Toward Automated Critique for Student-Created Interactive Narrative Projects

Shruti Mahajan[1], Leo Bunyea[1], Nathan Partlan[2], Dylan Schout[2], Casper Harteveld[2], Camillia Matuk[3], Will Althoff[4], Tyler Duke[4], Steven Sutherland[4], Gillian Smith[1]

[1]Worcester Polytechnic Institute, Massachusetts, USA {spmahajan, lrbunyea, gmsmith}@wpi.edu

[2]Northeastern University, Massachusetts, USA {partlan.n@husky.neu.edu, dylan.schouten@gmail.com, c.harteveld@neu.edu}

[3]New York University, New York, USA {cmatuk@nyu.edu }

[4]University of Houston Clear Lake, Texas, USA {althoffw7955@uhcl.edu, tduke09@gmail.com, sutherland@uhcl.edu}

Abstract

Automated feedback has the potential to provide significant assistance to student game creators. Here, we present a system for generating automated, critique-like feedback for students creating games in the StudyCrafter platform. We implemented a system that builds a personalized feedback report for students based on a templated format. This critique uses automated analysis of structural and interactive aspects of the game narrative and recommends alternate games for students to examine as inspiration. To test our system, we conducted a pilot study with 10 student groups developing narrative-based games. A key understanding from the study is that determining the appropriate depth of assessment and critique without overwhelming the student is important.

Introduction

Games are often created by designers who lack game design expertise, especially in contexts where games are being used for a purpose that is in addition to entertainment. For example, students may make games to explore a particular content area–so-called "constructionist gaming" (Kafai and Burke 2015). In such design contexts, game creators often use tools that are intended to be easier to use than more complex game engines (for example, Twine (Klimas 2009) or Scratch (Resnick et al. 2009)), in order to reduce the technical complexity of the game creation process. While these tools make it easier to *program* games, they do little to offer assistance in *designing* games (Gee and Games 2008). Designers who are operating in these contexts may have limited access to expert designers; for example, students who make games to learn about programming in their computer science classes may not have a teacher with game design experience in the room.

How do we better support novice designers in their design process? While previous work in AI-based design support has focused largely on co-creation (Liapis, Smith, and Shaker 2016) or automated playtesting (Holmgård et al. 2018), we propose using AI to simulate a different part of the iterative design process: expert, in-process critique. Critique is commonly used in classroom settings (Costantino 2015) across a variety of art and design disciplines. Critique can be conducted by either an expert (such as a teacher or practicing professional) or by peers (such as fellow students). It is typically process-oriented, offering commentary on the current iteration of the creative project and suggestions for either changes or alternate directions the project might take in the future. Thus, an AI system that performs critique is a form of creativity support system that plays the role of a "coach" according to Lubart's taxonomy (Lubart 2005).

In this paper, we present a prototype for an automated critique system that operates in the domain of interactive narrative games, and insights into challenges faced in critique generation for a student audience. These insights are derived from both reflecting on the creation of the system itself, and from feedback on the system provided by student game creators in a pilot study. For this study, students were creating choice-based narrative games using a tool called StudyCrafter[1] (Harteveld et al. 2017), as part of an undergraduate psychology course focused on experiment design. StudyCrafter supports creating narrative-based games with 2D graphical elements drawn from an existing library of assets; the design tool is similar in complexity to Scratch or Twine. The games students created in the course were virtual experimental *scenarios*, often taking the form of small narrative-based scenes or vignettes. The students did not have any opportunity for expert feedback on their projects from a game or narrative design perspective. As a prototype, the critique generator is not integrated into the StudyCrafter user-interface. This allows for easier iteration, but means that student feedback is not instantly available, and must be generated by the research team.

Our automated critique system incorporates two primary elements of critique: a) feedback on the current status of the project, as it compares to other projects the system has analyzed, according to different aspects of interactive narrative design; and b) suggestions for other projects students might look to that are both highly related to and extremely distant from the project that they are creating. The system was tested twice: once for early-stage projects after students had spent only a week on game creation, and again on the same projects at a much later stage of creation. Students provided feedback on how helpful they perceived the feedback to be,

[1]https://studycrafter.com/

and what they wished to see in the future.

We begin this paper with a description of how automated critique generation fits into a larger landscape of AI-based creativity support in game design contexts, and automated evaluation of creative artifacts. We then describe the approach we take to generating the automated feedback, and provide a partial illustrative example of the type of feedback that is generated. We describe the methods and results from a pilot study with 10 student groups creating narrative games in a psychology course, and close with a discussion of how we intend to improve this work in the future as well as broader challenges to face in automated critique generation in educational contexts.

Related Work

We position our research relative to other work in AI-based creativity support in game design contexts. Much work in this space has focused on co-creativity, in which an AI agent creates portions of game content (often levels) alongside the human creator (Yannakakis, Liapis, and Alexopoulos 2014; Shaker, Shaker, and Togelius 2013; Guzdial et al. 2019). Of these tools, Sentient Sketchbook (Yannakakis, Liapis, and Alexopoulos 2014) and EDDY (Baldwin et al. 2017) may be closest to our aims, as they both provide a visualization of properties of the generated content in realtime–thus providing designers with a vocabulary to reason about differences between design variants. Though our system has no co-creative component, we do aim to provide in-process feedback on the current design state.

Another area of work that we draw inspiration from is automated-playtesting. The goal with automated playtesting is typically to provide near-instant feedback on the qualities of a game from a player's perspective, simulating the kind of feedback usually received from players, as part of an iterative design process. Automated playtesting frames design feedback in terms of 'correctness' (removing bugs) or 'satisfiability' (making a game that meets experience goals). For example, Zook et al. (2014) propose a machine learning-based approach to both testing and tuning game parameters to achieve satisfactory results; Holmgard et al. (2018) develop "personas" that allow them to predict how different kinds of players will react and behave in designed game environments. Our goal with simulated critique is to similarly operate within the context of an iterative design process, but to provide feedback that can be incorporated into student reflection. It does not take the place of playtesting with human players, nor is it intended to.

System Overview

Figure 1 shows an overall architecture diagram for our approach to automated critique report generation. Our inputs are the in-progress scenario provided by the student, a database of complete scenarios, a metrics configuration file, and a template for the generated feedback. We apply a subset of previously derived metrics for StudyCrafter projects (Partlan et al. 2018) informed by an underlying theory of interactive narrative (Carstensdottir and Seif El-Nasr 2018), divided into three categories: narrative structure complexity, interaction point affordances, and interactive affordances, to both the input file and the scenario database. We then normalize the results, and cluster scenarios for each category of metrics. Finally, we use the results from this clustering process to expand a human-authored template to produce a customized feedback report for each project.

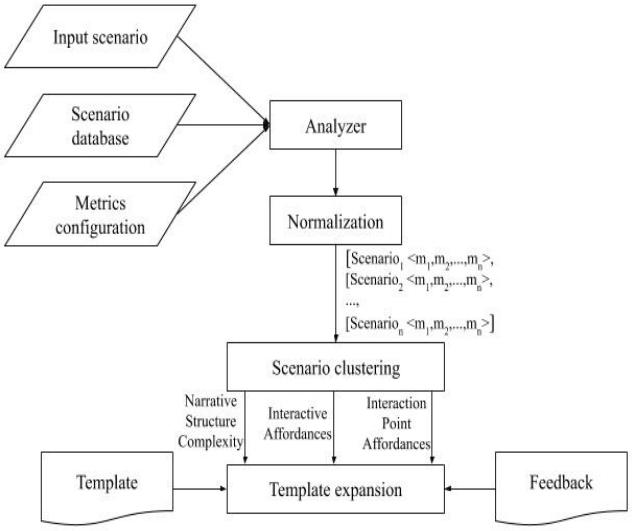

Figure 1: This diagram represents the stages of analysis, clustering, and template-based critique generation.

Input Scenario

Our system is intended to work at any level of project completion, including early stage drafts without a complete narrative. For example, one of our input games in this study involves a customer (the player) purchasing coffee from a barista at a cafe. In the first iteration of the project, the screen is entirely white. The player chooses their gender and an avatar to represent themselves. After the avatar is chosen, there is a black screen and no other actions can be performed. In iteration two, there are still no graphical elements to the game. However, there are dialogue boxes that now appear after an avatar has been chosen. These dialogue boxes play out the conversation between the customer and the barista, with the mood of the conversation fluctuating based on player choices.

Scenario Database

Along with the input scenario, the other input to our system is a dataset consisting of 23 complete scenarios created in StudyCrafter by students from previous iterations of both the course we are testing our feedback system in and another, similar course at a different university. These scenarios vary greatly from each other in terms of narrative structure, action space, length, number of characters, and how integral story is to the play experience. Because the scenarios in the database are created by a similar population of student authors, we expect that input scenarios are likely to fit somewhere within (or near) the spectrum of design possibilities encapsulated by the scenario database. Using a scenario

database lets us calibrate a "scale" for each metric we apply, and allows us to offer feedback to students relative to other scenarios that are of a familiar complexity and focus.

Analyzer

From the 24 metrics that were initially proposed by Partlan et al. (2018), we selected 9 of these metrics for use in feedback generation. See Table 1 for a list of the metrics we used. We reduced the number of metrics for two major reasons: to ensure that clustering would not disproportionately favor any single aspect of the narrative design, and to ensure that metrics can be run on projects that are incomplete.

To balance metrics and reduce bias during the clustering phase, we calculated which metrics are linearly dependent for the projects in the scenario database, and excluded metrics that were linearly dependent on others. We chose which of the linearly dependent pairs to keep based on how interpretable we felt the metric would be to authors with a limited background in computer science or narrative design. For situations where metrics were calculated for both the 'total' and 'average,' we selected only one of them to keep for critique generation.

The metrics are designed to run on complete scenarios, yet we are giving feedback to students whose work is in-progress, and as a result are likely to have disconnected graph components. We classify each metric based on how well it is expected to perform for early and late stage projects, and only use those metrics that correspond to the stage of the project that is input. Early stage metrics are those that can be calculated on incomplete story graph representations. An example of an initial metric that we chose to exclude is one that calculates the number of strongly connected components in a graph; for an incomplete story graph, this metric will not report useful information that is likely to be applicable in future design iterations.

Clustering

We generate three different clusterings of scenarios, based on each grouping of metrics: narrative structure, interactive affordances, and interaction point affordances. The narrative structure clusters compare similarities in the story structure of the current project with other projects in the database. The interactive affordances clustering focuses on the opportunity for player action across the scenario. Interaction point affordances clusters based on what each of the individual choices looks like in the scenarios.

Algorithmic Description The purpose of clustering is to visualize types of scenarios in the database and to compare and visualize the proximity of the input scenario to existing scenarios.

Preprocessing. The clustering stage takes, as input, the metric values for the input scenario and all scenarios in our database. These metric scores are then normalized. The metric values in our database each have a large range, and these ranges differ per metric. For example, 'Edges Traversed Variance' ranges from 1.26×10^{-25} to 273485.3602, while 'average script nodes' ranges from 4 to 211. When clustering with such different score ranges, results can be skewed by larger scores. To avoid this, normalization is necessary. The data are normalized by converting all the scores in the scenario database to a scale from 0 to 1.

Hierarchical Clustering. We cluster scenarios using complete-linkage hierarchical clustering (Manning, Raghavan, and Schütze 2010). In this method, the distance between two clusters is defined by the distance between the two points that are farthest from each other in the clusters.

Initially every scenario is a separate cluster. By repetitive merging they all end up belonging to one big cluster. The intermediate steps and the clusters formed are particularly important. The clusters that have the least distance are merged; this distance is calculated by the above-mentioned complete linkage method. In order to visualize these clusters, their formation and how they merge at each step, we use dendrograms, a tree representation of this process. Lastly, once separate clusters are identified, we find the nearest and the farthest scenario from the scenario in question. The nearest scenario has the shortest Euclidean distance to the input scenario and is thus most similar within the cluster. The farthest scenario, calculated by the longest Euclidean distance from the scenario, is the most dissimilar one, found in the most distant cluster.

Template-Based Feedback Generation

Our primary goal in generating feedback for students is to prompt them to reflect upon their own design choices. We use a human-authored feedback template to guide the feedback generation process. Here we report the final version of the feedback template; an initial version was used for the first round of feedback in our pilot study, and modified as discussed in the Pilot Study section. For space reasons, here we provide the general structure of the document but not the specific text used.

Feedback Introduction An introductory paragraph serves to set the reader's expectations for both the expected accuracy and nature of the feedback. It does this by briefly explaining how the results were generated and how to get the most benefit from the document, and by clarifying that the results are automatically generated.

Categorical Introductions The introductions to each of the three categories for feedback (Narrative Structure, Interactive Affordances, and Interaction Point Affordances) with metrics results data start with an overview of their respective topic. They then suggest two reference projects for the reader to explore for creative inspiration. The first of these projects is the scenario from the scenario database that is closest to the input, depicting what the scenario might be similar to in it's completed stage, if the students follow the same design path. The second is a scenario from the furthest cluster from the input scenario, it is unrelated to the input scenario. The farthest scenario is suggested to introduce diverse ideas and possibilities. These paragraphs end with links to relevant subsections within the "Additional Information" subsection.

Metrics Results Each metric result is organized into a sentence that does not list the project's explicit numer-

Cluster	Metrics	Definition	Justification
Narrative Structure	Average Outdegree	Average number of edges leaving the nodes in the script graphs	Determines linearity of narrative
	Average Layout Nodes	Total layout nodes divided by number of scenes	Determines number of objects in the scene
	Average Script Nodes	Total script nodes divided by number of scenes	Determines average length of script in a scene
	Edges Traversed Variance	Variance (average squared deviation from the mean) of Average Edges Traversed	Determines the variation in experience length between playthroughs
	Number of Scenes	Number of vertices in the scene graph	Determines length of experience
Interactive Affordances	Average Number of Choices	Number of choice edges in the interaction maps, averaged over all maps	Determines average number of choices available to the player across each scene
Interaction Point Affordances	Choices Per Interaction Point	Average number of choice edges out of each interaction point	Determines how many choices are offered to a player at each prompt
	Feedback Per Interaction Unit	Number of feedback nodes divided by the number of interaction nodes	Determines the amount of visual feedback a player receives for making a choice
	Feedback Per Event	Number of feedback nodes divided by the number of event nodes	Determines the ratio of things that happen based on player action versus things that happen regardless of player action

Table 1: The metrics for each of the three categories: Narrative Structure, Interactive Affordances, and Interaction Point Affordances

ical score, but instead provides a comparison to other projects in the scenario database. There are four different terms that can be inserted into the metrics explanation sentence, determined by which quartile the project falls into for each metric. For example, *"Compared to other projects in our database, players are [extremely infrequently/occasionally/frequently/often] presented with a choice."* Our reasons for not providing a numerical score are discussed in the Pilot Study section.

Additional Information The Additional Information section is filled with subsections that provide detailed, practical explanations of the terminology and concepts that are used in the feedback report.

Generator Results

Figure 2 is an example feedback report for a scenario. Fields that vary based on the project are *italicized*. For brevity, we exclude the "Additional Information" section.

Pilot Study

We conducted a pilot study in order to evaluate how easy to understand and helpful the generated feedback would be to people developing scenarios in StudyCrafter. The participants, students in an undergraduate experimental design course taught in psychology by Dr. Sutherland (a co-author on this paper), worked in groups to develop a scenario that served as a behavior science research experiment. There were 10 groups in total. With this study we collected participants' response to generated feedback, and also asked students about how helpful feedback was at the end of the semester after their projects were submitted.

Methods

Groups created their projects over the course of a semester, with approximately 5 weeks dedicated to working in the software. There were two intermediate deadlines where they sent us their projects: the first at week 1, the second at week 4. The goal with this was to determine how useful the feedback was and how to improve the feedback generation process.

We asked for projects at these two stages to identify how the feedback may need to change over time, and which metrics could be included at each of these stages. Generating feedback would be more applicable at stages with enough content to evaluate. For each of these intermediate deadlines, we generated feedback for the students within a week. Note that for the first deadline, we were using an initial version of the feedback form that differs from the version reported in section "Template-Based Feedback Generation" in the following key ways:

- The first draft of the template reported the metric numbers.
- The first draft included detailed descriptions of what the metrics and categories mean alongside the feedback. The final version has links that students can follow to further detail and concept explanation, and keeps the portions unique to the project on the first page.
- Language used in the first template contained more technical and mathematical explanation in much more detail than the second.

Student Survey Results

Students were asked to provide an open-ended response to the usefulness of feedback after each round, and how they thought it might be improved in the future. In the first round of feedback, 9 out of 10 groups responded. In the second round, 6 out of 10 groups responded.

Stage 1: At this stage, there were two major themes in the feedback we received from students: a false impression of

Group *10 - "Bystander Effect"* Feedback
Below is a report generated by running analytics on your project in its current form. Because your project is incomplete, there are some metrics that we were unable to generate scores for. Some of the metrics that are included in this report may not be completely accurate or may change drastically as you continue to build upon your project. They are however, our best guess for providing you with meaningful feedback. This report consists of an analysis of the current state of your project based on a predetermined list of metrics. It is also important to note that this feedback is purely on your project's gameplay experience, not on the design of the experiment. Additional information about different aspects of your project, can be found on the next few pages.

Narrative Structure
The information in this section will give you an idea of how complex your story is. If you continue to design your project in the way indicated by the metrics, your final game will most resemble *"Color and Shapes."* A project that is structured entirely differently from yours in this area is *"Deserted Island: Cabinet Mystery."* Consider looking at either of these two projects for ideas as you continue to work. If any of the following statements are difficult to understand, please see the following sections on the following pages for more information: Branching vs. Linear Narrative, Number of Scenes, Scene Clutter, Script Length, and Experience Length.
Your current narrative is structured linearly.
Compared to other projects in our database, you have an *above average number of scenes*.
Compared to other projects in our database, scenes in your project are *highly populated*.
Compared to other projects in our database, the length of your project's script is *short*.
Compared to other projects in our database, the length of your game experience is *very short*.

Interactive Affordances
The information in this section will give you an idea of the opportunity for player action or choice across your entire project. If you continue to design your project in the way indicated by the metrics, your final game will most resemble *"Course Selection Frenzy."* A project that is structured entirely differently from yours in this area is *"Deserted Island: Cabinet Mystery."* Consider looking at either of these two projects for ideas as you continue to work. If any of the following statements are difficult to understand, please see the following sections on the following pages for more information: Choice Frequency.
Compared to other projects in our database, players are *infrequently* presented with a choice.

Interaction Point Affordances
The information in this section will give you an idea of what an individual choice looks like in your project. If you continue to design your project in the way indicated by the metrics, your final game will most resemble *"Course Selection Frenzy."* A project that is structured entirely differently from yours in this area is *"An Unusual Situation."* Consider looking at either of these two projects for ideas as you continue to work. If any of the following statements are difficult to understand, please see the following sections on the following pages for more information: Number of Choice Actions, Visual Responsiveness, and Game Reactivity.
Compared to other projects in our database, your project offers the player a *smaller* number of choices at each point of interaction.
Compared to other projects in our database, your project is *not visually responsive at all*.
Events in your game are *not influenced by player actions*.

Figure 2: An example feedback report for one of the scenarios.

being "graded" or "scored," and confusion in understanding the metrics. These themes implied to us that students misunderstood the purpose of the feedback as grade-based, not critique-based, hence we made changes in the template used in Stage 2 in an attempt to clarify this.

A false impression of a score scale. The instructors for the course reported that some students felt demoralized by what they perceived as low scores when they saw their metrics scores, and that they had to explain verbally to students that they were not being graded. We decided to not explicitly report scores, but instead use quartiles and keywords for descriptions. The intention was for the metric values to be interpreted as informational, rather than as a value judgment.

Confusion in understanding the metrics. As the students read the report for the first time, the explanation was too long and complicated for them to understand. In response to this, we changed the structure of our feedback report. The new structure had links for more information and detailed explanation, which avoided overwhelming students with information and allowed them to specifically look for more information in aspects of design that they were more interested in. To make the report easier to understand, we adopted a simpler writing style.

- "The feedback is lengthy. The calculations are hard to grasp even with an explanation. Perhaps, using concise and short explanations of every section would be better and should be said in layman's terms."
- "The feedback was a bit difficult to understand due to the fact the we did not understand what the numbers the were being presented represented until after a bit of explanation, but even then we were confused about how the numbers were supposed to help us."

Desire for formative, constructive suggestions. Students expected the feedback to contain more suggestions in the form of instructions on how to change their projects.

- "We did not find the feedback helpful given that most of our project wasn't done. It was provided by essentially an A.I. and the numbers did not make sense."
- "We feel as if the feedback only gives you numbers and statistics but not direction as where to go per se."
- "In regards to our feedback it is much appreciated, however we feel it wasn't able to be as helpful because we had not gotten very far on our project. We had very little content and we realize that is our fault."

Stage 2: Student responses in the second round were more approving. One group said, "We liked that the changes we

made based on your original feedback made changes on this feedback, specifically because we saw the impact of those changes reflected in the feedback." Another group said, "This feedback was much easier to understand therefore much more helpful. It gave us good examples of other games to view. Now we just need to finish our game to complete the overall picture."

Some students still found the feedback confusing and were expecting more advice and clear implications for what they could do next. One group said "There is a lot of information given in this feed back but we are confused if we should branch out our scenes more or should we take different steps." While another said, "Thank you for the feedback, but how can we make the game feel more responsive to the player? Also how do we shape the game to make it work with the players choices. We also need help timing the scenes to make the implicit bias more accurate." Since we did not try to address students' desire for suggestions in our second iteration of feedback generation, it is perhaps not surprising that students still felt this was lacking.

Final Student Discussion At the end of the class, after projects were submitted, students were asked again for open-ended discussion on the feedback generation system. Some students overall felt that the generated feedback was helpful to them:

- "The most helpful part of the feedback was being told that we were on the right track and that our game design had effective interactions."
- "The most useful part was them telling us that our game was too short or there were not enough choices or too many scenes. I wish we would have gotten information that was clearer. The only reason the useful part was so useful was because it was easy to understand. Everything else was not very clear."

Interestingly, some groups also reacted negatively to the automated nature of the feedback, describing it as "robotic" and worrying that it will result in homogeneity in the games/experiments.

- "...it just felt very generalized and from a robotic program but even with that feeling we did take the advice we did receive..."
- "it relies to heavily on comparing your project to other projects which I think homogenizes the experiments a bit. I definitely did not feel rewarded for trying something different in the feedback."

While they had mixed reactions, it is clear that students found some use in the feedback, but expected more instruction on what they should do to improve. There is still much work to do on making the generated feedback accessible and useful for this audience.

Discussion and Future Work

Conducting the pilot study yielded changes in our feedback generation template, but not the overall approach. Detailed and long technical definitions confused students. We changed this by using simpler and clearer language. For example, we described the metric 'Average Outdegree' to students as 'Branching vs. Linear Narrative.' Additionally, many students viewed our feedback as an evaluation report and misinterpreted the metric values to be a score scale. Changing this to a quartile and keyword system also had a positive effect. Students found the keywords more comprehensible.

The final results suggest that there is further scope for improvement. One of the groups responded, "Thank you for the feedback, but how can we make the game feel more responsive to the player? Also, how do we shape the game to make it work with the players' choices." Students expect recommendations to improve their projects. It is not currently clear how to best provide this kind of constructive feedback in an automated fashion; further, there is a need to balance asking questions that prompt students to reflect on their design and risking being misinterpreted as telling students what to do. Additionally, it is important to take the context of receiving feedback into account: our users in this pilot study were students creating a game for a graded project. Further work is necessary to determine to what extent their reactions were informed by this.

In future work, we intend to conduct more studies to determine at which development phases the feedback would be most beneficial. We analyzed projects at early and late stages, but not during the times where the projects were undergoing the most rapid changes. Other future steps include the automation of the decision-making process to determine whether a particular metric is relevant at a given stage of completion. And, we are interested in both adapting existing metrics (Purdy et al. 2018) and developing additional metrics that use natural language processing to critique the content of the scenario.

Conclusion

An AI system might not be able to produce feedback that is as beneficial as feedback given by a human. Nevertheless, on development platforms, automated systems present users with immediate suggestions that human experts cannot because of time and cost constraints. We attempted to provide such immediate suggestions to users on StudyCrafter. However, providing useful automated feedback to students is a challenging problem as it must provide critique without being inundating. Another complicated task is suggesting personalized points for improvement, which is essential to make the feedback useful. Given these challenges, this work contributes to assisting students with the generation of timely and helpful automated feedback.

Acknowledgments

We are grateful to all the students who participated in our pilot study and gave us valuable feedback on this first prototype of our critique generation system. Thanks also to participants at the NII Shonan Seminar No. 130 (Artificial General Intelligence in Games) for many conversations that influenced the direction of our research. This material is based upon work supported by the National Science Foundation under Grant No. IIS-1736185.

References

[Baldwin et al. 2017] Baldwin, A.; Dahlskog, S.; Font, J. M.; and Holmberg, J. 2017. Towards pattern-based mixed-initiative dungeon generation. In *Proceedings of the 12th International Conference on the Foundations of Digital Games*, FDG '17, 74:1–74:10. New York, NY, USA: ACM.

[Carstensdottir and Seif El-Nasr 2018] Carstensdottir, E., and Seif El-Nasr, M. 2018. Interaction Maps for Interactive Narratives. Technical Report NU-CCIS-TR-2018-001, College of Computer and Information Science, Northeastern University, Boston, MA.

[Costantino 2015] Costantino, T. 2015. Lessons from art and design education: The role of in-process critique in the creative inquiry process. *Psychology of Aesthetics, Creativity, and the Arts* 9(2):118.

[Gee and Games 2008] Gee, E., and Games, I. 2008. Making computer games and design thinking a review of current software and strategies. *Games and Culture* 3:309–332.

[Guzdial et al. 2019] Guzdial, M.; Liao, N.; Chen, J.; Chen, S.-Y.; Shah, S.; Shah, V.; Reno, J.; Smith, G.; and Riedl, M. O. 2019. Friend, collaborator, student, manager: How design of an ai-driven game level editor affects creators. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, 624:1–624:13. New York, NY, USA: ACM.

[Harteveld et al. 2017] Harteveld, C.; Manning, N.; Abu-Arja, F.; Menasce, R.; Thurston, D.; Smith, G.; and C. Sutherland, S. 2017. Design of Playful Authoring Tools for Social and Behavioral Science. In *Intelligent User Interfaces*, 157–160.

[Holmgård et al. 2018] Holmgård, C.; Green, M. C.; Liapis, A.; and Togelius, J. 2018. Automated Playtesting with Procedural Personas through MCTS with Evolved Heuristics. *arXiv:1802.06881 [cs]*. arXiv: 1802.06881.

[Kafai and Burke 2015] Kafai, Y. B., and Burke, Q. 2015. Constructionist gaming: Understanding the benefits of making games for learning. *Educational psychologist* 50(4):313–334.

[Klimas 2009] Klimas, C. 2009. Twine.

[Liapis, Smith, and Shaker 2016] Liapis, A.; Smith, G.; and Shaker, N. 2016. Mixed-initiative content creation. In *Procedural Content Generation in Games*, Computational Synthesis and Creative Systems. Springer, Cham. 195–214.

[Lubart 2005] Lubart, T. 2005. How can computers be partners in the creative process: classification and commentary on the special issue. *International Journal of Human-Computer Studies* 63(4-5):365–369.

[Manning, Raghavan, and Schütze 2010] Manning, C.; Raghavan, P.; and Schütze, H. 2010. Introduction to information retrieval. *Natural Language Engineering* 16(1):100–103.

[Partlan et al. 2018] Partlan, N.; Carstensdottir, E.; Snodgrass, S.; Kleinman, E.; Smith, G.; Harteveld, C.; and Seif El-Nasr, M. 2018. Exploratory Automated Analysis of Structural Features of Interactive Narrative. In *Proceedings of the 14th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*. Edmonton, AB, Canada: The AAAI Press, Palo Alto, California.

[Purdy et al. 2018] Purdy, C.; Wang, X.; He, L.; and Riedl, M. 2018. Predicting generated story quality with quantitative measures.

[Resnick et al. 2009] Resnick, M.; Silverman, B.; Kafai, Y.; Maloney, J.; Monroy-Hernández, A.; Rusk, N.; Eastmond, E.; Brennan, K.; Millner, A.; Rosenbaum, E.; and Silver, J. 2009. Scratch: programming for all. *Communications of the ACM* 52(11):60.

[Shaker, Shaker, and Togelius 2013] Shaker, N.; Shaker, M.; and Togelius, J. 2013. Ropossum: An authoring tool for designing, optimizing and solving cut the rope levels. In *Ninth Artificial Intelligence and Interactive Digital Entertainment Conference*.

[Yannakakis, Liapis, and Alexopoulos 2014] Yannakakis, G. N.; Liapis, A.; and Alexopoulos, C. 2014. Mixed-initiative co-creativity. In *FDG*.

[Zook, Fruchter, and Riedl 2014] Zook, A.; Fruchter, E.; and Riedl, M. O. 2014. Automatic playtesting for game parameter tuning via active learning. In *FDG*.