# Using Extracted Features to Inform Alignment-Driven Design Ideas in an Educational Game

**Erik Harpstead, Christopher J. MacLellan, Vincent Aleven, Brad A. Myers**
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA 15213
{eharpste,cmaclell,aleven,bam}@cs.cmu.edu

## ABSTRACT
As educational games have become a larger field of study, there has been a growing need for analytic methods that can be used to assess game design and inform iteration. While much previous work has focused on the measurement of student engagement or learning at a gross level, we argue that new methods are necessary for measuring the alignment of a game to its target learning goals at an appropriate level of detail to inform design decisions. We present a novel technique that we have employed to examine alignment in an open-ended educational game. The approach is based on examining how the game reacts to representative student solutions that do and do not obey target principles. We demonstrate this method using real student data and discuss how redesign might be informed by these techniques.

### Author Keywords
Educational Games; Alignment; Analytics; Game User Research

### ACM Classification Keywords
H.5.2 User Interfaces: Evaluation/methodology; K.3.1 Computer Uses in Education: Computer-assisted instruction (CAI); K.8.0 Personal Computing: Games.

## INTRODUCTION
The idea that well-designed games can possess powerful affordances for education has become well accepted by researchers and practitioners alike [1,7,11,19]. The original question of whether or not games are good for learning has given way to new questions of what particular aspects of effective games lead to good learning and out-of-game transfer, and how new games can be better designed to improve learning [6]. This change in orientation toward design issues requires a new suite of methods that can be applied within the design process of an educational game. Having rich analytics data earlier in the process can facilitate discussions around what does and does not work in a particular educational game's design.


*CHI 2014,* April 26 - May 01 2014, Toronto, ON, Canada

ACM 978-1-4503-2473-1/14/04…$15.00.
http://dx.doi.org/10.1145/2556288.2557393

Designing educational experiences is difficult because there are a number of different perspectives and concerns that need to be balanced [13,30]. Taking an educational design in one direction may cause some other aspect to swing out of balance. This issue is particularly salient when talking about educational games, which must balance the concerns of being an engaging experience, so students actually play them, and an educational experience, so students actually learn something that they can apply outside of the game. One important aspect of games as educational experiences is alignment, which is the idea that game success and educational success go hand-in-hand, meaning that the game rewards actions that are likely to be educationally effective.

To assist educational game designers in balancing these concerns, a number of researchers have put forth theoretical frameworks and patterns for good educational game design [1,11,19,21]. These frameworks are based on existing work in the learning sciences and input from practicing game designers. Having a theoretical grounding while designing can provide a context for discussing design issues and how they might impact the numerous components of an educational game's design; however, a theoretical framework alone will not guarantee that a game achieves its ultimate goals. Throughout the design process, factors inherent in normal design iteration can impinge upon the design and cause conflicts between different components [13]. While having a design framework can inform what redesign options educational game designers might have for addressing a given design conflict, it may not easily inform them which option is most valuable to explore.

Design decisions made within an iterative design loop are often based on intuitions. However, the notion that design intuitions can fail when subjected to empirical scrutiny is a well-documented phenomenon in HCI and game user research [4,16,17]. A better approach is to guide design by using theoretical frameworks, but even this approach may lead to undesirable outcomes when there are no clear methods for assessing design outcomes. Bringing analytical methods into the design process to assess the design of a game can help inform discussions by providing objective measures to compare alternatives.

We believe that combining theory based design thinking with early-stage analytic techniques can help designers to reason through the nuances of their game designs in an empirically informed way. In this paper, we demonstrate a specific technique which we have found useful for addressing alignment related design issues in the educational game *RumbleBlocks* [5]. The technique employs a clustering method that we have described previously [8] and situates it within an educational game design framework of our own design [1]. We believe these elements to be modular and in demonstrating this process we make two contributions to the educational game design and HCI literature:

- A novel analytical technique that uses log data from actual students to assess the alignment of an educational game's mechanics to its stated educational goals.
- A demonstration of how the EDGE educational game design framework [1] can be applied within the design process of a game, where prior work has only demonstrated the summative capacity of the framework.

## ANALYTICS IN GAMES

Analytics has become one of the principal research thrusts in game user research. Analytics research proposes a number of methods for understanding players' behavior within games [4,12,25,31]. All of these methods are driven by a desire to understand players' experiences and use that knowledge to reason about game iterations and redesigns. The vast majority of existing work on game analytics is concerned with measuring player *engagement*. A primary reason for this is the simple principle that players will not actually play a game if it is not fun. While this work is certainly important, within the realm of educational games, engagement cannot be the only measure applied. We argue that how closely a game *aligns* with the educational goals is an important metric that should also be considered.

## ALIGNMENT

Alignment is not something traditionally tested for as part of educational game design. In practice, alignment issues are normally caught during the course of standard efficacy testing, where students would take pre- and posttests to measure learning after playing a game, or during an A/B manipulation to test the impact of a particular game feature. Such experimental designs tend toward making conclusions about a game as a whole, making it hard to consider the impact of particular features of a game's design, unless a large number of experimental conditions are employed [20].

Evaluating alignment is important because there are a number of ways that an educational experience can be misaligned with its educational goals. One of the more common ways a game can be misaligned lies in how it provides feedback to a learner. If the feedback is inconsistent or poorly timed, then it makes it difficult for the learner to associate their actions with correct understanding. This focus on feedback differentiates our work from the related psychometric concept of construct irrelevant variance [18] which concerns itself with how well a test is measuring what it should be measuring.

One situation where feedback might be inconsistent occurs when the instructional task requires knowledge that the designer did not anticipate [15]. These unexpected demands can result in inconsistent feedback in situations where a learner is performing some task that requires some desired knowledge as well as some unanticipated knowledge. In these situations, learners may be told they are wrong even if they possess the desired knowledge simply because they are lacking the unanticipated knowledge. This results in the learner receiving negative feedback even when they may be correctly applying the desired knowledge. A classic example of such unanticipated knowledge comes from the work of Koedinger and Nathan when examining the student performance on different types of algebra problems [15]. The common wisdom among teachers was that algebraic word problems were more difficult than formal symbolic problems, such as $5x + 3 = 6$, because students had to intuit the underlying mathematical equation the story represented. Contrary to this common wisdom, Koedinger and Nathan found that students actually have more difficulty with the formal symbolic problems because there was a barrier in understanding what the mathematical symbols meant. As there was no feedback for how well a student understood the original problem, they could not learn to correct their understanding and improve on formal symbolic problems.

A different situation where feedback tends to become inconsistent is when feedback is delayed from student actions that express understanding of the concept or when students must infer the nature of the feedback from the complex state of the world. In educational technology literature this phenomenon is commonly referred to as situational feedback [24,26], where learners execute one or more actions and must then infer from the system's response whether or not they were correct in their thinking. These situations arise most prominently when learners must exercise multiple skills at once. If a learner applies one skill correctly and one incorrectly they would receive a single piece of negative feedback and be left to intuit which skill they failed to apply correctly. These situations make it difficult for learners to determine which of their actions are correct. This issue is especially common to games which fall into the "configure and run" design pattern [21], where players build up some complex structure as a potential solution to an in-game puzzle and finally ask the system to evaluate it by running a simulation and seeing if it achieves the desired result. While such puzzles can be engaging, their feedback leaves the learner having to reason about what exactly lead to their success or failure.

The issue of misaligned feedback is particularly important in educational games because they are feedback-rich environments that can provide many, often subtle, cues about player status. Players can take this feedback as guidance

**Figure 1. A screenshot from *RumbleBlocks*. Players must build a tower that is tall enough to reach the alien on the cliff, while also covering the blue energy balls.**

and adjust their behavior accordingly. When a game is being designed to convey an instructional message, it is important that the game provides feedback that aligns with the instructional goals of the system.

## THE EDGE FRAMEWORK

Issues of alignment represent complex problems that exist in the interactions between multiple game design elements; so solving them is not a simple task. Employing a theoretical educational game design framework can assist in thinking across the elements of a design. The framework that we employ in our own work is called the EDGE framework [1]. EDGE, which stands for *Engaging Design of Games for Education*, employs three main components in looking at educational game design: educational objects, game design theories, and learning science principles.

The framework starts with advocating for rigorous educational objectives, drawing on inspiration from the educational theory of "backwards design" [32]. Backwards Design calls for starting from a set of goals, determining how you will assess students based on those goals, then considering how you will design an intervention, which "moves the needle" on those assessments.

The EDGE framework conceptualizes game mechanics mainly through the lens of the Mechanics, Dynamics, and Aesthetics (MDA) framework [10]. MDA is particularly applicable to the educational setting because of its bi-directional perspective on game design, considering both the designer's and the player's views. A game designer primarily has control over the mechanics of a game, or the base rules of the game's system. The player, on the other hand, interacts with a game as an aesthetic experience, which can only ever be indirectly perceived by the designer. In between these two perspectives are the dynamics of the game's system, which can be hard to anticipate and control from a design perspective. Instructional design can be seen through a similar set of perspectives: a teacher, or instructional designer, has control over the base mechanics of an educational experience while the student's learning of intended content takes place as indirectly perceivable events [14], with many messy dynamic factors in between complicating the process.

The learning principles covered in the EDGE framework are left intentionally modular, to allow for changes in theoretical perspective and context of use. The principles used in the framework can be taken from a number of sources that exist in the learning sciences literature [14,22,28]. Each collection of principles represents a summation of many years of research in the learning sciences, not unlike the usability heuristics commonly employed in the HCI community [27]. The specific collection of principles used will change based on the goals, context, and demographics of the game.

In general, the EDGE framework could be seen as trying to combine the three wisdoms of instructional design practitioners, game design practitioners, and learning science researchers to allow for a reasoned way of integrating multiple perspectives while designing educational games.

## RUMBLEBLOCKS

The game we will be primarily discussing is called *RumbleBlocks*, which 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) [5]. 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. The game follows a narrative of the player helping a group of stranded aliens on a number of foreign planets. In each level the player encounters an alien who is stranded on a cliff with their deactivated spaceship left off to the side of the world (see Figure 1). Players must build a tower out of blocks that is tall enough to reach the alien so that they can give the alien back his ship. In the process, they must also cover a series of energy dots with their tower so that the ship will receive power. Once the player has placed the ship on top of the tower, an earthquake is triggered when the ship powers up. If the earthquake topples the tower, or knocks the ship off the top, then the player must restart the level; if the tower remains standing, with the ship on top, then the player succeeds and moves on to the next level.

Previous analysis of *RumbleBlocks* [8,9] suggested that this domain possesses some issues with alignment between its mechanics and educational goals; however, this prior work could not provide enough detail to inform redesign. Previously, we showed that principle-relevant metrics, i.e. metrics from game log data which measure a student tower based on adherence to a particular principle, predicted success in a logistic regression for the wide base and symmetry principles, but not for the center of mass principle. This suggests that students adhering to the center of mass princi-

| | Unprincipled | Principled |
|---|---|---|
| Successful | **Bad** | **Good** |
| Unsuccessful | **Good** | **Bad** |

**Figure 2. A matrix showing the possible alignment interpretations of student solutions based on how principled they are and success feedback that the game assigns to them.**

ple would be receiving inconsistent feedback on their tower designs [9]. We examined clusters of student solutions based on extracted features from student game states to see how well they aligned with designer expectations and found a number of levels which deviated strongly from the designers' vision for a level [8]. Our current analysis looks to draw from our previous techniques with the motivation of digging deeper into the issue of misalignment and providing a fine grain view of player data capable of providing actionable design recommendations, whereas prior work stopped at diagnosing the existence of a problem. We make use of the prior datasets to help ground the approach in existing work.

## ANALYSIS

At a high level, our approach 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 actual students in the target population into a series of representative solutions. Once a collection of representative solutions is gathered, each one is evaluated using a principle-relevant metric (PRM), which measures how closely the representative solution embodies a certain target principle normalized across all other representative solutions to the same challenge. 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.

Using this approach, representative solutions can arrive at one of four designations, best thought of as the 2x2 matrix shown in Figure 2. 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 a designation of successful, which would represent that the game is reinforcing target concepts to the student. Similarly, solutions that are unprincipled should be given a designation of failure, which would represent that the game is discouraging deviations from target concepts, allowing a student 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.

We analyze PRM alignment in terms of representative solutions because they are a more concise representation of the solution space. Our goal is to help identify causes, and potential fixes for, cases of misalignment. Clusters help greatly in achieving this goal. They make it easier to investigate the causes of misalignment because solutions within a cluster should all be aligned or misaligned in the same way. Without clusters, one could employ other methods to know that a particular level is problematic but then one would have to resort to sifting through each individual solution one-by-one to figure out what the reason for the misalignment was. Additionally, we do not consider frequency information, or how many users used each representative solution, at this stage because this information is likely to vary by playtest population.

### Data

The data we are using in this section comes from a formative evaluation of *RumbleBlocks* conducted in 2 local area schools using 174 students in the target demographic (5-8 year-olds). The evaluation took place in the context of a pre-post design over 4 days; it began with a pretest of target concepts on the first day, 2 sessions of roughly 45 minutes of gameplay each subsequent day, and ended on the final day with a counterbalanced posttest of the same target concepts. Throughout the study, the game was instrumented to log student actions at a high level of fidelity capable of being replayed through the game's engine. This replay approach allows us to perform a number of different analyses on the same dataset, allowing us to be more flexible and iterate on potential principle-relevant metrics [9].

Our novel analysis method assesses the degree to which success in playing the game aligns with educational objectives as the frequency with which solutions fall into the two desirable quadrants of Figure 2. This method starts by clustering student solutions. To generate student solution clusters with the *RumbleBlocks* data, we first generate a set of features using the conceptual feature extraction process, which we have reported on previously [8]. This process outputs a feature vector for each student solution that describes which structural patterns are and are not present within the solution. These patterns may correspond to individual blocks, pairs of blocks, other combinations of blocks, or even whole towers. We then use these feature vectors as input to standard clustering methods to yield a set of solution clusters for each level. We then view each cluster as a representative solution for that level.

For each representative solution we calculate the average value of each PRM on all the towers with that solution. The average PRM value is then interpreted as the canonical PRM for the representative solution. Finally, the PRM scores of all the representative solutions to the same level

are normalized to make it easier to compare between solutions.

In *RumbleBlocks,* the PRMs correspond to calculable metrics from a student's tower; see Figure 3 for a visual representation. For levels targeting the "objects with wider bases are more stable" principle, the width of the tower's base is used as the PRM. For the "objects with lower centers of mass are more stable" principle, we calculate the tower's overall center of mass and take its height from the ground. For the "symmetrical objects are more stable" principle, we create a ray that extends from the center of the base of the tower through the center of mass and calculate the angle of this ray and its absolute difference from 90°. While this is not a perfect representation of geometric symmetry, in practice this measure adequately captures the concept as it exists in the *RumbleBlocks* dataset.

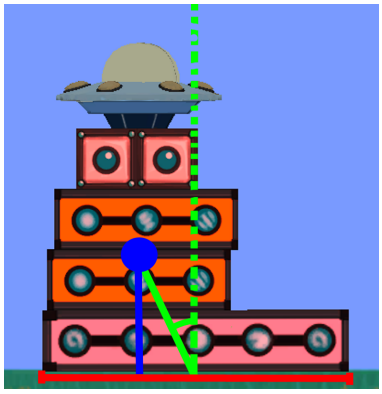
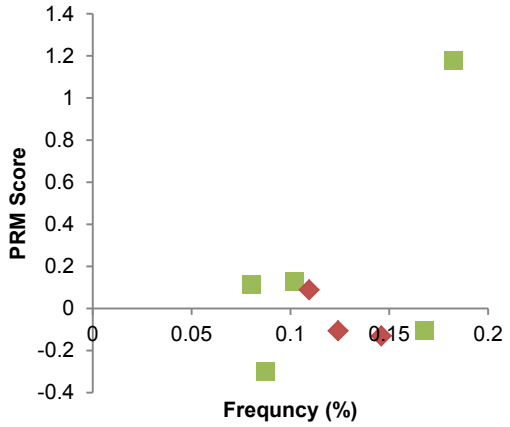


**Figure 3. Examples of the 3 Principle-Relevant Metrics (PRMs) used in *RumbleBlocks*. The red line represents the width of a tower's base. The blue line represents the height of its center of mass, and the measure of the green angle is used for symmetry**

### Interpretation

To interpret the data, we are interested in finding two particular patterns within the representative solutions. In the first pattern, we are looking for instances where the game is providing incorrect feedback, either negative feedback to a principled solutions or positive feedback to an unprincipled solutions. To quantify this pattern, we compute a measure of misalignment for each level. This measure is calculated by taking the average error of the representative solutions, where the error for a solution is defined by its total distance (in terms of PRM score) from where it should be based on its correctness designation. For example, correct solutions with a positive PRM score would have an error of 0, whereas correct solutions with a negative PRM score would have an error equal to -1 times their PRM score. Similarly, incorrect solutions that have a negative PRM score would have an error of 0 and incorrect solutions that have a positive PRM score would have an error equal to their PRM score. We did not weight the solutions' error by their observed frequency because we are interested in finding places where the game provides misaligned feedback in general and not where it provided misaligned feedback to the play-testing population, which may differ somewhat from the release population.

In addition to misalignment, which exhibits incorrect feedback, we were also interested in identifying situations where we might be giving inconsistent feedback. To identify these levels, we compute a metric we call *discrimination*, which we define as the absolute difference between the average of the PRM scores for the correct and for the incorrect solutions. This measure is useful for identifying levels that have correct and incorrect solutions that receive similar



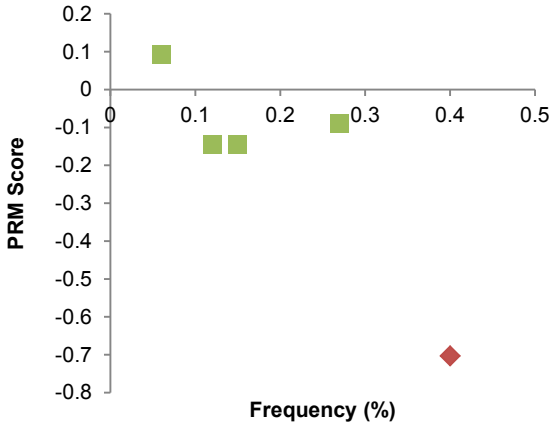




**Figure 4. Two examples of the PRM score vs. Frequency charts. The x-axis shows the frequency with which students used each solution. The y-axis shows the PRM-score of each solution on the target PRM for that level. Green squares denote solutions that are >50% successful while red diamonds denote solutions which are ≤50% successful. The example on the left contains a prominent positive cluster while the example of the right contains a prominent negative cluster. Both of these examples show what would be considered good alignment to educational goals where highly principled solutions tend to succeed and highly unprincipled solutions tend not to succeed.**

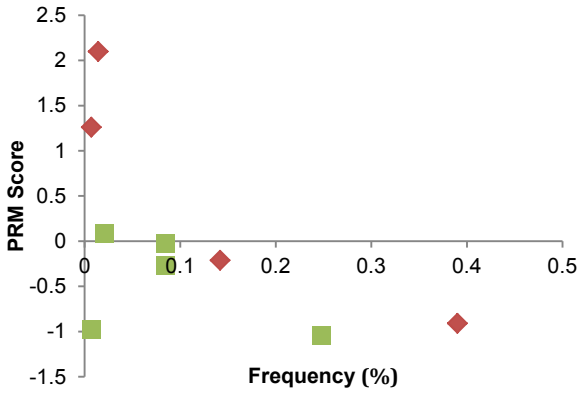



**Figure 5. A plot of frequency of solution (as a percentage) vs. PRM score for all of the clusters on the Symmetry_07 level of RumbleBlocks.**

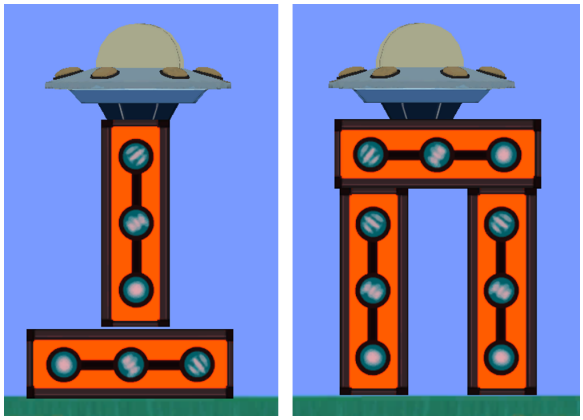


**Figure 6. Two example student solutions to the Symmetry_07 level. The solution on the left comes from a majority unsuccessful cluster while the solution on the right comes from a majority successful cluster.**

PRM scores. This measure suggests places where some additional missing knowledge, in addition to the PRM, is determining success and failure.

Once levels of interest were identified, we plotted their data on charts like the ones show in Figure 4. In these charts each point represents a single representative solution, as determined by the clustering process. The y-coordinate of each point denotes that solution's normalized PRM score, while the coloring and shape of the point denotes the majority success designation of the representative solution. The x-coordinate shows what percentage of users in the testing population used a solution similar to the representative solution. We re-introduce usage frequency at this stage to help prioritize attention when grounding the data with corresponding screenshots. As per the 2x2 matrix in Figure 2, the ideal pattern to be observed is for successful points to appear high on the graph, and unsuccessful points to appear low on the graph.

For the purposes of this paper we will focus on two levels: one with high misalignment and another with low discrimination. For each of these levels, we visually inspected the PRM score vs. frequency plot. The first less desirable level is shown in Figure 5. In this example, there are two highly frequent solutions, the 2 points farther to the right, where one is mostly successful and the other is mostly unsuccessful, however they do not differ strongly in their PRM scores. When we examine screenshots of student solutions to this level we see 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 is considered a failure while the other is considered a success. This represents *RumbleBlocks* giving inconsistent feedback to players about the symmetry principle.

Another anomalous example can be seen in Figure 7, which shows a plot of the different solutions to the level CenterOfMass_10_PP. This level is used as part of an in game pre-post design. Because the level was used for pre- and posttest, it omits the energy ball 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 ball 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 2 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 8, 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 or not the spaceship falls off of 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 2, 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.

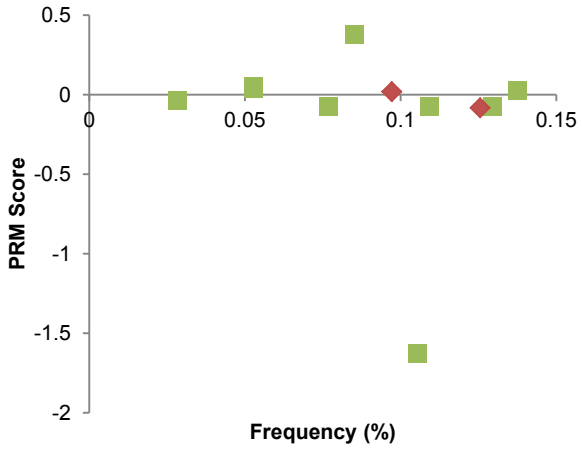



**Figure 7. A plot of frequency of solution (as a percentage) vs. PRM score for all of the clusters on the CenterOfMass_10_PP level of RumbleBlocks.**

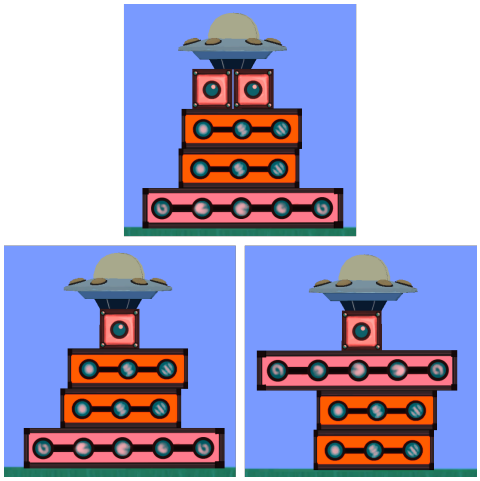


**Figure 8. Examples of student solutions on the CenterOfMass_10_PP level. The solution at the top comes from one of the majority successful clusters while the two at the bottom come from majority unsuccessful clusters.**

The patterns we 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, we wanted to see if there was further support in the structural data to support our conclusions. To do this we used the structural features generated through the conceptual feature extraction process and used a Chi$^2$ analysis to identify which structural features present in student solutions were more predictive of success. As the feature engineering process generated 6,010 potential features that could be present in a student solution, we had to apply Bonferroni correction to the analysis to control for the number of statistical tests. Fifteen substructures were found to be statistically significantly correlated with success after correction, however eleven of those substructures corresponded to the same grounded structure (referred to as "NT37") ultimately leaving five significant substructures, shown in Figure 9. Four of the structures, shown in the red region to the left of the figure, were negatively correlated with success while the remaining one, shown in the green region on the right, was positively correlated with success. At a high level, the same issue can be seen in Figure 9 as in Figures 6 and 8, where putting the spaceship on top of a single square block seems to lead to failure more commonly than putting it on top of a wider platform. This analysis would suggest that there is a more widespread issue with the mechanics of *RumbleBlocks* that goes beyond mere level design.

## DESIGN IMPLICATIONS

As demonstrated by our analytical results, there is a clear misalignment between the feedback provided to students in *RumbleBlocks* and the principles that the game is trying to teach. This misalignment appears both in how discriminating the principles are in terms of success, as in the Symetry_7 example, and in potential troubles contributed by the secondary success criteria of the spaceship having to remain on top of the tower, shown in the CenterOfMass_11 example and the Chi$^2$ analysis. The results of our analysis can be filtered through the theories of the EDGE educational game design framework to reason about what the designers of *RumbleBlocks* might do to bring their game back into proper alignment.

Providing feedback that students can use to evaluate their conceptual understandings, is an important principle of learning that is advocated by many theoretical frameworks in the learning sciences [7,14,22]. Examining the alignment results, we see that there are cases where adherence to the target principle of a level does not translate into success for players. This would mean that if we want to have players learn the current principles of the game, we are not providing feedback that will actually help them to attend to their errors in thinking about those principles. If we look at the case of Symmetry_7 we see a pattern where successful and

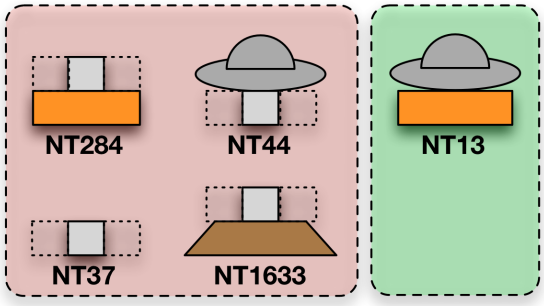



**Figure 9. Rendered results of a Chi$^2$ analysis of structural features in *RumbleBlocks* which predict the success of a tower in the earthquake. Student solutions which contained the features in the red shaded region to the left were more likely to be unsuccessful in the earthquake while solutions which contained the feature in the green region to the right were more likely to be successful.**

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 PRM score on both of the other two principles, wide base and low center of mass. This is because the principles used in *RumbleBlocks* are meant to be applied to single connected body structures, while the towers in the game are stacks of disconnected blocks. Using a design that involved connected structures was actually considered in the preliminary design phases of *RumbleBlocks* but initial prototype testing indicated that players found the disconnected structure to be more fun to play with. In terms of EDGE's Mechanics Dynamics and Aesthetics component, the designers also thought that having a disconnected block mechanic would allow for more interesting dynamics in the design. This situation highlights how difficult it can be to find a good balance between engagement and alignment and emphasizes the importance of analytics to help navigate this challenge.

One way that the designers of *RumbleBlocks* might address this issue is to consider new mechanics for the game that can better address the goals of the game. One such solution would be to introduce a mechanic that allows players to glue blocks together so that the blocks act more like the connected structures modeled by the principle-relevant metrics. As previously noted by the designers, this mechanical change would cause a resultant change in the dynamics experienced by the players because fully connected structures would 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 balls 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 entirely. This solution would address the alignment problems highlighted in the CenterOfMass_10_PP example as well as the $Chi^2$ 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 mechanical changes can be filtered through the EDGE framework's MDA component and weighed against what the designers' wish to emphasize in redesign.

As an alternative to mechanical alterations to the game, the *RumbleBlocks* designers might choose to instead alter the educational goals to fit the experience they already have. While changing the goals of a game's design is generally uncommon, it may constitute a valid way of solving the problem. This solution would necessitate an examination of what kinds of knowledge players are employing as they interact with the game. To facilitate this exploration, we could develop a new set of principle-relevant metrics and see how success in the game aligns to these new metrics as opposed to the old ones. Keeping the EDGE framework's educational objective's component in mind, any alteration to the goals of the game would require that the designers create new external pre-posttests to allow for the measurement of how effective the game is at teaching the new goals.

While the EDGE framework helps to frame the directions the designers might take in iteration, it cannot easily inform which solution would be best. Our approach, along with *RumbleBlocks'* log replay system, could be used to consider the implications of each change. For example, if the designers wanted to explore adding in a glue mechanic between the blocks, the logs could be replayed with the change in place to examine how the various solutions would react differently in the earthquake. Alternatively, if the designers wanted to entertain changing the goals of the game then they could implement a new metric calculation within the replay system and see how this new metric performs as a principle-relevant metric for alignment. Rigorously evaluating which option is most promising would require the designers to run new playtests with players in their target demographic, but our proposed method of utilizing the replay system is a much less expensive way to focus on the most promising design.

## DISCUSSION

We have demonstrated our approach to evaluating the alignment of an educational game against its stated educational goals. The technique relies on analytics of student gameplay data from which principle-relevant metrics can be calculated along with the structure of student solutions. Additionally, we have shown how these kinds of analysis can be employed while thinking across the components of the EDGE educational game design framework. It is our intention that this demonstration could serve to inspire the use of similar techniques to be employed in other educational game settings.

We believe that our process of evaluating the alignment of an open-ended educational game can be applied to other game contexts. For example one could imagine an educational game designed to teach the basics of planetary motion where the structural relations of student solutions would describe which bodies are orbiting which, rather than the simple rectangular adjacencies present in *RumbleBlocks*.

We also believe the process could be used for other educational games whose goals lie outside the domain of physics. Consider a hypothetical example of using McCoy's *Prom Week* [23], a social simulation game, to teach the nuances of the concept of social capital [29]. Students could be tasked with navigating a social simulation with the goal of

asking their ideal date out to prom, but to do so they must properly curate their social capital in order to be able to ask their friends for assistance. The structural elements of student solutions in this example would be the structure of the social graph that students form as they play. A principle-relevant metric could be a quantification of their avatar's social capital, using any of the methods that exist in the social capital literature. It would be pedagogically important that accruing more social capital in the game actually leads to success in the game; otherwise students might learn to incorrectly conceptualize social capital.

At its core, our approach requires a way of grounding the educational objectives of a game in some kind of measurement. This measure could be a simple metric calculable from a game state, as in our *RumbleBlocks* example, or it could be a more complex composite measure. Secondly, our approach requires a way of capturing the space of solutions players employ on in-game challenges. For some games, such a space is straightforward to calculate but for more open-ended games, such as RumbleBlocks, it is helpful to be able to characterize solutions in terms of structural features. The particular systems we have used in our work are not strictly necessary for the general process, though they have simplified the process greatly.

While we believe our approach is sufficiently general to be applied to many games, it is important to note contexts where it might be less applicable. The notion of a principle-relevant metrics may not be appropriate to contexts where the content of a game is less easily quantified, for example in a serious game designed to change players' attitudes toward international conflicts [2]. Another place where the approach may be less useful is in games where the structural composition of a solution is less relevant to the educational goals of a game than the sequence of task that players perform. In these cases it may be more useful to generate features for student solutions based on the paths students take to arrive at the final solution rather than the solution itself. This would be similar to the work of Andersen *et al.* who use Playtracer to map players' paths through a game space [3].

In future work, we hope to explore additional ways to help educational game designers detect, diagnose, and resolve issues of alignment between their game designs and educational goals. In particular, we might investigate how to use the metrics in our current approach in a way that would not require as much visual inspection. If such an approach were developed, it could greatly reduce the load on designers looking to evaluate their games.

## CONCLUSION

In this paper, we have demonstrated an approach to exploring how well an educational game's mechanics align to its stated target knowledge. Not only have we been able to measure whether or not an alignment issue exists, but we have also demonstrated how to employ an educational game design framework in thinking through how any alignment issues might be mitigated. It is our belief that the techniques we have shown here are generalizable to other contexts and would be a helpful aid in honing the design of other educational games. We hope others are able to find the techniques useful in their own situations.

## ACKNOWLEDGMENTS

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 express in this article are those of the authors and do not necessarily reflect the position of the sponsoring agency.

## REFERENCES

1. Aleven, V., Myers, E., Easterday, M., and Ogan, A. Toward a Framework for the Analysis and Design of Educational Games. *Proc. DiGITEL 2010*, IEEE (2010), 69–76.
2. Alhabash, S. and Wise, K. PeaceMaker: Changing Students' Attitudes Toward Palestinians and Israelis Through Video Game Play. *International Journal of Communication 6*, (2012), 356–380.
3. Andersen, E., Liu, Y., Apter, E., Boucher-genesse, F., and Popovi, Z. Gameplay Analysis through State Projection. *Proc. FDG 2010*, ACM Press (2010).
4. Andersen, E., Liu, Y., Snider, R., Szeto, R., Cooper, S., and Popovi, Z. On the Harmfulness of Secondary Game Objectives. *Proc. FDG 2011*, ACM Press (2011).
5. Christel, M.G., Stevens, S.M., Maher, B.S., et al. RumbleBlocks: Teaching Science Concepts to Young Children through a Unity Game. *Proc. CGames 2012*, IEEE (2012), 162–166.
6. Clark, D.B., Tanner-Smith, E.E., and Killingsworth, S. *Digital Games for Learning: A Systematic Review and Meta-Analysis (Executive Summary)*. Menlo Park, CA, 2013.
7. Gee, J.P. *What video games have to teach us about learning and literacy*. Palgrave Macmillan, New York, 2003.
8. Harpstead, E., Maclellan, C.J., Koedinger, K.R., Aleven, V., Dow, S.P., and Myers, B.A. Investigating the Solution Space of an Open-Ended Educational Game Using Conceptual Feature Extraction. *Proc. EDM 2013*, (2013).
9. Harpstead, E., Myers, B.A., and Aleven, V. In Search of Learning: Facilitating Data Analysis in Educational Games. *Proc. CHI 2013*, ACM Press (2013), 79.
10. Hunicke, R., Leblanc, M., and Zubek, R. MDA : A Formal Approach to Game Design and Game Research.

*Proc. of the AAAI Workshop on Challenges in Game AI*, (2004), 1–5.

11. Isbister, K., Flanagan, M., and Hash, C. Designing Games for Learning : Insights from Conversations with Designers. *Proc. CHI 2010*, ACM Press (2010), 2041–2044.
12. Johnson, M.W., Eagle, M., and Barnes, T. InVis : An Interactive Visualization Tool for Exploring Interaction Networks. *Proc. EDM 2013*, (2013), 65.
13. Khaled, R. and Ingram, G. Tales from the Front Lines of a Large-Scale Serious Game Project. *Proc. CHI 2012*, ACM Press (2012), 69–78.
14. Koedinger, K.R., Corbett, A.T., and Perfetti, C. The Knowledge-Learning-Instruction Framework: Bridging the Science-Practice Chasm to Enhance Robust Student Learning. *Cognitive Science 36*, 5 (2012), 757–98.
15. Koedinger, K.R. and Nathan, M.J. The Real Story Behind Story Problems: Effects of Representations on Quantitative Reasoning. *The Journal of the Learning Sciences 13*, 2 (2004), 129–164.
16. Kohavi, R., Deng, A., Frasca, B., Longbotham, R., Walker, T., and Xu, Y. Trustworthy Online Controlled Experiments: Five Puzzling Outcomes Explained. *Proc. KDD 2012*, ACM Press (2012), 786–794.
17. Kohavi, R., Henne, R.M., and Sommerfield, D. Practical Guide to Controlled Experiments on the Web : Listen to Your Customers not to the HiPPO. *Proc. KDD 2007*, ACM Press (2007), 1–9.
18. Leighton, J.P. and Gokiert, R.J. The Cognitive Effects of Test Item Features: Informing Item Generation by Identifying Construct Irrelevant Variance. *Proc. NCME 2005*, (2005), 1–26.
19. Linehan, C., Kirman, B., Lawson, S., and Chan, G.G. Practical, Appropriate, Empirically-Validated Guidelines for Designing Educational Games. *Proc. CHI 2011*, ACM Press (2011), 1979–1988.
20. Lomas, D., Forlizzi, J.L., Koedinger, K.R., and Patel, K. Optimizing Challenge in an Educational Game Using Large-Scale Design Experiments. *Proc. CHI 2013*, (2013), 89–98.
21. Maciuszek, D. and Martens, A. Patterns for the design of educational games. In F. Edvarsen and H. Kulle, eds., *Educational Games Design Learning and Applications*. Nova publishers, 2010, 263–279.
22. Mayer, R.E. and Moreno, R. Nine Ways to Reduce Cognitive Load in Multimedia Learning. *Educational Psychologist 38*, 1 (2003), 43–52.
23. Mccoy, J., Treanor, M., Samuel, B., Reed, A.A., Wardrip-fruin, N., and Mateas, M. Prom Week. *Proc. FDG 2012*, ACM Press (2012), 235–237.
24. Miller, C.S., Lehman, J.F., and Koedinger, K.R. Goals and learning in microworlds. *Cognitive Science 23*, 3 (1999), 305–336.
25. Nacke, L.E., Drachen, A., and Göbel, S. Methods for Evaluating Gameplay Experience in a Serious Gaming Context. *Journal of Computer Science in Sport 9*, 2 (2010), 1–12.
26. Nathan, M.J. Knowledge and Situational Feedback in a Learning Environment for Algebra Story Problems. *Interactive Learning Environments 5*, 1 (1998), 134–159.
27. Nielsen, J. and Molich, R. Heuristic Evaluation of User Interfaces. *Proc. CHI 1990*, ACM Press (1990), 249–256.
28. Pashler, H., Bain, P.M., Bottge, B.A., et al. *Organizing Instruction and Study to Improve Student Learning (NCER 2007-2004)*. Washington, D.C., 2007.
29. Putnam, R.D. Bowling Alone: America's Declining Social Capital. *Journal of Democracy 6*, 1 (1995), 65–78.
30. Rau, M.A., Aleven, V., and Rohrbach, S. Why Interactive Learning Environments Can Have It All : Resolving Design Conflicts Between Competing Goals. *Proc. CHI 2013*, ACM Press (2013), 109–118.
31. Seif El-Nasr, M., Drachen, A., and Canossa, A., eds. *Game Analytics*. Springer London, London, 2013.
32. Wiggins, G. and McTighe, J. *Understanding by Design*. Pearson, 2005.