post-processing of data beyond simple aggregation or statistical testing because the desirable information exists directly in game logs. However, a measure-then-record approach has limited capacity to inform reframing of a game design situation because what can be said about the player experience is only what can be inferred from the measurements that exist in the data, as any other context was lost. Because of their capacity for scale and general rigidity, measure-then-record approaches are usually employed for later stage monitoring of games post release.

Record-then-measure techniques take the approach of recording some kind of representation of the player experience from which measures are distilled. This style of game analytics is far more common in educational game work with several existing examples [52,54,60]. The trace-based recording approach common to intelligent tutoring systems is also an example of this kind of analytics [8,29]. The general benefit of these kinds of approaches is that they preserve some portion of the context of the play experience allowing for that experience to be reframed to some extent. The second distinction in my categorization of prior analytics approaches is within the group of record-then-measure approaches and asks whether an approach is designed to record-to-measure or record-to-capture.

Record-to-measure paradigms are typified by a focus on recording player experience in service of a future measurement that is planned. The quintessential example of this kind of approach is Evidence-Centered Design [34,45,55], which is focused on an extensive process of building evidentiary arguments about learner competencies before any measurement takes place. Record-to-measure techniques are generally concerned more with a paradigm of playtesting to prove[1] to stakeholders that a design is working rather than inform design refinement. While these approaches have strong validity for developing assessments I would argue they are less suited for informing iterative design because they are generally committed to the lens of their assessment, which can limit the ability to reframe the context beyond that measurement.

Record-to-capture, on the other hand, generally intends to capture the player experience as it happened and defers any measurement or characterization of the experience until after it has been captured. This is commonly where extremely high fidelity, labor-intensive techniques, like video recording will be employed to provide ground truth for other methods [51,53]. These forms of analytics are desirable early in the refinement stages of playtesting[1] because they are the closest to the traditional form of design evaluation via observation. However, the maintenance of so much context runs the risk of drowning in excessive detail leading context heavy analytics approach to be discouraged by the broader community of research [4,19,38].

Replay Analysis

To enable an approach that is both capable of dealing in the rich complexity of player experience and allows designers to carry on their conversation with the design scenario, I propose the Projective Replay Analysis paradigm. At its core Projective Replay Analysis is a record-to-capture serious game analytics technique that grounds itself in lived player experience by focusing on the interpretation of replays of players' sessions. It further enables the use of previously captured player replays as input to run through a next iteration version of the game. This allows designers to evaluate the potential of a design iteration from the perspective of a previously sampled player demographic without having to gather new game play data from that demographic.

This approach differs from prior metric and process based approaches by examining a player's sessions as a live instantiation of the game state within a running game engine. This allows analysis to consider a much deeper picture of a player performance, which can be analyzed from multiple perspectives, and provides access to more contextual properties of game elements than a designer may have thought to measure. Forms of replay have been used to analyse both educational technologies [1,7] and interactive systems [49] in the past but the truly novel aspect of Projective Replay is the ability to go beyond the original recording context to analyse new design iterations using old data.

The Projective Replay Analysis approach consists of three major components: a schema for logging player actions, and a system to replay recorded logs through the game engine, both of which I have demonstrated in my prior work [25,26]; and a computational player model that

[1] http://playtestingworkshops.com/about.html

simulates the decision making process of human play testers, which forms the bulk of my proposed work.

The logging schema is used to capture players' sessions as replay fidelity log traces. I define replay fidelity as detail sufficient to recreate the game state. In general this means capturing player actions as well as relevant game context at the level of a basic action, defined as the smallest unit of meaningful action that a player can exert on the functional game space. These actions are meant to be contextualized to the game world (e.g., picking up or dropping an object), rather than the raw input of the player (e.g., mouse down at position (x, y)). Additionally, each action is paired with a description of the state of the game just before the action took place to provide contextual information. The paired recording of state is important in situations where a game's state and behavior could change for reasons other than direct player action (e.g., a physics engine simulating the motion of objects, or an non-player character making its own independent decisions). The emphasis on contextualized action paired with state is meant to embody a record-to-capture paradigm of analytics and maximize the amount of information to be available for future reframing and move testing.

The second major component of the approach is a system for replaying actions, which I refer to as a Replay Analysis Engine (RAE). The RAE reads in a player's log file and reconstructs the player's play session action-by-action. For each action, the RAE first constructs the state in which the action took place and then enacts the player action to let the game engine resolve any consequences of the action, using the same code that would normally handle such an action. Analyses can then be performed by augmenting the replayed state to create new measures with full access to any state attributes that would have been present at playtime. These analyses represent an accurate reproduction of the player's own experience because the re-instantiated state is composed of exactly the same game elements, in terms of code. This allows analyses to consider each action within the context in which it took place, and to know the initial conditions of an action that may take time to broadly affect the game environment. Having paired states with each action also allows logs to be replayed accurately without having to interpolate prior actions.

The third component of Projective Replay Analysis, which I will build in my proposed future work, is a computational player model used to simulate the decision-making processes of playtesters when exploring future game design iterations. This player model comes in two forms Literal Replay, and Flexible Replay. In the literal form, this player model simply acts as the normal RAE in the new game context by re-enacted players' actions as they occurred in the replay file. If literal replay ever encounters states or actions that are no longer compatible with new game mechanics it will simply fail and move on. The second form of player model is one augmented with an AI design making process. This player model takes demonstrations from players' replays and learns to perform its own actions within the game, in a style similar to the respective player. This player model is implemented using TRESTLE, a model of human concept formation [41] that learns hierarchical concept trees given structured examples. Using an Apprentice Learner Architecture paradigm [42], an action planner can be used to translate demonstrations into generalized action sequences (i.e., how learning) and TRESTLE can be used to learn concepts corresponding to when those action sequences should be employed (i.e., when learning). These learned action concepts can then be used to perform model tracing similar to how intelligent tutoring systems employ production rules [3].

The benefit of Projective Replay Analysis that I plan to demonstrate in my proposed work is that it can perform similar analyses on both existing game versions as well as future iterations without having to gather new data. Additionally, the method allows analyses to make use of all of the game information that would have been available at the time of the original playtest enabling multi-faceted measurement that is free to evolve as designer proceed through reframing their understanding of the game context.

PRIOR WORK

In my prior work I have demonstrated the affordances of Replay Analysis within the context of analyzing the educational game *RumbleBlocks*, a game about building and physics [13]. The evaluation of *RumbleBlocks* was guided by the overarching question of whether the game, as currently designed, was well aligned to its instructional goals. The investigation to answer this question proceeded through a series of stages that built on each other over time. Each particular analysis used a different perspective to frame player experience but all of them derived from the same original collection of data. The whole thread of work serves as an initial case study into how the Replay Analysis approach enables this kind of broad exploration of game alignment.

In this section, I will describe the various analyses of *RumbleBlocks* that I have done in the past and how each of them was facilitated by the Replay Analysis approach. This work has highlighted some potential issues in the existing design of *RumbleBlocks*, and suggested possibilities for redesign that I plan to explore as a component of my proposed work.

Background

Before getting into the details of the different analyses I performed it is necessary to have some background on the design and goals of the educational game *RumbleBlocks* [13] as well as the structure of the formative evaluation study that yielded the majority of my data.

RumbleBlocks

RumbleBlocks (Figure 2) is an educational game designed to teach basic structural stability and balance concepts to children in kindergarten through grade 3 (5-8 years old). Its main focus is on three basic principles of stability:

1. An object with a **wider base** is more stable
2. An object that is **symmetrical** is more stable
3. An object with a **lower center of mass** is more stable

These principles are derived from goals outlined in the National Research Council's Framework for New Science Educational Standards [48] and other science education curricula for the target age group.

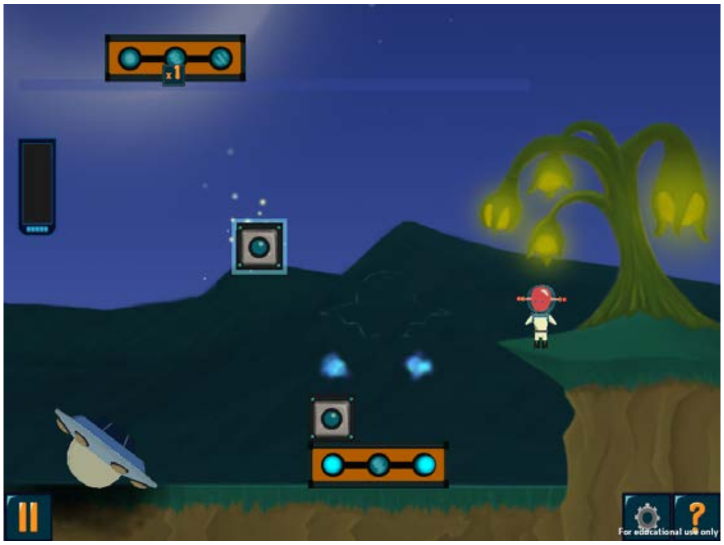

Figure 2. A screenshot of *RumbleBlocks*.

The game follows a sci-fi narrative where players help a group of stranded aliens on a number of foreign planets. Each game level starts with the player finding an alien stranded on a cliff and a deactivated spaceship left off to the side of the world (see Figure 2). The player's goal is to build a tower out of blocks that is tall enough to reach the alien so that they can give the alien's ship back. In the process, they must also capture a series of energy dots with their tower, which are captured in orbs on the blocks and are narratively used to provide the ship with power. Once the player has placed the ship on top of the tower, it powers up, and triggers an earthquake. If the earthquake topples the tower, or knocks the ship off the top, then the player fails and must restart the level; however, if the tower remains standing, with the ship on top, then the player succeeds and moves on to the next level.

Each set of levels in *RumbleBlocks* is designed to focus on a different principle of stability. The targeting of different principles is accomplished mainly through level design. The energy dots can be used to both scaffold and limit students' solutions to a level, forcing them to prioritize one principle over another. However, even with this scaffolded design, there are an unknown number of possible valid solutions to any given level because the earthquake mechanic relies on the dynamics of a real-time physics engine to evaluate the student's structure. That is, even though the level designer may intend for a particular tower design to be the solution, other designs may also work.

The unknown nature of a given level's solution space entertains the possibility that players could create solutions that ignore the target principle for the level. This would allow learners to complete the game without having to contend with the entire set of target principles. While one might expect that the physics engine takes care of most of these misalignment cases in *RumbleBlocks* (i.e., regardless of how well a solution embodies the target principle for a level it is still subject to the laws of physics), there are many design decisions that can affect this behavior. For example, changing the mass and friction properties of blocks can alter how likely a structure is to fall down under an earthquake. Additionally, altering the speed or magnitude of the earthquake can also affect game outcomes. This presents a challenge in anticipating whether or not the game is providing properly aligned feedback across the myriad of possible player experiences.

Formative Evaluation Study

The data I will be discussing was gathered as part of a formative evaluation study that I helped carry out. This formative evaluation took the form of a series of in-class playtests paired with out of game transfer tests. The study was done with 174 students in two Pittsburgh area public schools who were in the target demographic (ages 5-8). Testing took place over four sessions: an external pretest, two 40 minute sessions of play, and an external posttest.

Two sets of levels were selected to be used as in-game pre- and posttests counterbalanced across players. These levels were chosen out of the normal pool of levels, but were altered to remove the energy dot mechanic and to prevent players from retrying after a failed attempt. These special levels were placed after a short collection of tutorial levels, which explained the basic mechanics of the game, and at the end of the game. This design allowed us to get a sense of how players built before and after they had experience with the game. In addition to the in-game evaluations, players also took out-of-game paper and pencil tests, before and after playing the game. These tests contained items relating to stability and construction, based on the three principles of base width, low center of mass, and symmetry.

Throughout the formative evaluation study, *RumbleBlocks* was instrumented to record replay fidelity log files. In the context of *RumbleBlocks* this meant recording every time that a player picks up or places a block[2]. I built an RAE into the game in order to take these replay fidelity log files and reenact them within the game engine [26]. This engine enabled the various reframings of log data that I performed throughout the analysis that follows.

Replaying for Pre-Posttest Measures

The formative evaluation study that my data comes from was performed using a pre-posttest design. As a first pass of evaluation, I leveraged this pre-posttest design to confirm whether *RumbleBlocks* was succeeding in getting players to improve in their understanding of its target principles. As I stated previously such an analysis would primarily be confirmatory with regard to alignment evaluation. If positive learning results were found it would be reasonable to assume that the game was well aligned but if results are not positive or conclusive then the interpretation with regard to alignment becomes unclear.

The results from the formative evaluation study were promising. The out-of-game tests showed a slight, yet significant increase in player's performance from pretest to posttest, using a paired-samples t-test, $t(173) = -2.13$, $p = .03$, $d = .16$. Looking at players' pass rates on the in-game pre-post levels would seem to suggest a similar conclusion, showing a significant, medium sized increate in performance using a paired-samples t-test, $t(173) = -4.96$, $p < .001$, $d = .51$.

While these initial pre-posttest results are encouraging, I wanted to look into the replay data to see if there was behavioral evidence that players not only got better at the game but did so because they were better instantiating the principles it targets. This would mean that before and after playing the game for some time, players would build towers in the unguided pre-posttest

[2] The game also logged any time that any two objects collided by this proved to be an excessive amount of detail.

levels that showed a better awareness that (1) a structure with a wide base is more stable, (2) a structure with a lower center of mass is more stable, and (3) a structure that is symmetrical is more stable. It is important to note that looking at a difference in metrics related to learning goals is different from looking at the difference in player success rate. If we entertain the possibility that the game is not necessarily well aligned, then it is possible that players could improve in their pass rate in the game for reasons other than following the principles that are central to the game's goals.

To find out whether players were better leveraging the physics principles targeted by the game in their solutions, I instrumented the RAE to calculate a variety of metrics based on each player's final state of each in-game pre-posttest level. I refer to these metrics as Principle-Relevant Metrics (PRMs) in that they are metrics that are intrinsically tied to one of the principles of stability that the game targets. These metrics were: the width of the tower's base (Figure 3A), the height of the tower's center of mass relative to the ground (Figure 3B), and a measure of symmetry defined as the angle formed by a ray from the center of the base to the center of mass and 90° (Figure 3C). Once I had a base of metrics I normalized the scores across players within each to have a mean of 0 and a standard deviation of 1. This was done to account for the nuanced difference in level design making it difficult to compare metrics directly between levels that possess different affordances for solution design.

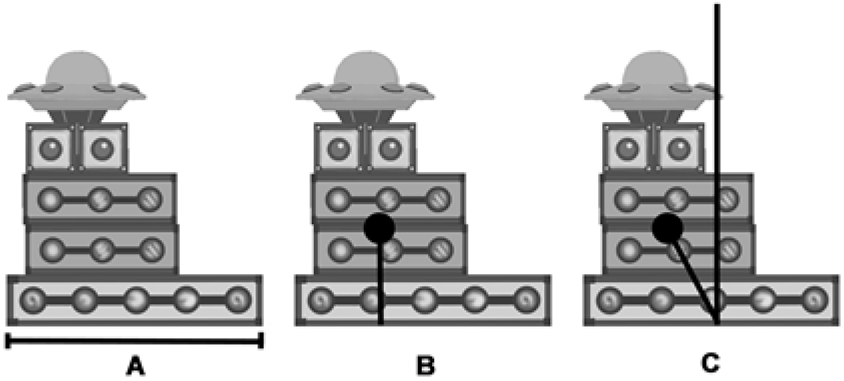

Figure 3. A visual depiction of each of the 3 Principle-Relevant Metrics used in analysis. (A) Base Width, (B) Center of Mass Height, and (C) Symmetry Angle.

To see if there was any improvement on the use of principles in players solutions from the pre- and posttest levels I compared each students averaged PRMs using a paired-samples *t*-test. Looking at the results in Table 1 I saw a significant improvement for the base width and symmetry metrics, meaning that at the end of playing the game, students were beginning to design towers that had wider bases and more symmetrical layouts. However, I did not see any significant difference in terms of center of mass height, meaning that students did not seem to attempt to lower the center of mass of their structures. This result suggests the possibility that the current version of the game may possess a misalignment in how it handles the low center of mass principle, however as stated previously a null result in pre-posttest comparison cannot reach this conclusion definitively.

	Pretest		Posttest				
Metric	***M***	***SD***	***M***	***SD***	***t*(173)**	***p***	***D***
Base Width	.60	.01	.64	.01	-2.77	.006	.30
Center of Mass Height	1.61	.02	1.63	.02	-.66	.501	.08
Symmetry Angle	5.98	.34	5.20	.27	1.98	.050	.19

Table 1. t-test results for average Principle-Relevant Metrics from pretest to posttest.

Feedback Alignment Analysis

Knowing from the pre-post level analysis that there were possibly some misalignment issues with *RumbleBlocks*, the next step in analysis was to determine if the game was properly incentivizing players to act in a way that corresponds to the goals for the game. If the game is knocking over towers that are principled or letting poorly designed towers remain standing, then players will not know what to make of the feedback they are given and improve toward better understanding. Such cases would be examples of misalignment.

To examine this question I wanted to test if there was a relation between the relative principled-ness of student solutions and the whether the game deemed the solutions successful. To do this I needed to establish whether the principled-ness of a tower (measured by the relevant PRM) could predict that a tower would stand or fall in the earthquake. To explore this question I employed a regression analysis. What we would expect from this analysis is that the principle which is targeted by a level has a strong predictive relationship with success on that level. Further, It is important to note that this analysis is concerned primarily with the behavior of the game and not with student performance. In this context students are merely providing the test data for my analysis of the game’s system.

To facilitate this analysis, I employed the RAE to calculate the same PRMs from the pre-post analysis, except this time to do it for all levels. I wanted to explore how well metrics that should indicate a well-constructed tower (i.e., a domain-principle-based separation function of solution space) actually corresponded to a player passing a given level (i.e., a feedback-based separation function of solution space). To do this, I created 3 groups of solutions by collecting together all player solutions to levels that target each of the three principles (i.e., all levels targeting the wide base principle together, all levels targeting the low center of mass principles together and all levels targeting the symmetry principle together). I performed a logistic regression using each of the metrics of the players’ towers (generated by replay analysis) to predict success on a level, with all metrics normalized within level to mean 0 and standard deviation to account for variation due to level design.

The results of the regression analyses can be found in Table 2. When looking at the PRMs for base width and symmetry levels, there is a significant relationship between the PRM and success on the level, which is what would be expected if those levels are appropriately incentivizing their target principle. The relationship for the center of mass PRM however was not found to be significant. This would mean that, counter to what the target principles suggest, players who build with lower centers of mass are not any more likely to succeed on levels that target the center of mass principle than players who build towers with higher centers of mass. This could not have been the *RumbleBlocks* designers' intent.

Group	**Coefficient**	***B***	***SE B***	***β***	***p***
Symmetry Levels (df = 1788)	(Intercept)	1.044	.061	17.250	< .001
	Base Width	.449	.054	8.368	< .001
	Center of Mass Height	.418	.089	4.700	< .001
	Symmetry Angle	-.205	.069	-2.969	.003
Center of Mass Levels (df = 2107)	(Intercept)	1.379	.063	22.042	< .001
	Base Width	.022	.066	.326	.745
	Center of Mass Height	-.046	.047	-.975	.330
	Symmetry Angle	-.165	.043	-3.803	< .001
Wide Base Levels (df = 1997)	(Intercept)	1.729	.074	23.463	< .001
	Base Width	.221	.069	3.229	.001
	Center of Mass Height	-.113	.097	-1.164	.245
	Symmetry Angle	.011	.078	.135	.893

Table 2 The results of a logistic regression of success of solution on principle relevant metrics for levels targeting each of the three principles. Note that for the Center of Mass and Symmetry Angle principles a lower coefficient estimate (B) is better.

Clustering Replayed Solutions

The logistic regression analysis agrees with the previous findings that players did not seem to be improving at the center of mass principle likely because they are not being given consistent feedback in terms of the principle. The next question that arises out of this is: if players are not getting consistent feedback on the center of mass principle, what is the game doing in these situations? Answering this question requires the ability to look at players' solutions in much closer detail than is provided by distilled metrics. The Replay Analysis Approach can provide the ability to look at the particular structure of solutions that the players create, and how the game reacted in those situations.

Attacking this question required wrestling with the issue that neither the designers of *RumbleBlocks* not I had a sense of how many ways there were to solve any of the levels in the game. Further among the various ways to solve each level, which ones are the young players of the games target demographic more likely to employ. Resolving these issues would enable looking at the kinds of feedback the game provides to common solutions and whether that feedback appears to make sense given the principles of the game. Rather than comb through myriad solutions to each level myself I made use of the RAE to reframe the log data I has as a detailed representation of a the solution space of each game level that was amenable to machine learning methods.

At a high level, this analysis involves capturing a picture of the space of solutions that students use to overcome in-game challenges. This solution space is generated by clustering the individual solutions created by students into a subset of representative solutions (because there are too many to view individually). Once a collection of representative solutions is gathered, each one is evaluated in terms of its PRMs. Finally, the PRM is compared to the positive or negative feedback designation that the game's mechanics assigned to the majority of individual solutions embodied by each of the representative solutions. This allows me to analyze the general principled-ness of a group of student solutions and how the game treated them as a means of evaluating alignment.

To perform these analyses, I first had to convert the solutions into a representation that captured their essential structural features. For example, many students might build a tower that uses an arch pattern, whereas others might build an inverted "T" shape. I wanted a representation that captured elements of these basic structural patterns. To build this new representation, I first instrumented the RAE to produce representations of student towers aligned to a two dimensional grid. This process makes use of a number of capabilities exposed by the game engine in the RAE, such as the physical properties of blocks (e.g., their collider dimensions) that would not have been possible to recover without a live engine.

Next, I employed two-dimensional grammar induction to learn a set of patterns that could be used to describe all of the student solutions in the entire dataset. A two-dimensional grammar consists of three components:

1. Terminal Symbols, which represent the blocks, spaceship, and empty space (in this context);

2. Non-terminal Symbols, which represent structural patterns consisting of more than a single block

3. Rules, which map non-terminal symbols to pairs of other non-terminal symbols oriented in a certain direction (horizontal or vertical), or non-terminals to terminal symbols (a unary relationship).

To help illustrate the concept, Figure 4 shows an example grammar (u, h, and v represent unary, horizontal, and vertical respectively), and the parses for two simple towers.

To learn a grammar, I employed an algorithm developed with colleagues called conceptual feature extraction (CFE) [24]. CFE first generates an exhaustive set of rules that describe every possible way to parse all of the solutions in the set. Next it computers all of the possible parses of each solution. Given the parses for each solution, it creates a vector for each solution which contains a 1 for every non-terminal present in the solution and a 0 for every non-terminal not present in the solution. This resulting feature vectors contains information about all of the structural patterns present in each solution. These patterns may correspond to individual blocks, pairs of blocks, other combinations of blocks, or even whole towers.

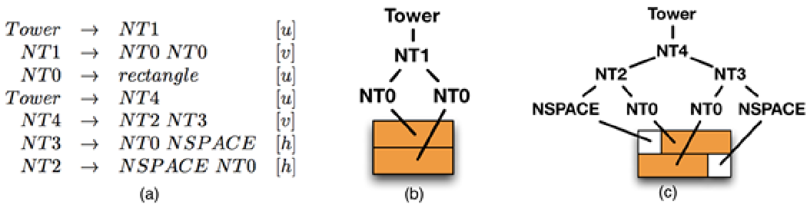

Figure 4. A simple two-dimensional grammar (a) and the parse trees generated by applying this grammar to two solutions (b and c).

For each level, I clustered the featurized solutions using g-means, a variant of the common k-means clustering algorithm that chooses a value for k optimizing for a Gaussian distribution within clusters [23]. This produced a set of different groups for each level, where each group represents solutions that share structural similarity. The resultant clusters can be summarized as representative solutions that embody the general trend within the cluster. For each cluster, I created a representative solution by averaging the PRM scores within the cluster and then assigning the success label that the game assigned to the majority of solutions within the cluster. This gives me the ability to think about common patterns of solutions through a single representative solution rather than individual solutions.

To get a sense of the general trends in how the game treats different solutions to a particular level I created representative plots of the clusters like the one show in Figure 5. These plots show each representative solution plotted with its frequency of use (as a percentage of all observed solutions for that level) along the x-axis and its relative principled-ness, in terms of a normalized PRM score, along the y-axis. The squares represent solutions that are mostly successful where the filled diamonds show solutions that are mostly unsuccessful. When examining these plots, two different patterns are primarily of interest: principled failures and unprincipled successes, which both represent the game generally giving feedback contrary to the target principle for the particular level. These cases can shed light on potential problems with a game's alignment. The analysis of *RumbleBlocks* highlighted several cases, but for the purposes of this paper, I will discuss two in detail.

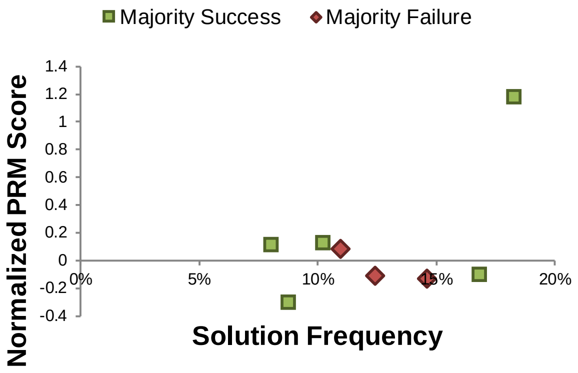

Figure 5. A plot of representative solutions' PRM score versus frequency.

The first problem level is Symmetry_7 meaning that the level is meant to be targeting the concept that a symmetrical structure is more stable. In this example, there are two highly frequent solutions: the two points farther to the right in Figure 6. One is mostly successful and the other is mostly unsuccessful, but they do not differ strongly in their PRM scores. When I examine screenshots of student solutions to this level, I saw the situation in Figure 6, where the tower on the left (an inverted T-shape) comes from the majority failure solution while the tower on the right (an arch shape) comes from the majority success solution. While it is clear from the examples that the left tower should fail (as it did frequently), it is important to remember that this level is designed to target the symmetry principle, which says a symmetrical structure should be more stable. Both solutions seen in these representative solutions are generally symmetrical, but one was considered a failure while the other is considered a success. This case represents *RumbleBlocks* giving inconsistent feedback to players about the targeted symmetry principle. An alternative interpretation is that this level should not be labeled as targeting the symmetry principle give that its two most frequently used solutions both embody a reasonable level of symmetry.

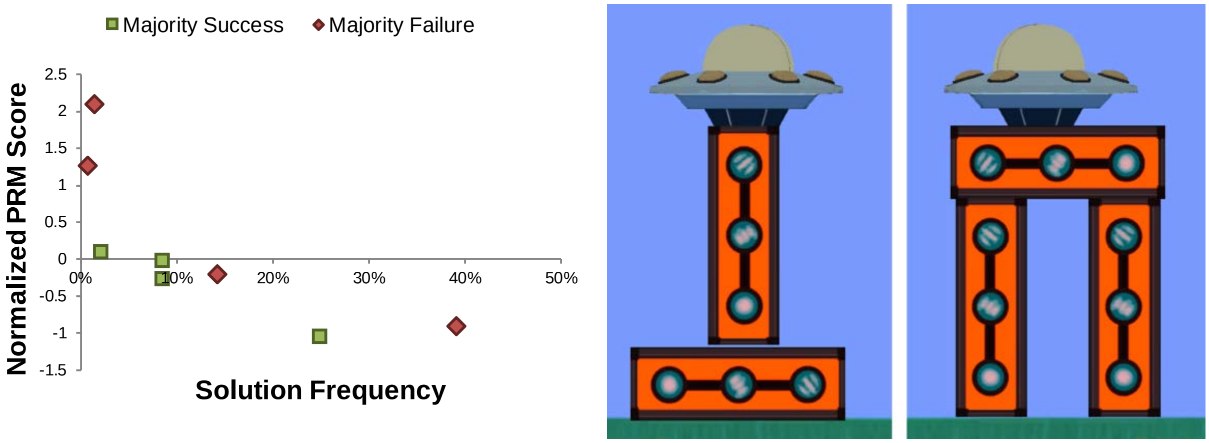

Figure 6. A plot of solution frequency (as a percentage) vs. PRM score for all of the clusters on the Symmetry_07 level of *RumbleBlocks* and Two example student solutions to the Symmetry_07 level. The solution on the left comes from a majority unsuccessful cluster

Another anomalous example is shown in Figure 7, which shows a plot of the different solutions to the level CenterOfMass_10_PP. This level was used as part of an in-game pre-post design meaning it omits the energy dot mechanic and is based on a level designed to target the low center of mass principle. It is harder to attribute patterns in the chart to elements of level design because it lacks the energy dot mechanic and thus does not restrict players as much as normal game levels; however, an interesting pattern develops nonetheless. The distribution of how many students created each solution on this level is more evenly spread out, but among groups of solutions that are all relatively equal in PRM score, we see two solutions, which are majority failure rather than success.

Visually inspecting the solutions students created to this level, we see the pattern that arises in Figure 7, where an example from one of the nearby successful solutions is shown on the top and an example from each of the unsuccessful solutions is shown on the bottom. The salient feature to note among the unsuccessful solutions is the presence of the alien's spaceship on top of a single square block. This points to a nuance in the game's mechanics, where an additional constraint on game success is whether the spaceship falls off the tower during the earthquake and not just that the tower continues to stand up. This opens the possibility, illustrated by the lower right quadrant of the matrix in Figure 1, that a student could build a perfectly reasonable tower that is judged as unsuccessful by the game because the spaceship falls off. This is an example of the more nuanced kind of alignment failure where a task requires an extra piece of unexpected knowledge to complete successfully. While the spaceship should also be subject to the same stability principles this level seemed to suggest it was a major determining factor in success.

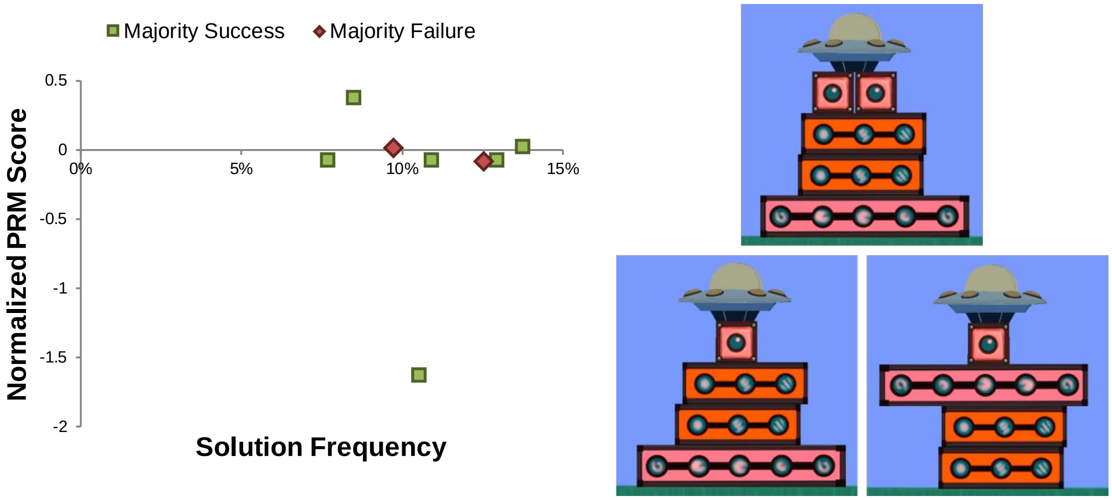

Figure 7. A plot of solution frequency (as a percentage) vs. PRM score for all of the clusters on the CenterOfMass_10_PP level of *RumbleBlocks* and examples of student solutions on the CenterOfMass_10_PP level. The solution at the top comes from one of the majority unsuccessful clusters.

The patterns I observed in our analysis of the Symmetry_7, and CenterOfMass_10_PP data were present in a number of other levels as well. As a pattern of salient features emerged, I wanted to see if there was further; evidence in the structural data to support the conclusion the *RumbleBlocks* might have an issue in certain structural features exerting too much influence on the success of a tower. To do this I used the structural features generated through the CFE process and used a χ^2 analysis to identify which structural features present in student solutions were more predictive of success. I performed a χ^2 test of each of the 6,010 symbols against solution success to see which patterns were most strongly related to success of a tower. Because this constitutes a large number of statistical tests and increases the possibility of Type I error, I applied a Bonferroni correction to the results to account for the number of statistical tests. This correction divides the usual cutoff for considering a result to be significant (.05) by the number of tests performed (6,010) and uses the result (8.32e-6) as the new bar for significance.

Overall, 19 grammar symbols were significantly related to success. However, representing the grammar symbols with grounded game objects resulted in only five distinct structures[3]. Once I had a selection of significant features I performed a logistic regression of those features against solution success to understand the direction of the relationship (i.e., does each feature predict

[3] The CFE process uses a set of recursive rules to represent space allowing it to create multiple symbols that would look visually the same but differ in how the negative space around the tower is handled. More details on this limitation can be found in the paper on CFE [24].

success or failure). I only present the directional results of this regression and not the actual coefficients. The results of this process are visually rendered in Figure 8.

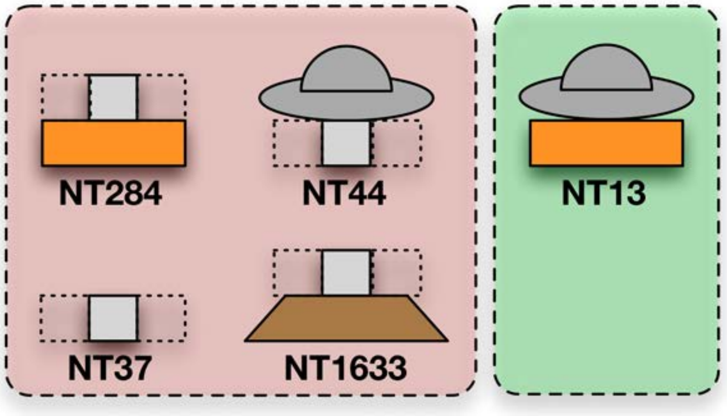

Figure 8. Rendered results of a χ^2 analysis of structural features in *RumbleBlocks* which predict the success of a tower in the earthquake. Student solutions that contained the features in the failure region to the left were more likely to be unsuccessful in the earthquake while solutions that contained the feature in the success region to the right were more likely to be successful.

Our original question asked: if players are not getting consistent feedback on the center of mass principle, what is the game doing in these situations? The pattern that arises from the χ^2 analysis demonstrates that the game tends to focus more on points of weakness with a lone square block without supports. This would mean that the game is generally punishing more nuanced sub structural faults of towers. The principles targeted by *RumbleBlocks* are generally meant to apply to whole structures and so do not necessarily accounted for these kinds of smaller structural problems. This would suggest some misalignment between the stated goals of the game and its feedback mechanisms as instantiated in the earthquake mechanic.

Further, the analysis demonstrates a strong difference between the width of the platform the spaceship is placed on and the eventual success of the tower (i.e., placing the spaceship on a single block is more likely to lead to failure and placing it on a wide block is more likely to lead to success). This points to the importance of the spaceship remaining on top of the tower as a secondary success criterion. While the designers were aware that the spaceship served such a purpose in the design, they probably did not think it would be such a strong determining factor to the potential detriment of other learning goals. When pursuing iteration, the designers of *RumbleBlocks* will have to consider if this result represents a flaw in the game's mechanics, which contradicts the message, or an opportunity to teach a nuanced aspect of stability and balance with some new feedback.

Design Recommendations

Examining the alignment results, I saw that there are cases where adherence to the target principle of a level does not translate into success for players. This would mean that the game is not providing feedback that will actually help students attend to their errors by thinking about the domain principles. If we look at the case of Symmetry_7 we see a pattern where successful and unsuccessful solutions are essentially the same in terms of PRM score. What is interesting about this particular example is that the unsuccessful solution is equivalent or better in terms of the score on both of the other two principles, wide base and low center of mass. When combined with the evidence form the χ^2 analysis that micro faults in a towers design are more likely to be a cause of failure than broad features this highlights the possibility that game's feedback structure is more complex than it should be to foster its desired principles.

One way of addressing this problem of micro faults is to explore a version of the game were blocks act together as a single rigid body, making it so feedback is likely to be better aligned with the broad principles that the game is meant to target. Using a design that involved connected structures was considered in the preliminary design phases of *RumbleBlocks*[4], but initial prototype testing indicated that players found the disconnected structure to be more fun to play with. The designers also thought that having a disconnected block mechanic would allow for more interesting dynamics in the design. However, if we introduce a mechanic that allows players to glue blocks together so that the blocks act more like the connected structures, the level may be better modeled by the principle-relevant metrics. As previously noted by the designers, this mechanical change would cause fully connected structures to react less to the in-game earthquake. A number of mechanical options could be considered to account for the drop in dynamics, such as adding in negative energy dots or more interesting terrain features (such as ravines between the alien and their ship).

Another possibility for changing the mechanics could be to remove the spaceship as a factor entirely. This solution would address the alignment problems highlighted in the CenterOfMass_10_PP example as well as the χ^2 analysis by removing the secondary success criteria of keeping the ship on the tower. Removing the spaceship mechanic could also preserve the interesting dynamics of having disconnected structures, which were valued by the designers; however, it adds other mechanical difficulties, such as removing the mechanic for how players submit a solution – placing the ship on top of the tower – as well as damaging the narrative aesthetic of the game by no longer having the player trying to return the ship to the alien.

Each of these analyses has contributed to an evolving understanding of the state of *RumbleBlocks* while also serving as a case study in the use of Replay Analysis to drive the alignment evaluation of an educational game. With the ability to reframe the recorded game data, I would not have been able to pivot between looking at trends in metrics to trends in structural patterns. Being able to follow the evolving thread of analysis allowed for the recommendation of several concrete design alternatives. What remains now is to see whether though new alternatives fix the problems that I have described.

PROPOSED WORK

My body of prior work has demonstrated the power of Replay Analysis to explore the alignment of an educational game using observed player behavior. The approach enabled several different kinds of analysis of *RumbleBlocks* and highlighted several potential directions for future designs. These conclusions alone would be enough to suggest several new iterations and subsequent formative evaluations. I propose to perform a close-the-loop [15] study that can prove that not only was Replay Analysis able to yield alignment focused design recommendations, but acting

[4] http://www.etc.cmu.edu/projects/illuminate/?page_id=221

on those recommendations will result in a better game. In addition to this study also propose to explore the potential for Replay Analysis further by demonstrating the ability to use Projective Replay Analysis to better understand the potential future directions a game design could take without having to gather new data.

While the paradigm of Projective Replay Analysis has potential benefit for guiding closed-loop design iteration, there is one fundamental concern with the approach; namely, if the game is different then players will play it differently. This would seem to imply that the instant a game's mechanics are changed, any replay recordings taken from the old version of the game instantly become invalid. I do not dispute this notion in general; however, I would argue that any change in player behavior from mechanical change is not absolute but rather a matter of degrees and dependent on the nature of the particular change made. It is this relationship between types of game design change and changes in player behavior that I plan to explore in my proposed work.

To frame this question of the effect of changing a game on the behavior of players I want to return to notion of alignment as a property of a game's solution space. Any change to a games mechanical structure will have one of five possible effects on its solutions space:

1. The new solution space could remain the same. This would mean that the affordances available to players to explore the solution space are no different than they were before. While space remains the same, the mechanical change could cause existing solutions to receive different feedback leading to players to explore portions of the solutions space more or less often than before.

2. The new solution space could be a subset of what it was before. In this case previously reachable solutions are no longer possible. This would be desirable in cases where specific solutions or mechanical uses were found to foster misconceptions.

3. The new solution space could be a superset of what it was before. This would mean that the game now affords more variation to players than it originally did. In such cases it is necessary to evaluate the alignment of this new territory of the solution space to ensure educational goals are still being met.

4. The new solution space could be a shift of the original space. That is, the space shrinks in some areas and grows in others. I would expect this form of solution space change to be more common than the subset and superset ones as it is rare that a mechanical change would be so localized.

5. The new solutions space could be a disjoint set from the original. This is the case where a mechanical change was so fundamental that it created an entirely new game.

To facilitate my exploration of the relationship between game design changes and player behavior I propose to implement two forms of Projective Replay Analysis. The first form, which I call Literal Replay (Literal PRA from now on), takes a players replay trace and reenacts it literally into the new game space as it was recorded. This first approach has the benefit of being cheap, as it is essentially how the RAE currently works, but it is also maximally confounded. If a design change renders a previously observed action impossible then a literal replay paradigm would simply fail. Alternatively, if a design change broadened the solution space by opened up strategic possibilities that were impossible before, then literal replays would not be able to handle the additional nuance.

To combat the limitations of Literal PRA I propose the second form of projective replay that I plan to explore: Flexible Replay (Flexible PRA from now on). In a Flexible PRA system, rather than take players' actions and re-enact them directly, the system instead uses them as training data for a player model. The player model simulates human behavior by approximating humans' knowledge acquisition and problem-solving processes [42]. The model is designed to acquire skill knowledge using demonstrations from an expert, but in this case the "expert" takes the form of individual replay traces of prior players. This in essence allows for the creation of virtual play testers that model the skills observed in prior playtesting populations.

The use of AI players models to test game designs has be suggested before; however, these earlier efforts have generally been in the style of computational caricatures [61] in that they are computational models designed to exaggerate some aspect of human behavior in order to make an argument. Holmgård *et al.* demonstrated the use of AI models for playtesting a puzzle game and found that there was little difference between models based on human playtesters data and the models designed by the designer [27]. The models in this work were based on a relatively simple game domain (navigating a maze where input is restricted to the four cardinal directions) and did not model the incremental learning processes of players. Further work has also been done to virtually playtest sketch-based and formal representations of games [62]. While these formalized approaches have shown much promise they require a designer to form their ideas in a representation that may not be readily natural to them.

To explore the potential of these two forms of Projective Replay Analysis I plan to undertake three studies. The overall question of these studies will be to validate the projective approach through a series of progressively more stringent tests. The first study will seek to demonstrate that the mechanical variations suggested by my prior replay analyses result in measurably better alignment using a Literal PRA analysis paradigm. The second study will probe the limitations of Literal PRA by comparing it to a Flexible PRA of the same game variations. Finally, the third study will look at fully closing the loop on design variations suggested in previous studies by performing a new formative evaluation with real students in a classroom setting. The results of this final study can serve as a ground truth comparison for the virtual playtests to establish the overall validity of virtual playtesting. In the following subsections I describe each study in more detail as well as an expected timeline.

Study 1

In the first study, I seek to demonstrate the initial feasibility of Literal PRA as well as answer the question of whether the recommendations that arose from prior alignment analyses actually lead to a better-aligned game. To answer this question, I plan to implement the design recommendations from the prior analyses of *RumbleBlocks* as two alternate variations of the game. I will then apply a paradigm of literal replay analysis using the same log data from my prior work.

First, I will modify *RumbleBlocks* to produce two variations. In the first variation, I will introduce a mechanic where placed blocks stick together as if they were glued together (with 'gluing' happening as soon as the player places a block). This variation is a response to the potential alignment issue with *RumbleBlocks* where success in the game seems overly dependent on micro faults in a tower's design rather than broader structural patterns as suggested by the game's target principles. The second variation is a response to the χ^2 analysis, which highlighted that the spaceship remaining on top of the tower is too important as a secondary success criterion in the success of the tower. To remove this issue, I will alter the mechanics of the game so that while the earthquake is happening, the spaceship simply floats above the tower, and is not affected by the shaking of the tower. This will require an alteration to the success mechanics of the game, changing the goal so that rather than keep the spaceship on top of the tower players will instead try to keep the blocks of the tower from shifting too much in the earthquake. This can be implemented by staying that the energy dots must remain activated by the tower at the end of the earthquake (which is not currently always the case).

Once new game variations are implemented, I will perform a Literal PRA with the log data from the original formative evaluation(s) of *RumbleBlocks* on both variations of the game. Because the game will now behave differently, player solutions are likely to be assigned different feedback than they were in the original version, causing a shift in alignment due to changing the game's feedback function. I predict that the alignment that results from these new versions of the game will be better (i.e., when re-running regression analyses all metrics will have a stronger and better-directed statistical relationship with feedback) than the current version of the game. Specifically, I think that the first variation (gluing blocks together) will result in better-directed metrics, and the second variation (changing the spaceship) will reduce the strength of the secondary success criterion.

Study 2

As I discussed previously, a strong caveat to the use of the Literal PRA paradigm is the possibility that game variation changes alignment by fundamentally altering the solution space rather than merely relabeling existing solutions. The main research question behind Study 2 is to investigate the extent to which this is a concern using Flexible PRA.

To explore this question, instead of replaying student's literal logs in the newly designed game variations, I will use them as training data for a player model. This player model will receive students' actions as demonstrations, which they will use to learn skill concepts for the game. The player models will then be used to play the game variations in order to generate new game actions. The solutions created by these actions will then be evaluated as if they had been generated by the students themselves. These solutions will be similar in character to the ones the students originally created, but they will have a chance to be distinct and responsive to the new dynamics created by the variation in game mechanics.

To analyze Flexible PRA, I will apply the method used in Study 1, comparing the metrics generated by the models' solutions to feedback provided by the game, and looking at the common occurrences of structural patterns. Additionally, I will compare the amount of solution space overlap between the original students and the flexible player models. At this point it is difficult to anticipate how the conclusion with regard to alignment will change using Flexible PRA rather than Literal PRA. Literal PRA is likely to more strongly conclude that the proposed design variations show better alignment between domain principles and game feedback because those variations are based directly off of the play experience recorded in the literal logs. However, Literal PRA would be confounded by not taking into account how players will react differently to new game dynamics. This would suggest that even if Flexible PRA reaches a weaker conclusion with regard to alignment than Literal PRA, that conclusion is likely to be more valid by better accounting for player variation. Definitively answering the question of which approach is more valid will require the results from study 3.

Study 3

The ultimate goal of the educational game design process is to create a game that new players can learn from. Therefore, the only true way to know if a game has become better is to test it with a new population of learners. In my third study, I plan to perform a closed-loop evaluation [15] of the game design modifications of *RumbleBlocks* implemented in Study 1. This will allow me to answer the two main research questions of my work, as well as validate the answers to these questions found in Studies 1 and 2.

The ideal context of this study would be to recruit a new classroom population of students. However, previous analysis of the game [12] and informal observations suggest that, while the game has been evaluated with a broad age range demographic (K-3), it is generally better suited to the higher-end age range of the demographic. In order to simplify the recruitment process, I will only focus on finding populations within the Grades 2-3 range.

The structure of the study will follow a similar design to the original formative evaluation of *RumbleBlocks*. Students will first take the pretest designed for the original formative evaluation with a few modifications based on previous psychometric evaluations [12]. Then students will be given two class periods worth of time to play the game with the modifications introduced in Study 1. Finally students will take the posttest (as used in the previous evaluation). The pretest and posttest have counterbalanced items, which will be used to evaluate students' general understanding of balance and stability concepts before and after playing the game.

Once the study is complete, I will perform the same analysis as was used in the past two studies to determine whether the game modifications led to an improvement in alignment. Further, I will compare the overlap in solution spaces between the new human data, the original literal replay data, and the flexible replay data to see how faithfully the virtual versions simulate the new player population.

The ideal outcome of these three studies would be a demonstration that the new iterations of the game show better alignment between domain principles and instructional feedback than the original game. If such a correspondence were found, it would provide evidence that the more lightweight virtual playtesting paradigms used in projective replay analysis can be used as a suitable proxy for real human playtesting.

Schedule

The schedule that I propose to follow for this work is as follows:

Event	Date
Propose	May
Development Work, Studies 1 and 2	Summer
Study 3	Fall
Defend	Spring 2017

REFERENCES

1. Vincent Aleven, Bruce Mclaren, Ido Roll, and Kenneth Koedinger. 2006. Toward meta-cognitive tutoring: A model of help seeking with a cognitive tutor. *International Journal of Artificial Intelligence in Education* 16, 2: 101–128. http://doi.org/10.1.1.121.9138

2. Vincent Aleven, Eben Myers, Matthew Easterday, and Amy Ogan. 2010. Toward a Framework for the Analysis and Design of Educational Games. *Proc. DiGITEL 2010*, IEEE, 69–76. http://doi.org/10.1109/DIGITEL.2010.55

3. Vincent Aleven. 2010. Rule-Based Cognitive Modeling for Intelligent Tutoring Systems. In *Advances in Intelligent tutoring Systems*. 33–62.

4. Mike Ambinder. 2014. Making the Best of Imperfect Data: Reflections on an Ideal World. *Keynote to the 1st Annual Symposium on Computer-Human Interaction in Play - CHI PLAY 2014*.

5. Alan Amory. 2006. Game object model version II: a theoretical framework for educational game development. *Educational Technology Research and Development* 55, 1: 51–77. http://doi.org/10.1007/s11423-006-9001-x

6. Leonard A Annetta. 2010. The “I’s” have it: A framework for serious educational game design. *Review of General Psychology* 14, 2: 105–112. http://doi.org/10.1037/a0018985

7. Ryan S J Baker, Albert T Corbett, and Angela Z Wagner. 2006. Human Classification of Low-Fidelity Replays of Student Actions. *Proceedings of the Educational Data Mining Workshop at the 8th International Conference on Intelligent Tutoring Systems*, 2002: 29–36.

8. Ryan S J D Baker and Kalina Yacef. 2009. The State of Educational Data Mining in 2009 : A Review and Future Visions. *Journal of Educational Data Mining* 1, 1: 3–16. http://doi.org/http://doi.ieeecomputersociety.org/10.1109/ASE.2003.1240314

9. Sasha Barab, Michael Thomas, Tyler Dodge, Robert Carteaux, and Hakan Tuzun. 2005. Making learning fun: Quest Atlantis, a game without guns. *Educational Technology Research and Development 53*, 86–107. http://doi.org/10.1007/BF02504859

10. Marion Boberg, Evangelos Karapanos, Jussi Holopainen, and Andrés Lucero. 2015. PLEXQ: Towards a Playful Experiences Questionnaire. *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY ’15*, 381–391. http://doi.org/10.1145/2793107.2793124

11. Richard Buchanan. 1996. Wicked Problems in Design Thinking. In *The Idea of Design*, Victor Margolin and Richard Buchanan (eds.). MIT Press, Cambridge, MA, 3–20.

12. Catherine C Chase, Erik Harpstead, and Vincent Aleven. Inciting transfer of learning beyond the game: adapting contrast-based instruction for educational games. *In preperation*.

13. Michael G Christel, Scott M Stevens, Bryan S Maher, et al. 2012. RumbleBlocks: Teaching science concepts to young children through a unity game. *Proc. CGAMES 2012*, IEEE, 162–166. http://doi.org/10.1109/CGames.2012.6314570

14. Re Clark. 1994. Media will never influence learning. *Educational technology research*

and development 42, 21–29. http://doi.org/10.1007/BF02299088

15. Doug Clow. 2012. The learning analytics cycle: Closing the loop effectively. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge - LAK '12*, ACM Press, 134. http://doi.org/10.1145/2330601.2330636

16. S Alan Cohen. 1987. Instructional Alignment: Searching for a Magic Bullet. *Educational Researcher* 16, 8: 16–20.

17. Girlie C Delacruz, Gregory K W K Chung, and Eva L Baker. 2010. *Validity Evidence for Games as Assessment Environments (CRESST Report 773)*. Los Angeles, CA.

18. Anders Drachen and Alessandro Canossa. 2009. Towards gameplay analysis via gameplay metrics. *Proc. MindTrek 2009*, ACM Press, 202. http://doi.org/10.1145/1621841.1621878

19. Anders Drachen, Magy Seif El-Nasr, and Alessandro Canossa. 2013. Game Analytics - The Basics. In *Game Analytics*, Magy Seif El-Nasr, Anders Drachen and Alessandro Canossa (eds.). Springer London, London, 13–40. http://doi.org/10.1007/978-1-4471-4769-5

20. James Paul Gee. 2003. *What video games have to teach us about learning and literacy*. Palgrave Macmillan, New York.

21. Chaim Gingold and Chris Hecker. 2006. Advanced Prototyping. *Game Developers Conference*.

22. M. P. Jacob Habgood and Shaaron E. Ainsworth. 2011. Motivating Children to Learn Effectively: Exploring the Value of Intrinsic Integration in Educational Games. *Journal of the Learning Sciences* 20, 2: 169–206. http://doi.org/10.1080/10508406.2010.508029

23. Greg Hamerly and Charles Elkan. 2004. Learning the k in k-means. In *Advances in Neural Information Processing Systems 16: Proceedings of the 2003 conference*, Sebastian Thrun, Lawrence K. Saul and Bernhard Schölkopf (eds.). MIT Press, 281–288.

24. Erik Harpstead, Christopher J Maclellan, Kenneth R Koedinger, Vincent Aleven, Steven P Dow, and Brad A Myers. 2013. Investigating the Solution Space of an Open-Ended Educational Game Using Conceptual Feature Extraction. *Proc. EDM 2013*, 51–58.

25. Erik Harpstead, Christopher J. MacLellan, Vincent Aleven, and Brad A Myers. 2015. Replay Analysis in Open-Ended Educational Games. In *Serious Games Analytics*, Christian Sebastian Loh, Yanyan Sheng and Dirk Ifenthaler (eds.). Springer International Publishing, Cham, 381–399. http://doi.org/10.1007/978-3-319-05834-4_17

26. Erik Harpstead, Brad A Myers, and Vincent Aleven. 2013. In search of learning: facilitating data analysis in educational games. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, ACM Press, 79. http://doi.org/10.1145/2470654.2470667

27. Christoffer Holmgård, Antonios Liapis, Julian Togelius, and Georgios N. Yannakakis. 2015. Evolving Models of Player Decision Making: Personas versus Clones. *Entertainment Computing*. http://doi.org/10.1016/j.entcom.2015.09.002

28. Robin Hunicke, Marc Leblanc, and Robert Zubek. 2004. MDA : A Formal Approach to Game Design and Game Research. *Proc. of the AAAI Workshop on Challenges in Game*

AI, 1–5.

29. Kenneth R Koedinger, Ryan S J d Baker, Kyle Cunningham, Alida Skogsholm, Brett Leber, and John Stamper. 2010. A Data Repository for the EDM community: The PSLC DataShop. In *Handbook of Educational Data Mining*, Cristobal Romero, Sebastian Ventura, Mykola Pechenizkiy and Ryan S.J.d. Baker (eds.). 43–55.

30. Kenneth R Koedinger, Albert T Corbett, and Charles Perfetti. 2012. The Knowledge-Learning-Instruction Framework: Bridging the Science-Practice Chasm to Enhance Robust Student Learning. *Cognitive Science* 36, 5: 757–98. http://doi.org/10.1111/j.1551-6709.2012.01245.x

31. Kenneth R Koedinger and Mitchell J. Nathan. 2004. The Real Story Behind Story Problems: Effects of Representations on Quantitative Reasoning. *The Journal of the Learning Sciences* 13, 2: 129–164.

32. Ron Kohavi, Alex Deng, Brian Frasca, Roger Longbotham, Toby Walker, and Ya Xu. 2012. Trustworthy Online Controlled Experiments: Five Puzzling Outcomes Explained. *Proc. KDD 2012,* ACM Press, 786–794.

33. Robert B. Kozma. 1994. Will media influence learning? Reframing the debate. *Educational Technology Research and Development* 42, 2: 7–19. http://doi.org/10.1007/BF02299087

34. Ld Landau. 2014. Pyschometrica Considerations in Game-Based Assessments. *White Paper Released by Glasslab*: 160.

35. Dylan Lederle-ensign and Noah Wardrip-fruin. 2016. What is Strafe Jumping? *ToDiGRA* 2, 2: 123–148.

36. Jacqueline P Leighton and Rebecca J Gokiert. 2005. The Cognitive Effects of Test Item Features: Informing Item Generation by Identifying Construct Irrelevant Variance. *Proc. NCME 2005*, 1–26.

37. Christian Sebastian Loh, Yanyan Sheng, and Dirk Ifenthaler (eds.). 2015. *Serious Games Analytics*. Springer International Publishing, Cham. http://doi.org/10.1007/978-3-319-05834-4

38. Christian Sebastian Loh and Yanyan Sheng. 2015. Measuring Expert Performance for Serious Games Analytics: From Data to Insights. In *Serious Games Analytics*. Springer International Publishing, Cham, 101–134. http://doi.org/10.1007/978-3-319-05834-4_5

39. Yanjin Long and Vincent Aleven. Educational game and intellignet tutoring system: A classroom study and comparative design analysis. under review.

40. Collin Lynch, Kevin D Ashley, Niels Pinkwart, and Vincent Aleven. 2009. Concepts , Structures , and Goals : Redefining Ill-Definedness. *International Journal of Artificial Intelligence in Education* 19: 253–266.

41. Christopher J Maclellan, Erik Harpstead, Vincent Aleven, and Kenneth R Koedinger. 2015. TRESTLE: Incremental Learning in Structured Domains using Partial Matching and Categorization. *Proceedings of the 3rd Annual Conference on Advances in Cognitive Systems - ACS 2015,* Cognitive Systems Foundation, 28–31.

42. Christopher J Maclellan, Erik Harpstead, Rony Patel, and Kenneth R Koedinger. The Apprentice Learner Architecture: Closing the loop between learning theory and educational data. *In submission*.

43. Andrea Martone and Stephen G Sireci. 2009. Evaluating Alignment Between Curriculum, Assessment, and Instruction. *Review of Educational Research* 79, 4: 1332–1361. http://doi.org/10.3102/0034654309341375

44. Pejman Mirza-Babaei, Lennart E Nacke, John Gregory, Nick Collins, and Geraldine Fitzpatrick. 2013. How does it play better? exploring user testing and biometric storyboards in games user research. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, ACM Press, 1499. http://doi.org/10.1145/2470654.2466200

45. Robert J Mislevy, Russell G Almond, and Janice F Lukas. 2003. A Brief Introduction to Evidence-centered Design. July.

46. Lennart Nacke and Anders Drachen. 2011. Towards a Framework of Player Experience Research. *Proc. EPEX 2011*.

47. Lennart E Nacke and Craig A Lindley. 2008. Flow and Immersion in First-person Shooters: Measuring the Player's Gameplay Experience. *Proceedings of the 2008 Conference on Future Play: Research, Play, Share*: 81–88. http://doi.org/10.1145/1496984.1496998

48. National Research Council. 2012. *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. The National Academies Press.

49. Alan S Neal and Roger M Simons. 1984. Playback: A method for evaluating the usability of software and its documentation. *IBM Systems Journal* 23, 1: 82–96. http://doi.org/10.1147/sj.231.0082

50. A. Imran Nordin, Alena Denisova, and Paul Cairns. 2014. Too Many Questionnaires: Measuring Player Experience Whilst Playing Digital Games. *Seventh York Doctoral Symposium on Computer Science & Electronics*, October: 69–75.

51. Jaclyn Ocumpaugh, Ryan S J d Baker, and Ma Mercedes T Rodrigo. 2012. Baker-Rodrigo Observation Method Protocol (BROMP) 1.0 Training Manual version 1.0.

52. V Elizabeth Owen, Dennis Ramirez, Allison Salmon, and Richard Halverson. 2014. Capturing Learner Trajectories in Educational Games through ADAGE (Assessment Data Aggregator for Game Environments): A Click-Stream Data Framework for Assessment of Learning in Play. *American Educational Research Association Annual Meeting*: 1–7.

53. Elizabeth Rowe, Jodi Asbell-Clarke, and Ryan S. Baker. 2015. Serious Games Analytics to Measure Implicit Science Learning. In *Serious Games Analytics*. Springer International Publishing, Cham, 343–360. http://doi.org/10.1007/978-3-319-05834-4_15

54. Elizabeth Rowe, Ryan S J d Baker, Jodi Asbell-clarke, Emily Kasman, and William J Hawkins. 2014. Building Automated Detectors of Gameplay Strategies to Measure Implicit Science Learning. *Proceedings of the International Conference on Educational Data Mining - EDM '14*, 337–338.

55. André a. Rupp, Matthew Gushta, Robert J. Mislevy, and David Williamson Shaffer. 2010.

Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *The Journal of Technology Learning and Assessment* 8, 4: 3–41.

56. Katie Salen and Eric Zimmerman. 2004. *Rules of Play*. MIT Press.

57. Jesse Schell. 2008. *The Art of Game Design: A Book of Lenses*. Morgan Kaufmann, Burlington, MA.

58. Donald A. Schon. 1982. *The Reflective Practitioner: How Professionals Think in Action*. Basic Books, New York, New York, USA.

59. Magy Seif El-Nasr, Anders Drachen, and Alessandro Canossa (eds.). 2013. *Game Analytics: Maximizing the Value of Player Data*. Springer.

60. Valerie J Shute and Lubin Wang. 2015. Measuring Problem Solving Skills in Portal 2. *E-learning Eystems, Environments and Approaches: Theory and implementation*: 11–25. http://doi.org/10.1007/978-3-319-05825-2

61. Adam M Smith and Michael Mateas. 2011. Computational caricatures: Probing the game design process with ai. *Artificial Intelligence in the Game Design Process - Papers from the 2011 AIIDE Workshop*: 19–24.

62. Adam M Smith, Mark J Nelson, and Michael Mateas. 2009. Computational Support for Play Testing Game Sketches. *Proceedings of the Fifth Artificial Intelligence for Interactive Digital Entertainment Conference - AIIDE 2009*, 167–172.

63. Anders Tychsen. 2008. Crafting User Experience via Game Metrics Analysis. *Proc. NordiCHI 2008*, 20–22.

64. Grant Wiggins and Jay McTighe. 2005. *Understanding by Design*. Pearson.

65. Michael F. Young, Stephen Slota, Andrew B. Cutter, et al. 2012. Our Princess Is in Another Castle: A Review of Trends in Serious Gaming for Education. *Review of Educational Research* 82, 1: 61–89. http://doi.org/10.3102/0034654312436980