Engagement or Distraction? Examining the Impact of Narrative Elements and Player Audience on Experience of Logic Grid Puzzles

Fiona Shyne, Kaylah Facey(✉), and Seth Cooper

Northeastern University, Boston, MA 02115, USA
{shyne.f,facey.k,se.cooper}@northeastern.edu
https://www.khoury.northeastern.edu/

Abstract. In this work, we explore different narrative modes for logic grid puzzles. We test these environments with a user study recruiting two audiences: crowd-workers on the Prolific platform and volunteers from social media groups related to mysteries and puzzles. While volunteers found puzzles easier, they enjoyed them less than the Prolific workers. Across both audiences, an increase in narrative increased the time taken on puzzles and the challenge of the puzzles. However, while some participants found the narrative immersive and enjoyable, others did not want any story or did not like the increased challenge.

Keywords: Logic Grid Puzzles · Narrative Puzzles · User Study

1 Introduction

Narrative is an important component of many games. Narrative can enhance feelings of immersion and engagement [3,12]. However, narrative elements can also be seen as a nuisance [12].

In this work, we use *logic grid puzzles* as a testbed for studying how narrative impacts the experience of puzzle solving. Logic grid puzzles are a popular form of pen-and-paper puzzle where players use natural language clues to mark relationships between entities on a grid (Fig. 1). Traditionally, they include a small amount of narrative, but the solving process does not meaningfully incorporate it. We extend this form to include increasing levels of narrative interaction. The "base clue" mode models the traditional form, the "paragraph" mode wraps clues in prose, and the "interactive fiction" mode provides a text-based environment to navigate. The base logic puzzle is procedurally generated using a system based on previous work [16,17].

We tested how these three narrative modes impact gameplay through a user study in which participants, either paid crowd-workers from the Prolific platform or volunteers from social media, solved a variety of puzzles. After each puzzle, participants were asked about difficulty, narrative quality, and enjoyment.

M. C. Reyes and F. Nack (Eds.): ICIDS 2025, LNCS 16375, pp. 296–313, 2026.
https://doi.org/10.1007/978-3-032-12405-0_16

Prolific workers found puzzles more challenging but also more enjoyable. Both groups found that increased narrative resulted in increased challenge. The impact of narrative on enjoyment was less clear. Some participants enjoyed the sense of immersion that increased narrative added, while others found it confusing or unnecessary.

2 Related Work

2.1 Stories in Games

There is debate over the role of narrative in games. Frasca [6] argues that "the potential of games is not to tell a story but to simulate... an environment for experimentation [by the player]." Mateas and Stern [11] disagree that an authored narrative should be abandoned, arguing instead that player actions should drive a plot structure that changes with each playthrough.

A story can be embedded in a game in myriad ways. Fernandez suggests that narrative should be told through objects in the game world, and Bizzochi [2] extends this idea to include UI elements. Particular attention has been paid to "narrative" or "fiction" puzzles that are integrated into the story of a game and drive its plot [5,9,20].

Evidence suggests that even exposure to a pre-game story can increase players' sense of presence in a game [13,19]. On the other hand, inclusion of a narrative may result in a *worse* experience for some players; Miller et al. [12] suggest that players be able to include or exclude narrative, and Siu and Riedl [18] find that narrative rewards were commonly either the most- or the least-favored rewards. In addition, engagement with a story matters. Immersion is not increased as much if players ignore the story [3], or if they are less interested in its genre [12].

2.2 Paid vs Volunteer Participants

Studies comparing the performance of unpaid volunteers with paid crowd-workers have had mixed results. For simple tasks, Siu and Riedl [18] found that crowd-workers outperformed volunteers on both time and accuracy. For complex tasks, most comparisons between crowd-workers and volunteers have found that crowd-workers may complete more tasks, but at a lower quality [7,10,15]. As our task is a complex puzzle game, we hope to contribute to the literature on the suitability of crowd-work for complex tasks.

3 Puzzle Design

We designed a set of puzzles that varied in narrative interaction: the extent to which the player has to interact with narrative content in order to progress. Design for these puzzles was iterative; modes with less narrative interaction were used as the basis for modes with more. Each logic grid puzzle was first generated

using a genetic algorithm [16,17]. Then for the "base clue" mode, we minimally edited the generator's output for clarity and grammar. For the "paragraph" mode, each clue was expanded into a short narrative paragraph. Lastly, for the "interactive fiction" (IF) mode, the paragraphs were used as the basis for an IF game. Examples are given in Fig. 1.

As a running example, we will discuss how we created the puzzles for one scenario: "The Wild Rose Train." In "Train", the player is a PhD student who must figure out who on a train stole their research.

3.1 Logic Puzzle Generation

The core of these puzzles is logic grid puzzles, a kind of pen-and-paper logic puzzle consisting of a grid and a list of natural language clues. To solve them, players must deduce entity relationships either given explicitly by the clues or inferred using logical reasoning. Previous works [16,17] presented a system that generates these types of puzzles given a set of categories and entities, which we selected based on the imagined scenario. The generator uses a *genetic algorithm*, an artificial intelligence algorithm modeled after biological evolution.

3.2 Base Clue Design

The first "base clue" mode operates directly on the output given by the generator.

The generator outputs a list of human-readable, but not necessarily grammatically correct, clues. To create the base clue puzzles, we edited the generated clues for grammar and clarity only. For example, "the station Forest Hills is the suspect Madeleine Baker" became "the person who departed at Forest Hills Station was Madeleine Baker."

In this mode, we also wrote a short introduction and conclusion to the chosen narrative. The conclusion is given to the player after they solve the puzzle. In "Train," the introduction tells the player that they are a PhD student who is trying to determine who stole their research, and the conclusion reveals the culprit.

3.3 Paragraph Design

Base clue puzzles can be expanded into "paragraph" mode puzzles. For each base clue, we invented a narrative reason the player would discover the information described by the clue. We then wrote a paragraph describing the clue in context. We used the same introduction and conclusion as in the base clue puzzle.

In "Train," "the person who departed at Forest Hills Station was Madeleine Baker" became a witness statement from "Chef Gardner," who says that Baker mentioned she wanted to find a flower in Forest Hills.

3.4 Interactive Fiction Design

Interactive fiction (IF) is a genre of games that are conveyed mostly or completely through text [1]. For each puzzle, we turned the paragraph clues into an IF game, using Ink [8]. In this version, users are presented with IF "nodes", each containing a block of text and a list of choices that lead to other nodes.

Creating an IF game requires a substantial expansion of the paragraph clues. Objects must be discovered by exploration, and witness statements are transformed into interactive dialogue. Also, while the paragraph mode only includes text with relevant information, in the IF game extra text is included to entertain or even mislead—taking care not to give any information that would directly contradict the logical clues.

In "Train," the player must navigate to the dining car to interview Chef Gardner. Gardner can talk about several topics but only has useful information about the suspect Madeleine Baker, requiring several follow-up questions.

The introduction and conclusion text were also modified to fit the IF format. In "Train," players must first report the theft to the conductor before they can search the train and interview other characters. Once the puzzle is solved, the conclusion is unlocked, which leads the player to the discovery of evidence pointing to the culprit.

4 User Study

To test the impact of different narrative modes, we performed a user study. In this study participants (either paid crowd-workers or volunteers), played up to four puzzles. After each, they were given a short survey. After playing all the puzzles they were asked to rank them based on difficulty, enjoyment, and narrative.

The study methods had approval from our IRB. All user study data, along with the puzzles, are available on the Open Science Framework[1].

4.1 Puzzles

Participants could play up to four puzzles. The first puzzle (the primer) acted as a basis for comparison with the experimental puzzles. Additionally, puzzles for three scenarios were manually authored in each of the three interaction modes (as described above), for a total of nine puzzles. Participants could play the primer and up to three of the (randomly assigned) experimental puzzles.

Primer. From a preliminary study, we found that the puzzles presented here were quite challenging to lay users. To account for this, we included a primer puzzle that had a 40% solve rate in the previous study [17], using the scenario of figuring out a school schedule. It was given in the base clue mode only. Participants completed the primer first and then up to three experimental puzzles.

[1] https://osf.io/4sae8/.

Fig. 1. Screenshot from the puzzle interface. The first three screen shots show the different narrative modes, including "base clue" (upper left), "paragraph" (upper right), and "interactive fiction (IF)" (bottom left). The bottom right screenshot shows the logic puzzle grid. These are all from "The Wild Rose Train."

Experimental Puzzles. We wrote puzzles for three scenarios. The first author designed "The Wild Rose Train" ("Train"), the scenario used as an example in the previous section. They also designed "The Great Chili Competition" ("Chili"), in which the player must reconstruct a recipe by interpreting family statements, recipe notes, and emails. The second author designed "Lady Rose Ellington's Chrysanthemum Ball" ("Ball"), in which the player must uncover the theft of a necklace by interpreting witness statements.

4.2 Recruiting Populations

We recruited from two separate populations: crowd-workers from the Prolific website and volunteers from social media. This was done because in previous

work we found that crowd-workers found puzzles very challenging, even when we did not [17]. This made us interested in investigating the differences between the crowd-working population and people with a specific interest in logic puzzles or mysteries (the target audience for this type of game).

Prolific. Crowd-workers were recruited and paid through the Prolific platform. Since these users were not expected to be particularly interested or skilled in logic grid puzzles, we based recruitment numbers on how many people successfully solved the primer puzzle. We hoped those participants' responses to the experimental puzzles would be based primarily on narrative mode. We continued recruiting on Prolific until each narrative mode was played by 20 users who had successfully solved the primer. However, data was captured for all users regardless of whether they successfully solved the primer. Prolific workers were paid $2.50 per puzzle they completed a post-survey for (including the primer puzzle), regardless of whether they solved the puzzle or the amount of time they spent.

Volunteers. We also sought out participants who were likely to enjoy this type of puzzle, by recruiting through various social media forums related to puzzles, games, and mysteries. These participants were volunteers and were not paid. We did not target a specific recruitment goal for volunteers, recruiting as many as we could in the study time (2 weeks and 5 days). We also did not exclude volunteers that failed to solve the primer, as there were far fewer volunteers than crowd-workers.

Randomization. Participants were given up to three experimental puzzles. If they chose to play all three puzzles, they were randomly given one puzzle from each mode and one from each scenario. No mode or scenario was duplicated, and the puzzles were given in random order.

Quantitative Measures. While participants played the puzzles, we captured a variety of quantitative measures such as time, number of attempts, and correctness. After each puzzle, participants were asked to fill out the narrative and enjoyment subscales of the Game User Experience Satisfaction Scale (GUESS) [14] and the cognitive subscale of the Video Game Demand Scale (VGDS) [4]. These scales were chosen for their relevance and brevity. We chose not to include complete GUESS or VGDS scales, as they each contain several irrelevant subscales that might confuse or fatigue participants. Participants who played all three puzzles were asked to rank them in terms of difficulty, narrative quality, and enjoyment.

Qualitative Analysis. Participants could leave open-ended comments on both individual puzzles and the final rankings. To analyze these comments, the first

two authors performed an open coding process. They initially reviewed the text responses independently. Both authors first read through the responses, without adding any codes, to get a sense of the data. Then they individually coded each response, disregarding comments irrelevant to the puzzles. Finally, the first author synthesized the codes ("enjoy," "fun," "easy," "straightforward," "hard," "dislike hard," "liked hard," "good challenge," "narrative length," "sifting through information," "not enough information," "narrative confusion," "dislike narrative," "liked narrative," "tabs," "interactivity," "interface," "balance," "emotion," "suggestion," "rude"). In total, Prolific workers left 79 relevant comments, and volunteers left 33.

5 Impact of Recruiting Population

5.1 Quantitative

We examined how the two groups of participants (volunteers and Prolific workers) reacted to the puzzles based on type (primer vs experimental). To do this we performed a two-way ANOVA with recruitment and puzzle type as the independent variables, and subjective difficulty, enjoyment, narrative, and time as the dependent variables. For post-hoc evaluation, we used Dunn's tests. To test interactions, we performed Dunn's test on the four groups that are formed from the combination of recruitment and puzzle type. These tests included all participants regardless if they solved the primer, and we report on significant effects (summary statistics given by Table 1).

Participants. In total 142 Prolific workers were recruited. Of these participants, 76 (54%) did not solve the primer, 43 (56%) of whom did not play any more puzzles. Of the 66 (46%) who did solve the primer, 2 (3%) did not play any experimental puzzles, 19 (29%) played one, and 45 (68%) played two or more. The median total time Prolific workers spent on puzzles was 17.5 minutes, and they were paid a mean of $6.20. We also calculated a median pay rate of $18.80 per hour, not including time spent on surveys.

In total 34 volunteers were recruited. Of the volunteers, 6 (18%) did not solve the primer while 28 (82%) did. Of those who did not solve the primer, all but 1 did not play any more puzzles. Of those who did solve the primer, 2 (7%) did not play any more puzzles, 12 (43%) played one experimental puzzle, and 14 (50%) played two or more experimental puzzles. The median total time volunteers spent on puzzles was 10 min.

The differences between Prolific workers and volunteers in total time spent and total number of puzzles attempted were not found to be significant by a Dunn's test.

Challenge. There was a significant interaction of recruitment and puzzle type on the challenge scale ($F = 24.325845, p < 0.0001$). The post-hoc test showed that all four groups were significantly different from each other, shown in Fig. 3.

Prolific workers found both the primer and the experimental puzzles more challenging than the volunteers did. Volunteers found the experimental puzzles much harder than the primer, while Prolific workers found the experimental puzzles equally as hard as the primer.

There was a significant interaction of recruitment and puzzle type on correctness ($F = 0.80, p = 0.0371$). The vast majority (93%) of volunteers completed the primer, and they solved a majority (67%) of experimental puzzles they attempted. Meanwhile, just under half (46%) of Prolific workers completed the primer, and they solved a minority (16%) of experimental puzzles. There was also a significant interaction of recruitment and puzzle type on the percentage of incorrect marks. Prolific workers solving experimental puzzles have a higher percentage of incorrect marks than those solving the primer, and Prolific workers have a higher percentage of incorrect marks than volunteers. Correctness and incorrect marks are shown in Fig. 2.

Table 1. Volunteers vs Prolific workers on the primer and experimental puzzles. Reporting mean (std).

	Primer		Experimental	
	Volunteers	Prolific workers	Volunteers	Prolific workers
Challenge Scale	3.06 (1.33)	5.75 (1.04)	4.64 (1.08)	5.75 (1.07)
Narrative Scale	2.59 (0.89)	3.99 (1.39)	3.60 (1.06)	4.23 (1.48)
Percent Incorrect	0.01 (0.04)	0.12 (0.24)	0.06 (0.13)	0.27 (0.25)
Time (s)	268.28 (201.26)	557.35 (408.26)	891.14 (508.26)	763.51 (651.36)

Narrative and Enjoyment. There was also a significant interaction of recruitment and puzzle type on the narrative scale ($F = 3.90, p = 0.0494$), shown in Fig. 4. Prolific workers found both the primer and experimental puzzles more narratively interesting than the volunteers did. Both Prolific workers and volunteers found the experimental puzzles more narratively interesting than the primer puzzles.

The ANOVA did not find a significant interaction for the enjoyment scale but did find a main effect of recruitment ($F = 10.37, p = 0.0014$), shown in Fig. 5. Prolific workers found the puzzles more enjoyable ($m = 4.90, std = 1.65$), than did volunteers ($m = 4.14, std = 1.14$).

Time. There was a significant interaction between recruitment and puzzle type on the time participants spent on each puzzle ($F = 6.89, p = 0.0091$), shown in Fig. 6. Both Prolific workers and volunteers spent less time on the primer than on the experimental puzzles. However, volunteers spent less time on the primer than Prolific workers but more time on the experimental puzzles than Prolific workers.

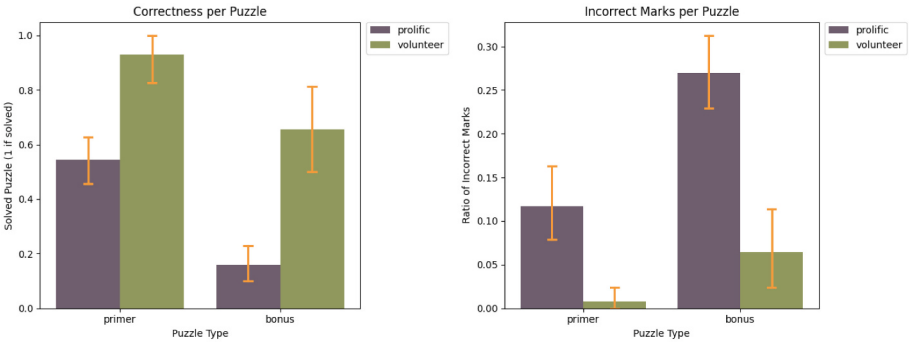

Fig. 2. The solve rate (left) and percentage of incorrect marks (right) by audience and puzzle type. The error bars show a 95% confidence interval.

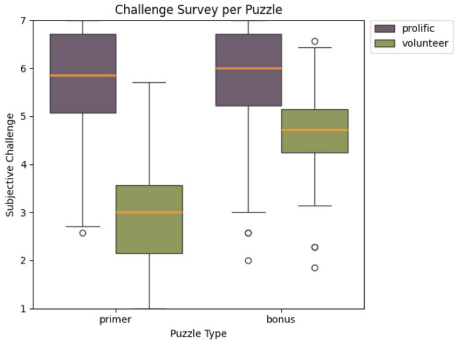

Fig. 3. Effect of audience and puzzle type on perceived challenge (orange bar represents median). (Color figure online)

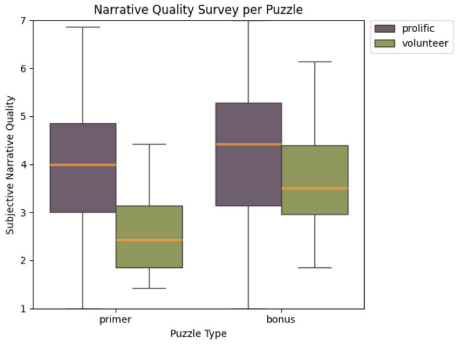

Fig. 4. Effect of audience and puzzle type on narrative quality (orange bar represents median). (Color figure online)

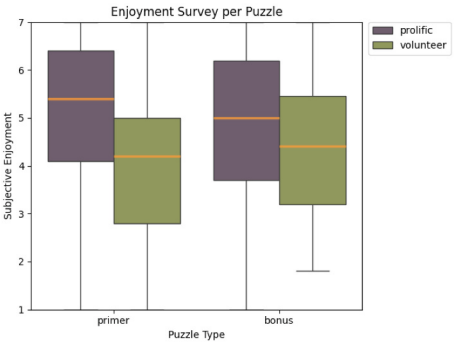

Fig. 5. Effect of audience and puzzle type on enjoyment (orange bar represents median). (Color figure online)

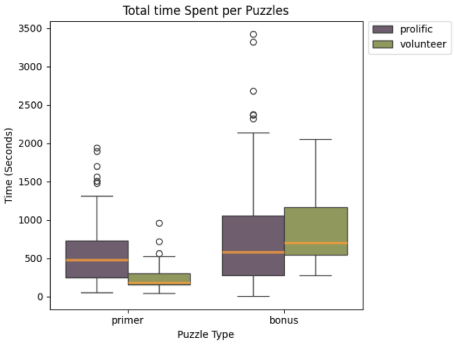

Fig. 6. Effect of audience and puzzle type on time spent (orange bar represents median). (Color figure online)

5.2 Qualitative

A major difference between Prolific workers and volunteers was the quality of comments given. While both groups included text responses at about the same rate (just under 40% for both groups), volunteer responses were more detailed. Volunteers left an average of 25 words in post-puzzle surveys, while Prolific workers' comments averaged only 6 words. The difference was closer in the ranking comments, with volunteers again commenting about 25 words per response in comparison to 17 words on average for Prolific workers. Prolific workers left many comments that amounted to simple statements that the puzzle was "good" (13) or "hard" (12), while volunteers left none of that type. A couple Prolific workers even gave rude comments such as "people do that for fun? Yikes." There were no rude comments from volunteers.

Prolific workers were more likely to comment that puzzles were hard (Prolific: 23, 29%, volunteers: 1, 3%). Many Prolific workers (8; 32%) even stated that their final enjoyment rankings were solely determined easiest to hardest. Similarly,

volunteers were more likely to comment that puzzles were easy (Prolific: 2, 2%, volunteer: 5, 15%).

Many codes were found consistently among the different populations. Both volunteers and Prolific workers were confused about particular parts of the narrative (Prolific: 7, 9%, volunteers: 3, 9%), thought the puzzles did not have enough information (Prolific: 17, 22%, volunteers: 6, 18%), and complained that they did not like sifting through narrative (Prolific: 9, 11%, volunteers: 6, 19%). This demonstrates that many of the codes we identified in the qualitative analysis could have been found if only one of the populations had been sampled from.

6 Impact of Narrative Modes

6.1 Quantitative

To test the effect of narrative mode on behavior and perception of the logic puzzles we performed Kruskal-Wallis tests with Dunn's post-hoc tests for the experimental puzzles, for all 34 volunteers and the 66 Prolific workers who solved the primer. As for the impact of recruiting population, all measures were tested, but we only report the significant results.

Volunteers. Summary statistics for significant results are given by Table 2. There was a significant effect of narrative type on the challenge scale ($S = 11.59, p = 0.0030$). Volunteers found the base clues significantly easier, than the paragraph or IF, though there was not a significant difference between IF and paragraph. Similarly, there was a significant effect of the percentage of incorrect marks in the submitted puzzle. The post-hoc test shows that the only significantly different groups are IF and base clues. The challenge scale and percent incorrect results are shown in Fig. 7.

Table 2. Volunteers by narrative mode. Mean (std), *italicized values* were not associated with significant results.

	Base Clues	Paragraph	IF
Challenge Scale	3.86 (1.12)	5.13 (0.68)	5.23 (0.65)
% Incorrect	0.02 (0.06)	*0.05 (0.10)*	0.17 (0.18)
Time	659.58 (461.89)	984.46 (497.95)	1139.11 (490.45)

There was also a significant effect on total time spent ($S = 9.15, p = 0.0103$). Again, base clue puzzles took significantly less time than paragraph or IF puzzles, while IF and paragraph were not significantly different from each other, shown in Fig. 8.

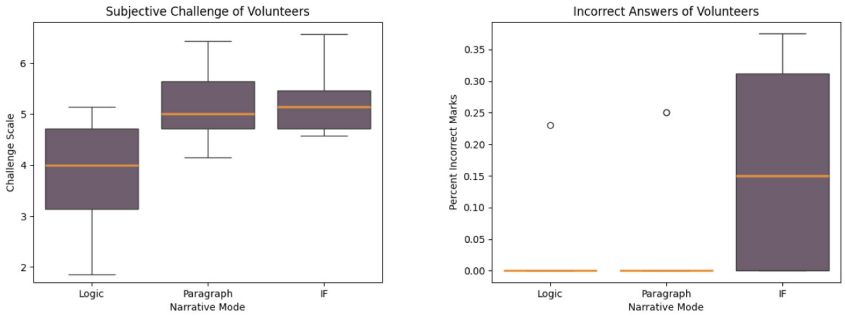

Fig. 7. Perceived challenge and the percentage of incorrect answers for the volunteers, by narrative mode

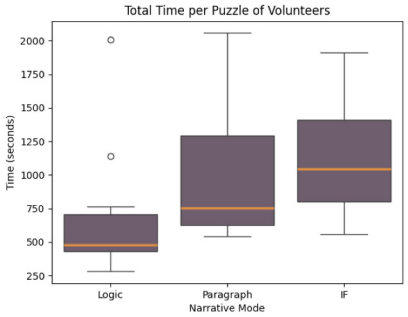

Fig. 8. Time spent on puzzles by volunteers based on narrative mode

Prolific Workers Who Solved Primer. Summary statistics for significant results are given by Table 3. The Kruskal-Wallis tests found significant results for the number of correct marks ($S = 7.79, p = 0.0204$), percent of incorrect marks ($S = 23.12, p < 0.0001$), and whether the puzzle was correct ($S = 9.49, p = 0.0087$). The post-hoc test found that Prolific workers had significantly more incorrect marks for IF puzzles, followed by paragraph, and base clue puzzles. The post-hoc tests for correctness found that significantly more Prolific workers got the base clue puzzles correct (36%), than for paragraph (14%), or IF (7%), though there was not a significant difference between paragraph and IF. Percentage incorrect and percentage correct are shown in Fig. 9.

The challenge rankings also had significant results ($S = 16.13, p = 0.0003$). Base clues were ranked significantly lower in challenge, than paragraph, or IF, though paragraph and IF were not significantly different from each other. The rankings are shown in Fig. 10.

There was also a significant effect on time spent ($S = 16.28, p = 0.0003$). Prolific workers spent significantly more time on the IF puzzles than paragraph

Table 3. Prolific workers who solved primer by narrative mode. Reporting mean (std).

	Base Clues	Paragraph	IF
Challenge Ranking	1.52 (0.80)	2.07 (0.62)	2.41 (0.80)
% Incorrect	0.10 (0.15)	0.24 (0.19)	0.34 (0.21)
Time	634.63 (485.57)	837.42 (561.62)	1285.86 (817.17)

and base clue puzzles, though there was not a significant difference between paragraph and base clue (Fig. 11).

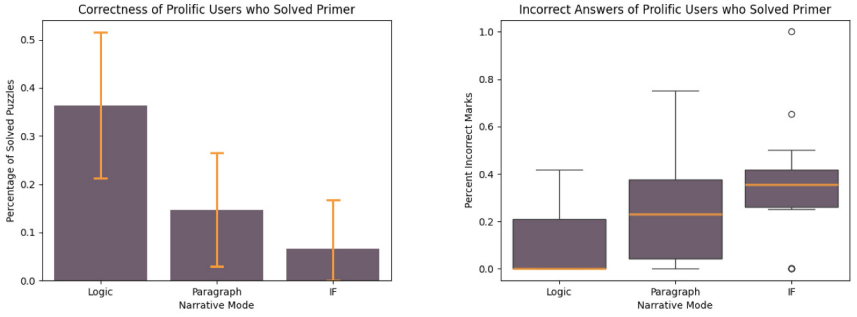

Fig. 9. Percentage of Prolific workers who got the puzzle correct and percent of incorrect marks by narrative mode.

6.2 Qualitative

Comments on Narrative. From the open-text responses, we received a wide range of views about the narrative in these puzzles. Some comments (5) talked about how participants did not like the narrative, because they do not like narrative in general or did not like the narrative in the puzzle. In contrast, just as many comments (5) suggest that participants enjoyed the narrative elements. Some participants (10) were confused about different parts of the narrative. These participants stated that to solve the puzzles they had to "take some leaps" or that different parts had "inconsistenc[ies]."

Participants also commented on the length of the puzzles. Most often participants thought the puzzles were too long (11), particularly the "Ball" puzzle in IF form (9): "the fun aspect of the puzzle was lost trying to go through and make sure you read everyone's statement." One participant did enjoy the length of "Ball" in interactive fiction, ranking their overall enjoyment based on how long it took them to read through. Some participants also wanted a balance in narrative length: "the train [IF] had too much narrative ... [t]he competition ["Chili", base clue] could have had a little more."

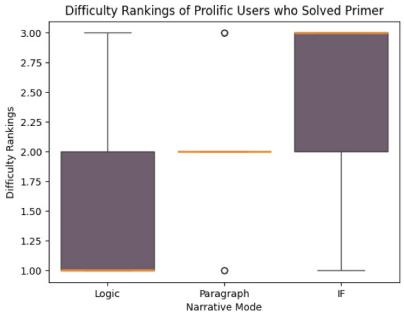

Fig. 10. Rankings (1-3) of challenge from Prolific workers who completed the primer.

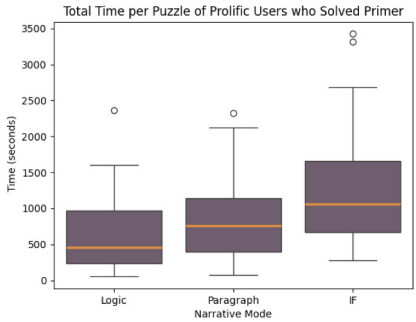

Fig. 11. Total time spent per puzzle of Prolific workers who solved the primer

Comments on Difficulty. The majority of comments about difficulty stated that the puzzles were too hard (24). Other participants found that they enjoyed the challenge of the puzzles (6). Some participants found puzzles to be easy (7), most often volunteers (5) and participants doing the primer (5).

A common complaint was that there was not enough information to solve the puzzle (24). This occurred across all puzzle types. Most often participants said something to the effect of "it seemed" like there was not enough information, while others more assertively stated that puzzles required guesswork or were "poorly designed" [primer]. However, all ten puzzles do not require guesswork and were solved by at least one participant.

Opinions about whether challenge is desirable also varied. More participants (9), mostly Prolific workers (8), stated that they did not enjoy challenging puzzles. These participants thought that easier puzzles made them "feel smarter" or were "satisfying." In contrast challenging puzzles were "confusing," "frustrating," or "annoying." However, some participants (4) said they enjoy challenge or that easy puzzles are boring.

Gameplay Format. Narrative mode had a varied effect on participants. Most often (15 comments) participants said they did not like this process of examining text for clues. This was because they "had to [go] back and forth" to find relevant information, or they had to go through text that was "less useful" to the puzzle. Some participants also felt that the "fun part" of puzzle solving was lost in IF.

On the other hand, some participants said they liked the process of sifting through information (5), or the inclusion of interactivity (5). Participants appreciated the process of exploration, finding clues for themselves, and deciphering information from a narrative. To one participant it was "like reading [the story] in real time."

Three participants stated their preference for paragraph mode, as it provides a balance between narrative elements and ease of accessing clues as "having the clues in separate tabs was much more convenient."

7 Discussion

7.1 Effect of Audience

Recruitment had an interesting effect. Volunteers were more successful in solving puzzles, found them easier, spent less time on the primer, and enjoyed the puzzles less. They were also more likely to leave detailed open-text responses and spent more time per experimental puzzle.

There are a couple of possible explanations for our outcomes. It is possible that volunteers were more interested in the experimental puzzles than Prolific workers, and therefore spent more time on them. However, it is also possible that Prolific workers were more incentivized to move on quickly, as that would increase their hourly pay. Additionally, volunteers come from forums for people asking for input on their projects, so they may be better equipped at giving critical feedback. On the other hand, Prolific workers without significant experience in the genre and incentivized to move quickly might be less critical.

Overall both audiences provided valuable insights. The majority of codes we found were in both populations. However, there is a tradeoff between quantity and quality of data. We were able to recruit many more Prolific workers, but their responses tended to be shorter and less descriptive. Another aspect to note is that recruiting volunteers may be more desirable for lower budget projects, as each Prolific user comes with a payment cost.

7.2 Effect of Narrative on Difficulty

It is clear that across both audiences, the increase in narrative in the paragraph and IF modes increased the difficulty of the puzzles. From the open-text responses, this is because the logic takes more work to interpret and there is more information to sort through.

There were mixed perspectives on whether this increase in difficulty is desirable. Many users, particularly those from Prolific, found this increase in difficulty frustrating. However, some participants appreciated the added challenge and

different methods of deduction. In particular there were participants who were strictly against clues being given directly or stated that they enjoyed extracting clues from narrative. It is interesting that opinions varied even among volunteers, who were recruited based on their interest in puzzles and mysteries. This shows the importance of tailoring the recruitment process to maximize the chances of reaching the target population.

7.3 Effect of Narrative on Enjoyment

The quantitative analysis did not find a relationship between puzzle format and enjoyment of narrative. However, open-text responses provide more insight into how narrative mode impacted experience of a puzzle. Some participants simply preferred the puzzle solving aspect, and wanted as little narrative as possible. Other participants were less negative to narrative in general, but thought the game was too long.

While some frustrations could be chalked up to personal preference, it also demonstrates design choices that can be improved on. The current version of the IF mode requires a significant mental load, as participants must remember all of the clues they found while also solving a difficult puzzle, which could be improved with interface updates or hint systems.

7.4 Limitations and Future Work

One major limitation of this work is the difference in sample size between audiences. Despite our best efforts in identifying relevant social media forums, we were not able to gather the same quantity of volunteers that we were able to from Prolific. Additionally, a greater depth of qualitative data could have been retrieved if participants were recruited for an in-person or video call study including a think-out-loud activity or interview.

This work was also limited in the design of the puzzles. Participants, especially from Prolific, found even the base clue mode of the puzzles fairly challenging. In retrospect, it may have been better to start with easier base puzzles, to investigate the impact of narrative more effectively. The narrative puzzles were also hand-authored by the first two authors of this paper. Therefore they were subject to the writing ability and style of these authors, regardless of the impact of narrative mode.

This work highlights several opportunities for future work. While the narrative in this work was hand-authored, future works can consider generating narratives along with the puzzles, either automatically or in a mixed-initiative fashion. Future work could also consider different ways to assist players, including solving with multiple people or with an integrated hint system.

8 Conclusion

In this work, we presented logic grid puzzles in three different modes that varied in how narrative was incorporated. We tested the impact of this across two

different audiences: volunteers from social media forums related to puzzles and mysteries, and paid crowd-workers from Prolific. Volunteers found the puzzles easier but enjoyed them less. Across both audiences, the increase in narrative increased the difficulty of the puzzle. The impact this had on the overall experience varied by participant. Some participants enjoyed the increase in difficulty and engagement, while others found it frustrating and confusing.

References

1. Leigh Alexander. 2014. The joy of text – the fall and rise of interactive fiction. The Guardian (2014). https://www.theguardian.com/technology/2014/oct/22/interactive-fiction-awards-games. Accessed 30 Aug 2024
2. Bizzochi, J.: Games and narrative: an analytical framework. Loading... **1**, 1 (2007)
3. Bormann, D., Greitemeyer, T.: Immersed in virtual worlds and minds: effects of in-game storytelling on immersion, need satisfaction, and affective theory of mind. Soc. Psychol. Pers. Sci. **6**(6), 646–652 (2015)
4. Bowman, N.D., Wasserman, J., Banks, J.: Development of the video game demand scale. In: Video Games, pp. 208–233. Routledge (2018)
5. Fernández-Vara, C.: From "open mailbox" to context mechanics: shifting levels of abstraction in adventure games. In: Proceedings of the 6th International Conference on Foundations of Digital Games, pp. 131–138. Association for Computing Machinery (2011)
6. Frasca, G.: Simulation versus narrative: introduction to ludology. In: The Video Game Theory Reader, p. 15. Routledge (2004)
7. Gandhi, K., Spatharioti, S.E., Eustis, S., Wylie, S., Cooper, S.: Performance of paid and volunteer image labeling in citizen science — a retrospective analysis. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, vol. 10, pp. 64–73 (2022). https://doi.org/10.1609/hcomp.v10i1.21988
8. inkle: Ink (2016). https://www.inklestudios.com/ink/. Accessed 30 Aug 2024
9. Karhulahti, V.-M.: Fiction puzzle: storiable challenge in pragmatist videogame aesthetics. Philos. Technol. **27**(2014), 201–220 (2014)
10. Krause, M., Kizilcec, R.: To play or not to play: interactions between response quality and task complexity in games and paid crowdsourcing. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, vol. 3, pp. 102–109. Association for the Advancement of Artificial Intelligence (2015)
11. Mateas, M., Stern, A.: Interaction and narrative. In: The Game Design Reader : A Rules of Play anthology. MIT Press (2006)
12. Miller, J.A., Buse, K., Dhaliwal, R.S., Siegel, J., Cooper, S., Milburn, C.: Wrapped in story: the affordances of narrative for citizen science games. In: Proceedings of the 18th International Conference on the Foundations of Digital Games, (FDG 2023), pp. 1–11. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3582437.3582443
13. Park, N., Lee, K.M., Jin, S.A.A., Kang, S.: Effects of pre-game stories on feelings of presence and evaluation of computer games. Int. J. Hum. Comput. Stud. **68**(11), 822–833 (2010)
14. Phan, M.H., Keebler, J.R., Chaparro, B.S.: The development and validation of the game user experience satisfaction scale (GUESS). Hum. Factors **58**(8), 1217–1247 (2016)

15. Sarkar, A., Cooper, S.: Comparing paid and volunteer recruitment in human computation games. In: Proceedings of the 13th International Conference on the Foundations of Digital Games. Association for Computing Machinery (2018)
16. Shyne, F., Facey, K., Cooper, S.: Generating solvable and difficult logic grid puzzles. In: Genetic and Evolutionary Computation Conference (GECCO 2024 Companion), July 14–18, 2024, Melbourne, VIC, Australia. Elsevier (2024a). https://doi.org/10.1145/3638530.3654337
17. Shyne, F., Facey, K., Cooper, S.: Procedurally puzzling: on algorithmic difficulty and player experience in QD-generated logic grid puzzles. In: Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, vol. 20, pp. 127–137. Association for the Advancement of Artificial Intelligence (2024b). https://doi.org/10.1609/aiide.v20i1.31873
18. Siu, K., Riedl, M.O.: Reward systems in human computation games. In: Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY 2016), pp. 266–275. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2967934.2968083
19. Troxler, M., Qurashi, S., Tjon, D., Gao, H., Rombout, L.E.: The virtual hero: the influence of narrative on affect and presence in a VR game. In: CEUR-WS, Affective Computing Context Awareness and Ambient Intelligence (AfCAI) (2018)
20. Wei, H., Durango, B.: Exploring the role of narrative puzzles in game storytelling. In: Proceedings of DiGRA 2019 Conference: Game, Play and the Emerging Ludo-Mix. DiGRA, Tampere (2019). https://dl.digra.org/index.php/dl/article/view/1101