It’s Dangerous to Go Alone: A Data-Driven Approach to Hint Design in a Puzzle Game

Anuj Pasricha
University of Illinois (U-C)
1406 W. Green St.
pasrich2@illinois.edu

Sinduja Subramaniam
University of Illinois (U-C)
201 N. Goodwin Av.
ssbrmnm2@illinois.edu

Helen Wauck
University of Illinois (U-C)
201 N. Goodwin Av.
wauck2@illinois.edu

ABSTRACT
Players typically find themselves quitting a video game in utter frustration, either because the game-play is difficult to understand, or because they find themselves helpless and cannot get to their goal state. In this paper, we focus on the latter reason for this frustration by presenting a framework for utilizing the data collected from a puzzle-based game, to give players hints which help them play more strategically and therefore enhance learning. We present our measure of strategic play (number of explored squares on the game board) and describe how this measure can be used to generate hints that assist the user. We did not notice a significant difference in the number of explored squares between the original version of the game and a version with hints. However, our main contribution was to come up with a measure of strategic play which displayed a consistent correlation with game outcome across all tests.

Author Keywords
educational games; game design; data-driven design; hints

ACM Classification Keywords
K.8.0 [Personal Computing]: General – Games; H.5.0 [Information interfaces and presentation]: General; K.3.1 [Computer and education]: Computer Uses in Education - Computer-Managed Instruction (CMI)

INTRODUCTION

Much research has already investigated the potential cognitive and learning benefits of video game playing, including increased response time [1], increased spatial skills, and if the game contains educational content, increased motivation and engagement in subject matter relative to traditional classroom lectures [2]. But the benefits of playing games are lost if the player finds them too frustrating, tedious, or hard, and gives up. We define this as a poor *player experience.* Games can be made easier by adding hints that give away information about how to progress through the game, but such hints are generally one size-fits-all and may be either too generic to be reduce the game’s difficulty or too specific, giving away information too readily and making the game too easy, reducing the amount of learning and boring players to the point that they may quit. Existing hint systems for educational games are typically text-based hints designed to be triggered only after the player performs a certain “wrong” action [1, 3], or recommender systems for choosing which mini game to play first [2]. Our research focuses on in-game hints, not recommender systems, and text-based hints lack the personalization that we aim to provide with our hints. In addition, we take the view that wrong actions are beneficial in the short term, as the player can potentially learn from them on their own.

This paper presents a case study of a data-driven method for adding hints to a game to strike a balance between too easy and too difficult, and between too generic and too specific, simultaneously enhancing player experience and learning. The problem we address with our hinting system is repetitive nonstrategic behavior that indicates that learning is not occurring and thus progress is not being made towards a solution. We postulate that learning and strategy are related problem-solving concepts; an increase in strategic problem solving (in this context, strategic game-play) indicates that something has been learned. Player experience is related to these ideas as well; strategic game-play may enhance player experience because the player has a better idea of how to approach the problem and is less likely to get frustrated and give up.

Given our time constraints for this project, we selected a simple, pre-existing game to analyze: a spatial puzzle mini game, which we will henceforth refer to as the “statue puzzle.” from the 2006 adventure game *The Legend of Zelda: Twilight Princess.* The game consists of a two dimensional “board” reminiscent of a chess board with certain squares missing. The player and two movable statues are on the board together and their actions are coupled; every time the player turns or moves, so do the statues. The player must guide the two statues to the two specially marked squares in order to win (Figure 1).

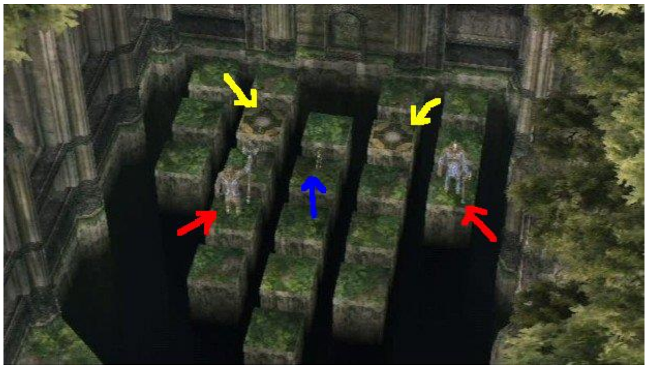


Figure 1. Statue puzzle. Player (blue) must guide statues (red) to the specially marked squares (yellow).

There are many puzzle games available today to choose from, but we selected this particular one for a few reasons. First, it is widely viewed by the gaming community as one of the most frustrating parts of *Twilight Princess*, and many players end up giving up in frustration and looking for a solution online. We wanted to see if adding data-driven hints to the statue puzzle could reduce this frustration in players. Second, there are many possible paths to victory in the statue puzzle minigame, so hints could be personalized for the player depending on the strategy each individual tries. Third, the game is simple, which reduces the complexity of the data we must analyze and allows us to better isolate relevant player data.

Based on our own intuition after playing the game ourselves several times, we developed the hypothesis that accidentally trying the exact same sequence of moves several times in a row was the main source of frustration for players. Therefore, we focused on this metric in the first round of our study, in which 10 volunteers played our implementation of the original statue puzzle game. After the results indicated that these repeated sequences were not the source of the problem, we analyzed the number of unique game states each player's first game had, as this metric has been used in previous work as a proxy for strategic behavior [3]. We discovered a slight correlation between the number of unique game states and whether a player was victorious on their first game, which motivated us to develop a hint system that focuses on increasing strategic behavior. We also discovered that a related metric, the number of unique board squares visited by the player, was even more correlated with player success than game states by visualizing player movement and turns in heat maps as in [5].

We were then able to run the second round of the study, which included the 10 original volunteers as well as 9 new players. All players in our study were assigned to one of four groups, allowing us to measure whether hints affected the difficulty of the game and related metrics like player engagement, frustration, entertainment, and boredom while controlling for the effects of learning through repeated practice.

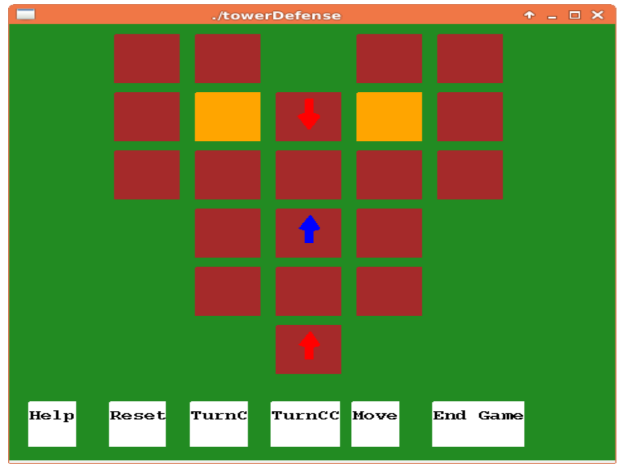


Figure 2. Our statue puzzle implementation. Player (blue) must guide statues (red) to the specially marked squares (yellow).

We found no significant difference between completion time, number of unique game states, or number of unique board squares visited for groups playing with and without hints. There was also a noticeable difference in the number of games each player decided to play among groups; groups playing with hints played more games than those playing without hints. Of the players that completed our post-game survey, most found the version with hints more fun, easier, not as boring, and less frustrating than the version without hints, and rated the helpfulness of hints an average of 2.3 on a scale from 1 to 5. These results suggest that using data to drive hint creation may be a valuable strategy for educational game designers, but more data is needed to statistically validate our results and to investigate whether our methods can be applied to different games.

APPROACH/CONTRIBUTION

This paper presents a data-driven approach that can be used to generate personalized hints and improve the experience of players in puzzle-based games. Our major contribution is that we provide a framework for generating hints from a combined input of user activity and generalized measures. Additionally, we use this framework to drive player motivation and provide the player with an enhanced, more effective experience. We have used measures (such as exploration, repeated sequences, states, basic statistics) that can be taken as a base to generate hints at different points in games, and at the same time, be extended and applied to other games.

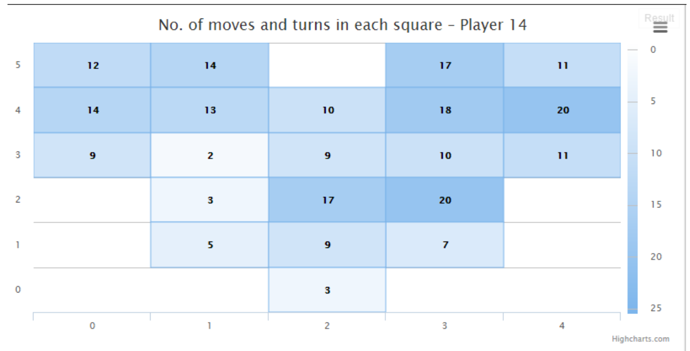
Since our main goal is to generate and use personalized hints to improve user experience, our evaluation methodology is focused on measuring the effectiveness of the hints we generate. Our experiments were conducted on our implementation of the statue puzzle game (Figure 2) to collect user activity data for generating hints and evaluation. We have divided our experiment into two phases - one phase was conducted in mid-October where players played the game without hints, and the other phase was conducted in mid-November where players played either a version of the game with hints or the original version. This was done to compare and study improvement when provided hints to play the same game. This can be used to measure the learning curve between the version without hints and with hints, and if it positively impacts player experience in games. We chose to use a four quadrant study to control for the effects of learning through repeated practice.

Our four quadrant study method consists of four groups within the two phases of development:

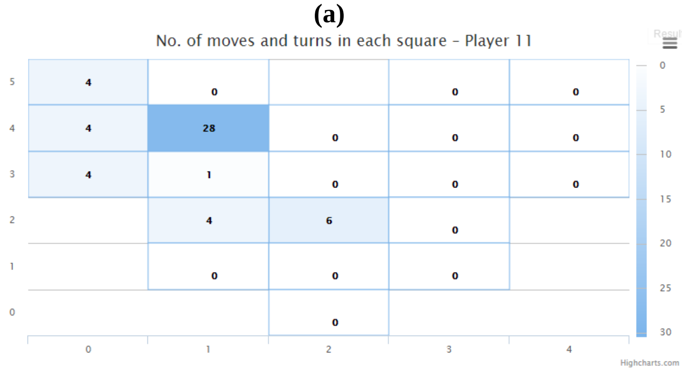
- **Orig(P1)+Orig(P2)**: Set of players who played the original version of the game (without hints) in both phases.
- **Orig(P1)+Hints(P2)**: Set of players who played the original version during first phase and version with hints in the second phase.
- **Orig(P2)**: Set of players who played only the version with hints and only in second phase.
- **Hints(P3)**: Set of players who played only the version with hints and only in second phase.

First, we used exploration of the game board as a measure of how strategically someone is playing. The more the player explores different game states, the more he/she is considered to be playing strategically (O'Rourke et al). We want to see if playing a version of the game with hints increases the number of unique squares a player visits, indicating an increase in strategic play. Our primary motivation behind using the number of explored squares as a measure for strategic play stems from our results from preliminary testing.

We initially looked into repeated sequences of data, based on the intuitive idea that the more a person gets stuck in a loop-like behavior, the harder it is for them to make progress towards a solution. However, we found after the first phase of our study that players did not get stuck in repeated sequences as often as we had thought, and when they did, the particular sequence repeated was generally unique to that player. We then decided to use the approach presented by O'Rourke et al - strategic behavior based on the number of unique game states a player traverses during the game. We noted that both the number of unique game states and the number of unique board squares visited correlated with the game outcome. We chose to focus on squares visited rather than game states since it was a simpler measure to calculate and easier to visualize. The striking difference in squares visited between a player who eventually gave up and a player who won the game can be seen in Figure 3 below.



(a)



(b)

Figure 3. Darker shades of blue indicate more moves from a square. (a) Heatmap for winning player shows more explored states (more overall blue spread). (b) Heatmap for player who eventually quit shows fewer explored states (less overall blue spread).

We implemented hints in the statue puzzle game as follows. Every 30 moves, if a player has not made sufficient progress in exploring the board (has visited fewer than 18 distinct squares and has not explored more than 1 additional distinct square since the last hint was given), the hint is displayed directly on the game board. The next time the player moves, the hint vanishes. We chose a value of 18 squares for the hint threshold based on the average difference between the numbers of squares explored by eventual winning players and eventual losing players, and based the rest of our constraints on our own experimentation with the game. The hint temporarily changed the color of all unvisited squares to white and all visited squares to black, and displayed text encouraging the player to try to get to the squares that hadn't been visited yet (Figure 4).

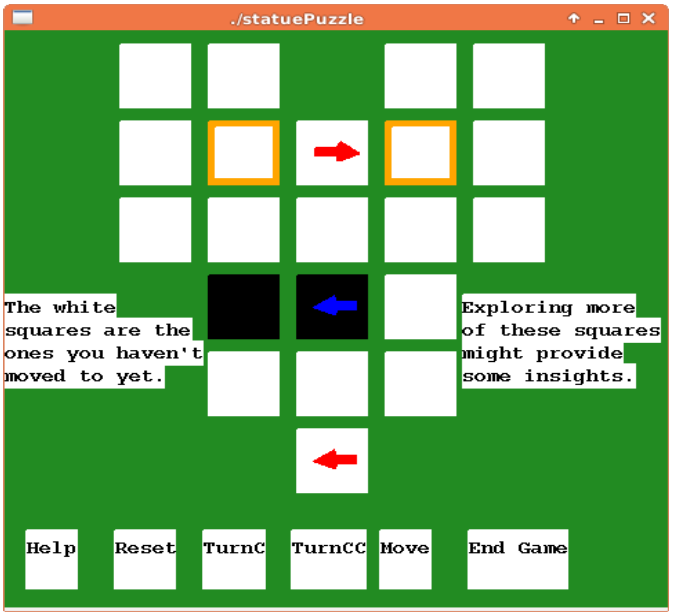


Figure 4. Example of a hint in-game. The player is shown which squares haven't been explored yet (white) and encouraged to explore them.

Secondly, we measure the time spent playing in each game. This gives us a proxy measure of engagement and motivation of the player. If the player is playing the game voluntarily with no outside pressure to keep playing, then presumably, the longer a player plays the game, the more motivated and/or engaged they are. We can use time spent playing as a means to gauge improvement in player experience and learning in addition to measures of strategic game-play. We also analyzed the series of attempts of the game made by each player to determine the effect repeated practice has on player experience and learning.

Finally, we asked players to complete a post-game survey that can gave us information about player opinions about each version of the game they played. This serves as another measure of improvement in player experience. We want to see if providing hints improves player experience as measured by the players' opinions on each version of the game.

RELATED WORK

In order to come up with a valid test setup and good measures for strategic play, we surveyed a number of papers on educational games and user interaction with games, in general.

Lee et al present a very simple solution for dealing with massive action spaces in open-ended educational games. Their data-driven approach learns player behavior in such games based on their previous interactions with the game itself. The generalizable and domain-expertise independent scheme provided a very strong basis for our initial path to understanding user interaction in games. However, our preliminary results did not show a very strong correlation between a player's previous interactions with the game and the outcome for that particular game. We are trying to measure how strategically a user interacts with the game, which does not necessarily correlate with how much a player "learns," which is what the paper was trying to predict.

Conati et al present an approach for making education games more effective, such that they actually trigger learning and not simply entertain the player. The approach generates probabilistic models for each player and uses these intelligent agents to provide personalized instruction to each student, while sustaining the "entertainment factor" of the game itself. We did not utilize probabilistic models due to project timeline constraints, however we gave a group of players hints to measure whether they were actually learning or not. Developing probabilistic models for each player might prove to be a worthy pursuit if we were given more time.

O'Rourke et al present an approach for understanding the demographic variance in player behavior across different age groups. Based on the concept of strategic behavior and non-strategic behavior, the authors conclude that younger players display less strategic behavior, while older players display more strategic behavior. The notion of strategic play used in our project is derived primarily from this paper. However, we are not concerned with differentiating between age groups - our test subjects belong to the college-level age group (18-21 years). Our primary concern lies in picking game board features to distinguish between less strategic play and more strategic play as a way to predict if a player will with the given game or not.

Overall, different aspects of our approach are inspired from different aspect of the papers described above. We utilized key ideas from each paper to develop a wholesome approach to suit our needs because not any single paper fit our needs perfectly, as described above.

EVALUATION AND RESULTS

Overall, our results show a high degree of variance. Because of this and our small sample size, we provide a mostly qualitative analysis of results here. We also remove outliers from the data (games where the player made less than 4 actions before quitting). While we saw correlations between the number of unique game states and game outcome and between unique board squares visited and game outcome, we did not observe a correlation between time spent playing and game outcome. Given the high correlation between unique game states and unique board squares visited, we focus our evaluation on unique board squares visited, the simpler measure. We found no significant difference between completion time, number of unique game states, or number of unique board squares visited for groups playing with and without hints. There were, however, some potentially significant differences among all four groups overall.

Number of Unique Squares Visited By Treatment Group				
	Orig+Orig	**Orig+Hints**	**Orig**	**Hints**
Mean	12.50	17.00	16.50	18.00
Median	13.33	15.44	14.25	15.73

Table 1: Basic statistics for number of unique board squares visited across all 4 groups.

Players in Orig+Orig visited fewer unique board squares than any other treatment groups (Table 1). This may be due to the fact that this group is the only treatment group that played the same version of the game in both study phases, and therefore there may be a familiarity effect at play. Orig's game times were much longer than those of other groups, and had much higher variance than any other groups (Figure 5). This may be simply due to the individual variance between players given our small sample size. However, of the other 3 groups, Both of the groups that participated in both phases of the study (Orig+Orig and Orig+Hints) showed lower game times than groups that did not (Orig and Hints). This may be an indication that either Orig+Orig and Orig+Hints were more efficient at winning the game due to repeated practice, or that these groups had less patience and motivation to keep playing because they had already played the game in the first phase.

There was also a noticeable difference in the number of games each player decided to play among groups; groups playing with hints played more games than those playing without hints; Orig+Orig and Orig had a total of 6 and 7 games, respectively, and Orig+Hints and Hints 8 and 10 total games, respectively. This may be an indication of increased player motivation in the hint groups.

Of the players that completed our post-game survey, most found the version with hints more fun, easier, not as boring, and less frustrating than the version without hints, and rated the helpfulness of hints an average of 2.3 on a 1 to 5 Likert scale. Average Likert scale ratings for fun, easiness, boredom, and frustration were the same across treatment groups.

DISCUSSION AND FUTURE WORK

We presented a data-driven approach for reducing the frustration players experience in a typical puzzle-based game and enhancing learning through increasing strategic gameplay. This approach relies on our metric for strategic play, the number of unique board squares visited, which displayed a consistent correlation with game outcome (win or loss) across both phases of our study. Game time differences between players who won the game and those who lost, another metric we used as a proxy for player learning, were not significant. This may indicate that time spent playing is a poor proxy for player motivation and

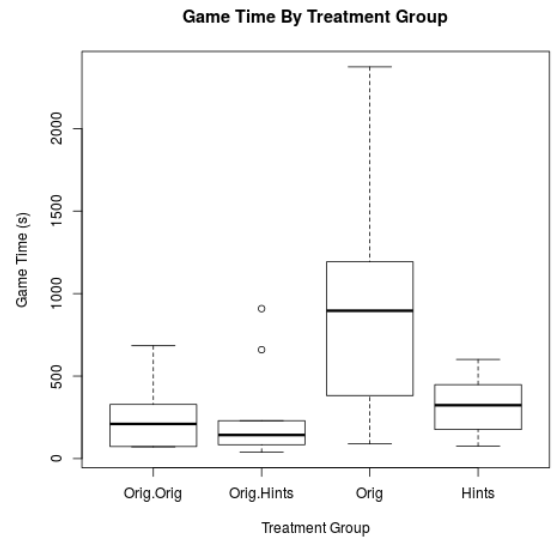


Figure 5. Orig shows longer game times and game time variance than other groups. Both Orig+Orig and Orig+Hints have shorter game times than Orig and Hints.

engagement; a short game may indicate either a non-engaged player or an efficient player. Additionally, conducting these tests on a larger scale would certainly help obtain more significant differences.

Our survey results suggest that players might prefer the version with hints, but only 5 players played both versions, so it is hard to draw definitive conclusions from their preferences. Players may also have been biased towards preferring the version with hints since they may have felt that it was the version that they were "supposed" to like better. Once again, we need more data to be sure and perhaps a better way of phrasing survey questions to reduce bias. Given the ambivalent response to our data-driven hints, it is necessary to do more experiments testing different hint triggering parameters to determine if these hints provide an improvement over the original game. A side by side comparison could also help determine how our data-driven, visual hints compare in effectiveness with other types of hints.

Our work has been limited to a semester of effort and includes a subset of measures that we can use to generate personalized data-driven hints. One future work could be to identify more measures that are general from data. In this context, a future study could also include a study of whether using a mixture of general and specific measures would work better or not. For example, in this study we focus on measures that are very general and can be applied to most games. Instead, we could form categories of games, and identify some general measure sand some measures that are more specific to the category to provide better hints [7].

Another future work can be to change and make the games more adaptive using the player activity data. In a fashion similar to computer adaptive tests, there could be a difference in game difficult as and when they player is playing the game too well to make it more challenging [6]. There could be subtle increase in difficulty level, introduced either by game design or difficulty level of comprehension of hints provided.

Our experiments were limited to single player games. Similar approach when applied to multi-player games might pose challenges such as “Which hint to provide to which user?”, “How to provide collaborative hints?”, “Do hints change depending on whether the players are against each other in the game, or working together to reach a goal against two other players who are together? How can same hints be given to some and different to others?”

To conclude, while there are many other avenues that can be explored for this project, our major contribution for coming up with a good metric for strategic play still holds consistently across all variations of the game.

ACKNOWLEDGMENTS

We thank all the participants for giving us valuable feedback on game design and helpfulness of hints. We would also like to thank Ranjitha Kumar for giving us the basic knowledge necessary to develop a useful data-driven system.

REFERENCES

1. Conati, Cristina, and Xiaohong Zhao. "Building and evaluating an intelligent pedagogical agent to improve the effectiveness of an educational game." *Proceedings of the 9th international conference on Intelligent user interfaces*. ACM, 2004.
2. Kalloo, Vani, and Permanand Mohan. "Personalized game based mobile learning to assist high school students with mathematics." *Advanced Learning Technologies (ICALT), 2010 IEEE 10th International Conference on*. IEEE, 2010.
3. O’Rourke, Eleanor, et al. "The effects of age on player behavior in educational games." *FDG*. 2013.
4. Lee, Seong Jae, Yun-En Liu, and Zoran Popovic. "Learning Individual Behavior in an Educational Game: A Data-Driven Approach.”
5. Drachen, Anders, et al. "Guns, swords and data: Clustering of player behavior in computer games in the wild." *Computational Intelligence and Games (CIG), 2012 IEEE Conference on*. IEEE, 2012.
6. Lütticke, Rainer. "Problem solving with adaptive feedback." *Adaptive Hypermedia and Adaptive Web-Based Systems*. Springer Berlin Heidelberg, 2004.
7. Etheredge, Marlon, Ricardo Lopes, and Rafael Bidarra. "A Generic Method for Classification of Player Behavior." *Second Workshop on Artificial Intelligence in the Game Design Process*. 2013.