Proceedings of the Fifteenth AAAI Conference on Artificial
Intelligence and Interactive Digital Entertainment (AIIDE-19)

Is Your Game Generator Working? Evaluating Gemini, an Intentional Generator

**Joseph C. Osborn,[1] Melanie Dickinson,[2] Barrett Anderson,[2] Adam Summerville,[3]
Jill Denner,[4] David Torres,[4] Noah Wardrip-Fruin,[2] Michael Mateas[2]**
[1]Pomona College
[2]University of California, Santa Cruz
[3]California State Polytechnic University, Pomona
[4]ETR Associates

Abstract

Determining whether a game generator is working properly is challenging, since it entails conducting potentially many evaluations of generated games and synthesizing these into a net evaluation of the system. The problem is compounded when the generator has a human-centered goal: for example, that the generated games should be interpreted as having certain mechanics or as being *about* particular ideas.

In this paper, we examine the Gemini game generator and develop an evaluation instrument that tests the interpretability of its generated games' mechanics and higher-order proceduralist arguments. In the process we build empirical evidence for the claim that some amount of non-systems-based framing is required in order for arguments made by procedural rhetorics to be sensible to players. The tools we have assembled for this evaluation can be applied to game generators more broadly; game generators should be allowed to invent games which go beyond merely formally "good" or subjectively "fun."

Introduction

The automatic generation of games has been of interest to the AI community for at least 25 years (Pell 1992). As a term, *game generation* is extremely broad: what types of games? For what players? Generated by what method? Evaluated by which criteria? Desirable qualities of games change over time and for different audiences, and there is no right answer; moreover, the criteria by which a game is judged are themselves informed by the process of creating it.

Game generator designers usually begin by selecting some criteria—the length of games played, the fairness if all players play optimally, constraints on solvability, and so on—and refine these criteria after observing the generated designs. When we want to evaluate such a generator empirically to determine whether it is successful, we generally ask whether generated games score well on these metrics and whether players like to play these games. This circumscribes games as primarily either formal systems like logic puzzles or, respectively, as subjective experiences which can be summed up in a few numerical scores.

The authors of the Gemini system, rather than claiming that it could produce *high-quality* or *fun* games, instead argued that Gemini could generate *effectively interpretable* games (Summerville et al. 2018). Gemini-generated games, they asserted, could afford interpretations consistent with (and primed by) a responsive framing narrative to produce an effective game-with-a-purpose. This is a genre which itself often deals with helping players understand problems like healthcare reform, social justice, or climate change.

Only a human can evaluate the claim that a game is or is not about something (for that human, in that cultural context) authoritatively. We have therefore devised a human-centered instrument for evaluating game generators. Our approach can be generalized to other game generation criteria besides game quality, perhaps including inventiveness, visual or audio excellence, and so on. We have also combined the use of our formal instrument with more informal investigation via pilot tests (reported here) and traditional playtesting (not reported). In the process, we have discovered a minimal level of cultural communication that seems necessary to make games comprehensible and interpretable. While minimal, vague instructions regarding game controls and distinguishable signifiers for game characters are sufficient to communicate game rules, accessing wider cultural issues seems to require external framing: for example, a title and culturally-relevant colors.

Related Work

Gemini is an interesting case to evaluate because it is automatically generating games supporting certain *meanings*. We therefore need to situate our work not only in relation to game generators and how to evaluate them, but also relative how games can *mean* anything at all.

Procedural Rhetoric

The notion that games express ideas through the designs and operations of their systems is a long-standing one. For example, 25 years ago, writing of *SimCity,* Starr railed against the "built-in bias of the program against mixed-use development" (Starr 1994). More recently, discussion of how game systems express ideas has increasingly used the notion of "procedural rhetoric" as its foundation (Bogost 2007). Bogost defines procedural rhetoric as "the art of persuasion through rule-based representations and interactions, rather than the spoken word, writing, images, or moving pictures."

Many examples of procedural rhetoric combine the expression of ideas through rules and interactions with the expression of ideas through more traditional means. For example, *September 12th* portrays through its systems the arguments that missile strikes inevitably lead to civilian casualties and that civilian deaths inspire terrorism (Frasca et al. 2003). These systemic expressions are made apparent to players in a context that includes traditional visual expression (explosions flattening pedestrians, mourners huddled over them) and auditory expression (the sounds of explosions and weeping).

Some designers have, instead, attempted to explore a "strong" idea of procedural rhetoric, with a focus on what systems by themselves can express. Humble, for example, claimed that "the rules of a game can give an artistic statement independent of its other components" (Humble 2006). Humble's *The Marriage* is a game with no sound and with imagery limited to colored circles and squares. Many players do successfully understand the game to be communicating ideas about a human relationship; on the other hand, Juul argued that it only communicates these ideas "if the player understands that the game represents a marriage at all" (Juul 2007). The game's title does significant work in this regard, and even the colors employed in *The Marriage* are culturally loaded (Begy 2013). The same argument holds for many pieces of "abstract" media.

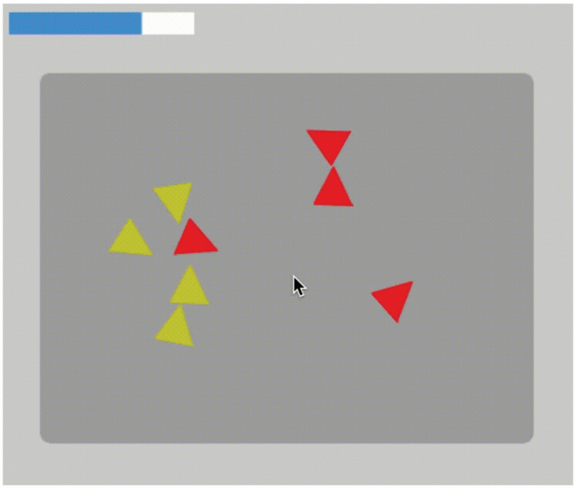
Besides these humanistic arguments, we are aware of one empirical investigation which took place after our evaluation. This compared player reactions to *September 12th, LIM,* and *Threes JS* (Anderson, Karzmark, and Wardrip-Fruin 2019; Kopas 2012; Vollmer et al. 2014). *LIM* is in the strong procedural rhetoric tradition, described as "a sparse, abstract game which conveys a powerful message to the player through its systems alone" (Allen 2014). *Threes JS* is a mathematical puzzle game, included as a control. The authors found that *September 12th* and *LIM* were understood to make arguments, while *Threes JS* was not. Surprisingly, they also found that *LIM*'s argument was widely misunderstood, possibly because it was removed from its "typical framing."

Clearly, games can communicate what they are about, and the behavior of their systems can be central to this communication. It remains unclear what non-system scaffolding players need to correctly interpret references to and arguments about cultural concepts.

Gemini

Gemini is an intent-based game generator that is given authorial intents (such as "involves resource maintenance" or "students increase [the resource] stress") and produces hundreds of different games that fulfill those intents with different combinations of mechanics (see Figure 1). The kinds of games that can be generated are simple 2D graphical one-screen games with entities and resources. Entities are represented by geometric primitives; they can move around, interact with one another, and cause things to happen. Resources are scalar values that increase or decrease when particular events occur, or over time. Players interact with generated games through mouse movement and clicks, each of which can be assigned many different possible in-game effects.

Gemini's development was guided by its conception as part of a larger project, *Emma's Journey*, a game that uses



```
required(hand_eye_coordination).
required(risk_reward).
:- not reading(good,resource(r(1))).
:- not reading(maintenance,resource(r(1))).
:- not action(mode_change(game_loss)).
:- 2 {action(mode_change(N))}.
touching_bad :-
  precondition(overlaps(E1,E2,true),0),
  reading(bad,0).
:- not touching_bad.
```

```
...
precondition(
  control_event(button(mouse, held)),
  outcome(click_to_spin_entity_e_1)).
result(outcome(click_to_spin_entity_e_1),
  rotates(e_1, ccw, 5)).
precondition(
  overlaps(e_1, e_2, true),
  outcome(o_1)).
result(outcome(o_1),
  decrease_over_time(resource(r_1), 1)).
precondition(
  le(resource(r_1), 1),
  outcome(o_4)).
result(outcome(o_4),
  mode_change(game_loss)).

precondition(tick, tick).
result(tick,
  moves(e_1, forward, resource(r_1))).
result(tick,
  moves(e_2, forward, 5)).
result(tick,
  look_at(e_2, e_1, nearest)).
```

Figure 1: A Gemini game, its generating design intent, and excerpt of generated code.

both game generation (via Gemini) and story generation to tell a story about a woman facing the different possible consequences of climate change (Samuel et al. 2017). Gemini was given a different set of design intents for each level of *Emma's Journey*, and each set yielded a pool of minigames meant to communicate an idea or feeling from the larger narrative. Our evaluation used those same games but removed them from the context of the larger game and narrative.

Gemini internally represents the *aboutness* of objects bumping into other objects and numbers increasing and decreasing, much like Game-o-Matic (Treanor et al. 2012). Unlike Game-o-Matic, it synthesizes its operationalization of proceduralist readings with the generation process via answer-set programming, a constraint satisfaction framework based on first-order logic. Instead of generating a game that hopefully matches the desired reading, Gemini derives readings and generates rules simultaneously (Summerville et al. 2019). For example, readings of the game shown in Figure 1 include that the goal is to produce resource `r_1`, that `r_1` is good for the player, that entity `e_2` consumes `r_1`, that the passage of time is bad, that entity `e_2` is bad, that the game involves hand-eye coordination and risk-reward tradeoffs, and that the game is about maintaining the level of `r_1`. Note that some of these conclusions build on each other (touching `e_2` reduces `r_1`, and `r_1` is good, so `e_2` must be bad). All of the interpretations demanded by the design intent are therefore satisfied, and indeed are proven to be satisfied in every generated game thanks to the constraint solver. Therefore, only games with internally consistent proceduralist readings can be generated, and our evaluation task is therefore to ensure that the system's reading comports with human readings.

Evaluating Game Generators

Game generation is a wide field, but here we will focus on three systems especially relevant to the topic of evaluating the interpretability of Gemini games. The first is Ludi, which produces symmetric territory-control games good enough to be sold commercially; we include it here because its evaluation was comprehensive and included both automated and human judgments (Browne 2008). The second is ANGELINA, which has created platformers that have competed in game jams and collected measurements of their subjective quality (Cook, Colton, and Gow 2017a; 2017b). The last generator we consider is Game-o-Matic, whose space of generable games is most similar to Gemini but which has never been subjected to a formal evaluation (Treanor et al. 2012).

Ludi evaluates games by gathering a few dozen game quality metrics through automated self-play. These signals feed the generation and refinement process, as successful games are selected and recombined to make new games. Do these arbitrary metrics agree with human judgments of game quality? Browne explored this question, first recreating well-known games in Ludi and asking humans to rank them pairwise, then mapping these rankings onto the engineered quality criteria. He found a subset of Ludi's metrics which correlated with human rankings and used these to produce a set of original games which a different set of humans proceeded to rank pairwise as before. The metrics' performance on this subset was also consistent with human rankings.

Browne's approach was well-motivated and appropriate for capturing the subjective play quality of abstract two-player territory-control games as a function of measurable properties of the games' dynamics. Unfortunately, it is not clear how to extend this approach to questions of interpretation and meaning, which can be hard to operationalize. Moreover, it relies on an effective automated game player to collect the game quality metrics.

ANGELINA is rooted in the literature of computational creativity and game procedural content generation, and encompasses many distinct generators with varying goals and capabilities. We focus here on two iterations, ANGELINA$_3$ and its 3D generalization ANGELINA$_4$. The former uses keyword searches and sentiment analysis to analyze news stories and automatically map images of real-world situations and personages onto game entities, yielding a game which is in some sense about the article. The game is paired with a title and a narrative commentary also generated by the system which justifies the design (Cook, Colton, and Gow 2017b). Interestingly, the game rules are held essentially constant across all ANGELINA$_3$ games, with the only rule differences concerning the function of powerups (which may affect three distinct game behavior variables). In this sense, it makes the opposite move to Gemini, which focuses on wide rule variation and minimal, abstract framing.

ANGELINA's evaluation has involved two main approaches: an informal calculation of a so-called "curation coefficient"—the proportion of the system's output which the designer would be happy to show someone—and game quality rankings obtained from the games' participation in game jams, with and without revealing the games' automated author. This latter ranking included dimensions such as "Fun," "Graphics," "Theme," and "Humour", which while addressing a broader range of human experience than Ludi's evaluation, stopped short of the question of interpretation central to Gemini's main argument. Cook further supposed that letting ANGELINA provide a mapping from objects to concepts could help players to "read" and understand its arguments, but this claim has not yet been evaluated.

Since Game-o-Matic was never formally evaluated, we can only note that it relied on a stronger human touch—human meta-designers picked nouns as concepts and connected them with verbs, and assigned graphics to the nouns—while limiting in some ways the scope of relationships between objects (it could only capture a fixed set of binary relations) (Treanor et al. 2012). Gemini's use of a design intent language enables more open-ended construction of meaning, while its abstract graphics and relative lack of framing give less interpretive support to players; Gemini's space of possible mechanics is also larger.

This also situates Gemini in a different space from recent work by Guzdial and Reidl (Guzdial and Riedl 2018). The conceptual expansion approach mines game mechanical relations of known types from game character and background graphics and then modulates these relations arbitrarily (or via direct human interaction) to create new games; on the other hand, Gemini starts from a rich knowledgebase of game interpretation and orients generation of a fixed set of mechanical relations around those interpretive goals.

Evaluating Gemini

Gemini's goal is to generate the playful components of meaningful, *effectively interpretable* games, akin to newsgames. Importantly, human authors are asked only to give *design intents*—constraints on the space of generated games. The Gemini game generator makes two key claims: first, that it understands the games it generates; and second, that human designers or players will interpret its games consistently with respect to its own understanding.

Effectively interpretable means two things: that players can understand the rules, and that players can confirm or even independently come up with interpretations consistent with the system's. We expect that the first is prerequisite to the second, and *comprehensible* game mechanics are not guaranteed with highly procedural game generators like Gemini and their large combinatorial spaces of possible mechanics. Thus, a crucial problem we had to solve in designing our evaluation was how to present the games in a way that didn't prime players with explicit instructions for how to play or what the game means, while also giving them enough information to reasonably discover how to play and interpret it on their own.

We undertook several pilot studies before our main experiment. In our preliminary studies we saw that without *any* instructional or interpretive information, new players unfamiliar with Gemini were lost, frustrated, and resistant to engaging with the frameless games. Unfortunately, following good game design practice by explicitly providing instructions and themed player goals would preclude meaningfully asking participants to interpret and explain the game's mechanics and meaning.

The balance we ultimately struck was to outline for the participants the space of possible mechanics or controls that might appear in Gemini games, without specifying which would occur in the game they were currently playing. After playing, participants were asked to select mechanics that were present from a list of possibilities.

This style of question is less amenable to higher-level thematic interpretations, since the space of subjective interpretation is much broader and more open. We took two general approaches to interrogating what participants thought games meant. In both initial pilot studies and the larger experiment, we solicited free text responses to questions like, "What do you think the game is about?" or, "What message (if any) is the game trying to convey?" For the larger formal experiment, we additionally gave a small set of specific interpretations in the post-play survey which participants were asked to choose between.

Participants had trouble answering open-ended interpretation questions, often leaving them blank or describing mechanics rather than meaning. We also found that some individuals are good at articulating interpretations of media—regardless of how well the media communicates its message, scaffolds interpretation of that message, or if it even has an underlying message to communicate. We therefore recommend both free text and multiple choice questions providing subsets of possible interpretations (in that order, to prevent priming).

Providing a possibility space of mechanics gave players the tools they needed to experience agency and successfully interpret mechanics. On the other hand, our approaches for evaluating higher-level interpretations of games could use further refinement. Moreover, this kind of deep human-centered evaluation is not meant to scale to large numbers of conditions or exhaustively cover a generative space. Instead, we hoped to closely investigate individual generated artifacts which are presumed to be representative of the generative space as a whole. It is therefore important to keep in mind that we are evaluating six specific games in these experiments as a proxy for the larger possibility space of Gemini games created by these design intents. We believe that Gemini's intention-oriented architecture helps ensure that these are indeed representative examples.

Pilot Studies

We conducted several pilot studies of Gemini, evolving our instrument over time. Our first tests asked participants to play a single game that had been generated for *Emma's Journey*, divorced from its larger context. We provided explicit, game-specific mechanical instructions without thematic flavor as participants played. Participants played as much as they wanted and then answered a questionnaire, identifying the game's mechanics and interpreting what the game might be trying to communicate. If the strong proceduralist hypothesis held, participants would identify the game's mechanics and dynamics, as well as the implicit, higher-level interpretive goals the author encoded in the design intent.

The results were mixed, with some of the mechanics and interpretations being consistently correctly identified, some that weren't consistent, and some that were consistently incorrect, for the specific Gemini game being played. Unfortunately, the correct answers were often just reproductions of low-level control instructions like the ones we provided players. Furthermore, the game was still entirely abstract, leaving participants to make significant interpretive leaps to understand it as representing more than simple shapes and colors. In the questionnaire for the first two pilot tests, player responses included saying they "don't know what anything represents," and that they thought the game was about "testing how people react to abstract games" or "nothing."

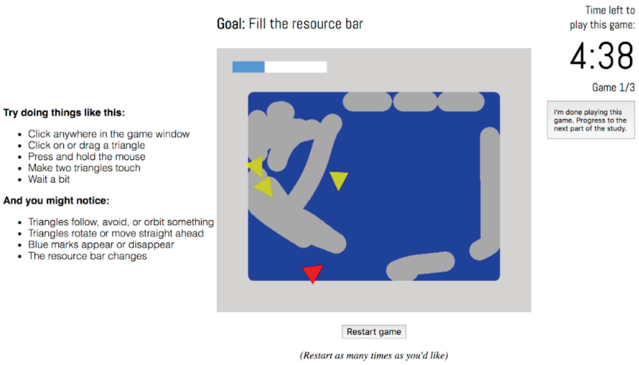
We saw improvement in participants' ability to correctly identify game mechanics and intended interpretations, as well as decreased confusion, when we added even minor thematic scaffolding in the instructions and by adding a thematic title (e.g., "Red triangles represent people trying to clean up a beach and save a crab population. They can be dragged around with the mouse," for a game called "Beach Cleanup"). This context grounded the game in a real-world domain, sparking more positive player reactions ("It made me proud that a series of small changes became something good"), and more sophisticated interpretations of what the game was about, such as "people working together to take trash away." There were, however, still reports of confusion and difficulty in interpretation from some of the participants. Hooking into priming fiction in this way is more consistent with the theory of proceduralist readings, which explicitly accounts for players' expectations of game object behavior in terms of real-world social knowledge about those objects' visual representations. The abstract visuals of these games were never meant to stand apart from the fiction, and removing the games from their context made it difficult for players to read them. This was a surprisingly important result to which we will return in the conclusion of this paper.

Experiment

Our formal experiment had three goals: to investigate players' mechanical understanding; to see whether players and Gemini agreed on game dynamics (such as "this game is high stakes" or "blue harms red"); and to check whether players' interpretations of games' high-level meaning was consistent with the designer's. It is important to note that Gemini reasons about this high-level meaning only to the extent specified by the designer in their intent file; for example, "`distraction` is bad for `student`" is what the designer might have written as one their design intents in the context of an (implicit) argument about the role of smart phones in the classroom; Gemini has no built-in knowledge of what a student, a distraction, or a smartphone is.

We decided to present different players with two different sets of generated games: In the first set, all three games had similar mechanics but one had a different intended meaning from the others; and in the second, two of the games had different mechanics and came from the same design intent, while the third had similar mechanics to one of those two and came from a different design intent. These six games were generated from the most thoroughly tested design intents and selected by hand from a list of hundreds of generated games according to the similarity/difference criteria just described, where the first candidate that seemed different or similar enough was selected. Participants were therefore split into two groups; three specific games were chosen for the first group and three for the second group. The first mini-experiment checked whether participants could distinguish different meanings despite similar mechanics, and the second verified that the same meanings could be recognized even if the mechanics varied.

Participants Seventy-six undergraduate university and community college students participated in the formal experiment. Each participant was compensated with a $25 Amazon gift card. Recruitment was done through announcements in classes, posted fliers, and social media. All recruitment material described the task as playing three games and filling out surveys for a "video game study," with a time estimate of 30-45 minutes, compensation information, and assurance that gaming experience is not required. Nothing in the recruitment material, protocol, or instrument informed the participants that the games were generated by a computer program, and none of the participants had prior exposure to Gemini or its games.



Goal: Fill the resource bar

Time left to play this game:

4:38

Game 1/3

Try doing things like this:

- Click anywhere in the game window
- Click on or drag a triangle
- Press and hold the mouse
- Make two triangles touch
- Wait a bit

And you might notice:

- Triangles follow, avoid, or orbit something
- Triangles rotate or move straight ahead
- Blue marks appear or disappear
- The resource bar changes

I'm done playing this game. Progress to the next part of the study.

Restart game

(Restart as many times as you'd like)

Figure 2: The final study interface.

Of the 76 participants, 57% identified as male, 42% identified as female, and 1% identified as non-binary. Most students (92%) were between 18 and 23 years old. Only 9.2% reported being in a games-related major (e.g., Computational Media, Game Design, and Digital Media). Ninety-five percent of the participants reported having played video games at least once in their life but only 74% reported that they currently play video games. Only 46% of participants reported having experience playing abstract, non-representational games like the ones in the study.

Instrument Data were collected using a website instrument (presented in Figure 2) for playing Gemini games and taking online pre- and post-surveys, with links between them. After the pre-survey, participants were taken to a site introducing the gameplay task, and the range of possible actions and events that might appear in each game (e.g., "press and hold the mouse" or "click on or drag a triangle"). That list of possible actions and effects persisted on the screen throughout gameplay, so participants did not need to memorize it.

After the introduction screen, participants would play three different games, each for up to five minutes or until pressing a "next" button, and complete a post-survey after each. The post-survey was identical for all games, and was designed to measure the participants' initial reaction to the game they just played and how well they understood it. The first questions measured self-reported stress, engagement, difficulty of playing the game, and difficulty of understanding it. The next set were free text questions about the rules of the game, the strategy the participant used to play, and interpretations they made of the game. The third set involved identifying mechanics that were present in the game via multiple choice questions. Finally, after a participant played the three games in their set and answered survey questions after each one, they were given a final question that asked them to select a high level interpretation from two possible interpretations for each of the games they played.

Procedure Data collection took place at the university and community college, in a classroom, library, and user study lab. Participants were provided either a laptop with a connected mouse or a desktop PC on which to play the games.

They were randomly assigned to one of two groups, as described earlier. We achieved roughly equal group sizes, with 53% of participants assigned to Group A and 47% assigned to Group B. The order of the games was randomized for each participant. During the study, an administrator walked around the room to answer questions and ensure that the participants were not talking or looking at each other's screens, but otherwise left all introduction and instructions to the instrument for consistency.

We developed an answer key for the multiple choice questions testing mechanics recognition by identifying code snippets in each generated Gemini game associated with particular mechanics. The answer key allowed us to compute a score for a participant's mechanical understanding of each game they played (the mean score of all mechanics multiple choice questions, marked either correct or incorrect).

To measure higher-level interpretations, we coded qualitative responses. For each Gemini game, independent raters categorized responses to the question, "In your own words, what message (if any) is the game trying to convey?" into those that included some form of interpretation. Across all games, inter-rater reliability indicated a very high level of agreement, with Cohen's kappa = 0.81, $p < 0.001$. Where necessary, ties were broken by a third independent rater. Participants provided interpretive responses to 22.4% of games (e.g., "patience is key" or "challenges are easier with a team"), explicitly indicated that they believed the games contained no underlying message for 21.3% of games ("No message," "None," etc.), and indicated some degree of uncertainty for 12.3% of games ("I don't know," "I'm unsure," etc.). The remaining participants provided either no response (26.1%) or responses that indicated failure to understand the question (17.9%). Due to brevity of responses and the variety of possible interpretations, no attempt was made to judge their relevance or quality.

Results According to our mechanics answer key, participants agreed with Gemini about the lower-level mechanics present in its games 82.1% (*SD* = 7.00%) of the time, suggesting that Gemini games are relatively easy to understand at the level of mechanics if players understand the space of possible mechanics.

However, participants were only able to articulate free text interpretations to less than a quarter (22.4%) of the games played. And when asked "What interpretation do you feel best applies to each game you played?" only 43.4% (*SD* = 21.1%) selected the designer-intended high-level metaphor for each game from a set of two intentionally contrasting metaphors ("Contemplatively completing a beach clean-up," "Working to keep student confusion from getting too high," and "Neither of these themes feel like they could match this game"). While the proportion of players who selected the designer-intended interpretation out of three options (including "neither") was higher than the null hypothesis of 33%, a binomial exact test indicated that it was not significantly better than chance, with p = 0.068 (1-sided), 95% *CI*: [32.1%, 55.3%]. In Group A, the three games were all oriented mechanically around removing or minimizing a "bad" color on the playfield; two games required dexterity and timing and were meant to describe something like "scrubbing away confusion" in a lecture setting while the third was the contemplative, low-stakes beach cleanup scenario described earlier. In Group B, two of the games were (mechanically distinct) beach cleanup games while the student confusion metaphor was realized as dodging distractions. In both cases, players were unable to consistently ascribe the designer-intended meanings to games. This result compares interestingly with our preliminary findings in the pilot study, where framing information in later pilots effectively brought players to the designer-intended interpretation.

From this we conclude it is difficult to interpret higher-level meanings from abstract, themeless games; this is compounded for the more metaphorical lecture setting versus the relatively concrete beach cleanup. This is consistent with work (published after and separately from our experiment) finding participants incorrectly interpreted messages in an abstract, metaphorical game designed by a human expert (Anderson, Karzmark, and Wardrip-Fruin 2019). The fact that our work duplicates those results is significant for future abstract game generators.

Participants did especially well at identifying the mechanics of two games, "beach_A2," with a mean score of 90.0% (*SD* = 7.09%), and "beach_A1," with a mean score of 85.7% (*SD* = 11.4%). These games were the simplest of the six, involving clicking and dragging otherwise static triangles around the screen, and making them touch to erase a small area of background color underneath them, with the goal of erasing most of that color. "Beach_A1" was almost identical to "beach_A2," but with one additional player verb (click to spawn new triangles in random locations, as opposed to new triangles spawning on a timer). We suspect the 5% drop in mechanics comprehension is due to player actions triggering multiple effects (players clicked both to drag triangles and to spawn new triangles).

In the other four games, unlike the beach setting, some of the game objects were not controlled by the player. The average mechanics understanding scores for these games was lower (*M* = 79.3%, *SD* = 7.58%), suggesting it becomes more difficult to understand what is going on in an abstract game as soon as entities move around autonomously. We believe that adding models of visual feedback or motion design to abstract game generators like Gemini could help scaffold readings of intentionality.

Interestingly, participants who reported having played abstract games before scored significantly better at mechanics understanding (*M* = 84.3%, *SD* = 6.18%) than those reporting no prior experience with abstract games (*M* = 80.4%, *SD* = 7.19%), $t(74)$ = 2.53, p = .014, 95% *CI*: [.84%,7.05%]. There was no observed difference between these groups in either the multiple-choice interpretation of Gemini games' higher-level meanings or in the free-text interpretations of the games' meanings. We also (surprisingly) observed no statistically significant relationship between mechanics understanding of a game and how well players could interpret meaning from it, either on the free-text or multiple choice interpretation measures.

Conclusion

It is now clear that interpreting abstract games, beyond the operations of their systems, requires a baseline level of framing. This is consistent with the theory of proceduralist readings's initial development, but helps delineate how far is *too* far to stretch a purely systems-based argument—and in the process, how successful the purely abstract version of Gemini is on its own. To get to this point, we needed to develop a new human-centered approach which we believe will generalize to the evaluation of other game generators: not only defining goals ahead of evaluation time, but ensuring that the system gives a strong internal justification that it has achieved those goals. While not all generative systems work in terms of explicit intents, most systems at least have implicit design goals defined by the context in which generated artifacts will appear or the parameters offered to authors; as far as possible, these should be made explicit and evaluated against.

In summary, Gemini assumes a distinction between mechanics, dynamics, and design intentions (which circumscribe a space of desired interpretations without specifying mechanics). Knowing how Gemini succeeds at each of these aspects is important to evaluating the generator as a whole. We found in our pilot studies that all three might be working—if players are provided with some thematic context—but we weren't able to confirm the third aspect in a larger experiment. Specifically, we found that it is at least true that players cannot produce these proceduralist interpretations retrospectively when provided with multiple-choice questions after play. The uninterpretability of the abstract games used in the evaluation reinforces the conclusions of related work on interpretations of abstract games (Anderson, Karzmark, and Wardrip-Fruin 2019).

Games are not just entertainment machines, and creators of game generators can aim for goals besides generating enjoyable games that people like to play. In many cases, these goals will establish criteria that require human deliberation. For Gemini, we asked whether design intents were recognized by players; for other systems, other questions must be asked, but these are rarely easy to measure purely automatically. As in PCG broadly, game generators must be evaluated based on whether they are generating what they are supposed to; if the generator cannot reason about its own processes, the creator should set those terms and goals before evaluation (Shaker, Smith, and Yannakakis 2016; Summerville 2018).

Acknowledgements

The authors would like to thank Max Kreminski for their help coding qualitative responses and contributions to data analysis. This material is based upon work supported by the National Science Foundation under Grant No.1409992 and by Seed Fund Award CITRIS-2017-0187 from CITRIS and the Banatao Institute at the University of California.

References

Allen, S. L. 2014. Video games as feminist pedagogy. *Loading...* 8(13).

Anderson, B.; Karzmark, C.; and Wardrip-Fruin, N. 2019. The Psychological Reality of Procedural Rhetoric. In *Procedings of FDG 2019.*

Begy, J. 2013. Experiential Metaphors in Abstract Games. *Transactions of the Digital Games Research Association* 1(1).

Bogost, I. 2007. *Persuasive Games: the Expressive Power of Videogames*. MIT Press.

Browne, C. B. 2008. *Automatic generation and evaluation of recombination games*. Ph.D. Dissertation, Queensland University of Technology.

Cook, M.; Colton, S.; and Gow, J. 2017a. The angelina videogame design system—part i. *IEEE Transactions on Computational Intelligence and AI in Games* 9(2):192–203.

Cook, M.; Colton, S.; and Gow, J. 2017b. The angelina videogame design system—part ii. *IEEE Transactions on Computational Intelligence and AI in Games* 9(3):254–266.

Frasca, G.; Battegazzore, S.; Olhaberry, N.; Infantozzi, P.; Rodriguez, F.; and Balbi, F. 2003. *September 12th: A Toy World*. Newsgaming.com.

Guzdial, M., and Riedl, M. 2018. Automated game design via conceptual expansion. In *Fourteenth Artificial Intelligence and Interactive Digital Entertainment Conference*.

Humble, R. 2006. Game Rules as Art | Can a Game Make You Cry? *The Escapist* (41).

Juul, J. 2007. A certain level of abstraction. In *Digital Games Research Association Conference*.

Kopas, M. 2012. *LIM*.

Pell, B. 1992. Metagame in symmetric chess-like games.

Samuel, B.; Garbe, J.; Summerville, A.; Denner, J.; Harmon, S.; Lepore, G.; Martens, C.; Wardrip-Fruin, N.; and Mateas, M. 2017. Leveraging procedural narrative and gameplay to address controversial topics. In *International Conference on Computational Creativity*.

Shaker, N.; Smith, G.; and Yannakakis, G. N. 2016. Evaluating content generators. In *Procedural Content Generation in Games*. Springer. 215–224.

Starr, P. 1994. Seductions of Sim, Policy as a Simulation Game. *The American Prospect* 5(17).

Summerville, A.; Martens, C.; Samuel, B.; Osborn, J.; Wardrip-Fruin, N.; and Mateas, M. 2018. Gemini: Bidirectional generation and analysis of games via asp. In *Fourteenth Artificial Intelligence and Interactive Digital Entertainment Conference*.

Summerville, A.; Martens, C.; Harmon, S.; Mateas, M.; Osborn, J.; Wardrip-Fruin, N.; and Jhala, A. 2019. From mechanics to meaning. *IEEE Transactions on Games* 11(1):69–78.

Summerville, A. 2018. Expanding expressive range: Evaluation methodologies for procedural content generation. In *Fourteenth Artificial Intelligence and Interactive Digital Entertainment Conference*.

Treanor, M.; Blackford, B.; Mateas, M.; and Bogost, I. 2012. Game-o-matic: generating videogames that represent ideas. In *Proceedings of the Third Workshop on Procedural Content Generation in Games*, 70–77. ACM.

Vollmer, A.; Wohlwend, G.; Hinson, J.; and Li, A. 2014. *Threes JS*.