Chapter # - will be assigend by editors

REPLAY ANALYSIS IN OPEN-ENDED EDUCATIONAL GAMES

AUTHOR(s) - Erik Harpstead, Christopher J. MacLellan, Vincent Aleven, Brad A. Myers - Carnegie Mellon University

Abstract: In this chapter we describe our method for analyzing student behavior in an open educational game. The method incorporates a logging schema and multiple data mining techniques to extract features from replays of student behavior. Student gameplay is captured at a fine enough level of detail that the captured data can be used to accurately reproduce a student's play session. These log traces are then played back through the game's engine, allowing designers and researchers to run additional metric calculations and data transformations in a post-hoc manner without disturbing the student's original play session. We have applied this replay analysis methodology to explore and evaluate the design of the open-ended educational game *RumbleBlocks*. Using data provided by our approach we have been able to map out the space of possible student solutions to each of the game's levels and evaluate how the game's incentive structure aligns to its stated educational goals. Each of these thrusts has led to actionable design recommendations that can be evaluated in follow up studies.

Key words: Open-ended Games, Alignment, Replay Analysis

1. INTRODUCTION

The field of serious games has grown to cover a diverse array of domains and subjects. Along with this growth in topic area there has been a similar broadening of the design space of serious games to include a number of different game structures. With these structures come many new questions. For example, initial work on serious games looked at whether games could promote learning. It has generally been found that serious games are capable of accomplishing their goals of fostering learning and so the new questions in the field are what features of games contribute to this success and how can they be made better (Clark, Tanner-Smith, Killingsworth, & Bellamy, 2013). More broadly, as the potential design space of serious games grows, the field of serious game analytics must also grow to accommodate new structures and questions.

One such issue that has seen less discussion in serious game analytics is what we refer to as alignment. In serious games, alignment is the idea that the rewards of your game should correspond to the goals you have in mind for it. One example of potential misalignment is *DragonBox*, a game designed to teach students algebra, where in a recent study it was found that while students succeed in the game, they do not improve on pre-posttests of algebra outside of the game (Long & Aleven, 2014). In dealing with misalignment, designers could benefit from having more ways of knowing whether or not their games are well aligned and where potential sources of misalignment are. The field of Serious Game Analytics is well suited to fill this gap.

Current approaches in serious games only tangentially address issues of alignment. Common techniques such as pre-posttests and AB testing can only provide designers insight into whether games are aligned or which of the tested features are better for alignment. These approaches require an explicit experimental design and the foresight of what features might be worth varying. While useful for advancing the science of serious games, these approaches are less helpful for informing design because they tend to come at the end of the game design process, rather than during the process. Serious game analytics would benefit from further focus on the formative context and the development of methods for identifying variables that would be worth AB testing.

Knowing if a game design is aligned is particularly difficult in open-ended games, whose structure provides a wide variety of actions for players to take. These games can easily become misaligned because when players have more freedom of action there are more chances for the game to present conflicting feedback. Providing players with freedom is a central design goal of using an open-ended structure and so we must rely on serious game analytics to create new methods that allow designers of open-ended serious games to understand the alignment of their game without compromising player freedom.

A persistent difficulty in working with open-ended games is to understand player actions in these broad spaces in terms of the context in which they took place. To address this we find power in the use of replay systems. The use of replay systems has become a common idiom in game design; with many commercial games providing access to state-based replay files that can be analyzed by researchers (Weber & Mateas, 2009). Additionally, many common game engines provide facilities for quick recreation of game entities and whole states with little programming effort (Harpstead, Myers, & Aleven, 2013). By having access to a full replay of player actions then all actions can be considered within the context which they took place.

In this chapter we will explore more about what it means for a game to be open-ended and the particular kinds of challenges games of this type can run into. Then we discuss replay analysis, a technique we have applied to help us in analyzing open-ended games. Finally we walk through a case study of how we apply this technique to the open-ended game *RumbleBlocks* (Christel et al., 2012). We conclude with a discussion of the remaining challenges in the space of open serious game analytics.

2. OPEN-ENDED SERIOUS GAMES

We begin by defining the notion of an open-ended serious game and explore the challenges that open-ended serious games pose for design and serious game analytics. First we shall explore how other researchers have conceptualized this space in prior work before presenting our own definition of openness. Finally, we recount the prior serious games analytics work that we believe is relevant to the open-ended serious games space.

Squire is one of the first researchers to use the term “open-ended games” in the context of serious games and to highlight their strong potential for serious applications (Squire, 2008). In his analysis he makes a distinction between two game genres that have potential for serious games. The first is persistent worlds, which have a large, socially shared, explorable world. Two widely known games in this category are *World of Warcraft* and *Everquest*. Squire mentions that some serious games have been designed to inhabit this space, notable *Quest Atlantis* (Barab, Thomas, Dodge, Carteaux, & Tuzun, 2005). The second genre that he discusses is open-ended simulation games, or sandbox games. Critical to this second category is the feature that there is no single correct pathway through the game to an end goal. Squire gives the examples of *Civilisation III* and *Grand Theft Auto: San Andreas*. Both games contain large, non-social, spaces which the player can explore. Further, part of learning and playing in these games is to gain an understanding of the game’s overarching system in order to use it to the player’s advantage. It is this systems focus, Squire claims, that makes open-ended games attractive for serious applications.

Spring and Pellegrino have also discussed the idea of open games (Spring & Pellegrino, 2011). Influenced by Squire, Spring and Pellegrino also cite the importance of players coming to an understanding of a game’s system in order to win being a crucial component of why open games are attractive for serious applications. They present two insights regarding open games: they lend themselves primarily to conceptual knowledge and they allow students to learn through failure. These insights imply that while open games maybe be primarily sandbox-like experiences they are not devoid of designer implemented feedback. Squire argued that that you cannot have a targeted open-ended game, but Spring and Pellegrino suggest that targeted feedback is still a component of open serious games.

While never explicitly discussing the openness of games, some of Gee's learning principles apply to open-ended serious games (Gee, 2003). Paralleling Squire, Gee highlights the importance of having multiple solution paths through a game's space so that students can learn through exploration. Similar to Spring and Pellegrino, Gee argues that players learn from the practice of probing open systems and receiving feedback, what he calls the probing principle.

In order to understand the number of possible solution paths, we can turn to Shell's concept of functional game spaces (Schell, 2008). In Schell's framing, a functional game space is the space in which a game *really* takes place rather than the physical or virtual space that might be seen by the player. For example, while the game of *Monopoly* is physically played on a two-dimensional board, *Monopoly*'s functional game space is a one-dimensional loop. The location of properties on the two dimensional board has no meaning beyond conveying property adjacency in the one dimensional loop. From this perspective, the space a player physically works in might be much larger than the functional space of a game. However, the complexity of the functional space is really what determines whether a game is open ended, an idea similar to problem-solving theories more generally (Newell & Simon, 1972).

Combining these prior conceptions, we define open-ended games as those with complex functional spaces containing many solution paths. In one regard, this definition is less inclusive than those proposed by prior researchers; i.e., games with limited functional spaces are less open ended. Persistent worlds, such as *River City* (Ketelhut, 2006), are often thought of as open ended because players are allowed to explore a large virtual world, however, from a functional space perspective these virtual worlds are composed of a few relevant non-player character interactions and events spread through a vast space that contains little game meaning. In another regard, our definition is more inclusive than those proposed by others. For example, our definition includes games such as *Refraction* (Andersen, Gulwani, & Popovic, 2013), which have many functional solution paths, but do not allow for free exploration, which is a key feature of open-ended games according to prior conceptualizations. While others' definitions are focused heavily on the sandbox nature of open-ended serious games because of their pedagogical affordances, our definition centers more on the challenges that a games structure imposes on serious games analytics.

Taking the perspective that open-ended serious games are characterized by large functional solution spaces, the issue of alignment becomes a key challenge. As the complexity of the solution space increases it becomes more important for designers to understand how their game is behaving and giving feedback in all portions of the space. As serious games are meant to guide players toward a set of learning objectives designers need more insight into how this is working in practice. Are students meeting the objectives? What parts of the design are out of alignment with the objectives? We view providing such insight as a grand challenge for serious game analytics.

2.1 Prior work in open-ended serious game analytics

A number of researchers have looked into the application of analytics to serious games in ways that we believe are relevant to the particular problem domain of serious open-ended games. While these researchers may not have chosen to describe their work in this way, we find it relevant to the current discussion.

Games that we would call open-ended have been in a number of ways in the past. Spring and Pellegrino used an approach of counting the number of positive and negative plant interactions players used in the game *Hortus* and used the pattern of how this usage changed to infer player learning (Spring & Pellegrino, 2011). In a similar vein Shute and colleagues used a classifier in the game Newton's Playground to determine the kinds of simple machines players were using to solve physics puzzles. They found that players that had higher pretest scores also had higher usage of simple machines, suggesting the game is likely well aligned. We could describe both of these approaches as traditional metric styles or feature counting as both approaches use a system in the game itself to report the presence or count of a feature. While these approaches are easy to work with, because they log small straightforward metrics, they are also limited to the features that were deemed important at design time. If new features are desired by analysts then new playtests must be run to gather the data.

The work on Playtracer by Andersen, Liu and colleagues is highly relevant to open-ended serious games (Andersen et al., 2013; Liu, Andersen, & Snider, 2011). The Playtracer approach takes player log traces and aggregates them into a state graph for each game level. This allows designers to see the different ways that players move through the space of their game, including when they return to states and arrive at incorrect final states. The original method for Playtracer was to aggregate states that are exactly equal but further work examined an aggregation based on common game-relevant features that allowed for even better aggregation. While this feature-based projection approach was an improvement, it still had trouble on very large continuous spaces such as the one present in *Foldit* (Liu et al., 2011). Similar work in visual analytics techniques using state transition diagrams has been done by Eagle, Johnson and colleagues in the field of intelligent tutoring systems (Eagle, Johnson, Barnes, & Boyce, 2013; Johnson, Eagle, & Barnes, 2013). In terms of alignment analyses these approaches have benefits in helping designers seen the functional solution space of their game but in themselves do not provide strong guidance on how to explore that space in considering whether it accomplished design goals.

Another relevant body of work looks at using procedural content generation to create level designs in open-ended games that verifiably have no short cut solutions (Smith, Andersen, Mateas, & Popović, 2012; Smith, Butler, & Popovi, 2013). This work looks at the game *Refraction*, and explores framing the level design process as an answer set programming task. This approach allows designers to specify a series of constraints that all levels must hold and then asks a procedural content generation system to create level designs that are known to hold to those constraints. this attacks the issue of alignment from the perspective of making misalignment impossible. While this work is a very promising approach to the issue of alignment in educational games it has limitations in that it requires designers to adopt a very particular formalization of their game design. It also has the limitation that the constraints the designer wants to hold must be known in advance for the system to work, a limitation shared by common AB testing paradigms (Lomas, Patel, Forlizzi, & Koedinger, 2013).

The Assessment Data Aggregator for Game Environments (ADAGE) framework is another analytics approach that has been applied to a number of serious games in the past (Owen & Halverson, 2013). The ADAGE schema has a layer that captures current play context on top of standard game metric telemetry. This context tracking can greatly benefit the analysis of open-ended games as it can provide insight into the particular experience of an individual player. One potential drawback in the current design of ADAGE is that the contextual information must be specified by a game designer ahead of time. This means that deciding what information is relevant to analysis must happen near the beginning of the design process and any future analyses will be constrained by the information that was deemed relevant at the time. While this can potentially be problematic, the paper describing ADAGE provides guidance on how to approach this schema design process.

3. REPLAY ANALYSIS

Replay systems are becoming a more common design element in games broadly, with many commercial games providing access to player replay files (Weber & Mateas, 2009). By replay we are referring to a system included within a game that re-enacts player actions, recorded in a transaction log, so as to reproduce a player's session. This is distinct from a video replay in that the analyst has full access to running game code, opening up many avenues for analysis that would not be possible from a video recording or simple metrics logging.

We have developed an approach for using a replay system to aid in serious game analytics research which we broadly refer to as Replay Analysis. Replay Analysis is an approach to serious game analytics that tries to address some of the challenges of open-ended game design by logging repayable traces of players' sessions rather than a predefined set of metrics. Using this approach, player performance can be considered in the context in which it took place. Additionally, it allows researchers and analysts to prototype new measures of learning without having to commit early, smoothing the design process. In this section, we briefly describe the components involved in performing a replay analysis, for more in-depth implementation details on a replay analysis library that we have developed for the Unity game engine see: (Harpstead, Myers, et al., 2013).

The approach entails both a particular schema for logging data and a system to replay logs through the game engine. Logs of student behavior are captured at the level of a basic action, defined as the smallest unit of meaningful action that a player can exert on the game world, what Schell would call operant actions (Schell, 2008). 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). The reason for recording the smallest possible actions is to allow for analysis at various grain sizes by capturing data at the finest grain size. Additionally, it is easy for game developers to see where logging calls need to be inserted into game code because the basic actions make up the base mechanics of the game. Each action is also paired with a description of the state of the game just before the action took place. 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 for logs to be replayed accurately without having to interpolate prior actions.

The second major component of the approach is a system for replaying actions, which we 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 constructs the state in which the action took place and then enacts the player action to let the game engine resolve the consequences of that action, using the same code that would normally handle such an action. Analyses can then be performed by running calculations on the results, with full access to any state attributes that would have been present at play time. 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.

One of the major benefits of the replay approach is that analyses are free to evolve along with the questions of the design team because no potential data is lost in logging. This paradigm also allows game design to proceed without having to agree on set of metrics to capture beforehand, reducing friction between designers and analysts during the development process and allowing analysts to explore multiple candidate metrics without having to commit to one.

The full benefit of this approach is felt most strongly in the context of classroom playtests. Because of the administrative and logistical processes involved in securing large populations of students in a classroom setting, such playtests are comparatively rare making the data captured in them all the more valuable. The high signal fidelity of replay logs ensures that datasets captured through classroom playtests throughout the design process can continue to benefit analysis and iteration.

4. RUMBLEBLOCKS

To demonstrate replay analysis, we describe a number of analyses that were facilitated by this technique on a game called *RumbleBlocks* (Christel et al., 2012). *RumbleBlocks* is an educational game designed to teach basic concepts of structural stability and balance to children in grades K-3 (ages 5-8 years old). The primary educational goals are for players to gain an understanding of three main principles of stability: objects with wider bases are more stable, objects that are symmetrical are more stable, and objects with lower centers of mass are more stable. These principles are derived from goals outlined in the National Research Council’s Framework for New K-12 Science Education Standards (National Research Council, 2012), and other science education curricula for the target age group.

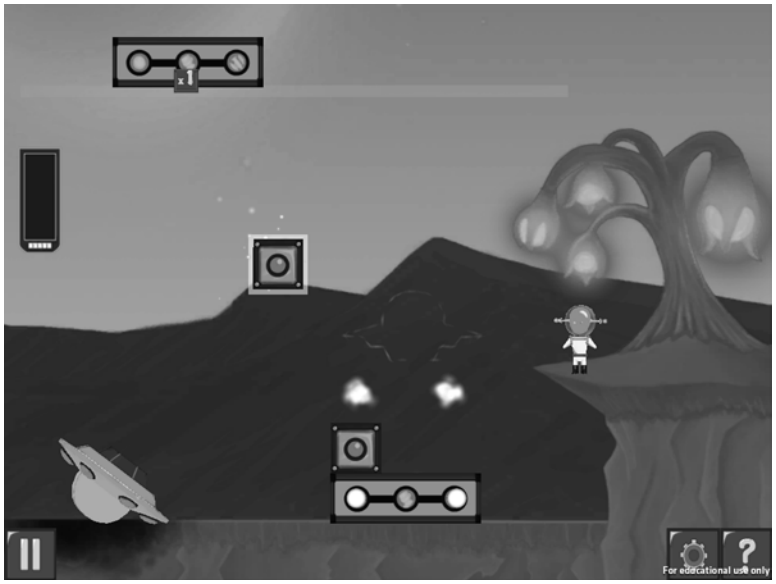

Figure 1. A screenshot of *RumbleBlocks*.

The game follows a sci-fi narrative where players help a group of stranded aliens on a number of foreign planets. Each 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 1). 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 cover a series of energy dots with their tower, which are captured in orbs on the blocks and provide the ship with power. Once the player has placed the ship on top of the tower, the ship powers up triggering an earthquake, if the earthquake topples the tower, or knocks the ship off the top, then the player 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.

We consider *RumbleBlocks* to be an open-ended game because, while its design may appear simple, it actually possesses a large functional solution space. Players are allowed place blocks in free space leading to effectively infinitely many possible solutions. The space could alternatively be considered as a grid, based on the size of the cube and implied by the energy dot placement, but even then there are many situations where there is a combinatorial number of possible solution configurations. Another element worth pointing out it that the designers of *RumbleBlocks* had no *a priori* knowledge of how many solutions can successfully satisfy the constraints of each level. While it may have been possible to enumerate all possible solutions, in a method similar to (Smith et al., 2012) such analysis was beyond the resources of *RumbleBlocks*' design team.

RumbleBlocks was outfitted to use our replay logging framework and replay analysis engine (RAE) to log player actions (Harpstead, Myers, et al., 2013). The basic actions involved in the game correspond to players' actions with the blocks (i.e. picked up, rotated or placed). The states being captured include a description of the position, orientation, and velocity of every block in the world as well as the spaceship. Logs can then be played back through the RAE and produce metrics, or other data encodings, as desired by our evolving analyses.

5. ANALYSES

Our analysis of RumbleBlocks progressed through a series of investigations, each facilitated by a different interpretation of players' log data. The data we discuss here were gathered as part of in-class playtests done with 174 students in two Pittsburgh area public schools, with players taken from the target demographic (ages 5-8). The goals of this playtest were to evaluate the current state of the game's design with a large group of players and to attempt to ground measures of learning within the game against out-of-game assessments of the game's educational objectives. 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 levels, but were altered to remove the energy ball 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 allows 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.

5.1 Learning results

The first question we sought to answer about *RumbleBlocks* was whether or not players were improving at the game's target concepts of structural stability and balance after playing the game. Turning first to the out-of-game tests, using a paired-samples t-test we measured a slight, yet significant, increase in performance from pretest to posttest, $t(173) = -2.13$, $p = .03$, $d = .16$. In looking at the difference in raw performance, i.e. the pass rate of the in-game pretest and posttest levels, a paired-samples t-test showed that there is a significant, medium sized, improvement in the passing rate from pre to post, $t(173) = -4.96$, $p < .001$, $d = .51$. While these results are encouraging, they can only tell us that players are getting better; it does not give us a sense of what they are getting better at.

We wanted to explore this question further by using the RAE and in-game log data to see if players were actually getting better at the specific principles targeted by the game. This would mean that from pre to post, players would build towers 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 but not in metrics related to the game's goals.

To find out whether players are learning the physics principles targeted by the game, we instrumented the RAE to read logs from the in-game pre-post levels and calculate a variety of metrics based on each player's final state of each level. We refer to these metrics as Principle-Relevant Metrics (PRMs). These metrics were: the width of the tower's base (Figure 2A), the height of the tower's center of mass relative to the ground (Figure 2B), 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 2C). These measures were then compared to values calculated across all other players for that same level in order to create a relative score for each player. This was done to account for the nuanced difference in level design making it difficult to compare metrics across levels. In the case of base width, the relative score was calculated relative to the maximum observed width for that level; for center of mass position, the score is relative to the lowest observed position for that level; and for symmetry a score is already relative to 90 degrees, which would represent perfect symmetry.

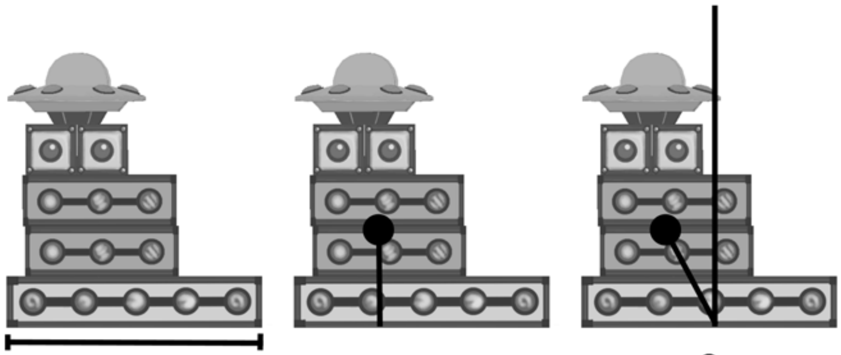

Figure 2 A visual depiction of each of the 3 Principle-Relevant Metrics used in analysis. (A) Base Width, (B) Center of Mass Height, and (C) Symmetry Angle.

To see if there was any improvement on the use of principles between the pre- and posttests we compared the related pre and posttest metrics for each student (averaged across the levels of the pre and posttest) using a paired-samples *t*-test. Looking at the results in Table 1 we 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, we 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 would suggest that the current version of the game may possess a misalignment in how it handles the low center of mass principle.

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

	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

5.2 Metric alignment analysis

Knowing from the pre-post level analysis that there were likely some design issues with RumbleBlocks, our next goal became understanding what those design issues were and how they might be addressed. 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 well designed 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 facilitate this analysis we augmented the RAE to calculate the same PRMs from the pre-post analysis, except this time to do it for all levels. We wanted to explore how well metrics that should indicate a well-constructed tower actually corresponded to a player passing a given level. To do this, we created 3 groups by collecting together all student 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. We performed logistic regressions using each of the metrics of the players' towers to predict success on a level, with all metrics normalized to mean 0 and standard deviation 1 to aid in interpretation of the relative strengths of their coefficient. 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. 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 our analysis of the game's system.

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

Table 2. The results of a logistic regression of success of solution on principle relevant metrics for levels targeting each of the three principles.

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

5.3 Solution clustering analysis

The logistic regression analysis agrees with the previous findings that players do not seem to be improving at the center of mass principle 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, then what are they doing? Answering this question requires a different view of the players' performance.

Rather than looking at how well players creations do in terms of the games' rules or the pedagogical principles, we wanted to get a more qualitative sense for what players were creating and how different features of these creations affected in-game performance (as measured by success on each level - does the tower stand or fall?). We hoped that analyzing properties of individual student solutions might provide insights into possible new design directions. However, there were over 6000 student solutions and combing through them all by hand would be impractical. To make understanding players' patterns of play more manageable, we wanted to cluster the solutions into groups that essentially embody the same solution. This way our analysis could proceed at the group level, which has many advantages. For example, we could determine whether students were creating the solutions originally envisioned by the designer. Furthermore, we could determine the specific structural features of a solution that contributes to its success or failure. This allowed us to explore how certain mechanics of the game affect the kinds of solutions that get rewarded, subsequently affecting what students do and learn. This type of analysis is useful for reflecting on educational goals and game mechanics on subsequent design iterations.

To perform these analyses, we 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. We wanted a representation that captured elements of these basic structural patterns. To build this new representation, we 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 collider information and individual block dimensions.

Next, we employed two-dimensional grammar induction to learn a set of patterns that can be used to describe all of the student solutions (Harpstead, Maclellan, et al., 2013). A two-dimensional grammar consists of three components: terminal symbols, which represent the blocks, spaceship, and space (in this context); non-terminal symbols, which represent structural patterns; and, rules which map non-terminal symbols to pairs of other non-terminal symbols oriented in a certain direction (horizontal or vertical). Rules can also map non-terminals to terminal symbols (a unary relationship). Figure 3 shows an example grammar (u, h, and v represent unary, horizontal, and vertical respectively), and the parses for two different towers. To learn a grammar, we generated an exhaustive set of rules that describe every possible way to parse all of the solutions. We then computed all the possible parses of each solution. Given the parses for each solution, we then created a vector for each solution which contained 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.

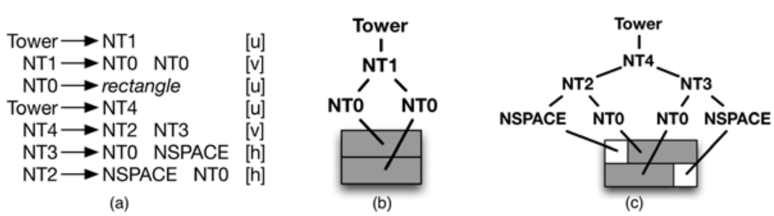

Figure 3. A two-dimensional grammar (a) and the parse trees generated by applying this grammar to two solutions (b and c).

For each level, we clustered the featurized solutions using g-means, a variant of the k-means algorithm that chooses a value for k optimizing for a Gaussian distribution within clusters (Hamerly & Elkan, 2004). This produced a set of different groups for each level, where each group represents solutions that share structural similarity. The resultant clusters allow us to get a picture of what the solutions space to the game looks like and begin to tease apart the different patterns.

As a first pass we wanted to see how often players conformed to the designers' expectations for a level. When designing levels, game designers often have a particular solution in mind, even if they intend the level to allow for a number of different solutions. We had a member of the game design team create a player trace that represented the designer envisioned solutions to each of the levels. We compared this trace to the clusters of other player solutions to identify which cluster each envisioned solution belonged to. Finally, we looked at how many players fell into the envisioned cluster as opposed to any other cluster. This analysis resulted in the pattern visible in Figure 4. A number of levels have a high degree of agreement with designers; these levels are mostly simple early levels or tutorial levels. Other levels have very little agreement with the designers' vision; these tend to be later levels that have a higher level of complexity.

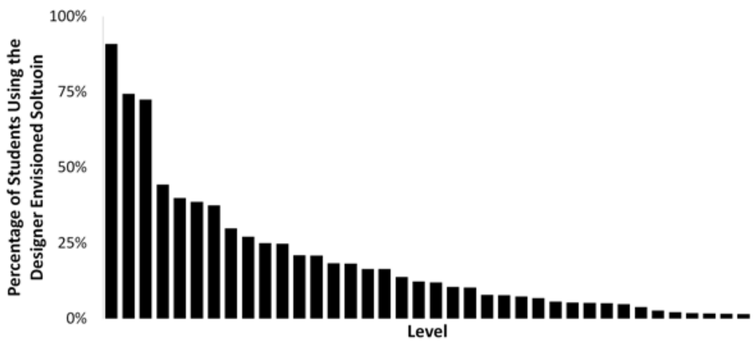

Figure 4. The percentage of students who used the solution envisioned by the designers on each level (sorted by percentage).

The pattern shown in Figure 4 helps to illustrate how open-ended *RumbleBlocks* can be, as the levels in the tail of the graph have a wider space of possibility than the designers may have envisioned themselves. It should be noted that this analysis does not necessarily result in a value judgment of the game, as players, particularly in an open-ended game, are likely to diverge from the designers' perspective. Instead it helps designers to understand where their own intuitions are likely to be the weakest.

Another way of using this clustering formulation is to explore the alignment of individual clusters allowing designers to have a picture of specific examples of misalignment (Harpstead, MacLellan, Aleven, & Myers, 2014). To do this we looked at the metrics, calculated for the previous regression analysis, within each cluster and looked at the patterns between those metrics and success on levels. For each cluster, we created a representative solution by averaging the PRM scores within the cluster and assuming the majority success value. This gives us the ability to think about common patterns of solutions through a single representative rather than a group. We then examined levels by looking at plots such as Figure 5. These plots show each representative solution plotted with its frequency percentage along the x-axis and its relative principled-ness, in terms of a normalized PRM score, along the y-axis. The lighter squares represent solutions that are mostly successful where the darker diamonds show solutions that are mostly unsuccessful.

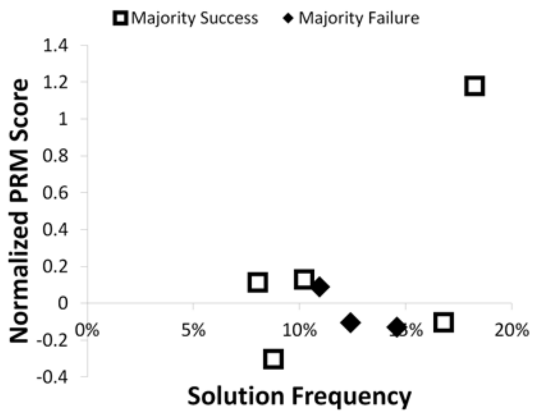

Figure 5. A plot of representative solutions' PRM score versus frequency.

When examining plots like Figure 5 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 level. When visually inspecting these misaligned cases across levels a pattern started to emerge; it appeared that failure seemed related to towers with small platforms for the spaceship to sit on top. Such a pattern is likely to lead to the spaceship falling off the tower, regardless of how well the design of the rest of the tower adheres to the target principles. While, there is a component of stability we were concerned that the secondary success criterion of the spaceship staying on the tower may be overshadowing the rest of the features of a tower, leading students to focus on a small element of their design rather than the design as whole.

To test this hypothesis we turned again to RAE and the features that were produced for clustering. These symbols represent binary relationships of different block types and so encode different sub-structural patterns present in player towers. We 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. We applied a Bonferroni correction to the χ^2 results to account for the number of statistical tests (6,010). Results for the χ^2 analysis can be found in Table 3, and visually rendered in Figure 6. Overall we found 19 symbols that were significantly related to success, after correction. Due to idiosyncrasies in our grammar learning algorithm it happens that 14 of these symbols would ground out to be the same, NT37 in Figure 6, and 2 of the symbols NT329 and NT44, share a similar relationship. This happens because the grammar learning process learns multiple rules to represent adjacent empty space and we have omitted these for clarity. Once we had a selection of significant features we performed a logistic regression of those 19 features against solution success to understand the direction of the relationship, i.e. does each feature more predict success or failure. We only present the directional results of this regression and not the actual coefficients.

Table 3. χ^2 results of students' solutions' structural features against success, with Bonferroni corrected *p*-values and logistic regression direction.

Symbol	χ^2	***p***	***direction***
NT 1633	20.95	0.028	−
NT 319 †	21.00	0.028	−
NT 1083 †	22.10	0.016	−
NT 154 †	26.27	0.002	−
NT 2458 †	26.65	0.001	−
NT 1166 †	30.22	< .001	−
NT 1537 †	30.25	< .001	−
NT 284	30.43	< .001	−
NT 329 ‡	33.35	< .001	−
NT 10897 †	35.83	< .001	−
NT 37 †	37.07	< .001	−
NT 13	38.55	< .001	+
NT 1538 †	39.18	< .001	−
NT 40 †	39.53	< .001	−
NT 11914 †	39.55	< .001	−
NT 150 †	39.55	< .001	−
NT 1541 †	68.50	< .001	−
NT 44‡	71.13	< .001	−
NT 151 †	73.11	< .001	−

Note. † solutions share a visual grounding to NT37; ‡ solutions share a visual grounding to NT44

The pattern that arises from the χ^2 analysis confirms that towers containing the spaceship on top of a lone square, or substructures containing squares without other supports, are more likely to fail than towers containing the spaceship on top of a wide platform. This analysis demonstrates that the spaceship remaining on top of the tower represents a secondary success criterion. While the designers were aware that the spaceship served such a purpose in the design, they 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.

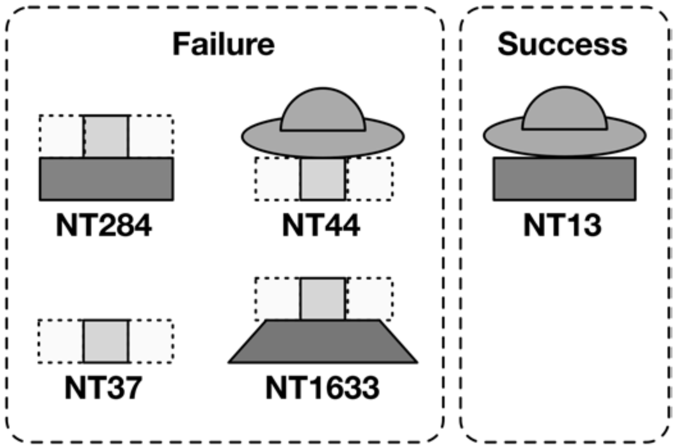

Figure 6. Visual renderings of the symbols from Table 3 found to be significantly related to success and failure.

6. DISCUSSION

In this chapter, we have demonstrated a number of analyses of the design of the game *RumbleBlocks*, all of which were facilitated by the use of replay analysis. Each of these analyses has contributed to an evolving understanding of the design of *RumbleBlocks* and how well it accomplishes its stated goals. As work on the game moves forward, similar assessments will be made of future versions to ensure existing design problems have been properly addressed.

While we have described a particular case study here we believe that the issue of alignment is an important one for serious game analytics to consider more broadly. Players approach games from different perspectives than their designers (Hunicke, Leblanc, & Zubek, 2004). When designing games for serious purposes it is important that designers be equipped with tools and techniques that help them understand whether their goals are being met. We have presented replay analysis as one means of filling this need but we view this as just a first step into what we expect will be a larger space within serious game analytics.

Another use of replay analysis, which we have not discussed here, is to use replayed sessions as a means of exploring the implications of alternative designs. For example, the designers of *RumbleBlocks* might want to entertain the idea of having the spaceship stick to the top block of a tower to reduce its effect as a secondary success criterion. The RAE could be instrumented to recreate players' final towers with this system in place and run the earthquake to see how the game reacts differently. While this particular use of replay does can provide some interesting perspective it does have limitations. For example, it could be argued that players would have played differently had certain rule changes been in effect. Additionally, exploring such alternatives with log files from different iterations of the game requires more attention to detail in version control to ensure replay results reflect the version of the game they were recorded with.

Another issue with the use of a replay paradigm is that it produces comparatively large log files. The entire *RumbleBlocks* dataset that we have discussed here (containing approximately 80 minutes of play from 174 players) is roughly 850MB in size. This does not include the residual files output by the RAE for use in statistical analysis or clustering. Additionally, while the replay process can be sped up, it does require some amount of real time to re-simulate the game environment. When taken to the scale seen in some other serious games work that involves millions of players over the course of months, this could quickly become intractable (Andersen et al., 2012; Lomas et al., 2013). A possible way of addressing this could be to only record replay fidelity logs for a subset of the overall player population. Another approach might be to use replay logs in the earlier stages of design, before a final set of metrics has been decided, as a way of prototyping measures of learning or other desirable outcomes.

7. CONCLUSION

The increase in the number of open-ended serious games is an exciting trend in the field of serious game analytics. While this trend is encouraging, it is important to keep in mind how the open-ended nature of some games can undermine their intended serious purposes. Replay analysis is just one of what we hope will become many analytics techniques for studying how players are navigating open-ended domains and interacting with content. We hope that other researchers can find utility in our approach and look forward to what new challenges the future of the field brings.

REFERENCES

Andersen, E., Gulwani, S., & Popovic, Z. (2013). A Trace-based Framework for Analyzing and Synthesizing Educational Progressions. In *Proceedings of the 31st International ACM SIGCHI Conference on Human Factors in Computing Systems - CHI '13* (pp. 773–782). ACM Press. doi:10.1145/2470654.2470764

Andersen, E., O'Rourke, E., Liu, Y., Snider, R., Lowdermilk, J., Truong, D., … Popovic, Z. (2012). The impact of tutorials on games of varying complexity. In *Proceedings of the 30th International ACM SIGCHI Conference on Human Factors in Computing Systems - CHI '12* (pp. 59–68). doi:10.1145/2207676.2207687

Barab, S., Thomas, M., Dodge, T., Carteaux, R., & Tuzun, H. (2005). Making learning fun: Quest Atlantis, a game without guns. *Educational Technology Research and Development*. doi:10.1007/BF02504859

Christel, M. G., Stevens, S. M., Maher, B. S., Brice, S., Champer, M., Jayapalan, L., … Lomas, D. (2012). RumbleBlocks: Teaching science concepts to young children through a unity game. In *Proceedings of CGAMES'2012 USA - 17th International Conference on Computer Games: AI, Animation, Mobile, Interactive Multimedia, Educational and Serious Games* (pp. 162–166). IEEE. doi:10.1109/CGames.2012.6314570

Clark, D. B., Tanner-Smith, E. E., Killingsworth, S., & Bellamy, S. (2013). *Digital Games for Learning: A Systematic Review and Meta-Analysis (Executive Summary)*. Menlo Park, CA.

Eagle, M., Johnson, M., Barnes, T., & Boyce, A. (2013). Exploring Player Behavior with Visual Analytics. In *Proceedings of the 8th International Conference on the Foundations of Digital Games - FDG '13* (pp. 380–383).

Gee, J. P. (2003). What video games have to teach us about learning and literacy. *Computers in Entertainment*. New York: Palgrave Macmillan. doi:10.1145/950566.950595

Hamerly, G., & Elkan, C. (2004). Learning the k in k-means. In S. Thrun, L. K. Saul, & B. Schölkopf (Eds.), *Advances in Neural Information Processing Systems 16: Proceedings of the 2003 conference* (pp. 281–288). MIT Press.

Harpstead, E., MacLellan, C. J., Aleven, V., & Myers, B. A. (2014). Using extracted features to inform alignment-driven design ideas in an educational game. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14* (pp. 3329–3338). ACM Press. doi:10.1145/2556288.2557393

Harpstead, E., Maclellan, C. J., Koedinger, K. R., Aleven, V., Dow, S. P., & Myers, B. A. (2013). Investigating the Solution Space of an Open-Ended Educational Game Using Conceptual Feature Extraction. In *Proceedings of the International Conference on Educational Data Mining - EDM '13* (pp. 51–58).

Harpstead, E., Myers, B., & Aleven, V. (2013). In search of learning: facilitating data analysis in educational games. In *Proceedings of the 31st International ACM SIGCHI Conference on Human Factors in Computing Systems - CHI '13* (pp. 79–88). Paris, France: ACM Press. doi:10.1145/2470654.2470667

Hunicke, R., Leblanc, M., & Zubek, R. (2004). MDA : A Formal Approach to Game Design and Game Research. In *Proc. of the AAAI Workshop on Challenges in Game AI* (pp. 1–5).

Johnson, M. W., Eagle, M., & Barnes, T. (2013). InVis : An Interactive Visualization Tool for Exploring Interaction Networks. In *Proceedings of the International Conference on Educational Data Mining - EDM '13* (p. 65).

Ketelhut, D. J. (2006). The Impact of Student Self-efficacy on Scientific Inquiry Skills: An Exploratory Investigation in River City, a Multi-user Virtual Environment. *Journal of Science Education and Technology*, *16*(1), 99–111. doi:10.1007/s10956-006-9038-y

Liu, Y., Andersen, E., & Snider, R. (2011). Feature-based projections for effective playtrace analysis. In *Proceedings of the 6th International Conference on Foundations of Digital Games - FDG '11* (pp. 69–76). ACM Press. doi:10.1145/2159365.2159375

Lomas, D., Patel, K., Forlizzi, J. L., & Koedinger, K. R. (2013). Optimizing challenge in an educational game using large-scale design experiments. In *Proceedings of the 31st International ACM SIGCHI Conference on Human Factors in Computing Systems - CHI '13* (pp. 89–98). ACM Press. doi:10.1145/2470654.2470668

Long, Y., & Aleven, V. (2014). Gamification of Joint Student/System Control Over Problem Selection in a Linear Equation Tutor. In *Proceedings of the 12th International Conference on Intelligent Tutoring Systems* (pp. 378–387). doi:10.1007/978-3-319-07221-0_47

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. (H. Quinn, H. Schweingruber, & T. Keller, Eds.)*Social Sciences*. The National Academies Press.

Newell, A., & Simon, H. A. (1972). *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.

Owen, V. E., & Halverson, R. (2013). ADAGE (Assessment Data Aggregator for Game Environments): A Click-Stream Data Framework for Assessment of Learning in Play. In *Proceedings of the 9th Games+Learning+Society Conference - GLS 9.0* (Vol. 9, pp. 248–254). ETC Press.

Schell, J. (2008). *The Art of Game Design: A Book of Lenses* (1st ed.). Burlington, MA: Morgan Kaufmann.

Shute, V. J., Ventura, M., & Kim, Y. J. (2013). Assessment and Learning of Qualitative Physics in Newton's Playground. *The Journal of Educational Research*, *106*(6), 423–430. doi:10.1080/00220671.2013.832970

Smith, A. M., Andersen, E., Mateas, M., & Popović, Z. (2012). A case study of expressively constrainable level design automation tools for a puzzle game. In *Proceedings of the 7th International Conference on Foundations of Digital Games - FDG '12* (p. 156). ACM Press. doi:10.1145/2282338.2282370

Smith, A. M., Butler, E., & Popovi, Z. (2013). Quantifying over Play: Constraining Undesirable Solutions in Puzzle Design. In *Proceedings of the 8th International Conference on Foundations of Digital Games - FDG '13* (pp. 221–228).

Spring, F., & Pellegrino, J. W. (2011). The Challenge of Assessing Learning in Open Games : HORTUS as a Case Study. In *Proceedings of the 8th Games +Learning+Society Conference - GLS 8.0* (pp. 209–217).

Squire, K. (2008). Open-Ended Video Games: A Model for Developing Learning for the Interactive Age. In K. Salen (Ed.), *The Ecology of Games: Connecting Youth, Games, and Learning* (pp. 167–198). Cambridge, MA: MIT Press. doi:10.1162/dmal.9780262693646.167

Weber, B. G., & Mateas, M. (2009). A data mining approach to strategy prediction. In *2009 IEEE Symposium on Computational Intelligence and Games* (pp. 140–147). Ieee. doi:10.1109/CIG.2009.5286483

ACKNOWLEDGEMENTS

This work was supported in part by a Graduate Training Grant awarded to Carnegie Mellon University by the Department of Education #R305B090023 and by the DARPA ENGAGE research program under ONR Contract Number N00014-12-C-0284. All opinions expressed in this article are those of the authors and do not necessarily reflect the position of the sponsoring agency.

AUTHOR INFORMATION

Full Name: Harpstead, Erik
Affiliation: Carnegie Mellon University
Mailing Address: 5000 Forbes Avenue, Pittsburgh, PA, 15213
Email: eharpste@cs.cmu.edu
Website: www.erikharpstead.net
Bio: Erik Harpstead is a PhD student in the Human-Computer Interaction Institute at Carnegie Mellon University and a fellow in the Program for Interdisciplinary Education Research (PIER). His research interest is the design and evaluation of educational games. To this end, he works on developing novel tools and techniques for evaluating educational games in terms of their stated goals. He is also interested in better understanding ways in which game designers and learning scientists can collaboratively evaluate games in service of redesign.

Full Name: MacLellan, Christopher, J
Affiliation: Carnegie Mellon University
Mailing Address: 5000 Forbes Avenue, Pittsburgh, PA, 15213
Email: cmaclell@cs.cmu.edu
Website: http://www.christopia.net/
Bio: Christopher MacLellan is a PhD student in the Human-Computer Interaction Institute at Carnegie Mellon University and a fellow in the Program for Interdisciplinary Education Research (PIER). His work centers on the applying artificial intelligence and machine learning techniques to construct models of how humans perform open-ended tasks. Using this approach, he studies human learning, problem solving, and design in the context of intelligent tutoring systems and educational technologies.

Full Name: Aleven, Vincent
Affiliation: Carnegie Mellon University
Mailing Address: 5000 Forbes Avenue, Pittsburgh, PA, 15213
Email: aleven@cs.cmu.edu
Website: http://www.cs.cmu.edu/~./aleven/
Bio: Dr. Vincent Aleven, an Associate Professor in Carnegie Mellon's Human-Computer Interaction Institute (HCII), has over 20 years of experience in research and development of advanced learning technologies

grounded in cognitive science theory and theory of self-regulated learning. He has built intelligent tutoring systems (e.g.,http://mathtutor.web.cmu.edu) and educational games, as well as authoring tools to facilitate their creation (e.g., http://ctat.pact.cs.cmu.edu). He has used these systems as platforms to investigate student learning in classrooms and online courses. He led a team that created games for science learning for children in K-3 in collaboration with Carnegie Mellon's Entertainment Technology Center. Aleven and his students have done research on educational games for policy reasoning and intercultural negotiation skills. He is a co-founder of Carnegie Learning, Inc., a Pittsburgh-based company that markets Cognitive Tutor™ math courses. He is a member of the Executive Committee of the Pittsburgh Science of Learning Center (LearnLab). He is or has been PI on 7 major research grants and co-PI on 10 others. He has over 180 publications to his name. He is Co-Editor-in-Chief of the *International Journal of Artificial Intelligence in Education.*

Full Name: Myers, Brad, A
Affiliation: Carnegie Mellon University
Mailing Address: 5000 Forbes Avenue, Pittsburgh, PA, 15213
Email: bam@cs.cmu.edu
Website: http://www.cs.cmu.edu/~bam/
Bio: Brad A. Myers is a Professor in the Human-Computer Interaction Institute in the School of Computer Science at Carnegie Mellon University. He is an IEEE Fellow, ACM Fellow, winner of nine best paper type awards and three Most Influential Paper Awards. He is also a member of the CHI Academy, an honor bestowed on the principal leaders of the field. He is the principal investigator for the Natural Programming Project and the Pebbles Handheld Computer Project, and previously led the Amulet and Garnet projects. He is the author or editor of over 430 publications, and he has been on the editorial board of five journals. He has been a consultant on user interface design and implementation to over 75 companies, and regularly teaches courses on user interface design and software. Myers received a PhD in computer science at the University of Toronto, and the MS and BSc degrees from the Massachusetts Institute of Technology during which time he was a research intern at Xerox PARC. From 1980 until 1983, he worked at PERQ Systems Corporation. His research interests include user interfaces, programming environments, programming by example, and interaction techniques.