Evaluating Planning-Based Experience Managers for Agency and Fun in Text-Based Interactive Narrative

Alejandro Ramirez[1] and **Vadim Bulitko**[1] and **Marcia Spetch**[2]
Departments of Computing Science[1] and Psychology[2]
University of Alberta
Edmonton, Alberta, T6G 2E8, CANADA
{ramirezs|bulitko|mspetch}@ualberta.ca

Abstract

Artificial intelligence (AI) techniques have been applied to video games to make the overall experience more enjoyable. In games with interactive storytelling (IS), player actions can substantially affect plot events and plot characters. Therefore, AI planning techniques have been used to shape the plot in response to player actions that conflict with authorial goals. While such methods are poised to increase player fun and agency, two recent implementations (ASD and PAST) have not been formally evaluated to date. In this paper we do so via a series of user studies for the first time. We show that ASD significantly enhances fun and agency, whereas PAST gets mixed results with an interaction between effects of the experience manager and player prior gaming experience in one user study, and marginally significant results for increased agency in a study with a constrained story domain.

Introduction

Research on *interactive storytelling (IS)* is at the intersection of artificial intelligence (AI), interactive entertainment, and traditional storytelling. It is important both because storytelling is deemed to be cognitively rich and because implementing such a storyteller with AI is likely to make video games more interactive, narratively complex and customized for a particular audience.

IS systems have been usually implemented around the concept of a *drama manager*: "an intelligent, omniscient, and disembodied agent that monitors the virtual world" in order to intervene and improve "some model of quality of experience" (Riedl and Bulitko 2013). A generalization of this idea is the *experience manager*, which extrapolates this concept to non-dramatic situations in a wide range of interactive environments (Riedl 2012). In the context of interactive narrative, *fun* (i.e., how entertaining and amusing an interactive experience is) and *agency* (i.e., the ability of the player to affect the virtual world) are frequently deemed important goals targeted by IS systems.

While some experience managers can improve fun and agency in an interactive narrative, they do so by relying heavily on story annotations and extra game content, therefore increasing the authorial burden on the developers.

To alleviate the authorial burden, AI planning techniques have been used to automatically anticipate ruptures in the original narrative due to player actions and repair it, aiming to increase players' agency. Of particular interest is the system called *Automated Story Director (ASD)*, which represents stories as formal plans and computes narrative repairs to retain story goals *a priori* specified by the author (e.g., an evil wolf eats granny in *The Little Red Riding Hood* folk story) (Riedl et al. 2008).

More recently, this approach was extended with a player model which enabled automatic customization of such repairs to the particular player type as detected during gameplay (Ramirez and Bulitko 2012). The resulting system, called *Player-Specific Automated Storytelling (PAST)*, requires the author to additionally specify a mapping between narrative events and player types (e.g., killing the wolf likely indicates a fighter inclination).

Despite their promise, neither of the two experience managers (ASD and PAST) has been formally evaluated with a user study. In this paper we do so in the context of text-based narratives centered on *The Little Red Riding Hood* folk story. The results indicate that ASD can increase player fun and agency compared to a fixed-narrative baseline. On the other hand, the use of player modelling within PAST did not consistently improve fun or agency in one evaluation with a broad story, but improved agency in another evaluation with a more constrained story.

The rest of the paper is organized as follows. We first define the problem formally and then review related work. We then describe the two experience managers in detail and present the empirical test bed employed in the user studies. Finally, we present and discuss the results and conclude the paper with future work directions.

Problem Formulation

The problem tackled in this paper is to determine if planning-based experience managers can deliver on their promise of increasing player fun and agency in the context of an interactive narrative experience. Specifically, the scope is set around planning-based managers that allow player actions outside the exemplar narrative, and that use automated planning to repair the narrative to a consistent state. This can be measured with a *post hoc* instrument (a psychologically validated questionnaire) in the context of a user study.

The standard setup is to have an experimental and a control condition which are different only in the use of the experience manager being evaluated. One important consideration here is the balance of story content across these conditions. This is fundamental to prevent fun and agency measured in the experimental condition from being affected merely by player's experience of a particular influential story content, absent from the control condition. Such balance is usually achieved by having players in the control condition play content experienced by random players from the experimental condition (known as "yoking" in psychology). Consequently, players in both conditions cover similar "story ground" in approximately the same proportion.

Related Work

While several IS approaches have been proposed in the last forty years, from choose-your-own-adventure books by Edward Packard (Kraft 1981) to modern experience and drama managers (Riedl and Bulitko 2013), formal evaluations of their performance have been scarce.

One notable exception is PaSSAGE (Thue et al. 2007; 2010; 2011) which was evaluated via a series of user studies accumulating over one thousand participant hours. While significant effects of using a player model to select content were demonstrated, the experience manager within PaSSAGE was manually scripted and did not use automated content generation techniques.

User studies have also been carried out for performance metrics other than fun and agency. For instance, a suspense manager was evaluated in a user study (Cheong and Young 2008) and significant benefits were found. Data-driven approaches (Yu and Riedl 2012; 2013) and "interestingness" ratings (Sharma et al. 2007) have been used for player modelling. The results indicate that it is possible for an experience manager to model a player cognitively. However, none of these studies evaluated planning-based experience managers for improving player fun and agency.

Planning-Based Experience Management

In this section, we briefly present the foundation of the two experience managers evaluated in the paper. Both ASD and PAST are based on representing narratives as formal plans. To illustrate, we will use an interactive extension of the classical narrative of *The Little Red Riding Hood* in line with the previously published work on ASD and PAST. In the story, a child (`red`) travels through a forest to deliver food to her grandmother (`granny`). She is confronted by an evil wolf (`wolf`) and the player decides in an interactive fashion how the narrative unfolds from there on. Any event in this interactive experience is a planning operator, denoted here with a Lisp-like notation. For instance, the event of the wolf eating the granny is encoded as the planning operator `(eat X Y)` which has the preconditions `(alive X)` and `(hungry X)` and the post-condition `(eaten Y)` with `X` bound to the wolf and `Y` to the granny.

In this approach, the author specifies the so-called *exemplar narrative*. This is done by listing some game events and causal and temporal dependencies among them. Note that the exemplar narrative needs not be linear. For instance, with the original exemplar narrative of ASD (Riedl et al. 2008), the player can experience the events in different temporal orders: the wolf can eat Little Red first or he can eat her granny first (Figure 1).

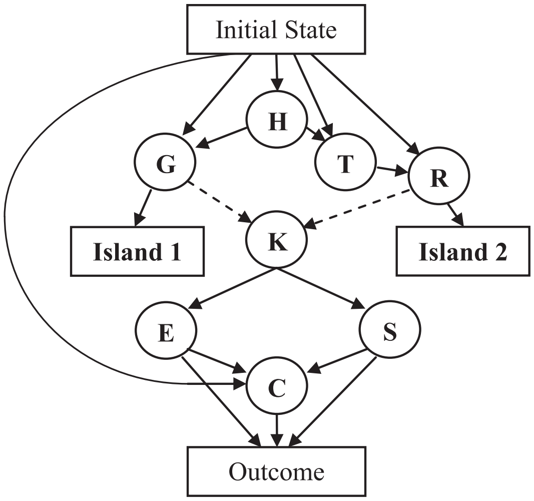
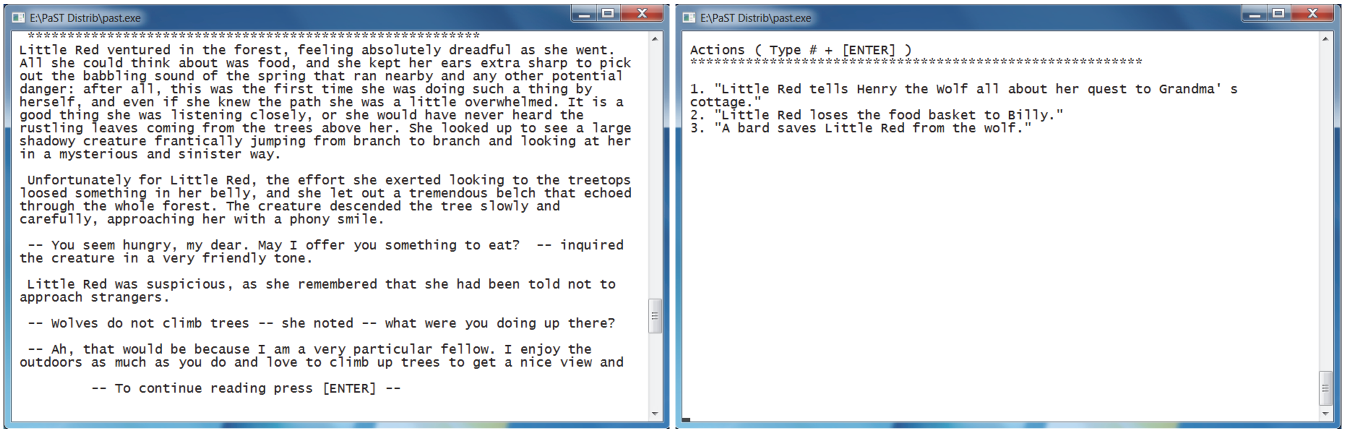


Figure 1: An interactive version of *The Little Red Riding Hood* employed in the studies: solid lines are causal links that can be threatened by ruptures; dotted lines are temporal links. Events: H `(greet red wolf)`, T `(tell-about red wolf granny)`, G `(eat wolf red)`, R `(eat wolf granny)`, K `(kill hunter wolf)`, E `(rescue hunter red)`, S `(rescue hunter granny)`, C `(give granny cake)`. Reproduced from (Ramirez and Bulitko 2012).

The agency the player has in the interactive story can, however, lead to *ruptures* where a precondition of an event is invalidated. For instance, the player could choose to `(kill wolf)` thereby establishing the post-condition of `(:not (alive wolf))` and thus invalidating the authorial goal `(eaten granny)`. Such ruptures can be foreseen by the AI planner and repairs to the narrative can be precomputed in response. The rupture above can be repaired by having a fairy resurrect the wolf. The authorial goals are shown as narrative "islands" and the outcome in the figure.

Frequently, a given rupture can be repaired in multiple ways, necessitating a method to select among them. Whereas ASD used an *ad hoc* method of selecting between contingency narratives, PAST uses a player model to do so. For instance, if the player has demonstrated a fighting inclination then PAST could fix the rupture resulting from `(kill player wolf)` by introducing a fiery pack of wolves. On the other hand, if the player is deemed to be more of a method actor then PAST could have a magic fairy peacefully resurrect the wolf.

In PAST, the player model is a vector of five numbers, each indicating an inclination towards a certain style of play: *fighter*, *power gamer*, *storyteller*, *method actor*, and *tactician*, based on Robin Laws' player types (Laws 2001). This approach was previously found effective within PaSSAGE



```
 **********************************************************
Little Red ventured in the forest, feeling absolutely dreadful as she went.
All she could think about was food, and she kept her ears extra sharp to pick
out the babbling sound of the spring that ran nearby and any other potential
danger: after all, this was the first time she was doing such a thing by
herself, and even if she knew the path she was a little overwhelmed. It is a
good thing she was listening closely, or she would have never heard the
rustling leaves coming from the trees above her. She looked up to see a large
shadowy creature frantically jumping from branch to branch and looking at her
in a mysterious and sinister way.

 Unfortunately for Little Red, the effort she exerted looking to the treetops
loosed something in her belly, and she let out a tremendous belch that echoed
through the whole forest. The creature descended the tree slowly and
carefully, approaching her with a phony smile.

 -- You seem hungry, my dear. May I offer you something to eat?  -- inquired
the creature in a very friendly tone.

 Little Red was suspicious, as she remembered that she had been told not to
approach strangers.

 -- Wolves do not climb trees -- she noted -- what were you doing up there?

 -- Ah, that would be because I am a very particular fellow. I enjoy the
outdoors as much as you do and love to climb up trees to get a nice view and

         -- To continue reading press [ENTER] --
```

```
Actions ( Type # + [ENTER] )
************************************************************

1. "Little Red tells Henry the Wolf all about her quest to Grandma' s
cottage."
2. "Little Red loses the food basket to Billy."
3. "A bard saves Little Red from the wolf."
```

Figure 2: Sample screenshot from the PAST test bed, showing an event description on the left `(greets red wolf)`, and three choices on the right: `(tell-about red wolf granny)`, `(persuade bard wolf)` and `(:not (has red basket))`.

although without any automated planning (Thue et al. 2007). The model is initialized with all components being equal. Afterwards, the vector is adjusted based on annotations of the actions the player takes (e.g., killing the wolf when other options are available indicates a fighter inclination). Each action and rupture in the virtual domain has to be annotated by the author; PAST then uses these to calculate a global annotation for each contingency narrative.

We made minimal changes to PAST from the original design (Ramirez and Bulitko 2012) for our user studies. Specifically, the original PAST precomputed all contingency narratives for all possible ruptures during an off-line phase based on player archetypes. The version of PAST evaluated in this paper forgoes the off-line precomputation. As a result, it saves memory and time off-line and allows the manager to select a narrative repair more closely fitting the player model at hand (instead of one of the five generic model archetypes). Additionally, the planning algorithm is invoked *ad hoc* only when a rupture takes place.

Evaluation

To evaluate ASD and PAST we constructed a text-based interactive adventure–inspired partly by RPG text-based adventure games–around the exemplar narrative of *The Little Red Riding Hood*, in line with the literature (Riedl et al. 2008; Ramirez and Bulitko 2012). The first version was used to evaluate ASD and had three possible ruptures and enough content for the AI planner to repair any of them. The second version was used to evaluate PAST and had five ruptures and three to five possible contingency narratives for each so that a repair suitable for different player types can be chosen. This second version also had a prologue with enough interaction for the system to populate the player model before the first rupture would take place. The third version was also created to evaluate PAST, and had only one rupture and five different repairs available.

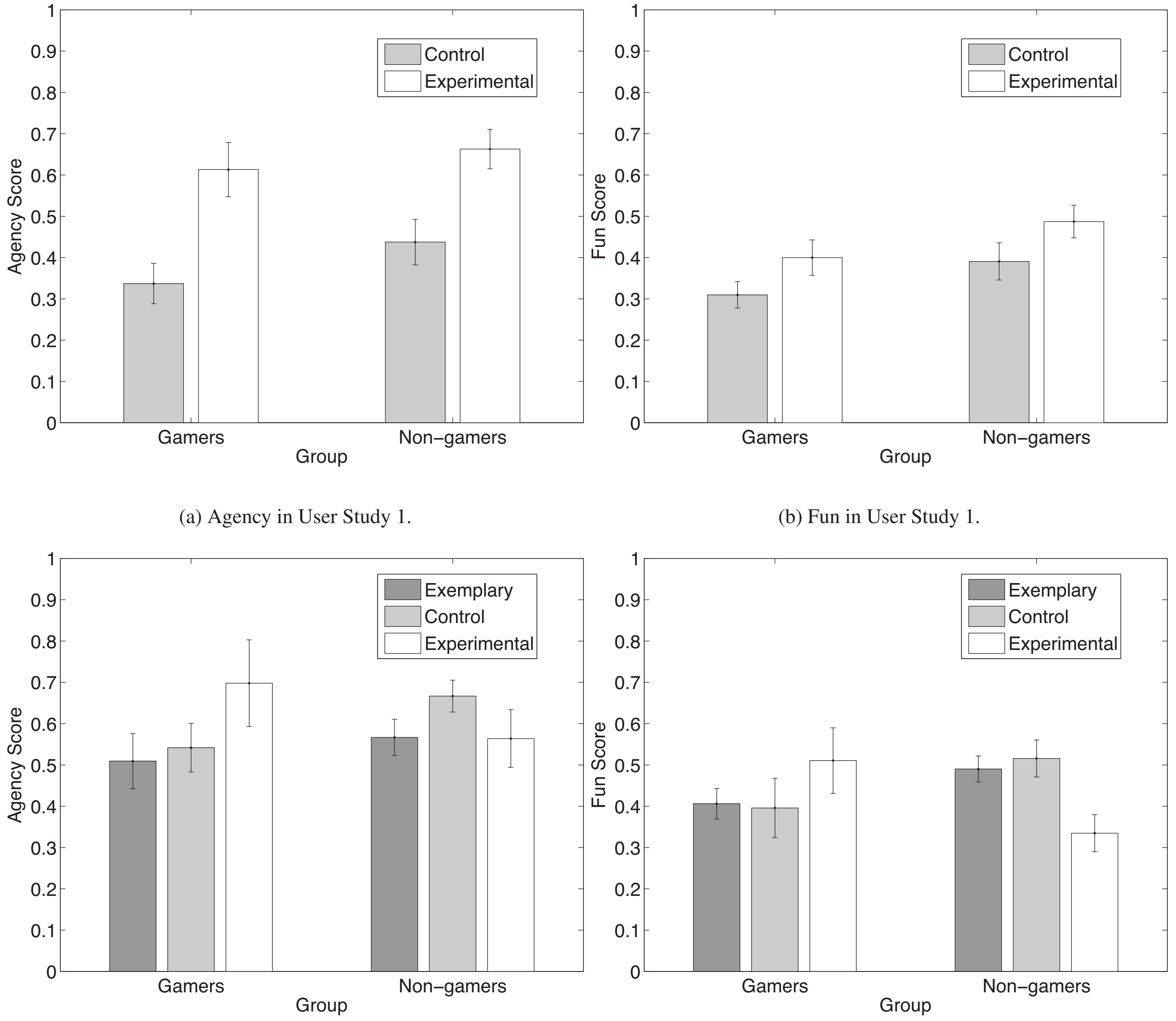
Methodology

Three user studies were conducted. Participants gave written consent and received a university course credit for their participation. These participants were tested in groups of ten to fifteen people, with sessions lasting approximately 30 minutes. The participants were briefed before the start, and then interacted with the story, selecting from on-screen choices (Figure 2). Upon completion, an evaluation survey was administered and a debriefing form was provided.

For the survey, we used a previously validated instrument (Vermeulen et al. 2010). The instrument had twelve questions aimed at measuring player fun and six questions aimed at their agency. Each question was a 5-point Likert scale from *completely disagree* to *completely agree*. The scores on all fun questions were added together and then normalized using the maximum and minimum possible scores. The result is referred to as the *fun score*. The same was done for the *agency score*. The instrument has been previously used to evaluate experience managers (Thue et al. 2011).

Because our experiment consisted of a reading assignment, there was an inherent risk that some participants might skim through the text without reading it. To reduce noise caused by such skimming we used task completion time to estimate the reading speed for each participant and compared it to average reading speeds reported in the literature. First-year college students are reported to skim text at 450 words per minute (Carver 1993). Given the 10% slow down due to reading off the screen (Zelfle 1998), we eliminated data from all participants who rather than that read above 405 words per minute. We then additionally removed data from all participants whose fun and/or agency score was an outlier according to Tukey's outlier filter (also known as interquartile outlier detection). The survey also measured participant gaming skills, in order to classify them into *gamers* (one hour or more of gameplay per week) or *non-gamers* (less than one hour per week).

The data for each metric were analyzed using a two-way ANOVA ($\alpha = 0.05$) with the factors being *experience man-*



(a) Agency in User Study 1.

(b) Fun in User Study 1.

(c) Agency in User Study 2.

(d) Fun in User Study 2.

Figure 3: Mean normalized scores and error bars for user studies 1 and 2 (agency and fun).

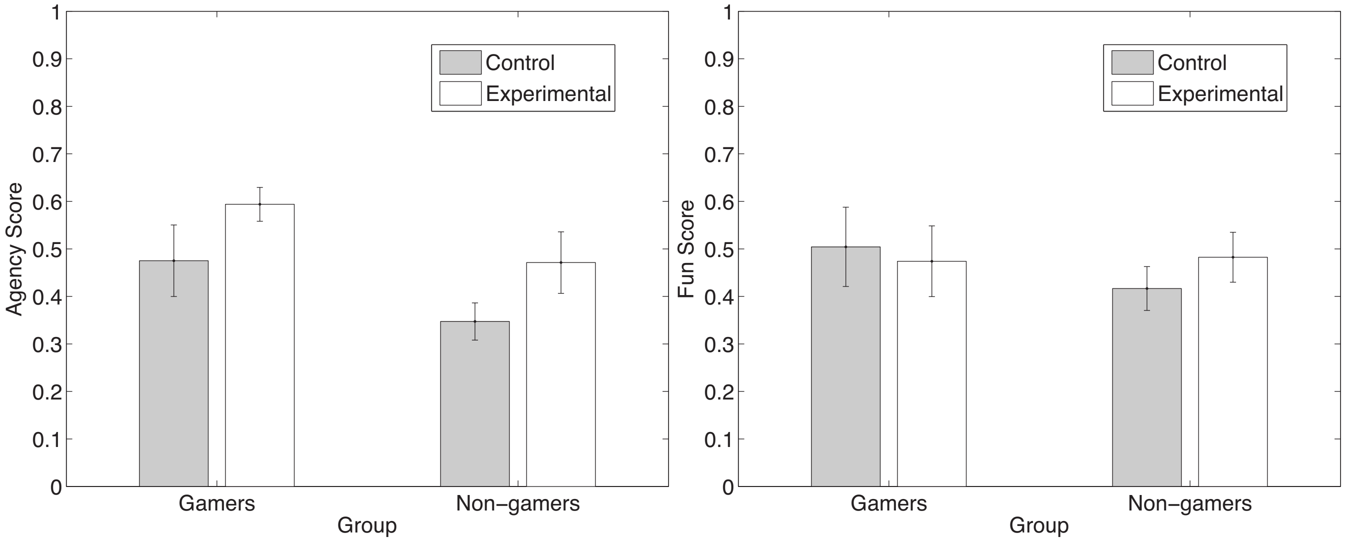
agement and *gaming skills*. The data did not violate the assumptions of independent observations, normality (assessed by the Kolmogorov-Smirnov test) or unequal variance (assessed by the Levine's and Welch's tests). *Post hoc* comparisons used Tukey's HSD test.

Results

User Study 1: ASD. The goal of the first user study was to compare ASD to an exemplar narrative. The control condition were participants who experienced the *Little Red Riding Hood* interactive narrative with choices limited to changing the temporal order of events. The experimental condition was additionally allowed to rupture the narrative at up to three plot points. The number of participants was $n = 81$ ($n = 98$ before outliers). There were 42 participants (21 gamers, 21 non-gamers) in the experimental condition and 39 (23 gamers, 16 non-gamers) in the control condition.

The experimental condition showed significantly higher scores for both fun ($F(1, 76) = 5.46, p = 0.02$) and agency ($F(1, 77) = 20.46, p = 0.001$) compared to the control condition. The interaction between gaming skills and experience management was not statistically significant. In the top row of Figure 3 means and standard errors are shown.

User Study 2: PAST. In the second user study, we evaluated the effects of player modelling in PAST. We had three



(a) Agency in User Study 3.

(b) Fun in User Study 3.

Figure 4: Mean normalized scores and error bars for user study 3 (agency and fun).

different conditions for the main factor of experience management: control, experimental and exemplar; players had the same set of choices in all three. In the experimental condition, narrative ruptures were repaired according to the player model learnt for the actual player. In the control condition, the player model was taken from a randomly drawn player in the experimental condition (as a part of the yoking procedure). If the random model coincided with the player's actual model then the participants would be counted in the experimental condition. The exemplar condition contained all players who never deviated from the exemplar narrative (i.e., never caused a rupture).

In total, $n = 72$ participants ($n = 98$ before outliers) were evaluated. The experimental condition had 23 participants (8 gamers, 15 non-gamers), the control condition had 22 participants (6 gamers, 16 non-gamers) and the exemplar condition had 27 (8 gamers, 19 non-gamers).

No statistically significant results were observed in the main factors of experience management and gaming skills for either fun or agency. However, a significant interaction between these factors was found for fun ($F(2, 66) = 4.73, p = 0.012$). Agency showed a similar interaction but it failed to reach significance ($F(2, 67) = 1.93, p = 0.1538$). Means and standard error bars are depicted in the bottom row of Figure 3. Tukey's HSD *post hoc* test showed that non-gamers in the control condition had significantly higher fun ratings than non-gamers in the experimental condition; no other comparisons reached significance.

We attempted to replicate user study 2 later in the academic term with $n = 63$ participants ($n = 83$ before outlier removal). The participants were distributed as follows: experimental: 19 (2 gamers, 17 non-gamers); control: 19 (5 gamers, 14 non-gamers); exemplar: 25 (10 gamers, 15 non-gamers). This time, no significant effects were found.

User Study 3: PAST with Constrained Domain. In the third user study, we also studied the effects of player modelling and planning; however, we constrained the story space to only 5 possible different stories (down from 31) by permitting only one rupture at a fixed time for all participants. Only two conditions were present for the experience management factor: control (ASD) and experimental (PAST).

The total number of participants was $n = 34$ ($n = 41$ before outliers were removed), with 17 participants in control (5 gamers, 12 non-gamers) and 17 participants in experimental condition (4 gamers and 13 non-gamers). No statistically significant results were found for fun ($p < 0.05$); however, a marginally significant result ($p < 0.10$) for agency ($F(1, 67) = 5.46, p = 0.097$) was found. No statistically significant interactions were found.

The Pull-back Effect. Some participants appear to have experienced a *pull-back effect* wherein they felt that upon breaking the narrative through their choices they were pulled back into the original plot. To illustrate, here are two comments collected from the user studies:

> *The plot of events seems predetermined, anything the user chooses may eventually loop back to the static plot.*
>
> *(...) It was disappointing when I would try to go away from the typical Red Riding Hood story and then it would just inevitably go back to that (it made me feel like my input in what happened next didn't matter).*

Discussion

The results for the first user study indicate that ASD increases player fun and agency over the exemplar narrative. This is despite the pull-back effect reported by some participants in the comment's section.

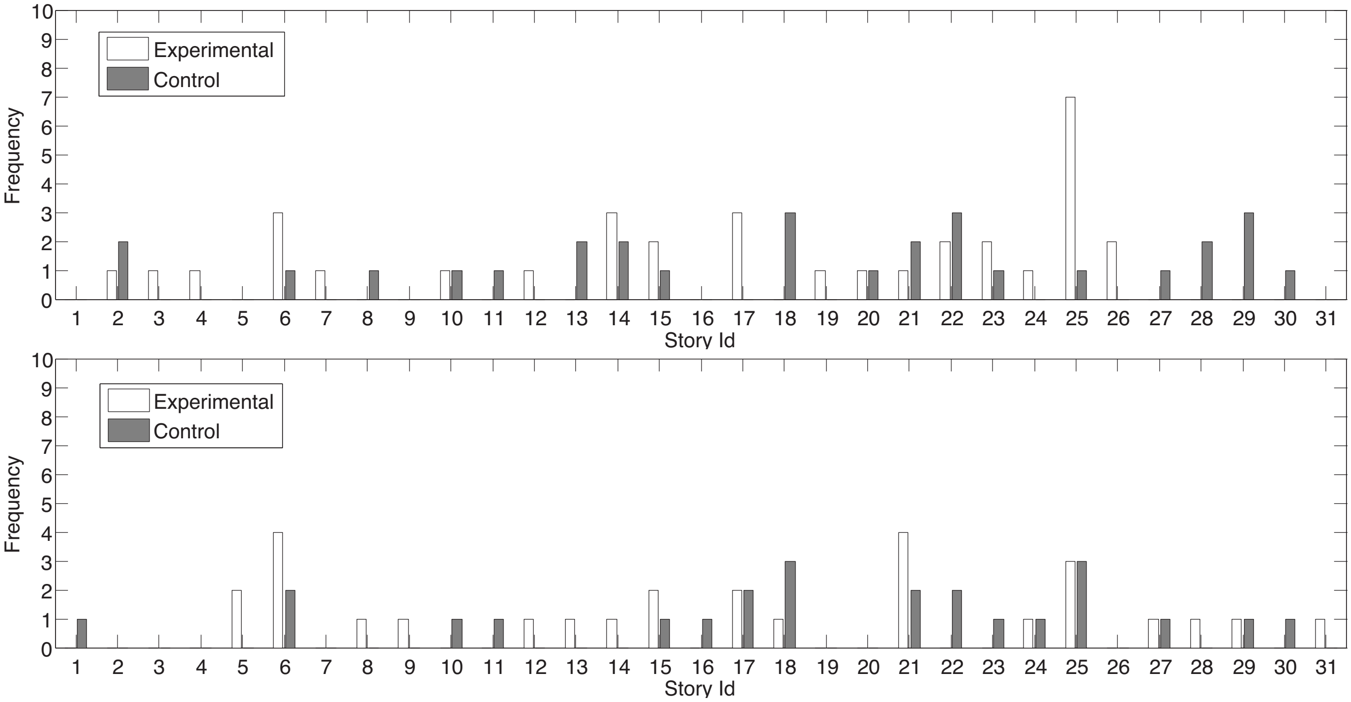


Figure 5: Distribution of players (experimental and control). Top: part 1, bottom: part 2 of the second study.

The results of the second user study did not show significant effects of player modelling in PAST for either player fun or agency. This is surprising since PAST is a combination of ASD and PaSSAGE, which in isolation showed a possitive effect. We believe several factors may cause this.

Firstly, the PAST user study had fewer participants than the ASD study split into more conditions, thus causing a loss of statistical power. Second, the latter part of the study was run at the end of the academic term, possibly affected by a lack of motivation in "last minute" sign ups, potentially adding noise. Third, the richer narrative space (i.e., more ruptures and more diverse repairs) could have induced a higher variance. To illustrate, Figure 5 shows the uneven distribution of players over different stories (i.e., each story being a collection of different events). Our attempt to control variance by randomly assigning player models from the experimental participants to control condition participants (yoking) accounted only for the types of repairs, hence not controlling for the actual ruptures taken. Lastly, agency and fun perception might differ between gamers and non-gamers in different experience management scenarios. However, since there is no statistical significance, it is impossible to tell if this is a fluke or an indicative result. If the latter is the case then PAST may increase fun only for gamers in this specific story domain.

The third user study, in a constrained story domain, showed marginally significant results that may indicate that agency is increased by PAST. A higher significance level might have been obtained with more participants.

Conclusions and Future Work

This paper presents the first formal evaluation of two recent planning-based experience managers in the context of IS. Both were evaluated with participants playing a text adventure based on a folk story on the basis of their self-reports of fun and agency. It appears that planning-based experience management can improve fun and agency over even a nonlinear exemplar narrative. On the other hand, when player modelling and planning is used, mixed results are obtained: when using a large story domain, positive results are seen only in gamers; however, we were not able to replicate them possibly due to different story coverage between conditions. When using a smaller story domain with a more even coverage, agency is increased for both groups. This highlights that the balance of content is of particular importance.

Results presented in this paper support previous conjectures that treating IS as AI planning is an effective way to improve player's experience while affording them more choice within a narrative. Adding player modelling seems promising, but further evaluations are required. Existing systems can be deployed in the field to foresee narrative ruptures and automatically compute repairs. These narratives can then be fleshed out with multimedia elements by game developers.

Acknowledgements

We would like to gratefully acknowledge the funding provided by the Natural Sciences and Engineering Research Council of Canada (NSERC), the help and consultations by David Thue, and the great work by Wayne DeFehr and Shane Riczu with different writing tasks for the protypes.

References

Carver, R. P. 1993. Reading rate: Theory, research, and practical implications. *Journal of Reading* 36(2):84–95.

Cheong, Y.-G., and Young, R. M. 2008. Narrative generation for suspense: Modeling and evaluation. In *Proceedings of the First Joint International Conference on Interactive Digital Storytelling (ICIDS)*, 144–155.

Kraft, S. 1981. He chose his own adventure. *The Day* 6.

Laws, R. D. 2001. *Robin's Laws for Good Game Mastering*. Steve Jackson Games.

Ramirez, A., and Bulitko, V. 2012. Telling interactive player-specific stories and planning for it: ASD+ PaSSAGE = PAST. In *Proceedings of the Eight AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 173–178.

Riedl, M. O., and Bulitko, V. 2013. Interactive narrative: An intelligent systems approach. *AI Magazine* 34(1):67–77.

Riedl, M, O.; Stern, A.; Dini, D.; and Alderman, J. 2008. Dynamic experience management in virtual worlds for entertainment, education, and training. *International Transactions on Systems Science and Applications* 4(2):23–42.

Riedl, M. O. 2012. Interactive narrative: A novel application of artificial intelligence for computer games. In *Proceedings of the 26th National Conference on Artificial Intelligence*, 2160–2165.

Sharma, M.; Mehta, M.; Ontanón, S.; and Ram, A. 2007. Player modeling evaluation for interactive fiction. In *Proceedings of the AIIDE Workshop on Optimizing Player Satisfaction*, 19–24.

Thue, D.; Bulitko, V.; Spetch, M.; and Wasylishen, E. 2007. Interactive storytelling: A player modelling approach. In *Proceedings of the Third AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 43–48.

Thue, D.; Bulitko, V.; Spetch, M.; and Romanuik, T. 2010. Player agency and the relevance of decisions. In *Proceedings of the Third Joint International Conference on Interactive Digital Storytelling (ICIDS)*, 210–215.

Thue, D.; Bulitko, V.; Spetch, M.; and Romanuik, T. 2011. A computational model of perceived agency in video games. In *Proceedings of the Seventh AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 91–96.

Vermeulen, I. E.; Roth, C.; Vorderer, P.; and Klimmt, C. 2010. Measuring user responses to interactive stories: towards a standardized assessment tool. In *Proceedings of the Third Joint International Conference on Interactive Digital Storytelling (ICIDS)*, 38–43.

Yu, H., and Riedl, M. O. 2012. A sequential recommendation approach for interactive personalized story generation. In *Proceedings of the 11th International Conference on Autonomous Agents and Multi Agent Systems*, 71–78.

Yu, H., and Riedl, M. O. 2013. Toward personalized guidance in interactive narratives. In *Proceedings of the Eight International Conference on the Foundations of Digital Games (in press)*.

Zelfle, M. 1998. Effects of display resolution on visual performance. *Human Factors* 40(4):554–568.