Growing a Puzzle Garden: Exploring Casual and Serious Features in a Mixed-Initiative Logic Puzzle Authoring Tool

Fiona Shyne, Kaylah Facey, Seth Cooper

Northeastern University, Boston, MA, USA
{shyne.f, facey.k, se.cooper}@northeastern.edu

Abstract

In this project, we introduce Puzzle Garden, a mixed-initiative tool for authoring logic grid puzzles. We present Puzzle Garden as a test bed for exploring how the inclusion of casual components can impact the design and use of creativity support tools. The "casual" components of Puzzle Garden are inspired by the design principles of casual creators, while "serious" components give users fine grain control over the end product. We investigate how this tool can be used to study how users interact with serious and casual elements in authoring tools, through a preliminary user study lasting three weeks. Through the outcomes of this study we compiled a list of suggestions for how researchers can approach similar study designs.

Introduction

Creators in the modern digital age have a plethora of options when it comes to computational tools to support their process. However, the decision of which support tool best supports an individual's needs can be tricky. On one hand, large scale software applications need to consider the needs of creative professions: implementing essential features, integrating into existing workflows, and high performance (Palani et al. 2022). On the other hand, "casual creators" (Compton and Mateas 2015) seek to de-emphasize the importance of results, instead seeking to enhance the joy of creation. Within the field of HCI, "casual" and "serious" designers are often looked at as independent populations. However, we seek to support creators who prefer to flow between spontaneous, casual exploration and serious result-driven creative design.

To explore how to accommodate such creators, we present a novel interface, Puzzle Garden, which supports creating logic grid puzzles (a type of pen and paper logic puzzle) with narrative components. This work is focused on investigating how the introduction of casual elements impacts creativity support tools, particularly mixed-initiative creativity support tools. To accomplish this, we created two versions of Puzzle Garden. In the "serious" version, the interface only contains features we determined to be "serious", which are focused on refinement of the end product. In the "hybrid" version, we added "casual" features, inspired by the design patterns of casual creators (Compton and Mateas 2015), which focus on allowing the user to easily explore the design space.

As an initial investigation of this environment, we randomly assigned participants to one of the two versions of Puzzle Garden and gave them up to three weeks to use it as much as they chose. In total, 18 participants used the interface, 10 with the hybrid version and 8 with the serious version. The feedback we collected from these participants allows us to provide a series of suggestions related to the design of long-term studies, the creation of creativity support tools, and the role of mixed-initiative tools for creation. Overall this work provides a test bed for studying casual components within a mixed-initiative creation tool, along with guidance on how to conduct similar studies.

The contributions of this project are: 1) the development of a test bed for studying casual and serious features in a mixed-initiative tool; 2) results of a long-term study to examine impacts of these features; and 3) recommendations for future long-term studies of this kind.

Previous Work

Creativity Support Tools

Computational interfaces that seek to aid in different parts of the creative process, called creativity support tools (CSTs), emerged as a sub-field of HCI in the early 2000s with the work of Ben Shneiderman (Shneiderman 1999, 2000, 2001, 2007). To this day, CST constitute a large body of research. Frich et al. (2019) attempted to compile a definition for CST from their literature review: "A Creativity Support Tool runs on one or more digital systems, encompasses one or more creativity-focused features, and is employed to positively influence users of varying expertise in one or more distinct phases of the creative process." However, the authors themselves admit that this definition is likely too broad to be of use. The Creativity Support Index (CSI) attempts to quantify user experience of CSTs using 6 factors: collaboration, effort leading to results, expressiveness, immersion, and enjoyment (Cherry and Latulipe 2014).

Research tends to focus on two populations when researching CST: novices and experts (Chung, He, and Adar 2021; Remy et al. 2020; Ledo et al. 2018). A design goal of CSTs is often to be able to support both. Remy et al. (2020) proposed several design principles for CSTs, one of

Phase	Click	Type	Description
Specification	Select Sample	Casual	User selects a sample scenario/category (*No Blank Page*)
	Edit Scenario	Serious	User creates or modifies their own custom scenario/category/entity
	Edit Grammar	Serious	User edits the grammar [or brainstorms] for a clue
Generating and Searching	Select Evaluator	Casual	User selects an evaluator's recommendation (*Entertaining Evaluations*)
	View/Select Similar	Casual	User views/selects puzzles that are similar to the selected one (*Mutation Shopping*)
	Filter by Clue/Solution	Serious	User partially or completely specifies the clues/solution they want to view
	Filter by Difficulty/Size	Neutral	User sets the range of difficulty/size they want to view
Refinement and Expansion	Request Brainstorm	Casual	User requests a brainstorm for a clue (*No Blank Page*)
	Edit Clue	Serious	User edits the text of a specific clue
	Edit Narrative	Serious	User edits the text of a narrative block or scenario text
Exporting and Sharing	Like Puzzle	Neutral	User saves a puzzle to be viewed later
	Open Link	Neutral	User gets a link for the puzzle that can be shared with others
	Save as PDF	Neutral	User saves a non-interactive version of the puzzle as a PDF
	Post Puzzle	Neutral	User posts a puzzle to the Puzzle Garden community to be played by others
	Add Comment	Neutral	User adds a comment to a posted puzzle

Table 1: How user actions are categorized as casual, serious, or neutral

which is the concept of a "low threshold, high ceiling, and wide walls." That is, interfaces should be easy for novices to pick up and use, have a large set of features for experts, and be adaptable to different types of projects. Despite this goal, many research projects are very small in scope, focusing only on simple tools (Frich et al. 2019) and evaluating their interfaces over a short period of time (most often less than a day) (Ledo et al. 2018). In this work we propose a tool with different levels of abstraction available and investigate how to design studies that last an extended period of time.

Casual Creators

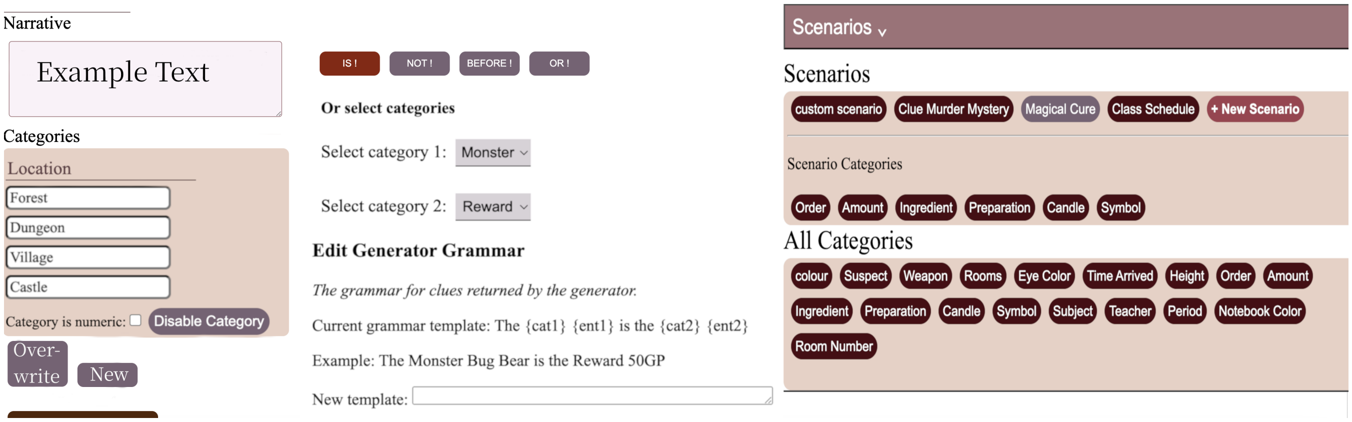
While research into CSTs for amateurs emphasizes improving the quality of the end result, more recently research into "casual creators" (Compton and Mateas 2015) that instead emphasize "autotelic" creativity has emerged as a sub-field. Casual creators are designed to encourage joyful exploration of a creative possibility space without a specific goal in mind. Notably the field of CSTs already acknowledges the importance of these factors with the Creativity Support Index (CSI) (Cherry and Latulipe 2014) factoring in both exploration and enjoyment. Compton and Mateas (2015) define a number of design patterns for casual creators to emulate. Tools should not intimidate users with a "blank canvas", and generated artifacts should include fast, "entertaining evaluations". To encourage exploration, actions should be limited to coarse changes that "modify the meaningful", and any changes users make should result in "instant feedback". To guide users' exploration of the possibility space, they should have a choice between similar options on a "chorus line", which they can use to "mutant shop" toward desired functional or aesthetic features. Finally, tools should facilitate interaction between users by providing support for "saving and sharing" or "hosted communities", and they should be easy to mod or hack so users can add their own features.

Despite the goals of casual creators, user studies show that often creators are frustrated by a lack of fine-grained editing abilities once they have generated an artifact that is close to what they want (Colton et al. 2020; Kreminski et al. 2020). We posit that some users start using a tool casually but become serious as they explore the possibility space and form clearer desires for an end product. Similarly, a serious user might benefit from tools that encourage exploration and joy.

Mixed-Initiative Procedural Content Generation

Since the late 90's, HCI researchers have been exploring the role of automation within user interfaces. Horvitz (1999) addresses the essential issue of wanting to ease interaction through automation, while keeping users' goals and control in mind. In tackling this, they address several principles for *mixed initiative* (MI) tools, where a computational agent and a human both take actions. A similar concept to MI is co-creation, where a computational agent and a human both contribute to the creation of an artifact. Co-creation can take a variety of interaction methods, as described in the framework by Rezwana and Maher (2023), with computational components working either in parallel or in serial with the human user, working on the same or divergent tasks, and contributing to creating, critiquing, or expanding artifacts.

MI is commonly used in procedural content generation (PCG), the automated creation of game materials (Lai, Leymarie, and Latham 2022). Lai, Latham, and Leymarie (2020) address the design requirements of MI-PCG tools for industry uses, including that interfaces must respect the control of the designer, must be fast enough to compete with manual feedback, and should be easily integrated into existing pipelines. Evolutionary computation techniques are common within MI-PCG, including mutant shopping and various forms of interactive evolutionary computation (Lai, Leymarie, and Latham 2022). Two prominent examples of evolution-based MI-PCG are the Sentient Sketchbook (Li-



Narrative: Example Text

Select category 1: Monster

Select category 2: Reward

The grammar for clues returned by the generator.

Current grammar template: The {cat1} {ent1} is the {cat2} {ent2}

Example: The Monster Bug Bear is the Reward 50GP

New template:

Figure 1: Specifying a puzzle: All users could edit scenarios (left) and grammar (middle). Only hybrid users had access to example scenarios and categories (right). Screenshots were modified and simplified for readability.

apis, Yannakakis, and Togelius 2013) and Evolutionary Dungeon Designer (Alvarez et al. 2018; Baldwin et al. 2017; Alvarez et al. 2021, 2019) projects. Both projects have a similar interaction design, where the user interacts directly with a 2D level, and an evolutionary component generates suggested alternative designs. Other methods have humans and machines contribute to separate tasks; for example Karavolos, Bouwer, and Bidarra (2015) generates game levels based on a mission graph provided by the user. Puzzle Garden uses a different approach, where the user first gives an initial specification (puzzle scenarios) and then can search for and refine artifacts (puzzles) generated using a non-interactive algorithm from previous work (Shyne, Facey, and Cooper 2024).

Puzzle Garden User Interface

Puzzle Garden is an interface to create logic grid puzzles. Each puzzle contains a *scenario*, which describes the setting for the puzzle, a set of *categories* that each contains a set of *entities*, and a list of natural language clues. The player must use the clues to determine which entities between categories are connected to each other. For example, a murder mystery puzzle might have the categories: suspects, locations, and weapons. The player must then determine which suspect was in what location and at what time.

Puzzle Garden allows users to create puzzles in four phases. In the *puzzle specification phase*, the user defines the scenario and categories of the puzzle. Using this specification, the evolutionary system generates a variety of possible puzzles which the user can explore, in the *generating and searching phase*. A generated puzzle enters the *refinement and expansion phase*, either by editing clue text or by adding narrative. Finally, the end puzzle can be exported to PDF or shared with a link or community post, in the *exporting and sharing phase*.

The features in Puzzle Garden are broken up into three categories:

1. **Casual Features**: These are features that allow users to easily and joyfully explore the possibility space. Each casual feature we include is an application of the design principles of casual creators.
2. **Serious Features**: As no existing framework exists, we define “serious” in contrast to “casual” creators. A serious feature is one that focuses on refining and improving the end product.
3. **Neutral Features**: These are features that follow one of the design patterns of casual creators but are also useful for a individual geared toward a high-quality end product.

The hybrid version of the interface has all features, while the serious version only has neutral and serious features. Code for the frontend[1] and backend[2] of Puzzle Garden is publicly available on GitHub. Full, unmodified screenshots of the interface are available on the Open Science Framework page[3].

Puzzle Specification

In the *Puzzle Specification Phase*, the user defines the possibility space by defining the scenario, the categories, and the entities within each category.

Serious Features A serious user first must create a new scenario (or select a previously saved one). When creating this scenario, the user can write a description (the scenario text) which will be provided at the top of any puzzle generated with this scenario. After specifying a scenario, the user must create at least two categories (each with at least two entities) to add to the puzzle.

The puzzle generator using its default grammar can produce clues that are awkward and hard to interpret. While clues can be edited manually after generation, the user can also tell the generator how to write clues by editing the grammar templates. For example, a user can tell the generator how to write an “is” clue (which says entities are

[1] https://github.com/fiabot/LogicPuzzleInterface/releases/tag/v1.0.0

[2] https://github.com/flaneuseh/logic_puzzles/releases/tag/v3.0.0

[3] https://osf.io/wx436/

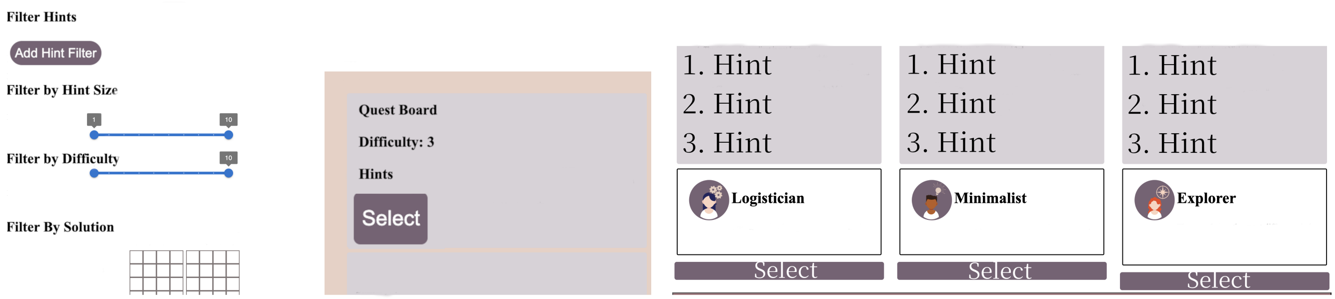


Figure 2: Viewing generated puzzles. Serious users (left) are only given the list of puzzles, while hybrid users (right) can also view AI expert recommendations. Screenshots were modified and simplified for readability.

connected) between the categories suspect and weapon with the string: "ent1 had the ent2", which could later be used to make the hint "Ms. Scarlet had the knife." On save, the existing template is overwritten.

In the same editor, the user can add narrative "brainstorms" for each type of clue. These have the same format as the grammar templates, but instead act as inspiration for writing narrative elements for each clue type. When a user is writing narrative text for a clue, they can view generated brainstorms that are available for that type of clue.

Casual Features Following the *No Blank Canvas* casual creator pattern (Compton and Mateas 2015), users in the hybrid category are provided with sample scenarios, categories, and entities to choose from.

There are three sample scenarios: "Clue Mystery" - a murder mystery based on the board game *Clue*, "Magical Cure" - creating a magical potion, and "Class Schedule" - figuring out a class schedule. Each scenario has a brief scenario text that explains the context of the puzzle, and between 5 and 6 sample categories with up to 6 entities are suggested to the user. The scenario text, the category name, and the entities can all be manually edited by the user before generation and saved for re-use.

We provide an updated grammar for all categories in each sample scenario. Users can select any sample category, even if it is not in the given scenario, but grammar is not provided between categories of different scenarios. Additionally, narrative brainstorms are provided for some hints within each scenario (but not all of them).

In the hybrid version of the interface, users can still manually create/edit grammar and brainstorms. However, by default the grammar editor is closed.

Generating and Searching

Once a puzzle is specified, the generator begins creating puzzles. This process is done in several cycles, and puzzles appear to the user as they are generated.

Generating puzzles Shyne, Facey, and Cooper (2024) describe a quality diversity evolutionary algorithm to generate logic grid puzzles. It is a combination of a constraint based algorithm (to include only solvable puzzles), and a quality diversity algorithm (to generate puzzles that vary in terms of difficulty and solutions). The difficulty is estimated by the solver, which was shown to be correlated with the perceived difficulty.

To adapt this algorithm for a user-facing interface, we generate puzzles in cycles. At first the generator only evolves for a short time, generally allowing puzzles to be presented to the user within seconds (if there are many categories or entities, it can take longer). However, to improve the quality and diversity of generated puzzles, the algorithm is prompted to continue evolving after sending the initial results.

For the first cycle, the generator evolves for 10 generations with a population size of 100. After the first cycle the generator evolves for 100 generations with a population size of 50. Puzzles are evolved for 10 cycles by default, but the user can increase the number of cycles at any time. When the user leaves the generation tab, any unsaved puzzles are discarded, and generation can no longer be resumed.

Serious Features As soon as puzzles are sent to the interface, they appear on the right side of the interface. For each puzzle, the user can see hints and the estimated difficulty rating: between 1-7 as determined by Shyne, Facey, and Cooper (2024). After opening a puzzle (either in the interface or in a new tab) the user can play the puzzle, like (save) it, edit the clues, or write a narrative. The interface provides several filters to search for puzzles:

- **Difficulty Range**: A range for the difficulty rating. This is considered a "neutral" feature and falls into the *Manipulate the Meaningful* pattern of casual creation.
- **Clue Size Range**: A range for the number of clues in the puzzle. This is also a "neutral" feature.
- **Clue**: Users can filter based on the clues in the clue list of the puzzle. This can be partially (e.g. any puzzles that have an "is" clue) or completely (e.g. any puzzle with the clue "Ms. Scarlet has the knife") specified by the user.
- **Solution**: Users can filter for a solution by partially filling out a sample grid.

Casual Features

There are two ways to casually navigate the design space: evaluator recommendations and mutation shopping. Additionally, the full list of puzzles and the filters are hidden by default, which implements the *Limiting Actions* design principle of casual creators.

The evaluator recommendations seek to implement the *Entertaining Evaluations* and *Limiting Actions to Encourage Exploration* patterns of casual creation. The user is presented with three "evaluators" that each select a puzzle based on a different goal. The "Logician" finds the hardest puzzle, the "Minimalist" finds the puzzle that is easiest and has the fewest clues, and the "Explorer" selects a puzzle that has the most different kinds of clues. Each evaluator is given an icon to make them seem like people making recommendations.

Mutant Shopping is a casual creator pattern that allows users to move about the possibility space by viewing artifacts similar to the one they are currently viewing. For each puzzle in Puzzle Garden, users can view the puzzles that are most similar in terms of shared hints and solutions. Users see up to 6 similar puzzles, divided into up to 2 each of equal, greater, and lesser difficulty.

Refinement and Expansion

After users find a puzzle that they enjoy, they can update it to suit their needs. Users can manually edit the clues or create a secondary version of the puzzle with increased narrative. The narrative version of the puzzle is presented as a series of paragraphs situating the clues in a story context.

While manually editing the clues or narrative is considered a serious feature, users can casually add narrative elements through brainstorms. For sample scenarios, we provide ideas for how a clue can be incorporated into a narrative. Where a brainstorm exists, it can be copied into the narrative block with one click. Users can then edit the narrative as usual.

Exporting and Sharing

Saving and Sharing is a principal of casual creators that emphasizes the importance of being able to export creations in an accessible format. However, exporting and sharing is important for both casual and serious creators, and therefore we consider these actions as "neutral". Users can export puzzles as PDFs, in either logical or narrative form. Additionally, puzzles include a shareable link for anyone (not just registered users) to play.

Hosted Communities is also a principal of casual creators that we consider neutral in this work. Puzzle Garden provides a community page where people can share and comment on their created puzzles.

Study Design

To test the impact of casual features in our mixed-initiative tool, we performed an exploratory long-term study over 3 weeks. Participants were sorted into two groups, the "hybrid" group which got both serious and causal features, and the "serious" group which only got the serious features. During the study period, participants were able to use the interface as much or as little as they chose.

Methods were approved by the authors' IRB. Participants consented to participate in research when signing up for the study.

Recruitment

Participants were recruited on social media forums related to puzzles, game generation, or interest in TTRPGs. The social media posts led to a Qualtrics form, which asked them for contact information (email), along with questions about their experiences and goals. They also ranked the six components of the Creative Support Index (CSI) (Cherry and Latulipe 2014). After submitting the survey, they were randomly assigned to the hybrid or serious group and given an anonymous username and password. This was used to create a custom account on Puzzle Garden, to which each participant could login and access personal content (e.g. liked puzzles or custom scenarios). The account information was emailed to the participant, along with information about the study and tutorials for the interface. The serious group got one text-based and one video tutorial, that explained all the features available. The hybrid group got two tutorials, each in both text and video form. The "beginner tutorial" explained the casual features of the interface, while the "advanced tutorial" explained all the features of the interface.

Recruitment posts were made periodically between May 5th and 8th 2025. Participants were given login information between May 5th and 14th 2025. All participants were allowed to access the interface until May 31st 2025.

All participants were volunteers and were not compensated for their time.

Study Period

Participants had access to the interface for up to three weeks, depending on when they were recruited. During this period they could use the interface as much or as little as they wanted. They could save scenarios, grammars, and puzzles to their account. Each group was given its own community page, where users could share and comment on generated puzzles. The researchers also had an admin account. The admin account posted one sample puzzle to both communities to encourage engagement, along with commenting on all posted puzzles. The admin account was clearly labeled "Admin" and shown in a different colored font.

Participants were encourage to fill out periodic surveys about their experience. The first part of this periodic survey was the Creative Support Index (Cherry and Latulipe 2014), to assess the overall ability of the interface to support creation. Participants were also asked what their goal was, if they accomplished it, what features they did or did not use, and what features they wished were included. All open text responses were optional. To encourage participants to fill out surveys, the interface gave them a "Research Score" based on how many surveys they filled out. This score has 7 levels from "Seed" to "Pollinator," in reference to the garden theme of the interface.

To encourage participation overall, all participants were emailed periodically (about once a week) to remind them to use the interface. This was also an opportunity to address confusions participants had about various parts of the interface. In response to questions, there was also a "Frequently Asked Questions" section of the home page.

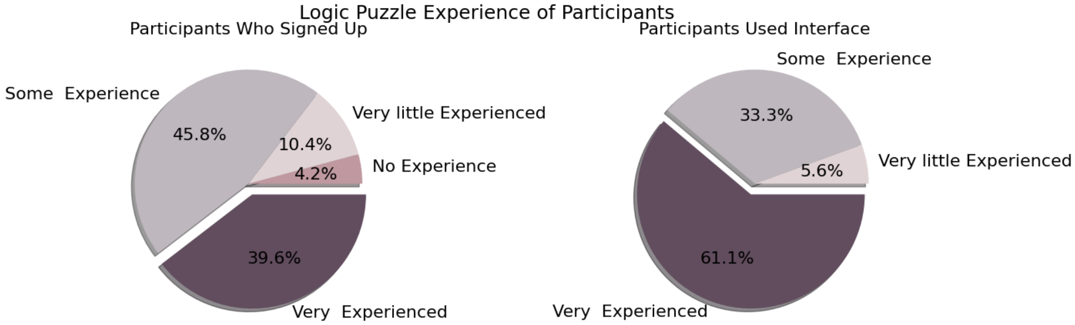


Figure 3: Experience playing logic grid puzzles of participants who signed up and who used the interface.

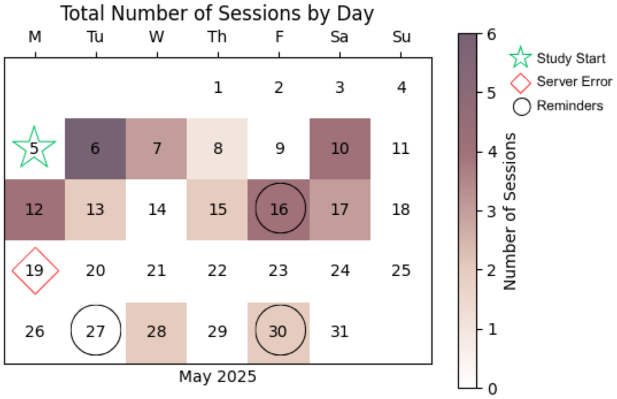


Figure 4: Annotated calendar of the study period.

Results

Preliminary Survey

Between May 5th and May 14th, 47 participants were given access to the interface. Only 18 of those participants used the interface long enough to make one of the clicks listed in Table 1. However, we report data from all participants who filled out the intake survey and consented to data collection, to investigate the factors that influence people to actually use the interface. The data from this study is available on the Open Science Framework [4].

Participants who signed up varied in terms of experience with logic grid puzzles, with just over one third (39%) having identified as being "Very Experienced" with logic grid puzzles, as shown in Figure 3. However, of the participants who actually used the interface, the majority (61%) of them identified as being "Very Experienced." Interestingly, we do not see the same effect in terms of experience with game design, which shows similar ratios for participants who did and did not use the interface. Overall, participants had less experience with game design than they did with logic puzzles, with 44% of participants who used the interface having little to no experience with game design.

Out of the 18 participants who used the interface, 10 participants were in the casual mode and 8 were in the serious mode. The highest amount of activity happened in the first two weeks of the study, with a couple participants using the interface in the last couple of days, following reminder emails (see Figure 4).

Impact of Mode

Despite the fact that casual features reduce the time required to create puzzles, hybrid participants did not spend less time on the interface, as shown in Figure 5. In fact, hybrid participants were more likely to spend more than 2 hours on the interface — 3 (30%) hybrid users compared to 1 (12.5%) serious user. One hybrid participant even spent over 9 hours in total on the interface. However, in both groups participants were likely to not spend much time on the interface, with 4 (40%) hybrid participants and 4 (50%) serious participants having a total time of less than 30 minutes.

We can look more at where participants spent their time, based on the clicks we tracked[5] (total counts provided by Figure 8). We used the clicks to determine what phase they were in at each point in their interaction. We then summed the number of times ("instances") participants were in each phase, as shown in Figure 7. Participants had the most instances in the early phases of the design process, and had fewer instances in later phases. Serious and hybrid users had similar numbers of instances in each phase.

Since hybrid users had a choice between serious and casual features, we can look at what mode they were more likely to use. Overall, hybrid participants were much more likely to use example scenarios and categories than create their own. Hybrid users were very unlikely to use the casual brainstorm features, but did sometimes edit narrative (likely the scenario text) or base clues. Hybrid participants occasionally liked (saved) puzzles, but were very unlikely to open a link, download, or post puzzles.

In both groups, participants were unlikely to like (save) puzzles they generated, as shown in Figure 6. A majority — 6 (60%) hybrid, 5 (62.5%) serious — of participants did not

[4] https://osf.io/wx436/

[5] Due to a tracking issue, we didn't capture "like" clicks so instead just used the number of liked puzzles.

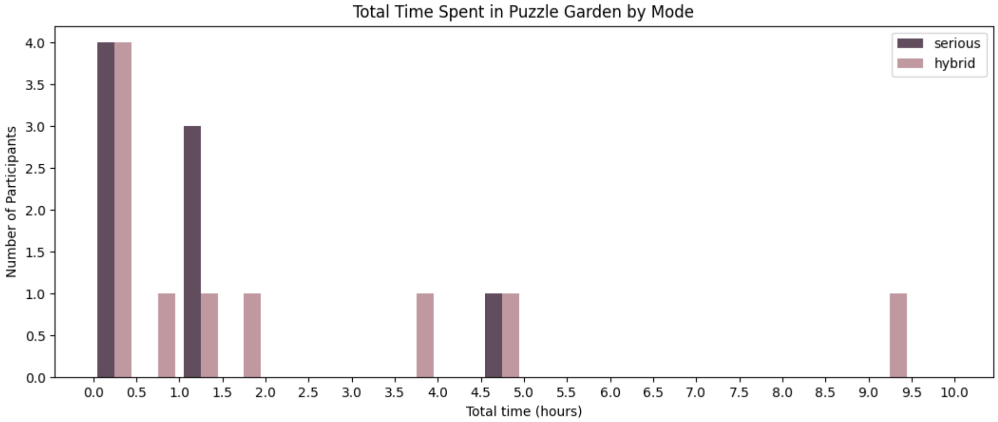


Figure 5: Total time participants spent on the interface.

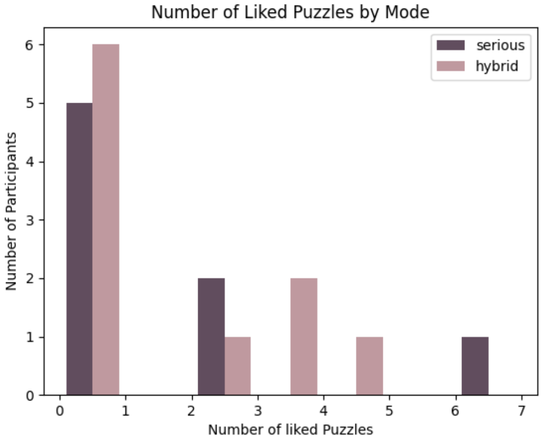


Figure 6: Total number of likes from participants who used the interface.

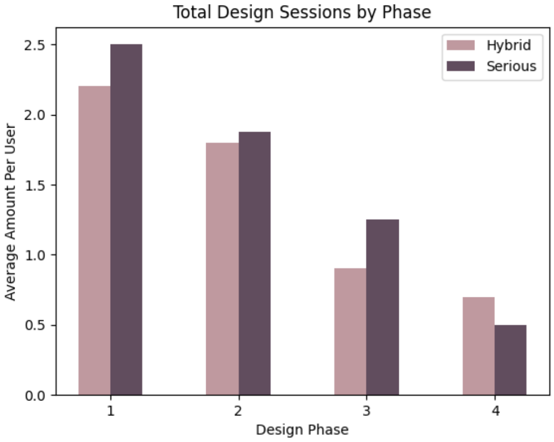


Figure 7: Number of times participants used each design phase.

like any puzzles. Combining liked puzzles and saved scenarios, we have puzzle specifications for 6 (60%) hybrid participants and 5 (62.5%) serious participants.

Hybrid participants had access to three example scenarios with categories and entities already created. From the hybrid participants, we found evidence for three methods of creating puzzle scenarios. The first method was using the example categories with minimal or no modification. For example, MaroonBeaver saved puzzles from Magic Cure unmodified and from Class Schedule where one category had entities from Clue Mystery. The second method we noticed is using the examples as inspiration, but modifying the entities themselves. The only example we saw of this was from PlumSparrow, who modified all the entities in the Clue Mystery scenario, for example having the suspects “Col. Ketchup” and “Mr. Bleu” and the weapons “Wet Noodle” and “Rubber Band.” The last method we recorded was participants coming up with completely original puzzle ideas. VioletOwl saved scenarios and puzzles relating to a tea party, and CoralCamel saved scenarios and puzzles related to a fairy-tale setting of a princess kissing a frog (these were the only puzzles saved). These puzzles included custom grammar rules, along with clues being manually edited after the fact. Participants only used one of the three methods, and did not switch between methods.

Serious participants did not have access to the example scenarios and therefore had to custom create any puzzle scenario they generated. Three serious participants saved a single scenario or puzzles from a single scenario. Two of these had no, or very minimal scenario text, with LimeCamel saving a scenario about breakfast and CyanCamel saving a puzzle about spies. OrangeCamel was more elaborate in their scenario text, describing children going into a labyrinth filled with monsters. Additionally, two participants saved puzzles using two different scenarios. For example, SilverDolphin had one puzzle about children and their favorite color, and another about an art project using different mediums.

No participants in either group used the narrative clue op-

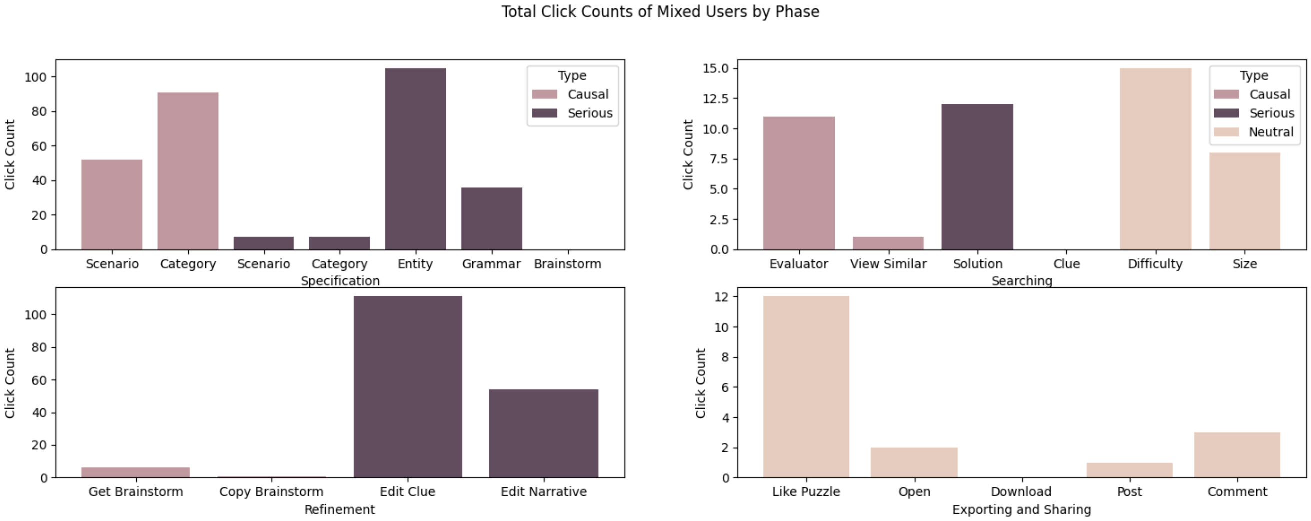


Figure 8: Total number of clicks between hybrid users, by design phase and click type.

tion, and most participants' base clues were straightforward logic, even if they were edited for clarity. However, SilverDolphin manually edited the base logic clues to be a little more creative then just conveying information. For example instead of saying "Hazel had a bird in their piece" they said "Hazel dreamed of feathers while creating her art." The community page was underutilized by both groups. In the hybrid group one participant (VioletOwl) posted a tea party puzzle, which was commented on by both the admin account and another participant. The admin account's posted puzzle was commented on by two participants (one being VioletOwl). All comments were generic and complementary. The serious community page had no posts or comments.

Five (62.5%) serious participants filled out periodic surveys, most (3) of whom filled out only one survey. Four (40%) hybrid participants filled out periodic surveys, all of whom only filled out one survey. For participants who filled out multiple surveys, scores were first averaged between all scores provided. Among these participants, the total CSI values did not vary by more than 1 point between any two surveys. There was no difference in CSI scores between participant groups, with both groups averaging around 50 out of 100. However, looking at the subscores, the the hybrid groups rated the interface slightly higher in terms of creativity and exploration. This shows that introducing casual elements might have a positive effect on some of the goals targeted by casual creators (see Figure 9).

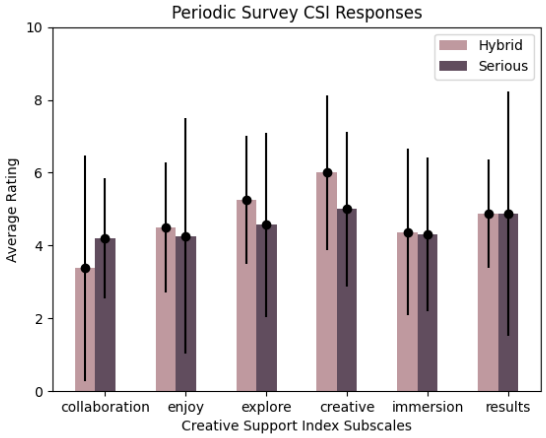


Figure 9: Creative Support Index sub-scale responses.

Open Responses

The periodic surveys had several open ended text responses that gave further insight to how participants engaged with the interface. We present comments grouped by the themes they addressed.

Problems with the interface: Participants commented on a variety of problems they had with the interface. This included small problems such as the "print was so tiny", the interface "didn't work properly in Firefox", or the "site wouldn't scroll to the bottom." Several wished the default clue logic was "less-ugly" (particularly that they should not include the category names) and complained that creating their own grammar templates was "too hard to do completely." Other participants did not see how to use some of the interface's features, including the "narrative clues", "how to save/share puzzles", and the "ability to set which entities go together" (which could be done by filtering by solution). Participants also complained about the flow of the interface, calling it "clunking" and "difficult to navigate." They suggested ways to fix this including not having a "popup asking me if I want to leave every time," and that "it would be nice to go back and continue editing the grammar." Finally, participants didn't like how the interface saved scenarios, such as wanting to "remove a category from a scenario."

Feature Suggestions: Participants suggested several features that could be implemented. We categorized these features as casual if they would make the creation process easier and faster, and serious if they would expand the current capabilities.

Casual features that were suggested include additions to the puzzle solving process, such as "auto-fill[ing] X's in the grid when a[n] O is input" or having the option to "see the completed grid without needing to solve it by hand." Another participant mentioned a method to make the creation of grammar templates easier by "sharing the template between" related clues.

Suggested serious features focus on different types of clues that could be implemented, and that "more varieties in logic available would be fun." Other participants specifically mentioned wanting to have a different type of numerical category "where all that matters is larger/smaller." For example one participant wanted to include "number of legs" and "I tried 2, 3, 6, 10 but then it wanted me to pick an increment which... does not work here."

Role of the Interface: Two participants commented how they were frustrated about how the interface's role is to generate the hints. One participant mentioned that "I like coming up with the clues myself." Another participant went further.

> This feels like research into having a computer do the interesting/creative part (generating the clues/puzzles) while having a human do all the boring paperwork parts (writing the clue templates, entering the category names and values) .... I feel like what I'd really want is a tool that collaborates with me on that: tries to generate clue text for me when I click in the grid, ..., let me easily ...undo/redo.. or drop in a clue that I tried earlier.... The filter-by feature is a nice touch, but... [w]hen I add a narrative criteria, I do not want to hope that it generated what I want.

Discussion

Study Design

We proposed an ambitious study design to track how users would interact with the interface over an extended period of time. To accomplish this we recruited individuals who would be most interested (from logic puzzle communities), let users use the interface as much as they chose, and didn't pay participants to incentivize participation. While this attempts to emulate the way participants would naturally react to such an interface, it resulted in a limited sample size (18 participants), which limits the types of analysis that we could accomplish. Below we provide several suggestions for researchers who want to perform similar studies.

One technique to increase recruitment would be to incentivize participation. The most common way to do this is through direct payment of participants. However, even setting aside budget concerns, researchers have to consider whether to pay a base rate or based on usage. Paying participants based on the total time spent can artificially increase the amount of time participants spend and inversely, a base pay could dis-incentivize long use times, as their hourly rate would go down. Research could consider other incentive strategies. For example, Puzzle Garden could host a competition in which the participant with the best puzzle would be deemed the winner and potentially given a prize.

Another method researchers could use is to allow the sample size of participants to remain small but to gather richer data from each participant. For example, conducting pre- and/or post-interviews with participants who engaged with the software could have given us a greater volume of qualitative data about participant perspectives and experiences (Guest et al. 2020). However, it is possible to collect richer types of data even when synchronous interviews are not possible. For example, researchers could have participants record their own screens and participate in think out loud activities while using the interface. Researchers could also conduct diary studies, where participants are asked to periodically reflect on different elements related to the research question.

Interface Improvements

Participants gave several pieces of feedback that suggest improvements that can be made to Puzzle Garden. From these responses, we can extract several recommendations that can be applicable to any interface design.

1. **Test across environments**: We tested Puzzle Garden on a small range of set-ups and screen dimensions. Our participants noticed several errors across different environments, including specific browser and screen sizes. When designing interfaces for a variety of users, consider the range of set-ups that are possible and be clear about which are supported.
2. **Make important features visible**: Our participants missed several important features that were included. While our participants highly valued the narrative aspect of puzzles, no participant used the narrative features of our tool. This is likely because "Narrative Clues" was hidden behind a toggle button whose default was "Logical Clues." Consider ways to ensure participants are aware of important features, such as by avoiding default options.
3. **Sensible defaults**: While you want to make important features obvious, less important configurations should be set up with defaults that nicely handle the most common cases. In Puzzle Garden, we used grammar templates that were robust to edge cases (different categories with same entities) but ugly in the average case. Consider the most common case first, and provide mitigation methods for edge cases.
4. **Flow between interface and processes**: Participants complained about the clunkiness of the interface. Participants were not easily able to flow between one process and other. Consider natural methods of moving between different design phases.

Impact of Interface Mode

Looking at the two different modes of interaction, there is limited evidence of casual features affecting the design process. The first observation is that while casual features allow users to create puzzles faster, we did not notice an effect on overall time spent in the interface. In fact, hybrid users were more likely to spend a long period of time using the interface. Being given sample categories to enable getting started quickly might allow an unsure user to transition faster to becoming interested in exploring the interface. However, we did observe that the example categories had an impact on creativity. Hybrid users often explored only the default situations with little modification. There was some evidence of the default categories inspiring creative thought (such as turning “Rope” into “Wet Noodle”). However, hybrid participants were much more likely to use the example scenarios and categories. Hybrid participants were also less likely to refine puzzles after generation, although they were more involved in exporting and sharing puzzles. Only hybrid participants participated in the hosted community, although that participation was minimal.

Overall, neither interface appeared to have the ability to transition users from casual explorers to serious and committed users. We had hoped including both casual and serious features would have this effect, but we do not observe evidence of this. This could be due to the interaction issues, as addressed in Section “Interface Improvements”. However, it could also speak to how just the *inclusion* of both kinds of features does not lead participants to transition from one type of mode to the other. Future investigation is needed to understand how interfaces can best support this transition from casual to serious (or vise versa).

Role of Mixed-Initiative Tools

Feedback from participants has led us to reflect on what the role of mixed-initiative tools like this should take on. What was particularly curious to us was the comment that the computational part (generating clues) was considered the “interesting” and “creative” task, while the task given to the human (creating the puzzle scenario/grammar templates) was the “boring” part. This was interesting to us, as Puzzle Garden was designed with the opposite in mind. We had wanted the human designers to be in charge of the creative and narrative elements, while the computer was in charge of the technical mechanics. This brings an interesting question: which parts of the designed process are considered creative and which are tedious? It is likely that this is dependent on the background of the individual. Perhaps it should not be the role of mixed initiative tools to *assume* which part to take over, but to leave that choice up to the user and take over whichever they are uninterested in. This is, of course, a tall order as this requires the interface to be skilled in many different types of creativity and be able to transition between them.

Our participants also gave us a suggestion of how the interaction method could be redesigned: allow the user to start creating and have the interface give suggestions of additions / alternate ideas. This interaction mode is certainly possible; it resembles the idea of Sentient Sketchbook (Liapis, Yannakakis, and Togelius 2013) and Evolutionary Dungeon Designer (Alvarez et al. 2018). However, it brings up a conflict between the philosophy of casual creators and these other creation tools. Casual creators are designed with the intention of exploration, often without direct manipulation of artifacts. That is how Puzzle Garden was implemented. While it is possible to filter for the inclusion of particular clue, the generator searches the entire possibility space without restricting to user constraints. However, our participants did not want to “hope” that their constraints would be met, and the best way to accomplish this would be to narrow the search space to only what the user asked for. This brings into question how to encourage the process of exploration, while still giving users agency.

Limitations

There are several limitations that effect the generality of our outcomes. The first limitation is that we only used one test environment: generating logic grid puzzles. This environment has a particular population it attracts, along with a specific design process that may not be applicable to other types of creative processes. We also chose a specific set of features to include, while one could imagine an unlimited number of causal or serious features that could be applicable to this design space. Future work can look at the impact of casual features across two or more environments.

Another limitation was the relatively small sample size of participants who used the interface, which we address in Section “Study Design”. Additionally, Puzzle Garden was made by a small team, during a short design period (around three months). This means that the interface is unpolished in a number of ways (discussed in Section “Interface Improvements”), which impacted the results beyond the test conditions we implemented. While we seek to improve the interface over time, it is unlikely that Puzzle Garden can compete with the usability of large scale commercial products that users are likely accustomed to.

Conclusion

We presented an exploratory user study of Puzzle Garden, a mixed-initiative puzzle authoring tool with casual and serious components. Participants were either given a version of Puzzle Garden with only serious components, or with both serious and casual components. Users across both interface groups behaved similarly in a variety of ways, but we noticed some effects of the casual components. The way participants interacted with the interface, along with the feedback they gave use, allowed us to provide guidance for similar projects going forward.

References

Alvarez, A.; Dahlskog, S.; Font, J.; Holmberg, J.; Nolasco, C.; and Österman, A. 2018. Fostering creativity in the mixed-initiative evolutionary dungeon designer. In *Proceedings of the 13th International Conference on the Foundations of Digital Games*, 1–8.

Alvarez, A.; Dahlskog, S.; Font, J.; and Togelius, J. 2019. Empowering quality diversity in dungeon design with interactive constrained map-elites. In *2019 IEEE Conference on Games (CoG)*, 1–8. IEEE.

Alvarez, A.; Grevillius, E.; Olsson, E.; and Font, J. 2021. Questgram [Qg]: Toward a Mixed-Initiative Quest Generation Tool. In *The 16th International Conference on the Foundations of Digital Games (FDG) 2021*, 1–10. Montreal QC Canada: ACM.

Baldwin, A.; Dahlskog, S.; Font, J. M.; and Holmberg, J. 2017. Mixed-initiative procedural generation of dungeons using game design patterns. In *2017 IEEE conference on computational intelligence and games (CIG)*, 25–32. IEEE.

Cherry, E.; and Latulipe, C. 2014. Quantifying the Creativity Support of Digital Tools through the Creativity Support Index. *ACM Trans. Comput.-Hum. Interact.*, 21(4): 21:1–21:25.

Chung, J. J. Y.; He, S.; and Adar, E. 2021. The Intersection of Users, Roles, Interactions, and Technologies in Creativity Support Tools. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference*, DIS '21, 1817–1833. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-8476-6.

Colton, S.; McCormack, J.; Berns, S.; Petrovskaya, E.; and Cook, M. 2020. Adapting and Enhancing Evolutionary Art for Casual Creation. In Romero, J.; Ekárt, A.; Martins, T.; and Correia, J., eds., *Artificial Intelligence in Music, Sound, Art and Design*, 17–34. Cham: Springer International Publishing. ISBN 978-3-030-43859-3.

Compton, K.; and Mateas, M. 2015. Casual Creators. In *International Conference on Innovative Computing and Cloud Computing*.

Frich, J.; MacDonald Vermeulen, L.; Remy, C.; Biskjaer, M. M.; and Dalsgaard, P. 2019. Mapping the landscape of creativity support tools in HCI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–18.

Guest, G.; Namey, E.; O'Regan, A.; Godwin, C.; and Taylor, J. 2020. Comparing Interview and Focus Group Data Collected in Person and Online. *Patient-Centered Outcomes Research Institute (PCORI), Washington (DC)*.

Horvitz, E. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 159–166.

Karavolos, D.; Bouwer, A.; and Bidarra, R. 2015. Mixed-Initiative Design of Game Levels: Integrating Mission and Space into Level Generation. *FDG*, 8: 2015.

Kreminski, M.; Dickinson, M.; Osborn, J.; Summerville, A.; Mateas, M.; and Wardrip-Fruin, N. 2020. Germinate: A Mixed-Initiative Casual Creator for Rhetorical Games. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 16(1): 102–108. Number: 1.

Lai, G.; Latham, W.; and Leymarie, F. F. 2020. Towards Friendly Mixed Initiative Procedural Content Generation: Three Pillars of Industry. In *Proceedings of the 15th International Conference on the Foundations of Digital Games*, FDG '20, 1–4. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-8807-8.

Lai, G.; Leymarie, F. F.; and Latham, W. 2022. On mixed-initiative content creation for video games. *IEEE Transactions on Games*, 14(4): 543–557.

Ledo, D.; Houben, S.; Vermeulen, J.; Marquardt, N.; Oehlberg, L.; and Greenberg, S. 2018. Evaluation Strategies for HCI Toolkit Research. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, 1–17. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-5620-6.

Liapis, A.; Yannakakis, G. N.; and Togelius, J. 2013. Sentient sketchbook: computer-assisted game level authoring. In *Foundations of Digital Games*. ACM.

Palani, S.; Ledo, D.; Fitzmaurice, G.; and Anderson, F. 2022. "I don't want to feel like I'm working in a 1960s factory": The Practitioner Perspective on Creativity Support Tool Adoption. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, 1–18. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9157-3.

Remy, C.; MacDonald Vermeulen, L.; Frich, J.; Biskjaer, M. M.; and Dalsgaard, P. 2020. Evaluating Creativity Support Tools in HCI Research. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, DIS '20, 457–476. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-6974-9.

Rezwana, J.; and Maher, M. L. 2023. Designing creative AI partners with COFI: A framework for modeling interaction in human-AI co-creative systems. *ACM Transactions on Computer-Human Interaction*, 30(5): 1–28.

Shneiderman, B. 1999. User interfaces for creativity support tools. In *Proceedings of the 3rd conference on Creativity & cognition*, 15–22.

Shneiderman, B. 2000. Creating creativity: user interfaces for supporting innovation. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 7(1): 114–138.

Shneiderman, B. 2001. *Supporting creativity with advanced information-abundant user interfaces*. Springer.

Shneiderman, B. 2007. Creativity support tools: accelerating discovery and innovation. *Communications of the ACM*, 50(12): 20–32.

Shyne, F.; Facey, K.; and Cooper, S. 2024. Generating Solvable and Difficult Logic Grid Puzzles. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, 699–702.