Effective player guidance in logic puzzles

Alice May Lynch

A thesis submitted for the degree of PhD
at the
University of St Andrews

2024

DOI: https://doi.org/10.17630/sta/705

For my parents, Penny and Tom

ABSTRACT

Pen & paper puzzle games are an extremely popular pastime, often enjoyed by demographics normally not considered to be 'gamers'. They are increasingly used as 'serious games' and there has been extensive research into computationally generating and efficiently solving them. However, there have been few academic studies that have focused on the players themselves. Presenting an appropriate level of challenge to a player is essential for both player enjoyment and engagement. Providing appropriate assistance is an essential mechanic for making a game accessible to a variety of players. In this thesis, we investigate how players solve Progressive Pen & Paper Puzzle Games (PPPPs) and how to provide meaningful assistance that allows players to recover from being stuck, while not reducing the challenge to trivial levels. This thesis begins with a qualitative in-person study of Sudoku solving. This study demonstrates that, in contrast to all existing assumptions used to model players, players were unsystematic, idiosyncratic and error-prone. We then designed an entirely new approach to providing assistance in PPPPs, which guides players towards easier deductions rather than, as current systems do, completing the next cell for them. We implemented a novel hint system using our design, with the assessment of the challenge being done using Minimal Unsatisfiable Sets (MUSs). We conducted four studies, using two different PPPPs, that evaluated the efficacy of the novel hint system compared to the current hint approach. The studies demonstrated that our novel hint system was as helpful as the existing system while also improving the player experience and feeling less like cheating. Players also chose to use our novel hint system significantly more often. We have provided a new approach to providing assistance to PPPP players and demonstrated that players prefer it over existing approaches.

ACKNOWLEDGEMENTS

First, and foremost, I would like to express my gratitude to my primary PhD supervisor, Chris Jefferson, for everything. Your insight, support and guidance have been and are invaluable.

Warmest thanks to my supervisor Bobby May for your advice and guidance, particularly in the areas new to me. I will always be grateful for your enthusiasm and encouragement, especially on surviving this process.

I would like to express my supervisor Uta Hinrichs, for your introduction to and guidance on qualitative research methods and approaches. And most of all for your support and inspiration.

And I would like to thank Ian Gent, both for your support throughout my PhD and, most especially, for joining my supervisory team at the end. This thesis would not have been finished without your support and extraordinary kindness.

Fearn, without your intervention I would not have ended up on this path. Thank you.

Thank you, to my amazing sister Constance, for rigorous typo detection and for being in my corner even when I wasn't.

To Abd Ardati, Saad Attieh, Iain Carson, Mun See Chang, Thomas Hansen, Gökberk Koçak, Daniel Koudouna, Thomas Martin, David Morrison, Julian Petford, Pireh Pirzada, Tomas Vancisin, Ryo Yanagida, Xu Zhu, Zac Chave-Cox and all the other amazing students I've met in St Andrews. Thank you for the support, the friendship, the inspiration, and all the fun.

Thank you to my parents, without whom none of this would have been possible.

Finally, to my husband Conor, thank you for your patience, the many cups of tea and unending belief that I could do this.

Funding

This work was supported by the University of St Andrews.

Research Data

Research data underpinning this thesis will be made available before the submission of the final version.

DECLARATION

Candidate's declaration

I, Alice Lynch, do hereby certify that this thesis, submitted for the degree of PhD, which is approximately 31,159 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for any degree. I confirm that any appendices included in my thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

I was admitted as a research student at the University of St Andrews in September 2017.

I received funding from an organisation or institution and have acknowledged the funder(s) in the full text of my thesis.

Date 7/1/2024 Signature of candidate

Supervisor's declaration

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree. I confirm that any appendices included in the thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

Date 7/1/2024 Signature of supervisor

Permission for publication

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand, unless exempt by an award of an embargo as requested below, that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that this thesis will be electronically accessible for personal or research use and that the library has the right to migrate this thesis into new electronic forms as required to ensure continued access to the thesis.

I, Alice Lynch, have obtained, or am in the process of obtaining, third-party copyright permissions that are required or have requested the appropriate embargo below.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

Printed copy

No embargo on print copy.

Electronic copy

No embargo on electronic copy.

Date 7/1/2024 Signature of candidate

Date 7/1/2024 Signature of supervisor

Underpinning Research Data or Digital Outputs

Candidate's declaration

I, Alice Lynch, understand that by declaring that I have original research data or digital outputs, I should make every effort in meeting the University's and research funders' requirements on the deposit and sharing of research data or research digital outputs.

Date 7/1/2024 Signature of candidate

Permission for publication of underpinning research data or digital outputs

We understand that for any original research data or digital outputs which are deposited, we are giving permission for them to be made available for use in accordance with the requirements of the University and research funders, for the time being in force.

We also understand that the title and the description will be published, and that the underpinning research data or digital outputs will be electronically accessible for use in accordance with the license specified at the point of deposit, unless exempt by award of an embargo as requested below.

The following is an agreed request by candidate and supervisor regarding the publication of underpinning research data or digital outputs:

No embargo on underpinning research data or digital outputs.

Date 7/1/2024 Signature of candidate

Date 7/1/2024 Signature of supervisor

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

CNF Conjunctive Normal Form

CPC Complete Particular Cell

CSP Constraint Satisfaction Problem

CXC Complete X Cells

MUS Minimal Unsatisfiable Set

PPPP Pen and Paper Puzzle Game

SAT Boolean Satisfiability

GLOSSARY

assignment A mapping of a variable to a value in that variable's domain.

binary PPPP A PPPP where there are only 2 possible values that the player complete a cell with.

candidate A candidate is a possible value which could be placed in a cell.

candidate PPPP A PPPP where there are more than 2 possible values that the player complete a cell with.

candidate filling Making a notation inside a cell of the possible candidate for that cell.

churn The rate at which users stop engaging with a product. For example, the rate at which players stop playing a game.

clue The information in the starting state of a puzzle. For example, the values present in cells in the initial puzzle state, as opposed to those filled in by the player.

complete Writing the final choice of value in a cell.

completing *See* complete

constraint A function that maps assignments of the variables in it's scope to true or false

constraint programming An approach to programming where the relationships (constraints) between entities are defined and a solution that satisfies all the constraints is found. Discussed in more detail in Section 2.2.2.1

dimension An individual row or column or (in the case of Sudoku) a 3×3 box, indicated by bold lines.

domain A defintion of values that a variable could contain. It can be a set of values or a function that maps the variable to the values it can take.

entry point Entry point meaning the first cell that a player can complete

error handling The approach a system or player uses to respond to and recover from an error.

error state A puzzle state that contains one or more incorrectly completed cell. The puzzle is in an incorrect state.

gamification The act of applying game-like aspects to a non-game activity. For example, introducing a points system to reward and motivate a user when they complete a daily habit.

isomorphic Two puzzles are "isomorphic" if there is a way of mapping one to the other which preserves the solving routes. For example, in a Sudoku, swapping all occurrences of the digits 1 and 2 or swapping the first and second row does not affect the solving paths. A formal description of isomorphic Sudoku is provided by Russel and Jarvis and discussed further by Chapman and Rupert [128, 21].

minimal unsatisfiable set An unsatisfiable set, C, is minimal if all subsets of C are satisfiable

minimum solve Minimum solve is the number of steps and the set of techniques required to solve a puzzle, if at each step the easiest technique (based on a chosen difficulty ordering) is chosen.

notation Any notes, anywhere on the page, that assist the player in storing information about the state of the puzzle but which are not the final choice.

overlapping dimension The combination of the row, column and 3×3 box that overlap a given cell.

procedural hint A hint which aims to move players forward to a new state that is closer to the solution [81].

progressive pen & paper puzzle game Puzzles which can be solved on paper, without any external knowledge, have a single solution, and can be checked for correctness by the player.

remedial hint A hint which aims to move players out of an erroneous state. In games other than PPPPs they are also used to move out of states where a solution cannot be reached. However, in a PPPPs the only state a puzzle can be in where a solution cannot be reached in an error state [81].

rXcY Identifies a cell in the puzzle. X indicates the row number and Y indicates the column number. Numbering starts at 1 and ends at 9 (i.e. the top left cell is r1c1 and the bottom right cell is r9c9).

scope The variables a constraint is applied to.

serious game A game which has an educational or external purpose beyond simple amusement

unsatisfiable set Any unsatisfiable subset of the set of constraints of an unsatisfiable constraint problem.

variable An unknown value in a constraint problem. It is associated with a domain. A value from the domain must be assigned for a solution to the constraint problem to be found.

CHAPTER ONE

INTRODUCTION

Pen & Paper Puzzle games are an extremely popular pastime that millions of people play on a daily basis. They are published in newspapers, websites, apps and books. They are played for pleasure, to improve mental acuity [113, 95], and as a learning tool; both for their intrinsic concepts [11] and as serious games that facilitate other learning goals [31, 151].

However, there has been limited research into how to provide meaningful assistance to players of of Pen & Paper Puzzle games. Assistance in games is important and is a key area of research and development for the majority of games. Assistance is used to mitigate the variation in players, and how challenging they find a particular element of a game. Too challenging and they will become stuck and give up, not challenging enough and they will become bored and give up. In this thesis we focus on a subset of Pen & Paper Puzzle games, which we refer to as Progressive Pen & Paper Puzzle Games (PPPPs). This category includes Sudoku, Binairo, and many more. An extensive definition is given in Chapter 2. The main research questions of this thesis are: How do people play PPPPs? and How do we provide assistance to players of PPPPs in a way that improves player experience compared to current systems?

The first stage of our research was to assess how accurate existing assumptions about players were. We conducted a qualitative, in person study, videoing participants solving Sudoku and using an open coding approach to analyse the videos. We found that participants did not make notation the way the literature and guides expected, they did not solve the puzzles in the order that was expected, and they made far far more mistakes than the literature and guides expected. This study made it clear that players do not follow a uniform pattern when solving

PPPPs and any attempts to provide assistance needed to be flexible and attempt to accommodate the variation within the players. The consideration that pointing players in only one direction would result in limited assistance to a small subset of players was a key consideration when designing the novel hint system. We focused heavily on ensuring that the system was as flexible and unintrusive as possible. Therefore, we designed an assistance approach that guided players towards cells that are likely to the easiest to solve, while offering alternatives if the players found those cells too challenging. This also ensured that players discoveries and deductions were their own, and therefore made it more likely they would remain engaged with the game.

We then conducted 2 online experiments to assess the efficacy of the novel hint system design via a prototype. The initial pilot study had focused on implicit assessment of the hint systems with a between group design and an experience assessment based on the Game Evaluation Questionnaire. This study showed some interesting results, but the participants were not forced to use the hint system and whether or not they used a hint system proved to have far greater impact than which study condition they were under. Therefore the experiment was redesigned to a within group design and focused on explicit assessment of the hint systems. The first study, using Binairo, found that participants preferred to use the Novel hint system and rated it as enhancing their experience more than the next cell system, and feeling less like it felt like cheating or reduced their enjoyment. The second study, using Aquarium, had some technical issues, due to unexpectedly high engagement but still produced substantial usable data, the results of which were entirely consistent with the first experiment. The key findings from both studies are summarised in Table 1.1.

Overall, in this thesis we demonstrated that there are significant flaws in current assumptions about players of PPPPs. We present a novel and generalisable approach to providing guidance and assistance to the players of PPPPs which could be used to significantly improve player experience and engagement, both in entertainment and educational contexts. Finally, we present the findings of set of experiments that demonstrate the efficacy of the novel approach we developed across multiple PPPPs.

	Binairo Study		Aquarium Study	
	β	T-value	β	T-value
I found it enhanced my experience	0.5084	2.068	0.5690	3.687
I found it helpful	-0.2714	-0.89	-0.0310	-0.176
I found it felt like cheating	Row 3	Row 3	Row 3	Row 3
I found these hints reduced my enjoyment	Row 4	Row 4	Row 4	Row 4
It gave me the type of help I wanted	Row 5	Row 5	Row 5	Row 5

Table 1.1: The results of linear mixed-effects model of the questionnaire results. The participants were asked to rate their agreement with the statements on the left regarding the novel hint system. A more positive β value indicates greater agreement when compared to the responses rating the traditional hint system. The t-value indicates whether the difference in ratings was significant, a t-value of >1.98 or <-1.98 is considered significant for these studies.

1.1 Context

PPPPs lend themselves to computational generation and solving. There has been extensive research into the automated generation and efficient solving of puzzle game levels, though comparatively little has focused on creating a challenging yet intriguing puzzle solving experience for a given human player [66]. The challenge a game presents to a player is an essential element of the amount of enjoyment and engagement that a player experiences. A game that is too hard is frustrating and the player will give up; too easy, and the player will become bored and, again, give up [73, 7].

The challenge a problem presents differs depending on whether a person or an AI is attempting to solve it. Some elements of a task that a person finds challenging will be computationally trivial, while others will be trivial to the person but computationally challenging [150]. This has a natural impact on difficulty assessments: if the difficulty of a problem is measured by how hard an Artificial Intelligence found it, will that reflect the challenges a person would face?

AI models of human behaviour are, by necessity, based on assumptions, frequently the assumption that people are systematic and rational. This assumption is present across a range of fields [104, 71, 79, 48]. This assumption is being challenged in AI models where the primary goal is to predict human behaviour or facilitate human understanding, such as studies examining the difficulty or challenge presented by some types of games. There is extensive research examining how to analyse how difficult a player found a level, the impact that has on their experience, and whether that can be used to construct AI which can either automatically classify levels or facilitate a dynamic difficulty system [125, 51, 6, 160, 7, 80, 4]. The growing areas of explainable AI and computational rationality relate to this and attempt to combine underlying cognitive processes with AI models; for example, Tabrez *et al* produced a system capable of identifying and expressing the disconnect between their underlying models and the mental models of the people interacting with them [148].

However, these assumptions persist in models of Pen & Paper puzzles. There has been extensive research into better, more efficient ways to generate valid puzzles and solve them. There is an assumption that the challenge presented when solving a particular level computationally reflects the challenge a person will be presented with [125]. Jarůsek & Pelánek noticed this phenomenon in Sokoban puzzles, as existing difficulty heuristics did not appear to reflect the experience of players. They examined both the time taken to solve the puzzle and the order in which people solved the puzzle and produced a model that partially predicted the difficulty of the puzzles [64].

Providing assistance within a game can allow the challenge level of the game to suit a wider range of players. One of the best examples of balancing a game for multiple skill levels comes from racing games, such as MarioKart where players that fall too far behind are provided with better 'power-ups'[1] than players that are doing better [19]. However, this type of assistance still requires some understanding of the player's experience. In the MarioKart example, the game assumes that the further behind you are, the more you are struggling; this is a fairly accurate assumption in a racing game. However, this can be harder to assess in puzzle games. Chris Campbell of Big Fish Games described an example of this issue when trying to build a hint system for the puzzle game "Drawn: The painted tower" - the hints failed to reflect the player's current understanding of the puzzle and therefore increased frustration rather than mitigating it [28]. They improved the system by making use of far more of the information in the game state to better assess and understand what the player knew and had worked out, and therefore what was likely to be useful to them [28]. Clearly, hint systems are more effective when they reflect the state of the puzzle from the player's perspective.

---

[1] Power-ups are in-game items that briefly give the player a significant advantage, for example by making them immortal or much faster than other players

1.2 Outline of this thesis

In this thesis, we focus on a subset of Pen & Paper puzzle games, referred to as PPPPs and defined in Chapter 2. We investigate how players solve PPPPs, and how to provide meaningful assistance that allows players to recover from being stuck while not reducing the challenge presented by a problem to trivial levels.

Chapter 2 introduces the key concepts from the relevant fields and explores related literature. We define and discuss PPPPs and how they are solved and generated. We discuss 'serious games', and the use of PPPPs as serious games. We also introduce DEMYSTIFY, a tool for producing human-understandable explanations of each step of a PPPP. Finally, we discuss the importance of providing appropriate assistance and how that is provided in digital games.

In Chapter 3 we present the results of a survey on people's puzzle solving habits and the results of a qualitative study of people solving Sudoku. The survey facilitated our design of the qualitative study. The qualitative study allowed us to examine existing assumptions about player behaviour and compare them with the observed player behaviour. We found all existing assumptions of player behaviour to be flawed. We also categorise different notation techniques used by players when solving Sudoku. This qualitative study provided key insights into how players approach PPPPs, which we used to facilitate our design of a novel hint system. The qualitative study also demonstrated that an interesting area of further research would be understanding player notation.

Next, in Chapter 4, we discuss the design and implementation of our novel hint system for PPPPs. The novel hint system focusses on the principle of guiding players towards the next step, rather than directly telling them the next step. A survey of existing hint systems in digital PPPPs, also presented in this chapter, demonstrated that the majority of current digital systems simply tell the player what the next step could be (or fill in a cell at random). The novel hint system we designed instead shows the player which cells are likely to be easier for them to solve.

The assessment of our novel hint system is discussed in Chapter 5, where we present four studies, comparing player experience of the novel hint system to player experience of a traditional 'fill in the next cell' system. The results of these studies demonstrated that players found the systems equally helpful, but the novel hint system felt less like cheating and enhanced their experience more than the more traditional system. The studies were conducted using Binairo and Aquarium, two quite different PPPPs, suggesting that the novel hint system is fairly generalisable.

Finally, in Chapter 6 we present our conclusions and possible directions for future work.

1.3 Contributions

The main research questions of this thesis are:

- How do people play PPPPs?
- How do we provide assistance to players of PPPPs in a way that improves player experience compared to current systems?

We consider the significant contributions of this thesis to be:

Contribution 1. Existing assumptions made by both models and guides about players of PPPPs are all flawed. Players are not systematic, consistent and effectively error free; they were found to be unsystematic, idiosyncratic and highly error prone. We found this via an in-person, qualitative study of how people play PPPPs This is discussed in Chapter 3 and has been published by Lynch *et al* [85]. This impacts both the design of AI models of players, and approaches to player assistance and game design.

Contribution 2. The design and implementation of a generalisable novel hint system for PPPPs, based on guiding players rather than telling them. This system was demonstrated to show significant improvements in player experience. This was done via a series of online experiments comparing the novel hint system to a traditional hint system across multiple PPPPs. The design of the system discussed in Chapter 4, the experiments are discussed in Chapter 5. This system could allow significant improvements in player experience and engagement across PPPPs, both in entertainment and more serious contexts.

CHAPTER TWO

BACKGROUND

2.1 Progressive Pen & Paper Puzzle Games (PPPPs)

"Puzzle game" is a somewhat poorly defined term. The following games are all described as puzzle games, despite being very different: "Portal": a complex, 3D, first person shooter style, computer game [152], "Drawn®: The Painted Tower": a point-and-click adventure game [12], and "Sudoku": a logic puzzle that can be solved using a pen (or pencil) and paper [140]. In this thesis, we focus on a subset of the latter category. We will refer to them as *Progressive Pen & Paper Puzzle Games (PPPPs)* throughout.

Definition 1 (Progressive Pen & Paper Puzzle Games) *PPPPs are puzzles which can be solved on paper, without any external knowledge, have a single solution, and can be checked for correctness by the player. Furthermore, every correct move a player makes adds information to the game.*

Crosswords are not covered by Definition 1; solving the clues requires some general or external knowledge. Furthermore, while the player may successfully fill in the grid with valid words, they cannot be sure, without consulting the puzzle designer, whether they have successfully solved the clues or instead merely found a word that fits the grid.

All PPPPs consist of some starting information, called clues, and a set of rules that constrain the solution. Variations in starting arrangements create different instances (often called levels) of a given game.

Johnson argues that digital versions of PPPPs represent a unique and important segment of the casual game market [66]. Casual games and their players comprise one of the largest shares of the digital game market [24, 78]. Despite this, they are often neglected in favour of 'core' games and their players [66]. Much of the research into casual games and their players has focused either on the economic aspects [127, 47, 59] or the demographic differences between 'casual' and 'hardcore' players [29, 159, 23] rather than their design. This issue is exacerbated for PPPPs, to the extent that Johnson argues that players of PPPPs represent 'the most academically neglected yet numerically vast demographic of "gamers" alive today' [66].

2.1.1 Examples of PPPPs

Our research focused on a generic solution for PPPPs. In this section, we introduce a number of PPPPs. Aquarium, Binairo, and Sudoku are used in the experiments described in this thesis. The remaining PPPPs are introduced to provide context and background.

Aquarium This puzzle game consists of an $n \times n$ grid, marked with arbitrarily shaped regions (aquariums), outlined with a darker border. Each row and column has a number at the end and top, respectively, indicating the total number of cells that should be water in the relevant row or column. The 'water' flows sideways and down within each region; every cell in the same row and region as a water cell must also be water, and every cell below a water cell and in the same region must also be water [115]. Example shown in Figure 2.1.

Binario This puzzle game consists of an $n \times n$ grid. It starts with a number of cells filled in with either a 1 or 0 (or a black or white circle in some variations). Every cell must be filled in with a 1 or 0 but there must be no more than two identical values adjacent in any orthogonal direction [116]. Example shown in Figure 2.2.

KenKen This puzzle game consists of an $n \times n$ grid with arbitrarily shaped regions (cages) marked out. Each row and column must contain all the digits $1 - n$ exactly once; there are no constraints on which digits can appear in a cage. Each area is labelled with a number and mathematical operator ($\times$, $\div$, +, -). The numbers

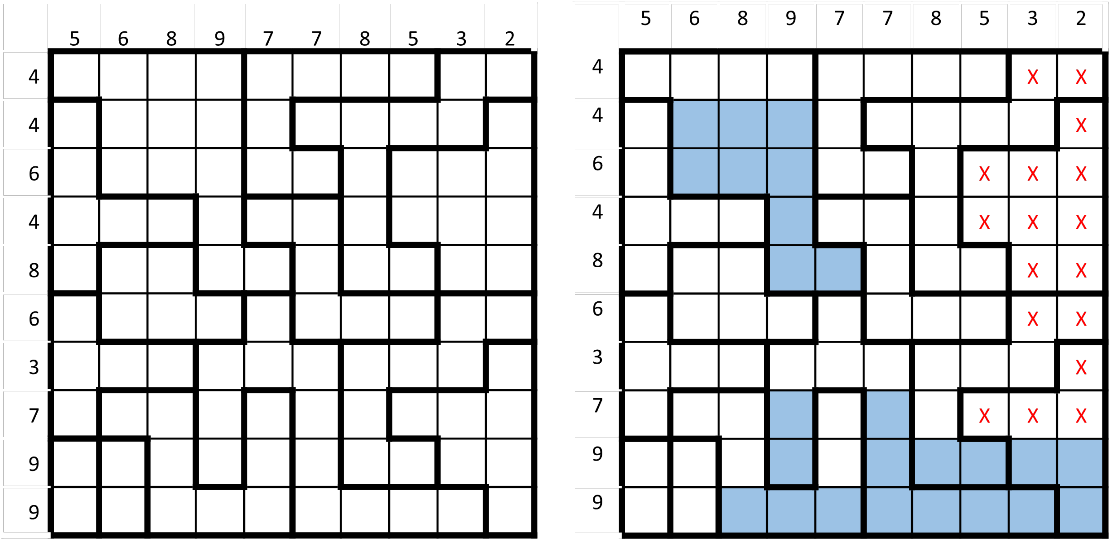
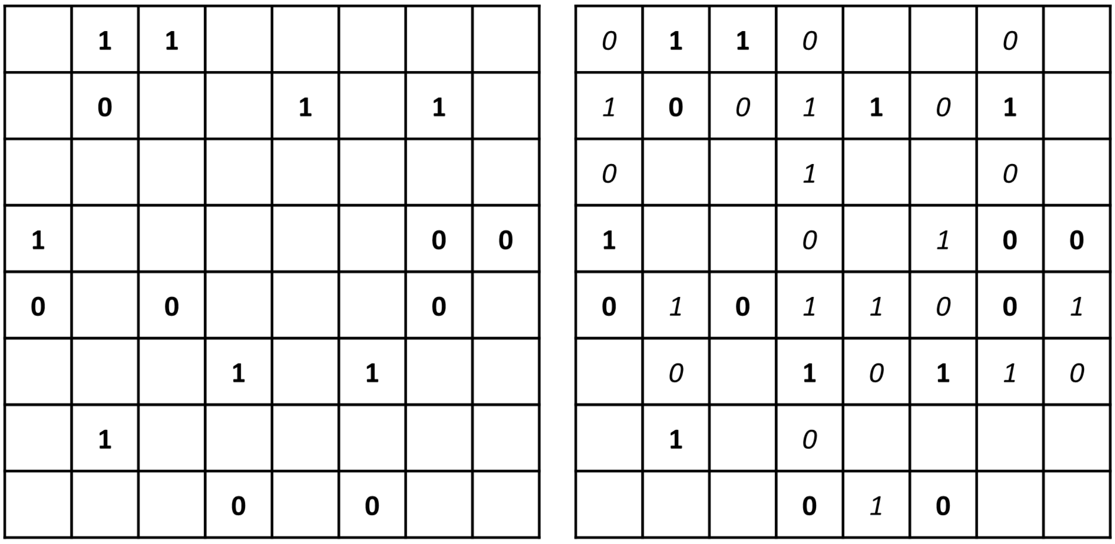


Figure 2.1: Example of Aquarium puzzle, left shows the starting state, right shows a partially completed puzzle [115]



	1	**1**					
	0			**1**		**1**	
1						**0**	**0**
0		**0**				**0**	
			1		**1**		
	1						
			0		**0**		

0	**1**	**1**	*0*			*0*	
1	**0**	*0*	*1*	**1**	*0*	**1**	
0			*1*			*0*	
1			*0*		*1*	**0**	**0**
0	*1*	**0**	*1*	*1*	*0*	**0**	*1*
	0		**1**	*0*	**1**	*1*	*0*
	1		*0*				
			0	*1*	**0**		

Figure 2.2: Example of Binairo puzzle, left shows the starting state, right shows a partially completed puzzle [116]

in the area must, when combined using that operator, produce the number [136]. Example shown in Figure 2.3.

SkyScrapers This puzzle game consists of an $n \times n$ grid. It starts with numbers at the end of some of the rows and columns. These indicate the number of 'skyscrapers' that can be 'seen' from that end of the row or column. Skyscrapers are represented by the numbers 1 to n: the number indicates their height, and

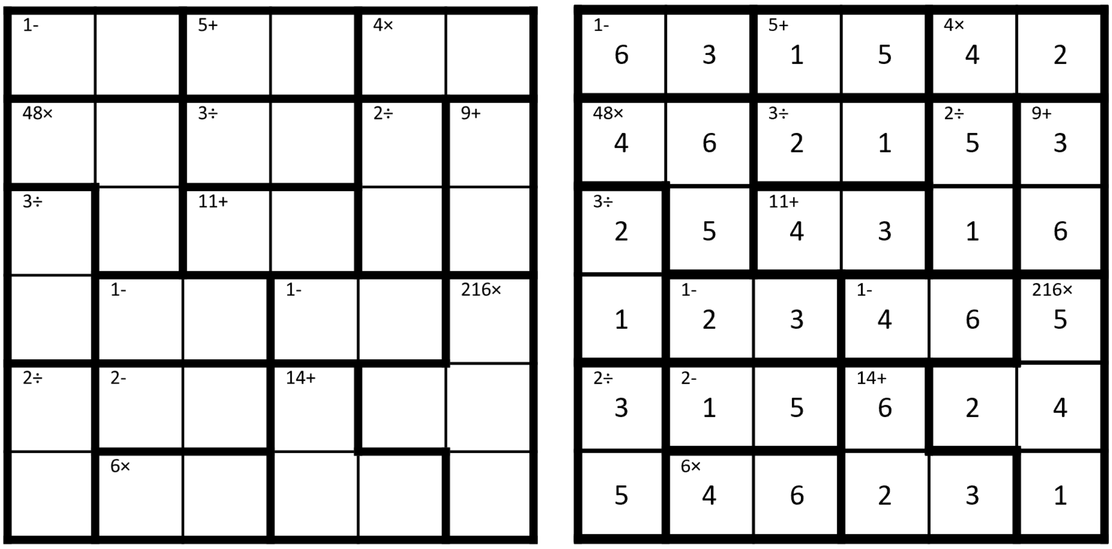


Figure 2.3: Example of KenKen puzzle, left shows the starting state, right shows a completed puzzle [136]

taller skyscrapers block shorter skyscrapers from being 'seen'. Each digit should appear exactly once in every row and column [117]. Example shown in Figure 2.4.

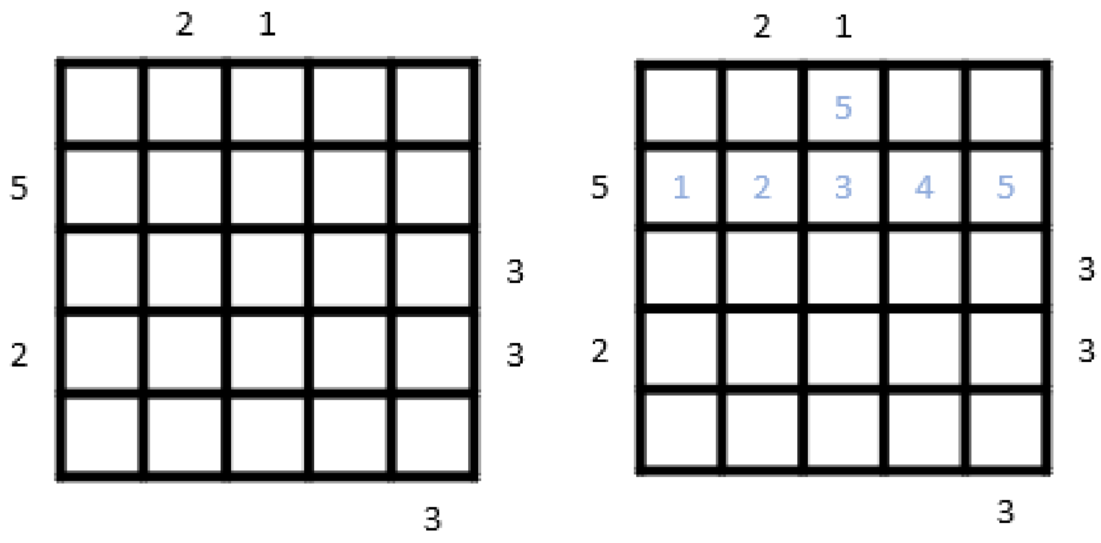


Figure 2.4: Example of SkyScrapers puzzle, left shows the starting state, right shows a partially completed puzzle [117]

Star Battle This puzzle game consists of an $n \times n$ grid marked with arbitrarily shaped regions (similar to KenKen), outlined with a darker border. Each region, row, and column must have y stars contained within it. It does not start with any clues [118]. Example shown in Figure 2.5.

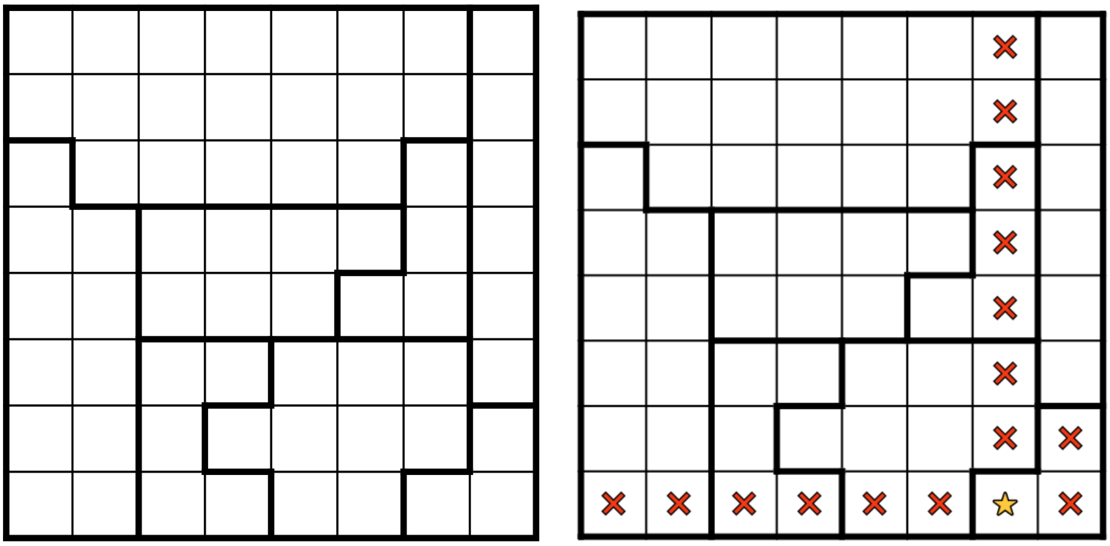


Figure 2.5: Example of Star Battle puzzle, left shows the starting state, right shows a partially completed puzzle [118]

Sudoku This very popular puzzle game has several variants, but the original consists of a 9×9 grid, split into nine 3×3 regions. It starts with a number of cells filled in with a digit between 1 and 9 (inclusive). Every 3×3 box, row, and column must be filled in with the digits $1 - 9$ without repetition of the digits within a given dimension (box, row, column). There are variants of Sudoku that add constraints, change the shape and size, or change the values to be filled in [140]. Example shown in Figure 2.6. We discuss Sudoku in detail in Section 2.1.2.

Tents and Trees This puzzle game consists of an $n \times n$ grid. The starting state consists of a number at the end of every row and column and an arrangement of trees within the cells of the grid. The object of the game is to place a tent orthogonally next to every tree, while having exactly the number of tents in a row/column as the number at the end of that row/column. Tents cannot be placed adjacent (including diagonally) to another tent [120]. Example shown in Figure 2.7.

We categorise PPPPs into binary PPPPs and candidate PPPPs. A binary PPPP is a PPPP where the player can fill only two values into each cell, for example Binairo where the only possible values are 0 and 1. Aquarium, Binairo, Star Battle, and Tents and Trees are all binary PPPPs. A candidate PPPP is a PPPP where the

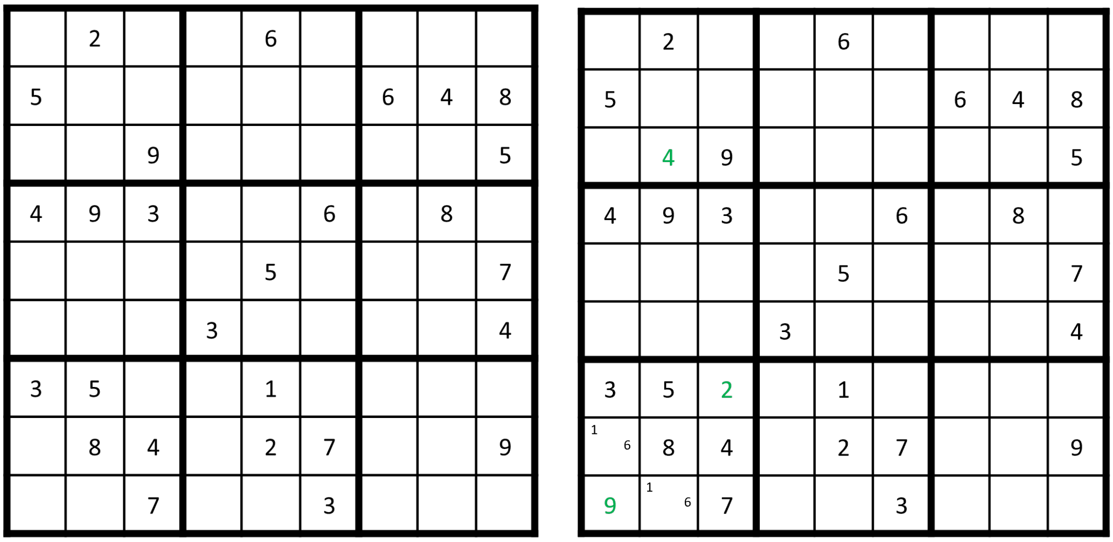


Figure 2.6: Example of Sudoku puzzle, left shows the starting state, right shows a partially completed puzzle [119]

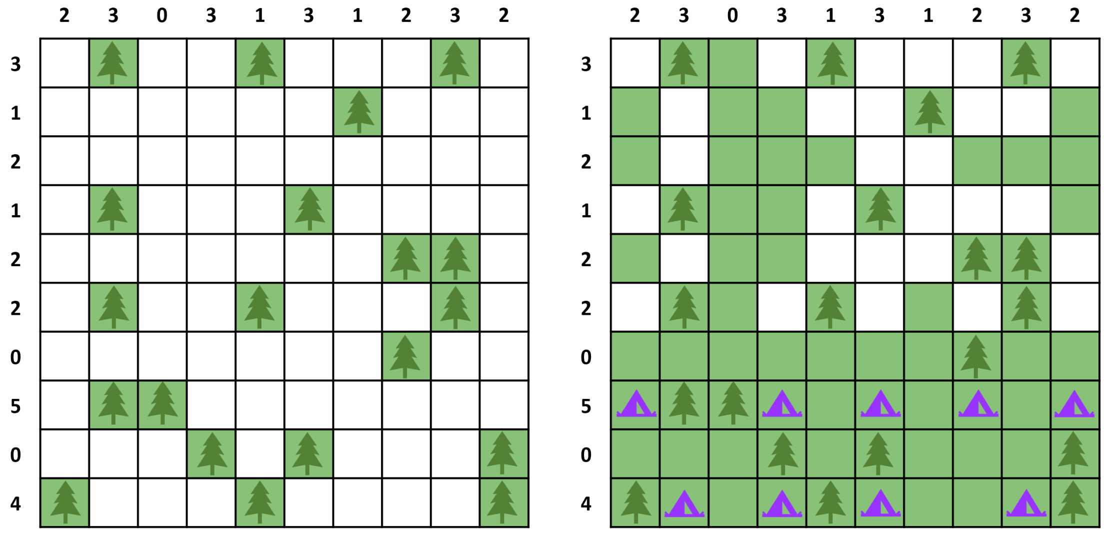
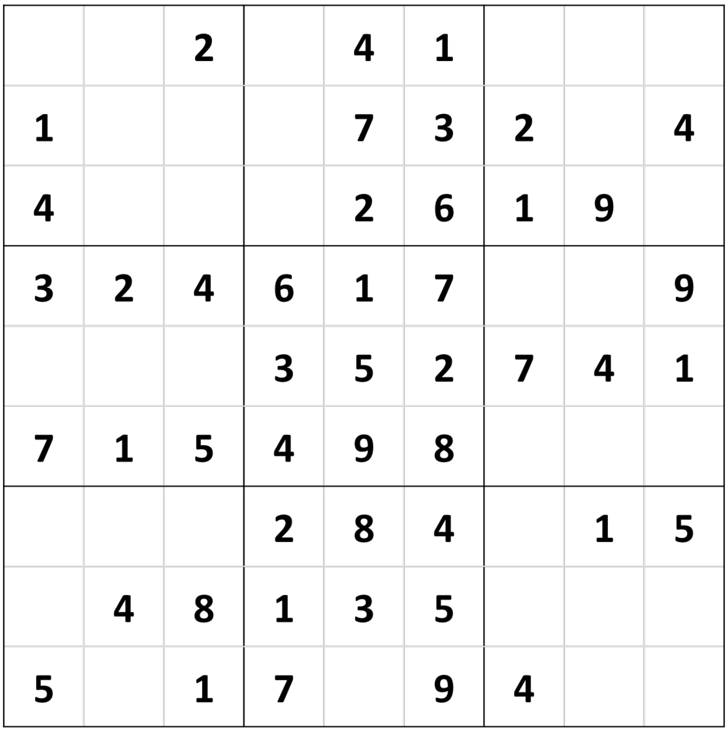


Figure 2.7: Example of Tents and Trees puzzle, left shows the starting state, right shows a partially completed puzzle [120]

player can fill in more than two possible values in a cell. This means that there is value for the player in making notes of possible candidates within a cell. KenKen, Skyscrapers, and Sudoku are all candidate PPPPs.

2.1.2 Sudoku in Detail

There has been extensive research on Sudoku directly and research that makes use of Sudoku in some way. As a result, we frequently use Sudoku as an example or



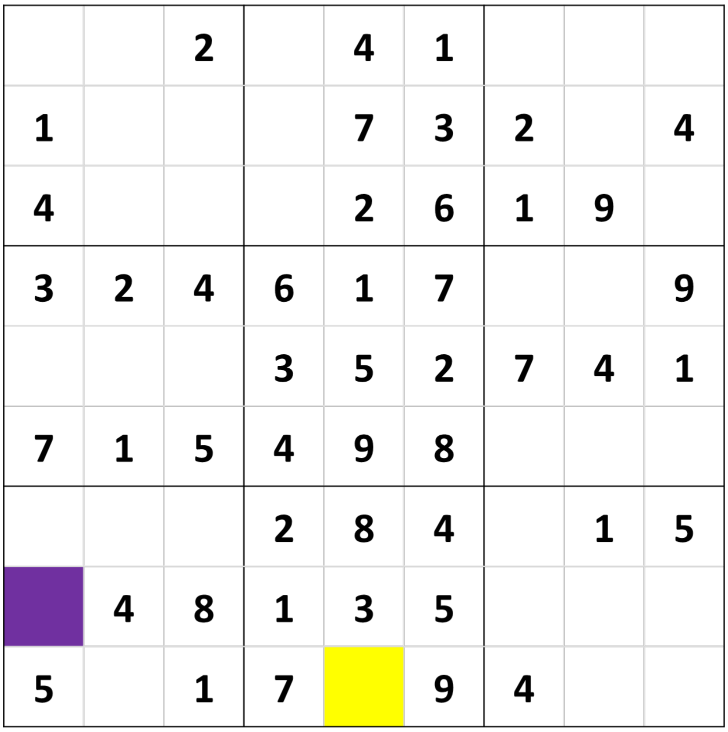
		2		4	1			
1				7	3	2		4
4				2	6	1	9	
3	2	4	6	1	7			9
			3	5	2	7	4	1
7	1	5	4	9	8			
			2	8	4		1	5
	4	8	1	3	5			
5		1	7		9	4		

Figure 2.8: Example Sudoku Puzzle

case study throughout this thesis. Our first study, discussed in Chapter 3, uses Sudoku. Therefore, we provide in this section a detailed explanation of Sudoku and the techniques that have been identified for solving Sudoku.

A Sudoku puzzle, shown in Figure 2.8, consists of a 9×9 grid divided into nine 3×3 boxes. Every row, column, and 3×3 box must contain exactly the digits 1-9 and the Sudoku has a unique solution. A Sudoku level starts with a number of cells in the puzzle filled in, referred to throughout this thesis as clues. The player then tries to use the puzzle rules to deduce the digits that go into the empty cells. Sudoku rose to popularity outside of Japan in 2004 when Wayne Gould sold mass-produced puzzles to the London times, and has become the most extensively studied PPPP, both by players and academics. There are extensive resources available on the techniques that can be used to solve Sudoku puzzles.

A non-exhaustive list of techniques follows; many of these techniques are known by multiple names. We use the names described in this section throughout this thesis.



		2		4	1			
1				7	3	2		4
4				2	6	1	9	
3	2	4	6	1	7			9
			3	5	2	7	4	1
7	1	5	4	9	8			
			2	8	4		1	5
🟪	4	8	1	3	5			
5		1	7	🟨	9	4		

Figure 2.9: Example of naked single, r9c5, highlighted in yellow and a hidden single, r8c1, highlighted in purple

2.1.2.1 Basic Techniques

The basic techniques identify a single digit that must be the solution of a particular cell. All other techniques focus on eliminating candidates. There are two approaches, the Naked Single and the Hidden Single.

Naked Single. A Naked Single occurs when eight of the nine possible digits are already present in the overlapping dimensions of a cell. Given that no value may be repeated in any dimension, the remaining digit must be the solution to that cell. [140] An example of this is the cell in row 9 and column 5 (r9c5[1]) of Figure 2.9, highlighted in yellow, which can only be 6.

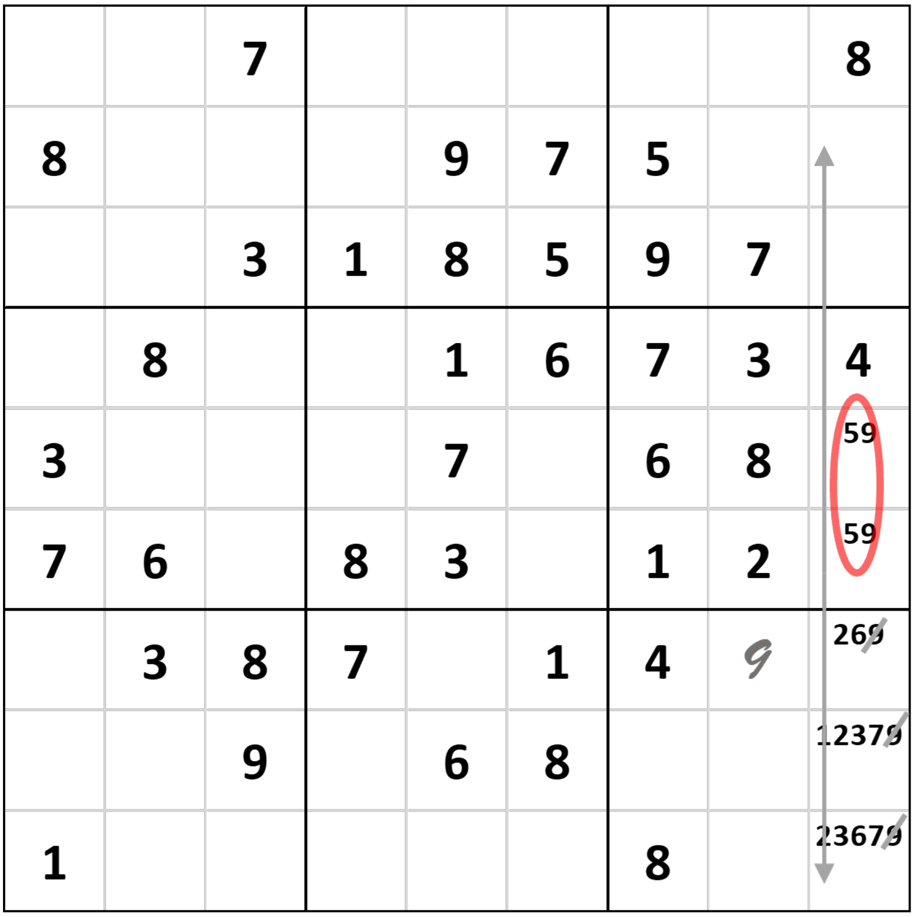
Hidden Single. A Hidden Single occurs when a digit, d, can occur only in one cell in a dimension. Given that all digits must occur once in every dimension, the solution to that cell must be d. [140] For example, in Figure 2.9, r8c1, highlighted in purple, contains a hidden single. The only cell 2 can go into is the cell in the bottom left 3×3 box. It is excluded from the three cells in row 7 (the top row of

---

[1] rX indicates the row in the Sudoku, numbered 1-9, cX indicates the column, numbered 1-9

the box) by the 2 in the same row (r7c4). 2 is excluded from the empty cell in row 9 (r9c2) by the 2 in the same column (r4c2). This allows the player to complete the cell with 2, despite 6 and 9 not being excluded directly from the cell.

Many Sudokus can be solved using only these techniques. Although candidate filling can be used to facilitate the use of Naked and Hidden Singles, it is not considered necessary.



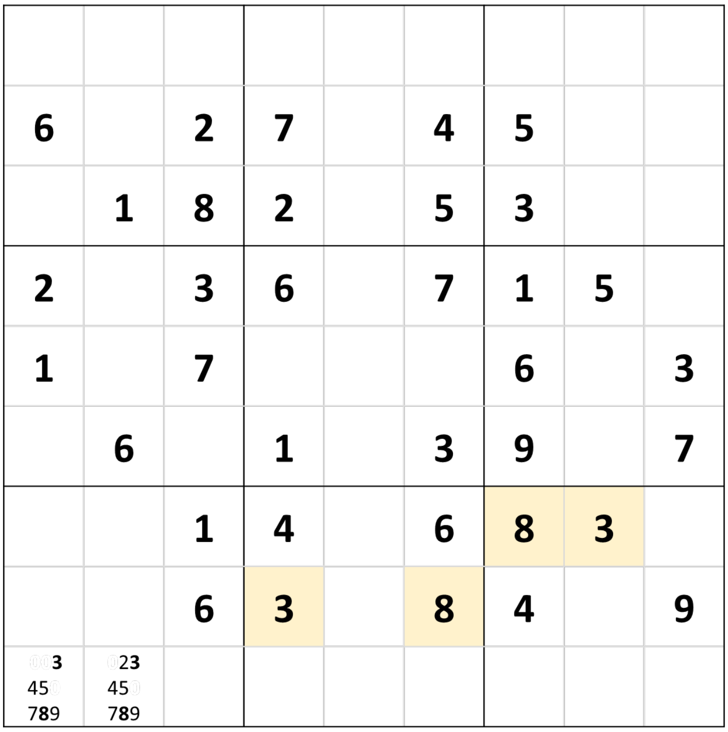
		7						8
8				9	7	5		
		3	1	8	5	9	7	
	8			1	6	7	3	4
3				7		6	8	59
7	6		8	3		1	2	59
	3	8	7		1	4	9	26~~9~~
		9		6	8			1237~~9~~
1						8		2367~~9~~

Figure 2.10: Example of a Naked Pair in r5c9 and r6c9. The Naked Pair results in the elimination of the candidate nines in c9. This reveals a hidden single in r7c8.

2.1.2.2 Subset-Based Techniques

For Sudokus that cannot be solved using the basic techniques, there are many more advanced techniques available. These techniques are generally considered more challenging than Naked and Hidden Singles [144, 140]. They all depend on tracking all possible values for every cell [140] and aim to eliminate possible values from cells until a Naked or Hidden Single is produced. The process of tracking all possible values is often referred to as candidate filling.

Naked Pair. A Naked Pair occurs when the only available candidates for two cells are the same two digits. This allows the player to deduce that the digits must be placed into the two cells; otherwise, there would either be a repeated digit or an empty cell. This allows the digits to be excluded from candidates elsewhere in the dimension. For example, Figure 2.10 shows a Naked Pair in r5c9 and r6c9, containing candidates 5 and 9. This allows candidate 9s elsewhere in the column to be removed. This in turn reveals a hidden single in r7c8, allowing the player to complete the cell and progress.

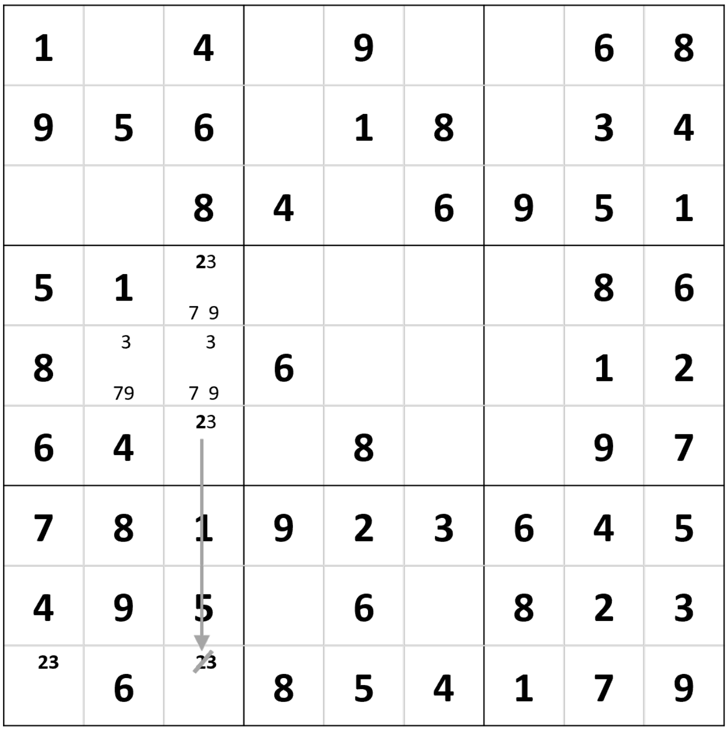


6		2	7		4	5		
	1	8	2		5	3		
2		3	6		7	1	5	
1		7				6		3
	6		1		3	9		7
		1	4		6	8	3	
		6	3		8	4		9
3 45 789	23 45 789							

Figure 2.11: Example of a Hidden Pair in r9c1 and r9c2. r9c1 and r9c2 are the only places that 3 and 8 can be placed in the bottom left 3 × 3 box. They are excluded from all other empty cells by the 3s in r7c8 and r8c4 and the 8s in r7c7 and r8c6. The hidden pair can be used to eliminate all candidates apart from 3 and 8.

Hidden Pair. A Hidden Pair occurs when two candidates are excluded from all but two cells in a dimension. This allows the player to deduce that no other candidate can be placed into these two cells; hence, other candidates can be eliminated from these two cells. An example can be seen in Figure 2.11 in r9c1 and r9c2. r9c1 and r9c2 are the only places where 3 and 8 can be placed in the bottom left 3 × 3 box. They are excluded from all other empty cells by the 3s in r7c8 and r8c4 and the 8s in r7c7 and r8c6. The hidden pair can be used to eliminate all candidates except for 3 and 8. This produces a Naked Pair which may allow further deductions.

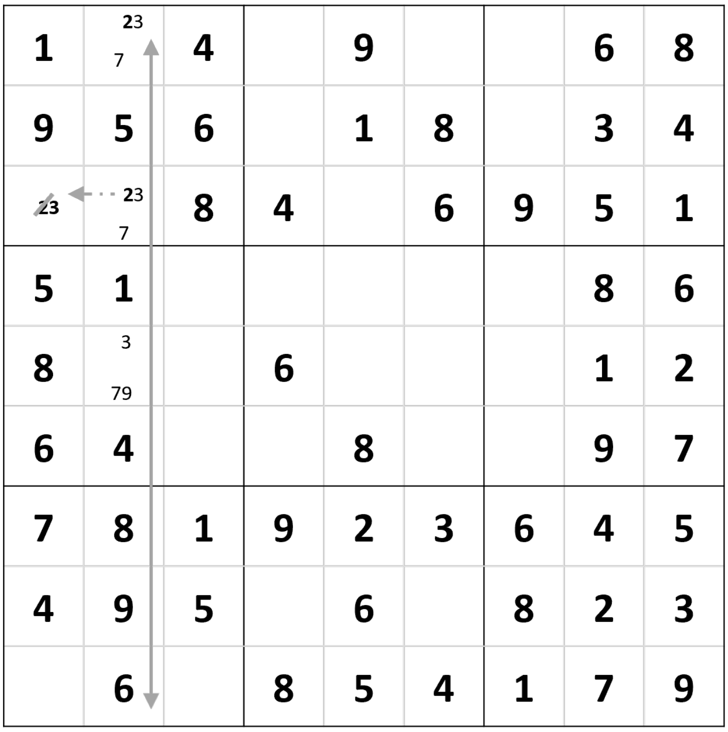
Naked and Hidden Pairs can be extended to Naked/Hidden Triples, which occur with 3 candidates and 3 cells, and Naked/Hidden Quads, which occur with 4 candidates and 4 cells . These are assumed to be more challenging than the pair versions, although there is no consensus on whether a Hidden Pair is easier or harder than a Naked Triple. It is worth noting that the digits comprising a naked triple or quad must be the only candidates available across all three/four cells. For example (1,2,3), (1,2), (2,3) or (1,2), (2,3), (1,3).



1		4		9			6	8
9	5	6		1	8		3	4
		8	4		6	9	5	1
5	1	23 7 9					8	6
8	3 79	3 7 9	6				1	2
6	4	23		8			9	7
7	8	1	9	2	3	6	4	5
4	9	5		6		8	2	3
23	6	~~2~~3	8	5	4	1	7	9

Figure 2.12: Example of a Pointing Pair in r4c3 and r6c3. r4c3 and r6c3 are the only places in the center left 3 box that a 2 can be placed. This allows 2 to be excluded as a candidate elsewhere in column 3.

Intersection Eliminations. Intersection Elimination is a generalisation, introduced by Stuart in The Logic of Sudoku [140], which covers techniques such as **pointing pairs** and **box/line reduction**. Stuart states the generalisation of this rule



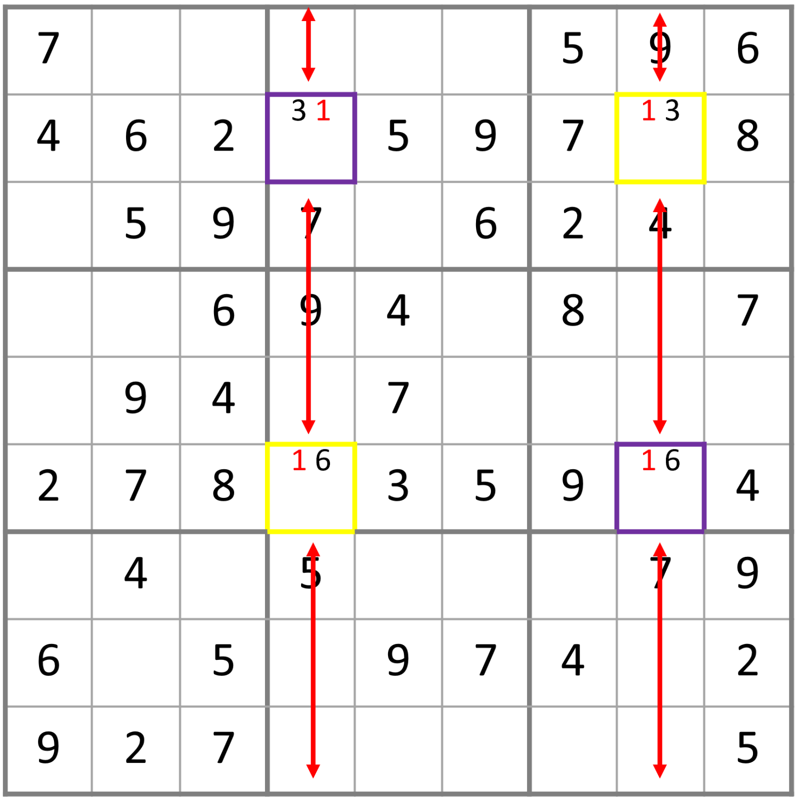
1	23 7	4		9			6	8
9	5	6		1	8		3	4
~~2~~3	23 7	8	4		6	9	5	1
5	1						8	6
8	3 79		6				1	2
6	4			8			9	7
7	8	1	9	2	3	6	4	5
4	9	5		6		8	2	3
	6		8	5	4	1	7	9

Figure 2.13: Example of a Box Line reduction. r1c2 and r3c2 are the only places in column 2 that the digit 2 can be placed. This allows it to be excluded from all cells elsewhere in the top left 3 × 3 box.

as "If any one number occurs twice or three times in just one [dimension], then we can remove that number from the intersection of another [dimension]." [140].

A pointing pair/triple occurs when all cells in a 3 × 3 box where a digit can be placed are in the same row or column. That digit can then be excluded from everywhere else in the row or column. There is an example of a pointing pair in Figure 2.12: r4c3 and r6c3 are the only places in the centre left 3 × 3 box where a 2 can be placed. This allows 2 to be excluded as a candidate elsewhere in column 3.

A box-line reduction occurs when the only places a digit can be placed in a row or column are within a 3 × 3 box. The digit can then be excluded from all other cells in that 3 × 3 box. There is an example in Figure 2.13: r1c2 and r3c2 are the only places in column 2 where 2 can be placed. This allows it to be excluded from all cells elsewhere in the top left 3 × 3 box.



7						5	9	6
4	6	2	3 1	5	9	7	1 3	8
	5	9	7		6	2	4	
		6	9	4		8		7
	9	4		7				
2	7	8	1 6	3	5	9	1 6	4
	4		5				7	9
6		5		9	7	4		2
9	2	7						5

Figure 2.14: Example of an X-Wing. The only place in the second and sixth rows that 1 can be placed is in the highlighted cells. Therefore either the yellow outlined cells must contain 1, or the purple cells must contain 1. As a result the player can eliminate 1 from everywhere else in the fourth and sixth column

2.1.2.3 Cycles

The Naked/Hidden Pair approach can be developed into a family of techniques known as 'fishy cycles[2]'. This is a large family of techniques which we only outline here. A full explanation of these techniques can be found in "The Logic of Sudoku" by Stuart [140].

X-Wing. The simplest 'fishy-cycle' is the X-Wing. An X-Wing consists of 4 cells, with at least one candidate in common, which line up along rows and columns to form a square, as shown in Figure 2.14. The cells in the square may be separated by intervening cells. In either the rows or the columns (not both) the candidate must only appear in the cells that form the X-Wing. This means that one of the two cells in that row/column will contain that candidate, which in turn means that the candidate can be eliminated from other cells in that row/column.

---

[2] Several of them are named after fish - Turbot, Swordfish, Jellyfish

The larger cycles work on the same pattern as the X-Wing but with a greater number of cells in the grid. The Swordfish occurs when 9 cells are arranged in a 3 by 3 grid arrangement and the Jellyfish requires 16 cells arranged in a 4 by 4 grid arrangement.

2.1.2.4 Esoteric Techniques

There are a variety of more advanced techniques, many of which are quite rare [140, 143].

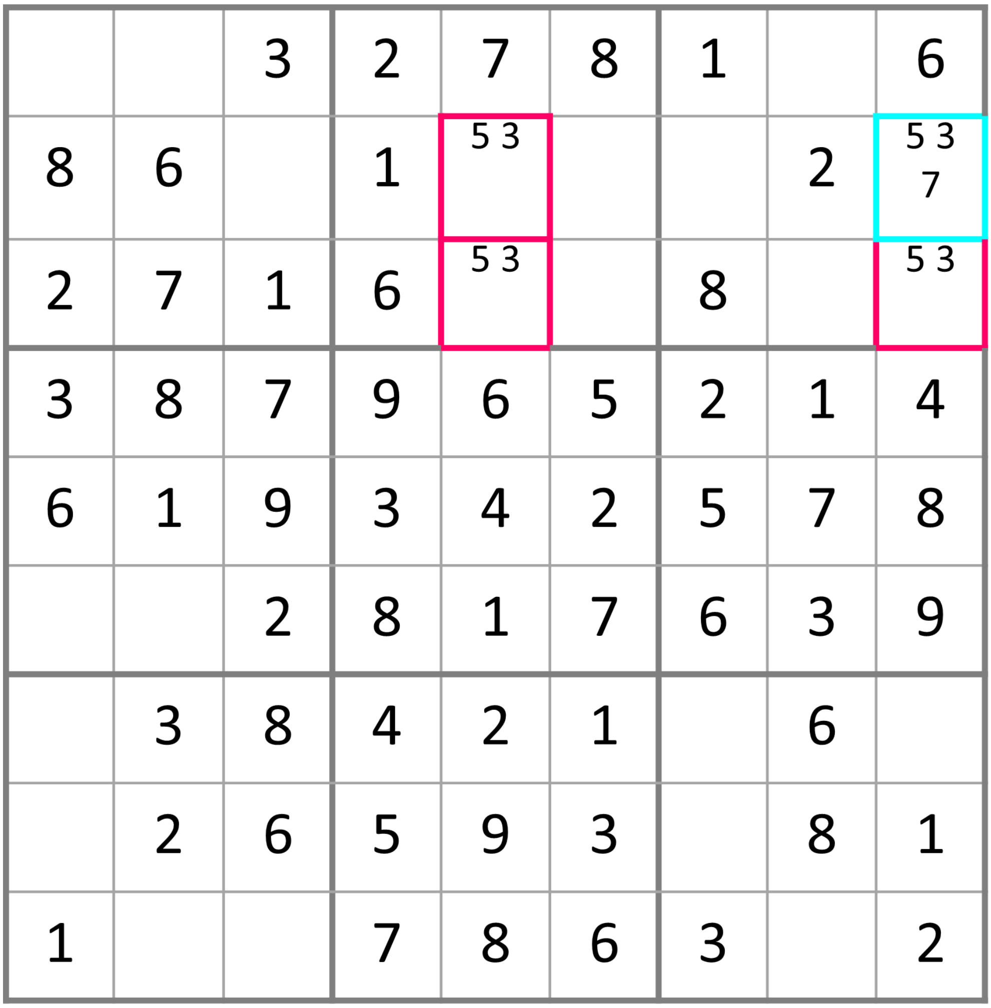
Unique Rectangle. This technique is predicated on Sudokus requiring a unique solution. It consists of 4 cells, which line up along rows and columns to form a square. All 4 of the cells should have two candidates in common and they should be the only candidates available in 3 of the cells. Those candidates can then be eliminated from the fourth cell (the one that contains alternatives) since, if it was completed to be either of the two candidates that all four cells have in common the Sudoku would have multiple solutions.

2.1.2.5 Chains

Chain techniques are effectively a formalised trial and error approach. They are assumed to be one of the last resort options, being considered both confusing and messy. The player guesses a digit and then follows the chain of deductions until they either complete the Sudoku, run out of possible deductions or reach a contradiction that can be used to eliminate all candidates involved in the chain. Players often use colours to track the chain and choose a starting cell with few candidates. There are several methods of performing “Chain” reasoning depending on where the player starts the chain and if they run multiple ones simultaneously [141]. Simple Colouring (used in Chapter 3) is a chain technique [141].

2.1.3 Serious Puzzle Games

"Serious game" is a term first defined by Abt [2], in 1970, as games which "have an explicit and carefully thought-out educational purpose and are not intended to be played primarily for amusement". It is a definition that is still largely applicable, however, it can be expanded to include games (and more often game mechanics [111]) that have been adapted to serious use, rather than designed from scratch.



		3	2	7	8	1		6
8	6		1	5 3			2	5 3 7
2	7	1	6	5 3		8		5 3
3	8	7	9	6	5	2	1	4
6	1	9	3	4	2	5	7	8
		2	8	1	7	6	3	9
	3	8	4	2	1		6	
	2	6	5	9	3		8	1
1			7	8	6	3		2

Figure 2.15: Example of a Unique Rectangle, shown in the four outlined cells. If all of them contained 5 and 3 then the ordering could be switched while still producing a valid solution. Therefore the cell outlined in blue must contain 7.

Susi *et al* expressed the core meaning of the term as 'games used for purposes other than mere entertainment' [147]. This difference is important, as one of the main criticisms of serious games has long been that they are not very good games [17, 9, 147]. Koster criticises gamification because the results are often a reward system shallowly layered over an activity with little to no actual game play, as they put it: "A reward structure alone does not a game make" [73]. The goal of serious games is often educational, but can also be to crowdsource access to hardware or people, [131], for example FoldIt, [27], which allowed players to contribute to protein folding research by playing the game.

Serious games represent a huge and growing market, with their global market size in 2022 estimated at US$ 9.1 Billion [135]. They have been shown to be effective when done well [26, 22] although they have been found to be less effective if the underlying game is considered ‘boring’ or lacking in challenge [17, 9, 147, 69].

PPPPs are frequently adapted into serious games, mostly for educational purposes. They are used both for their intrinsic concepts of logic and deduction, [11, 96, 124, 32, 100], and sometimes in adapted forms, to teach other skills [31, 13, 110, 151]. They are also sometimes recommended as a treatment or preventative for Alzheimer’s disease and dementia [5]. However the efficacy of their use in this context is highly controversial: some studies show some improvement [41], often combined with other factors [102, 76] and some suggest little improvement[139]. Despite their use beyond entertainment, there is very limited research on how to support players or encourage engagement.

2.2 Solving PPPPs

PPPPs are designed to be solved by an unassisted person. While digital versions may provide feedback, both to help with error handling and to assist the player in solving, the puzzles are solvable without that assistance. This limits the complexity of the rules and deductions required. People have limited working memory [30, 15], but generally have excellent pattern recognition and the ability to make complex logical deductions based on a number of contributing factors [122, 56]. Computers have extensive working memory and comparatively poor pattern recognition [74, 106, 138]. Therefore, people and computers take alternative approaches to solving PPPPs.

2.2.1 How do people solve PPPPs

There is very limited formal research on how people solve PPPPs. People are believed to solve PPPPs using logical deductions and pattern recognition [140]. This is supported by the many tutorials and guides available which describe the patterns that players should look for and the deductions that those patterns allow [140, 58, 114]. We use Sudoku (the most documented of the PPPPs) and describe examples of the techniques described in the guides and literature in Section 2.1.2.

The key assumptions we found in the literature about how people play PPPPs can be summarised as:

- Players solve puzzles using the techniques commonly described in the literature.
- The only annotation that players use is to list the allowable candidates for each cell.
- Players solve puzzles by repeatedly choosing the move of easiest difficulty.
- Players do not make mistakes when solving puzzles.

2.2.2 How do computers solve PPPPs

Computationally solving PPPPs has been extensively explored, especially for the ubiquitous Sudoku. Sudoku solving has been attempted computationally using many techniques, including backtracking algorithms [65, 87], ant colony optimisation [82, 89], genetic algorithms [90, 34], artificial bee colony algorithms [105], harmony search algorithms [45], flower pollination algorithms [1], integer programming, and simulated annealing algorithms [88]. Techniques applied to solve other PPPPs include: neural networks [42], black widow optimisation [98], and integer programming [52]. Computational solving techniques rarely resemble the literature on human approaches to solving, unless deliberately designed to [109].

A common approach used to solve PPPPs is constraint programming. Since Simonis modelled Sudoku as a constraint problem in 2005, [137], it has been extensively modelled as a Constraint Satisfaction Problem (CSP), [75, 84], and is regularly used a benchmark and example for CSP solvers [97, 77]. Sudoku solving using constraint programming has been explored with a range of different approaches [137, 49, 18].

Reeson *et al* used constraint programming in an early attempt to support human solvers [123]. Espasa *et al* presented DEMYSTIFY in 2021 [38], which used a constraint based approach to solve PPPPs in a human-like manner and produce human readable explanations. In the following sections we discuss constraint programming, Minimal Unsatisfiable Set (MUS), and DEMYSTIFY in detail.

```
letting D be domain int(1..9)

$The puzzle starting state
given starting_puzzle : matrix indexed by [D,D] of int(0..9)

$ The puzzle solution grid
find grid : matrix indexed by [D,D] of int(1..9)

such that

    $ Ensure that all cells that have a value in the starting state are set to that value in the solution.
    forAll row : int(1..9) .
        forAll col : int(1..9) .
            (starting_puzzle[row, col]!=0) -> (grid[row, col]=starting_puzzle[row, col]),

    $ States all digits in a row must be different
    forAll row : int(1..9) .
        allDiff(grid[row,..]),

    $ States all digits in a column must be different
    forAll col : int(1..9) .
        allDiff(grid[..,col]),

    $ States all digits in a 3x3 box must be different
    forAll i,j : int(1,4,7) .
        allDiff([grid[k,l] | k : int(i..i+2), l : int(j..j+2)])


```

Figure 2.16: A simple Sudoku model in ESSENCE [43]

2.2.2.1 Constraint Programming

Solving a problem using constraint programming consists of two stages. The first step is to build a model of the problem. The model of a CSP consists of variables, domains, and constraints. Variables are associated with a domain. The domain expresses the possible values the variable could be assigned to; it could be a set of values or a function that maps variables to the values they can take. A constraint maps possible assignments of variables within its scope to either true or false. The scope of a constraint refers to the variables to which it applies. The second stage uses a constraint solver to find a solution, where all variables have a value assigned such that all constraints resolve to true [126]. Solvers can also prove that a problem has no solution [126]. The solver is generally independent of the model, and can to a certain extent be treated as a black box. Solvers support all standard arithmetic and boolean operators, along with more domain specific operators, such as the allDiff constraint, which imposes that all variables passed to it must be different [46].

An example of Sudoku as a CSP, using ESSENCE [43], a high-level CSP modelling language is shown in Figure 2.16. Each cell is a variable with a domain containing the digits 1 to 9. They can be approached as a 9×9 matrix of variables, as shown on Line 7 of Figure 2.16 (the `find` statement is used to identify variables with unknown values that need to be assigned for a solution to be found). The starting puzzle is a 9×9 matrix of values, with 0 representing empty cells, as shown on Line 4 of Figure 2.16 (the `given` statement is used to identify inputs). The first constraint (constraints follow the `such that` statement) on Lines 12 to 14 ensures that every cell in the grid matrix where the corresponding cell in the starting_puzzle matrix is not zero (indicating an empty cell) is the same as the value in the starting_puzzle matrix.

The remaining constraints of the problem are the rules of the puzzle. The rule that each row must contain each digit exactly once can be expressed as nine allDiff [46] constraints, each applied to a row of the puzzle and shown on Lines 17 and 18 in Figure 2.16. An allDiff constraint expresses that every variable in a defined group must contain unique, distinct values [46]. The rule that each column must contain each digit exactly once can also be expressed as nine allDiff [46] constraints in a very similar manner to the row constraints and is shown on Lines 21 and 22 in Figure 2.16. The rule that each 3×3 box must contain each digit exactly once can also be expressed as nine allDiff [46] constraints applied to all cells in the 3×3 box, as shown on Lines 25 and 26 in Figure 2.16.

Boolean Satisfiability (SAT) problems can be considered a subset of constraint problems, expressed in Conjunctive Normal Form (CNF), where variable domains contain only True and False [126]. However, SAT solvers use different underlying techniques than constraint solvers [126]. Some tools can convert models to be compatible with either constraint solvers or SAT solvers, to make use of their different strengths [103].

2.2.2.2 Minimal Unsatisfiable Sets (MUSes)

MUSes are used by DEMYSTIFY [38], introduced in Section 2.2.3 when producing puzzle explanations.

Definition 2 (Unsatisfiable Set) *An* unsatisfiable set *is any unsatisfiable subset of the set of constraints of an unsatisfiable constraint problem.*

Definition 3 (Minimal Unsatisfiable Set) *An unsatisfiable set, C, is minimal if all strict subsets of C are satisfiable*

Unsatisfiable subsets can be found in unsatisfiable problems. Consider a problem consisting of the variables a, b, c, d.

Example Set 1 $\{a \wedge \neg b, a \wedge c, \neg c \vee b\}$ is a minimal unsatisfiable set as removing any of the constraints results in a satisfiable set. Table 2.1 shows the valid solutions for each strict subset of example set 1.

variable name	$\{a \wedge \neg b, a \wedge c\}$	$\{a \wedge \neg b, \neg c \vee b\}$	$\{a \wedge c, \neg c \vee b\}$
a	True	True	True
b	False	False	True
c	True	False	True

Table 2.1: Table showing the solutions for all strict subsets of Example Set 1

Example Set 2 $\{\neg d, a = d, a \vee d, a \wedge c\}$ is not a minimal unsatisfiable set as removing $a \vee d$ or $a \wedge c$ does not result in a satisfiable set. The former results in $\{\neg d, a = d, a \wedge c\}$ which cannot be satisfied as if $\neg d$ forces d to be False, $a = d$ then forces a to be False (since d is False and they must be equal), and $a \wedge c$ cannot be resolved to True as a is False. The latter (removal of $a \wedge c$) results in $\{\neg d, a = d, a \vee d\}$; $\neg d$ forces d to be False, $a = d$ then forces a to be False (since d is False and they must be equal), $a \vee d$ cannot be resolved to True as a and d are both False. Therefore, we have found two different unsatisfiable subsets; both are minimal, as if we removed any constraint from them they would be satisfiable.

2.2.3 DEMYSTIFY

DEMYSTIFY is a tool, written by Jefferson and presented by Espasa, Gent, Hoffmann, Jefferson, McIlree, and Lynch, for creating human-interpretable, step-by-step explanations of how to solve a range of PPPPs. [38] It consists of 3 parts: an extension of the constraint language ESSENCE [43], a Python library that produces step-by-step explanations, and a visualiser. [38] For this thesis, we used both the modelling language and the Python library to develop the novel hint system. We did not use the visualiser.

Espasa *et al* found the explanations were comparable to a range of guides and puzzles for Binairo, Futoshiki, Jigsaw Sudoku, Kakuro, Skyscrapers, Starbattle, Tents and Trees, Thermometers, X-Sudoku, and basic and tough Sudoku techniques [39]. The results for Sudoku diabolical techniques were less impressive, although this was proposed to be due to the exceptional complexity and size of very complex (and rarely used) Sudoku solving techniques, such as the Death Blossom. DEMYSTIFY did find similar, sometimes simpler, solving steps; but they were only accepted if they matched exactly [39].

The explanations were produced by creating an unsatisfiable problem by setting a cell to an incorrect value and then finding MUSes. This is done using a novel MUS finding algorithm, as existing tools to find MUSes were found to not produce suitable MUSes for explaining puzzles [39].

The size of the explanation sets was found to be heavily influenced by the design of the model used [39]. Espasa *et al* discuss the development of their Sudoku model in detail [38], and the final result is shown in Figure 4.11. This is discussed further, along with the models developed for this study, in Section 4.3.1 . There has been extensive research into how to write efficient constraint models, and it is not surprising that modelling puzzles to provide explanations has different requirements, which require careful development and further research.

2.3 Generating PPPPs

PPPPs require instances (levels) to be played, and for popular puzzles they need huge numbers of instances. These are either designed by a person or generated by a computer program. Human designers are generally seen to produce more rewarding and interesting games, while computer generation allows large numbers of levels to be produced cheaply and easily. Sudoku became ubiquitous outside of Japan only after Wayne Gould invented a computer program to automatically generate levels and successfully sold it to *The Times* [50, 137, 33]. However, Nikoli, one of the most popular publishers of Sudoku (supplier of Puzzler in the United Kingdom) and similar puzzles only sells human designed levels, and many other books and magazines advertise based on their 'handcrafted' puzzles [70, 149, 158].

Due to the popularity and success of Sudoku, the preponderance of research focuses on Sudoku. Therefore, as above, this section mostly refers to research on

Sudoku generation.

2.3.1 Computational PPPP Level Generation and Difficulty Grading

Sudoku puzzles have been generated using a range of algorithms, including hill climbing algorithms [40] and removal generating algorithms[86, 60].

Creating a PPPP level of specified difficulty is very challenging. Therefore, in general, computationally generated PPPPs are created, then a difficulty rating is assigned [40]. This grading is still sometimes done by a human puzzle solver, playing the level and rating it, however, this adds a human back into the loop and slows down content creation. The difficulty of a computationally generated PPPP can be calculated using a range of criteria. The majority of the research has focused on Sudoku; therefore, we focus our discussion on Sudoku grading research.

The simplest approaches to grading Sudoku involve counting the number of empty cells [86]. This approach has been shown to be a poor predictor of difficulty[68], as the complexity of the required deductions is often unrelated to the number of empty cells. More sophisticated approaches examine the strategies required to solve the puzzle and are sometimes extended to combine the difficulty of the deduction and the number of opportunities to apply it [108, 20, 142, 8, 93]. This is achieved by developing an AI player and using it to solve the puzzle. At each step, the next easiest deduction to make is found, and the model randomly chooses from all available opportunities to apply that deduction [68, 107]. The number of steps taken to solve the puzzle can be counted and mapped to a difficulty grading [68]. In more sophisticated models, each step is weighted by a heuristic based on the difficulty of each deduction [108]. This approach can be extended to combine the difficulty of the deduction and the number of opportunities to apply it, meaning that puzzles with lots of different ways to progress are considered easier than puzzles with fewer paths through the puzzle[142].

Assessing the effectiveness and validity of these models is a separate challenge. Many of the models are not validated against player data[20, 61], instead they are validated against difficulty ratings of published puzzles, which are often computationally calculated. It is hard to justify that a model is a good reflection of player experience and action, without comparing it against player generated data.

There have been studies that compare against player data. Stuart compared the time players took to finish the puzzle. This provided a rough match to the model, although a large (many empty cells) very easy puzzle can take the same length of time as a small (few empty) cells hard puzzle. Pelánek compared their model against the order in which players filled in the cells [108]. This could be an excellent measure of validity as long as the outliers are carefully investigated. PPPPs have a certain amount of inherent ordering due to the current information available in the puzzle, which may result in a high base level of matching.

All of the algorithms discussed above are very specific and carefully tailored to Sudoku. The models cannot be applied directly to other PPPPs or Sudoku variants.

2.3.2 Human PPPP Level Generation

Human designers approach puzzle design very differently from current computational approaches. Puzzles are not built, then graded; instead, the target audience is considered throughout the design process. The goal is not simply to produce a valid Sudoku; designers consider the narrative of the puzzle (what paths can the player take through the puzzle?) along with the visual appeal of the puzzle [134]. Nobuhiko Kanamoto's (chief editor of Nikoli; a major publisher of PPPPs), main criticism of computer generated puzzle games is the poor narrative path and the loss of a sense of communication between designer and player [70].

Human designers also often take into account the visual appeal of the puzzle. Snyder suggests starting to build a puzzle by choosing a pattern of clues; the pattern is then used as the basis for filling in values that facilitate the desired narrative [134]. Nikoli's Sudoku puzzle designers arrange the clues in a symmetrical pattern [70]. Designers may also select clues on visual/conceptual appeal, for example, by arranging a puzzle so that the digits have to be solved in order (e.g. all the 1s, then all the 2s) [134].

2.4 Providing Assistance

As discussed in Section 2.1, games are used extensively for leisure and increasingly for educational or serious purposes. Players find games most enjoyable when they present an appropriate level of challenge [73]; if a game is too easy, players become bored and stop playing [72], and if a game is too hard, players become frustrated by their lack of progress and give up [7].

2.4.1 Understanding Player Experience

Understanding the player experience of a game is an active and extensive research area. It is of both commercial and academic interest to understand what players react positively to and what causes negativity.

What is fun and how do we measure it? Koster, [73], believes that the key source of fun in a game comes from learning the patterns of play; the game ceases to be fun once all the patterns have been learnt or if it is too hard to find the patterns. He also discusses the importance of the game styling around the central mechanic [73]. PPPPs superficially appear to have very little window dressing overlaying the central mechanics; however, there is no reason for Sudoku to use numbers, and there are variants that use colours [3] or emojis [14] instead of the digits 1-9. There are variants with different symbols and styling for most PPPPs.

2.4.2 Approaches to Assistance in Games

There has been a range of research on providing assistance to video game players, ranging from games where the primary goal is entertainment to serious games with clear educational goals. The types of assistance provided to players cover a range of approaches, from tutorials, manuals, in-game assistance items, to in-game advice or hints. All of them attempt to improve player engagement and enjoyment, however, they fulfil different roles.

Tutorials are often the first thing that a player encounters and help the player understand the game [156]. Tutorials are, normally, not trying to explain every aspect of the game; instead, they aim to introduce enough to allow the player to easily engage with the game and learn the rest while playing [156]. In a PPPP, the tutorial would be the set of rules of the game, possibly combined with a very simple example of a first step. More detailed information about the game may be conveyed through a manual. Physical manuals are now very rare, although they used to be very common, [154, 62]. Games often still include a manual but it is digital and can be accessed through the game. Manuals provide detailed information that a player might struggle to glean from the game. Other games depend on Wikis, often made by users, to provide that information. Both tutorials and games provide assistance to the player by clarifying the rules and mechanics of the game [156].

Assistance items are a mechanic within a game that makes it easier, possibly for a short time, if the player is struggling. For example, players in the game MarioKart that fall behind are provided with items that allow them to quickly catch up or slow down the other players, allowing them to catch up [19]. PPPPs on paper do not use assistance items. Some digital implementations track how long the player has been stuck and after a while allow them the option of a hint [157]; however, we classify this as a subset of hint mechanics rather than assistance items.

Hints[3] or advice provide the player with guidance on their actions. Hints are more common in games with a puzzle element, although a range of games contain an advice element, for example, reminding a player of a fighting mechanic that is needed for a particular battle [112].

The focus of this thesis is how to provide meaningful hints, using effective feedback guidelines, to allow players to progress when they get stuck playing PPPPs.

2.4.2.1 Hint Systems in Games

Hints are received by players with mixed results. Players sometimes feel robbed if they are told a piece of information that they have already worked out. Wauck *et al*'s study [155] found that players seemed to prefer to choose when they wanted a hint, rather being automatically provided with hints, whether adaptively based on their playing style or every X seconds. Wauck *et al* also found that players with hints available on demand used less hints than they were presented with in either the adaptive or automatic condition [155]. Focus groups reported to Big Fish Games (a major publisher of casual games) that they felt that they were being punished if they needed a hint [25]. Big Fish Games changed the name from hint to advice, but it is not clear what the impact of this was [25].

In some cases, hints have been found to reduce engagement with a game and increase churn[4]. However, Sun *et al* propose that as long as the game is sufficiently entertaining, players will avoid using help systems excessively [146]. This could be seen as inconsistent with behaviours discussed by Koster [73], where players will choose the easiest, even if tedious, way to 'win' a game (for example, repeating an easy encounter 200 times to earn enough experience to progress, rather than a challenging encounter once or twice). However, players performing the latter

---

[3] Some descriptions of PPPPs use 'hint' for elements present in the starting state of the puzzle, such as the cells containing digits at the start of Sudoku, or the numbers at the ends of rows and columns in tents and trees. We use 'clue' for this purpose and exclusively use 'hint' to mean additional guidance provided to the player beyond the initial puzzle state.

[4] The rate at which users stop engaging with a product.

generally still see it as ‘playing the game’ whereas hints are often seen as ‘giving in’ or ‘cheating’.

However, hints, when done well, have been found to reduce player frustration, avoid players becoming ‘stuck’ and increase player enjoyment [53]. Hints have also been found to improve engagement and learning in user interactions outside of games, [130]. Schnepp and Rodgers found that incorporating a hint system into an online assessment tool improved learning outcomes over the course of a semester [132].

CHAPTER THREE

EXPLORATION OF SUDOKU

The research presented in this chapter was published in 2022 by Lynch, Jefferson, and Hinrichs [85].

Review of the literature revealed that there were numerous assumptions about the experience of people playing Sudoku. Furthermore, there were significant differences between how the AI community expected people to solve Sudoku and how people outside the AI community expected people to solve Sudoku.

In order to gain insight into people's problem-solving strategies and experiences while solving Sudoku puzzles, we conducted a short survey and an in-person study.

3.1 Sudoku Survey

In order to prepare for the in-person study of Sudoku solving, we conducted a small survey.

The purpose of the survey was to collect general information on people's Sudoku experiences and guide the design of the Sudoku solving study; see Section 3.2. The survey was not intended to collect in-depth or detailed information.

The survey focused on where people played Sudoku, what factors made a Sudoku fun, their educational background (with special attention to areas that might impact people's approaches to Sudoku, such as AI), and, in order to facilitate future

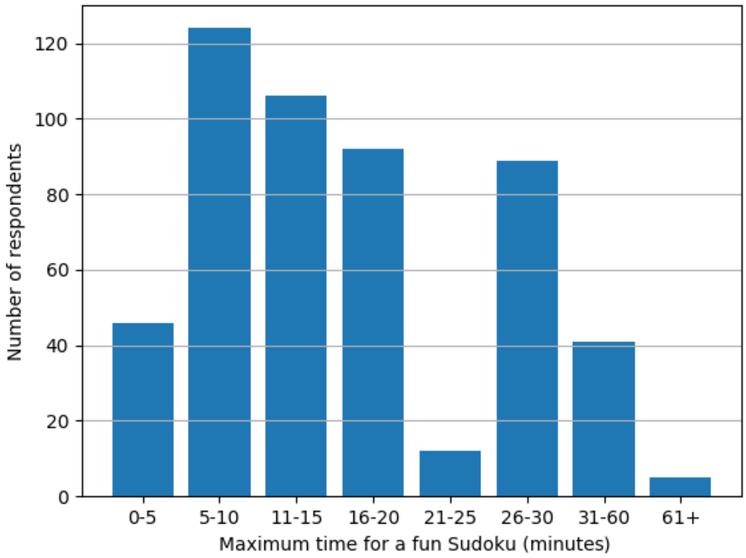


Figure 3.1: Histogram of survey results indicating maximum time that a "fun" Sudoku should take

research, what other puzzle games they enjoyed. We also asked the respondents if they played Sudoku competitively and if so what form they took. At the end of the survey, participants were given the option to leave an email address which would be used to invite them to participate in the subsequent Sudoku solving study. The full list of questions can be found in Appendix A.1.

The survey was distributed throughout the university, through social media, and through Sudoku-focused forums. It had a far greater response than expected; with 866 respondents, 681 completed the entire survey.

3.1.1 Survey Results

We excluded all incomplete responses and one complete but abusive response from the results.

Free text answers to 'how long should a fun Sudoku be?' were analysed and the result was grouped into time bins, as shown in Figure 3.1. Of the respondents 70% felt that the maximum time that a "fun" Sudoku should take was 20 minutes or less. We found that when solving Sudoku, 115 respondents only used digital tools, 187 had experience using both digital and paper tools, 204 only used paper, and 1 used neither (15 survey respondents had not played Sudoku, so were not presented

Newspapers/Magazines (paper)	158
Books (paper)	344
Apps (digital)	248
Websites (digital)	103
Other (paper)	3
Other (digital)	2
Other (Board game, game show)	3

Table 3.1: Number of respondents that play Sudokus via different media (they could select multiple options)

with this question). Therefore, 391 of the 522 respondents still solve Sudokus using paper and pen (e.g. in newspapers, magazines, and desk calendars; see Table 3.1). 122 of the respondents (24%) reported researching solving techniques.

3.2 Sudoku Solving Study

The following section discusses the motivation and design of a qualitative study exploring how people solve Sudoku.

3.2.1 Motivation for the In-Person Sudoku Solving Study

Our literature review highlighted that existing computational models of difficulty classification systems have been validated primarily against each other and/or the time taken to solve rather than the user experience, and are based on assumptions about the processes and experience of the players [64, 61]. The most sophisticated computational models of difficulty all rely on two key assumptions: that at every step players randomly choose one of the easiest available moves and that the definition of the easiest move is consistent amongst players. However, our survey found that most players (76% of the respondents) do not research Sudoku solving techniques, and therefore the play strategies they develop may not match the published approach, as defined by online tutorials and books on solving. Their perception of the easiest move is also more likely based on their own experience and therefore may also not reflect the "correct" approach as defined by online tutorials and books on solving. Furthermore, the definition of the easiest move is not always consistent between the guides [144, 145, 141]. For example, Sudoku Dragon's difficulty ordering was: the pointing pair technique (which they refer to as subgroup exclusion), followed by the hidden pair (referred to as hidden twin), and then naked pair (referred to as naked twin)[144]. In contrast, Stuart's ordering (which we use in Section 2.1.2) has the opposite ordering: naked pair, hidden pair, and then pointing pairs [141].

To understand how robust the assumptions underlying creation, difficulty classification, and help systems are, we needed to explore both the processes people use and the experiences people have when solving Sudoku.

We therefore designed an exploratory, qualitative study with the following question in mind:

How accurate are the following assumptions about Sudoku players, which are used as the foundation of the current best Sudoku models?

- Players solve puzzles using the techniques commonly described in the literature.
- The only annotation that players use is to list the allowable candidates for each cell.
- Players solve puzzles by repeatedly choosing the move of easiest difficulty.
- Players do not make mistakes when solving puzzles.

3.2.2 Participants

Participants for our study were recruited directly from the pool of participants who responded to the online survey and university members who expressed interest (3 participants were recruited this way). All participants were required to complete the online survey prior to participating in the study. We recruited a total of 31 participants (20 female; 11 male) with ages between 18 and 74 years, with a median age bracket of 25-34, see Figure 3.2.

All participants were required to participate in person and were compensated with a £5 book token (for a local bookshop) per session (a total of £10 vouchers if both sessions were attended).

The educational background of our participants may not be representative; 27/31 had completed an undergraduate degree. We did not require the participants to inform us if they were a member of the university.

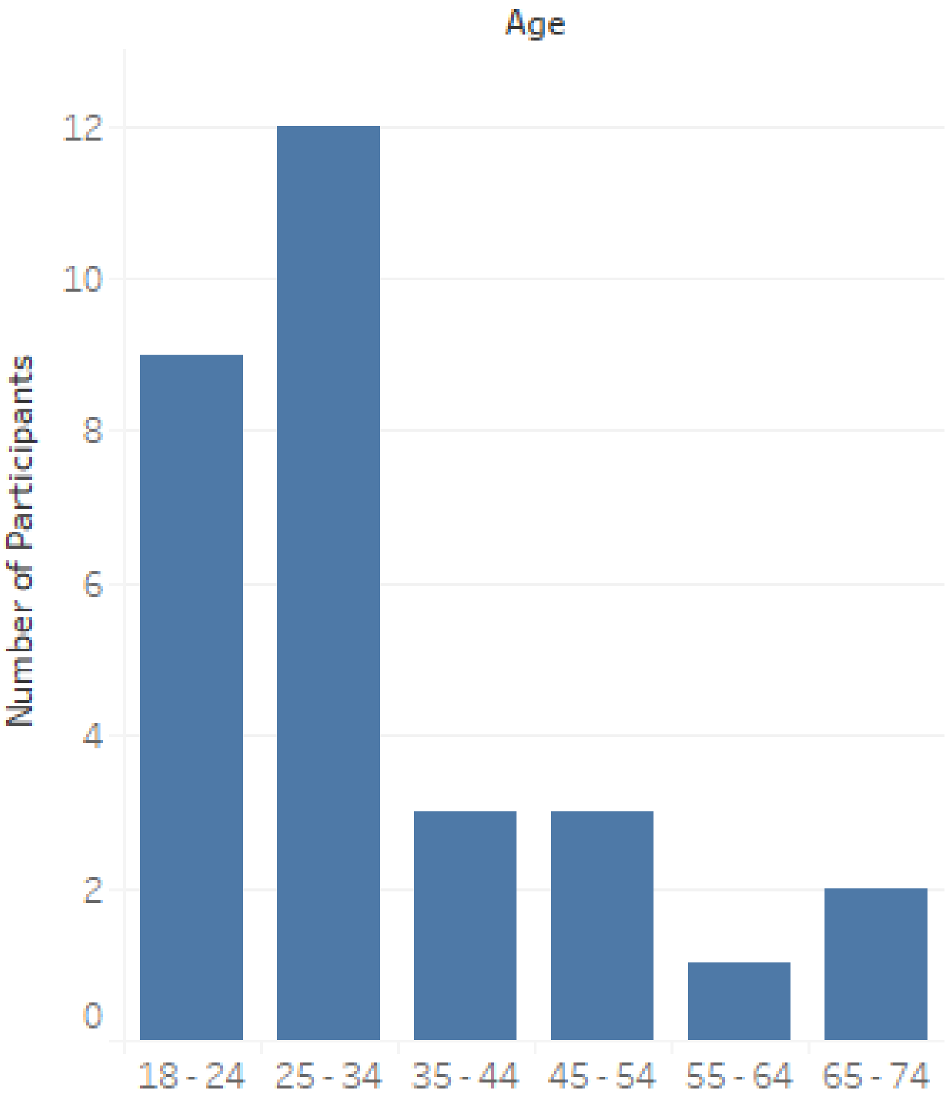


Figure 3.2: Age distribution of Sudoku solving study participants

We had more than 31 responses and chose respondents that, when combined, provided the best range of expertise levels. Unfortunately, there were very few respondents at the advanced and novice levels. Of the participants who had previous solving experience (29), six had only used digital tools prior to the study.

Participants were asked to rate their Sudoku expertise at the beginning of the study session, as there was a gap of several months between the survey being conducted and their participation in the study; we used this rating during data analysis. The expertise of our participants, as taken in the pre-study questionnaire for session 1, was 2 complete novices, 7 beginner players, 17 intermediate players and 5 advanced players, see Figure 3.3.

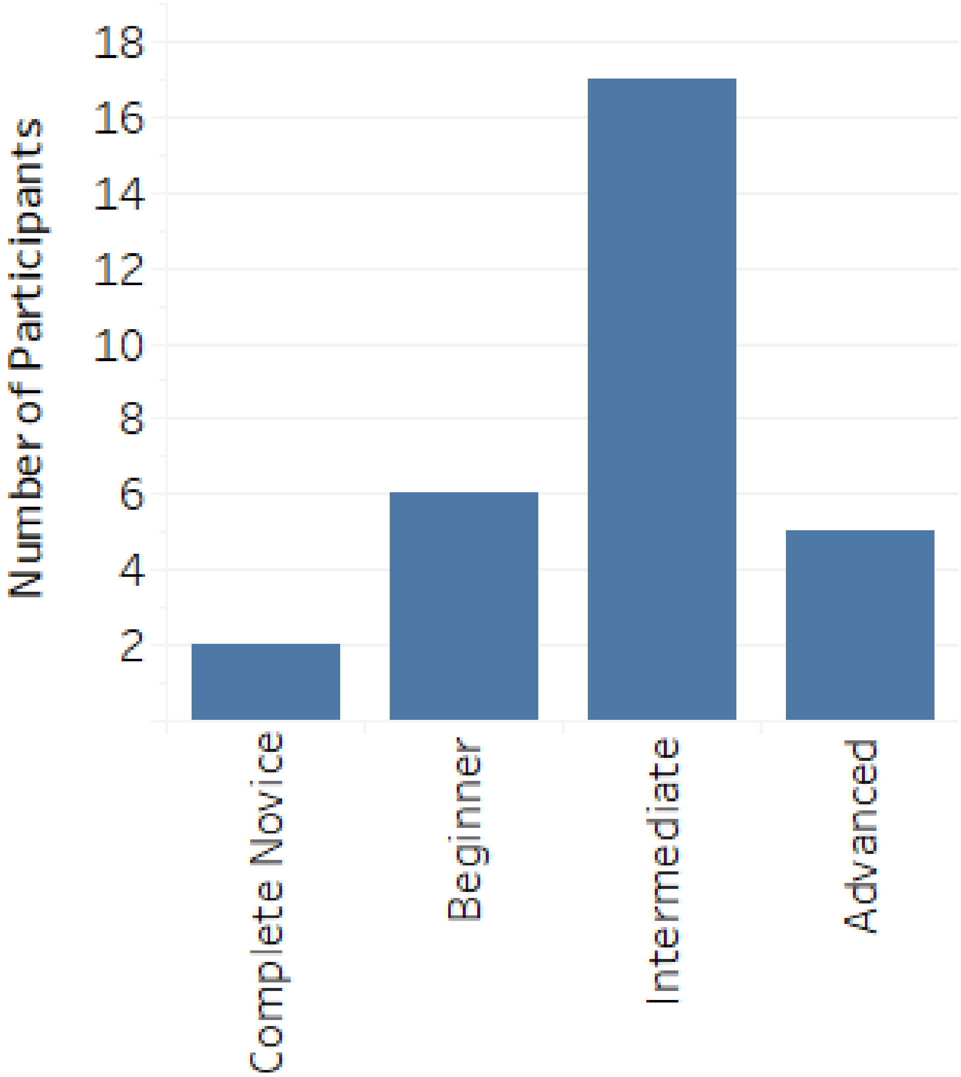


Figure 3.3: Expertise level distribution of participants, as taken in the Session 1 pre-study questionnaire

3.2.3 Study Design & Procedure

The study consisted of two sessions, both approximately one hour and fifteen minutes. The study took place in a private office; each participant had individual sessions during which they and the primary researcher were the only people present. The sessions were video and audio recorded, as discussed in Section 3.2.4.3.

Each study session (both 1 & 2) started with participants filling out a pre-session questionnaire. The questionnaire for session 1, available in Appendix A.2.1 asked participants to rate their expertise (on a scale of “completely new” to Sudoku, beginner, intermediate, advanced, or very advanced), their experience at solving Sudokus (on a scale of having solved no Sudokus, less than 10, between 10 and 150, between 150 and 500, or greater than 500), how recently they had solved a Sudoku (today, last week, last fortnight, last month, in the last six months, in the last year and more than a year ago) and if they had prepared for the study. The questionnaire for session 2, available in Appendix A.2.2 was similar: They were asked to rate their expertise again (in case they had reconsidered their expertise based on their experience in the first session), how many Sudokus they had solved since the first session (on a scale of no Sudokus, less than 10, between 10 and 30, between 30 and 50, or over 50) and how recently they had solved a Sudoku (today, yesterday, 3 days ago, more than 3 days ago). They were also asked if they had researched Sudoku since session 1. The aim of the pre-session questionnaires was to provide an overview of a participant's Sudoku solving experience.

In both sessions, the participants were then given the first puzzle (and a spare copy in case they wanted to start over) to solve. The experimenter briefly explained the Sudoku rules again and then left the participant alone to complete the puzzle. When the participant finished the task or wished to give up, they alerted the experimenter. The experimenter waited outside the office while they were solving, to avoid any additional stress or self-consciousness on the part of the participant. They were then asked to fill in a questionnaire which asked them to describe (or name) the solving techniques they had used, to provide ratings of how challenging they found the Sudoku (based on a 10-point scale, where 1 indicates no challenge and 10 indicates much too challenging) and how enjoyable and frustrating they found it (using 7-point Likert scales where 1 indicated no enjoyment/frustration and 7 indicated extreme enjoyment/frustration). They were also given the option to provide further comment on their Sudoku solving experience. This questionnaire is available in Appendix A.2.3. The experimenter then provided a new puzzle and left the room again.

In both sessions, participants were provided with Sudokus for the duration of the allotted hour, at which point participants were given the option to continue with the current puzzle. After 1.5 hours the study session was ended even if the participant had not finished. If they had not finished, they were given the option of taking a photocopy of the puzzle home. This was offered to avoid an issue that arose during the pilot study in which a participant refused to stop solving. Once they had finished the solving element of the study, a short, semi-structured interview was conducted. We chose to have an interview after task completion over a think-aloud study due to concerns about the additional cognitive load the think-aloud study imposes interfering with the solving process, particularly since the impact on working memory could bias participants towards a more note-based approach than they would otherwise employ [16, 57, 153].

The interviews asked participants to discuss their approach to solving the puzzles, discuss the puzzles they enjoyed the most, and those that frustrated them the most. They were then asked to explain any notation they used when solving the Sudoku and if they had come up with any approaches they had not used previously. The interview guidelines used are included in Appendix A.2.4.

We did not impose a time limit on individual Sudoku puzzles. Sudokus were provided to the participants on sheets of A4 paper. Paper was used instead of a digital tool, both due to the findings of the survey that most respondents solved on paper and to avoid limiting the participants' potential approaches. A digital interface, by necessity, makes assumptions about how participants will interact with the puzzle. Participants were provided with pencils, pens, coloured pens and pencils, a rubber and a pencil sharpener. This ensured that participants had the necessary stationery to use any methods with which they felt comfortable.

3.2.3.1 Session 1.

The primary goal of this session was to gain insight into the participant's overarching approach to solving Sudoku and the challenges they face. Therefore, participants were asked to completely solve the Sudokus provided. The Sudokus that participants were asked to solve were selected based on how participants rated the previous Sudoku they had solved. This approach was chosen due to the wide range of skill levels among the participants. Beginner players are unable to solve Sudokus that require advanced techniques, so giving such players several difficult puzzles would provide no useful information and be extremely frustrating for the participants. This calibration approach was designed to reduce the likelihood that participants were presented with Sudokus that were too easy or too challenging.

3.2.3.2 Session 2.

The primary goal of this session was to examine whether participants used the technique the literature expected to reach the next state of the Sudoku. It also explored the impact on the participant’s enjoyment, frustration, and challenge of the ’easiest’ next step requiring a particular technique.

The Sudokus were designed to require a particular technique (based on Stuart’s difficulty ordering and solver [141]). The design ensured that the Sudoku couldn’t be solved using any techniques that Stuart rated as easier than the technique being tested. Participants were asked to complete only the part of the Sudoku that required the technique being tested (as described in Section 3.2.4.2). This allowed more puzzles to be completed in the available time, allowing more techniques to be checked. The techniques were not described to the participants or named, the participants were not aware which technique was being tested by each puzzle.

All participants had the puzzles presented one at a time, in the order listed in Table 3.3. In session 2 each puzzle set consisted of an isomorphic pair of Sudokus[1], of which the participants were asked to partially complete one. Some puzzles asked participants to complete a particular cell, while others asked them to complete a given number of cells. These isomorphic puzzles were intended to explore whether there was a difference in difficulty between isomorphic problems. The participants were free to choose which half of each pair they wished to solve and were only expected to solve one of the two Sudokus presented to them in each set of puzzles.

3.2.4 Design of Study Sudokus

In this section, we discuss the decisions made when designing Sudokus for the study.

3.2.4.1 Session 1 Sudoku Design

Session 1 needed to provide participants with Sudokus of appropriate difficulty, therefore, it required Sudokus of various difficulties. The Sudokus also needed to be solvable in a timely fashion to allow participants to solve several during the session. To facilitate this, each Sudoku was built (either from scratch or by substantially adapting an existing puzzle) and then tested, initially using the online solver at SudokuWiki [141], in order to establish the minimum solve needed for

---

[1] the puzzles were identified with a letter and are described in Table 3.3, the isomorphic versions were identified by the numbers 1 & 2, e.g. A1 and A2 were the two puzzles presented to the participants in puzzle set A

each Sudoku, using the solver's built-in difficulty ordering (which is consistent with Stuart's in the Logic of Sudoku [140]).

We created four difficulty classes based on the techniques required to solve the chosen Sudokus. We based these classes on the expected order of difficulty of techniques as described in the literature discussed in Section 2.1.2. Naked Singles are considered the simplest technique, so Sudokus that only required Naked Singles were placed into the easiest difficulty class ("Very Easy"). Hidden Singles are considered slightly more challenging, so we placed Sudokus that required Hidden Singles as the hardest technique in their minimum solve [2] in the next difficulty class ("Easy"). Puzzles that required simple subset-based techniques (see Section 2.1.2.2) as the most challenging technique in their minimum solve were in the second most challenging difficulty class ("Medium"). All Session 1 puzzles requiring techniques considered more advanced than simple subset-based techniques were placed in the hardest difficulty class ("Hard").

The researchers then tested the time required to solve the Sudokus by solving them and measuring the completion time. The goal was to achieve a selection of Sudokus that required a range of techniques and allowed participants to complete approximately 2-4 Sudokus in the available time. 14 Sudokus were selected for session 1 and are listed in Table 3.2.

3.2.4.2 Session 2 Sudoku Design

Session 2 required Sudokus where the minimum solve for the next step required a particular technique and a way to highlight to the participant when they had completed the required stages, without over-influencing their process.

Two approaches were proposed: one was to ask participants to solve the Sudoku until they could complete a particular highlighted cell (CPC). The alternative was to ask the participants to solve the Sudoku until they could complete a given number (x) of cells (CXC).

The first approach (CPC) risked leading participants to focus on a particular cell. This would be advantageous as it avoids participants spending too much time on search; however, it could also heavily influence their entry point into the puzzle. To avoid the latter problem, all puzzles that used this approach required at least

---

[2] Minimum solve is the number of steps and the set of required techniques to solve a puzzle, if at every step the easiest technique (based on a given difficulty ordering, in this case Stuart's ordering) that can be used to progress with the puzzle is employed.

ID	Level	Empty Cells	Techniques Required (Min. Solve)
R	Very Easy	23	Naked Singles
D	Very Easy	36	Naked Singles
Q	Very Easy	19	Naked Singles
K	Very Easy	28	Naked Singles
F	Very Easy	41	Naked Singles
V	Easy	24	Hidden & Naked Singles
Y	Easy	24	Hidden & Naked Singles
All Sudokus from here include Hidden and Naked Singles			
M	Medium	27	1 Naked Pair
X	Medium	27	2 Naked Pairs
T	Medium	32	1 Naked Triple
S	Medium	38	1 Naked Triple
Z	Hard	32	1 X-Wing, 1 Naked Pair
L	Hard	22	1 Simple Colouring, 2 Naked Pairs
AA	Hard	33	1 Simple Colouring, 1 Swordfish

Table 3.2: Puzzles included in Session 1, with a difficulty class based on existing literature, the number of empty cells in the puzzle and the techniques required for a minimum solve of the puzzle.

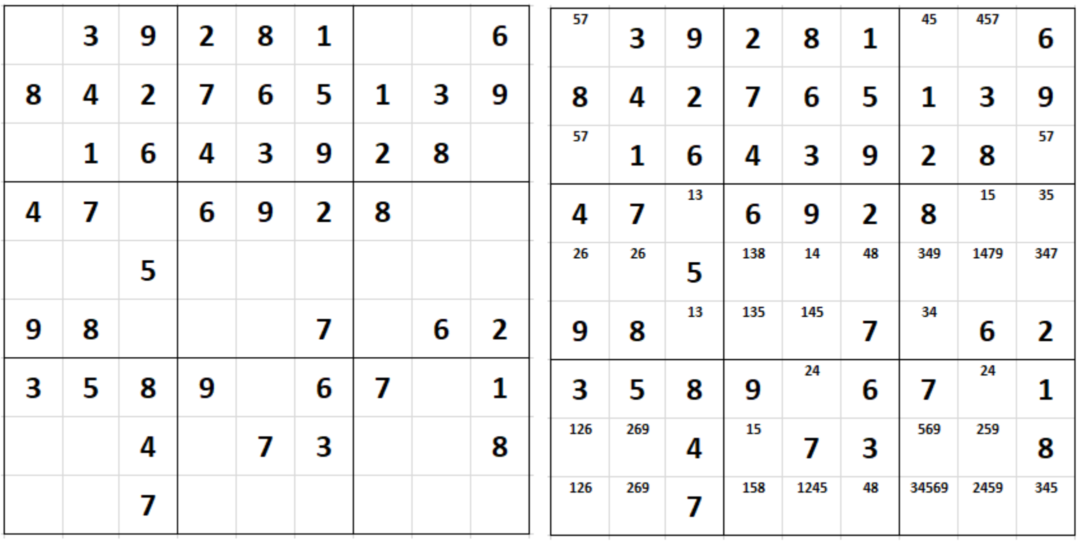
ID	Techniques being tested
A	Required the participant to complete a row/column missing with one empty square, followed by a Naked Single. (Complete Particular Cell (CPC))
B	Required a Naked Single, then another Naked Single (CPC)
U	Required two Hidden Singles (Complete X Cells (CXC))
C	Requires a Naked Pair, then a Hidden Single (CXC)
O	Requires a Hidden Pair, then a Hidden Single (CPC)
E	Requires an X-wing, then a Naked Single (CXC)
P	Requires a Unique Rectangle, then a Hidden Single (CXC)
W	Requires simple colouring (CXC)

Table 3.3: Puzzle types included in Session 2, with letter IDs, in the order they were presented to participants. (CPC: complete particular cell, CXC: complete X cells)

one additional step, beyond the required technique, in the minimum solve of the indicated cell.

The second approach did not risk guiding the participant as heavily, but made it more likely that they might find an alternative route through the Sudoku (using a more challenging technique, under the chosen difficulty ordering) and avoid using the desired technique entirely. We were unable to find any software that checks all possible routes through a Sudoku.

We decided to use a mix of these two approaches. The final Sudokus chosen are listed in Table Table 3.3. For puzzles E, P & W, which required techniques that require systematic candidate filling (see Table 3.3), together with the standard copy, we gave the participants a copy with the candidates filled in, but without all impossible candidates eliminated, as shown Figure 3.4. This was done in case it helped participants by reducing the mechanical effort involved in finding the solution, allowing participants to focus on the deduction element.



	3	9	2	8	1			6
8	4	2	7	6	5	1	3	9
	1	6	4	3	9	2	8	
4	7		6	9	2	8		
		5						
9	8				7		6	2
3	5	8	9		6	7		1
		4		7	3			8
		7						

57	3	9	2	8	1	45	457	6
8	4	2	7	6	5	1	3	9
57	1	6	4	3	9	2	8	57
4	7	13	6	9	2	8	15	35
26	26	5	138	14	48	349	1479	347
9	8	13	135	145	7	34	6	2
3	5	8	9	24	6	7	24	1
126	269	4	15	7	3	569	259	8
126	269	7	158	1245	48	34569	2459	345

Figure 3.4: Example of a Sudoku with no candidates filled in (left) and candidates filled in, but without all impossible candidates eliminated (right)

3.2.4.3 Data Collection and Analysis

The video and audio data captured the participants from two different angles, shown in Figure 3.5. The video cameras were arranged to capture participants' Sudoku sheets and all annotation activities in detail.

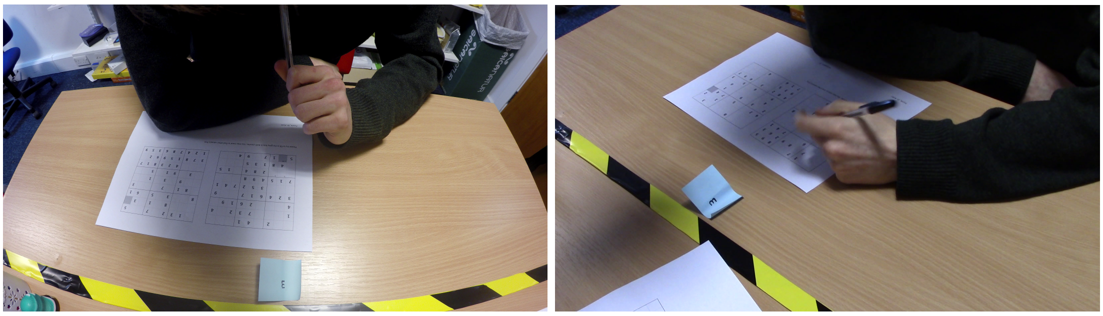


Figure 3.5: Primary (Left) and Secondary (Right) Camera Angle

Participants received paper questionnaires and their responses were transcribed. Interviews were conducted after each Sudoku solving session. The interviews were recorded with video-audio equipment to capture the participants' descriptions in detail and any gestures used to illustrate their points. The video data during puzzle solving were cropped and rotated prior to analysis to allow the participants' progress to be more easily observed, and the video taken during the short interview was just rotated.

We conducted a broad-stroke analysis of all participants' videos. This involved going through the videos, noting time taken per puzzle, the order in which the Sudoku was filled in, whether or not they used annotation, and mistakes when completing cells. We used an open-coding approach, in which the concepts of the coding scheme were iteratively developed during the analysis and then finalised and categorised among the 3 researchers [129].

We then selected 9 participants whose video we analysed in greater detail. These participants were chosen to provide a broad representative sample of the different techniques we observed in our earlier coarse-grained analysis. 3 participants from each self-selected competence level (beginner, intermediate, advanced) were analysed in greater detail. [129, 91, 54].

This more in-depth qualitative analysis conducted on these 9 participants' puzzle solving focused on the finer details of a participants' process: the type of annotation used, techniques employed while solving, and investigation into why they made mistakes and what type of mistake they made.

The analysis of the results of the questionnaire provided an overview of the experience of the participants. Analysis included plotting the results of the Likert scale to extract the median responses and coding the textual responses.

The interviews were analysed using a similar open coding approach, in which the concepts of the coding scheme were iteratively developed during the analysis and then finalised and categorised among the 3 researchers. The interviews did not provide significant insight beyond the questionnaire free text boxes. Frequently the questionnaires provided greater insight as they were conducted immediately after each puzzle was solved.

3.3 Findings of In-Person Sudoku Solving Study

In this section, we discuss our observations of the interactions of participants with the study tasks. We first focus on the participants' process when solving Sudokus and then look at the types of errors made by participants. Finally, we discuss the participants' perception of challenge. In the next section, we will discuss the implications of these findings.

3.3.1 Processes of Solving Sudokus

We found that annotation approaches varied widely between participants. Although many participants noted potential candidates in some cells, only 1/9 of the participants we coded in detail were mostly systematic about writing down all potential candidates in every empty cell. No participant always systematically removed candidates at every stage as they progressed through the puzzle and started completing cells.

3.3.1.1 Notation

We found that most participants (29/31) made some kind of annotation on at least some of their puzzles, as shown in Figure 3.6; interestingly, this trend was inverted for puzzle U [3] where very few participants used annotation. We observed a range of approaches to annotation and categorised them into the following:

Systematic Candidate Filling. This is the approach commonly described in the literature, systematically filling in all potential candidates in all cells in the grid, and systematically checking after every step if any could be eliminated. This approach was rarely used by the participants. Interestingly, 5 participants,

---

[3] The puzzle IDs are defined in Table 3.2 and Table 3.3

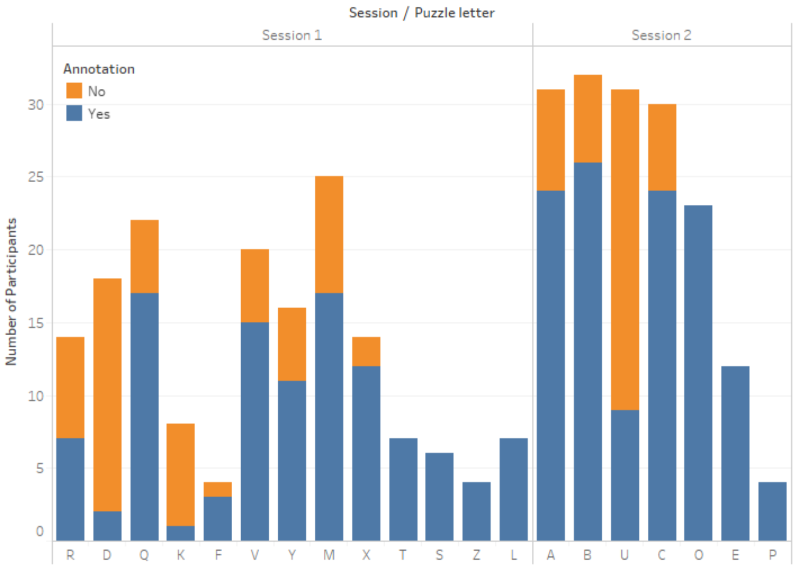


Figure 3.6: Use of annotation by participants (NB: the B, U, and C puzzle pairs had at least one participant do both of them and annotate one and not the other)

notably participant 14, believed they were using this approach despite it not being consistently observed in their solving.

Local Candidate Filling. An approach where all possible candidates in all the cells in a particular row, column, or box were filled in, as shown in Figure 3.7. This was rarely described by participants but was frequently observed in the video data.

As & When Digit filling. An approach in which the participant made notes of a particular digit without a system and without completing all potential candidates in a cell. Similarly to local candidate filling, this approach was rarely described but was frequently observed. Participant 10 acknowledged this approach, describing it as “...it goes with stream of thoughts, because I don’t want to be selective”.

Small Set Candidate Filling. An approach in which the participant only made note of a particular candidate’s possible positions in a row/column/box if there are only two possible positions (as shown in Figure 3.8). 7 participants explicitly described this process, with participant 9 describing this approach as “you’d have a look to see if there’s any way with only two numbers, so then you’d put in what they were, so that’s a 1 and 5, so you’d write in a little one and a little

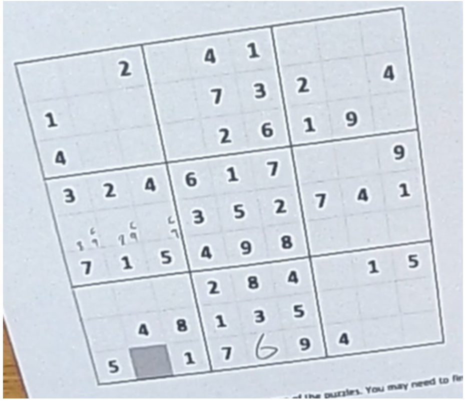


Figure 3.7: Example of local filling, the participant has filled in candidates for all cells in row 5, though nowhere else

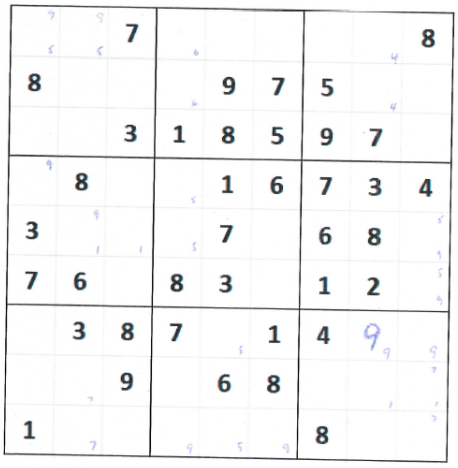


Figure 3.8: Example of participant using Small Set Notation - they have only noted candidates when they appeared twice in a dimension

five". Participant 32 [4] clarified that they preferred this notation to filling in all possibilities, as "otherwise it gets very messy and just confuses me".

Dimension Candidate Notation. This is an approach that differs from the approaches discussed above in that rather than noting candidates for a cell, the participant notes all candidates for a dimension, as shown in Figure 3.9. For example, participant 16 described their process as "I go through each rule, normally bottom to top and then I'll go, is there a one there [in the row] and write it at the side".

Highlighted Cell focused Notation. Shown in Figure 3.11, this phenomenon is the result of the experiment design and is not generalisable.

Other Approaches. Some participants used unique alternatives or extensions to the above that were not categorised (an example is shown in Figure 3.10).

Some participants eventually filled all cells with candidates, but they filled in clumps, completing a single cell, row, column, or 3x3 box. These participants tried to complete cells in between flurries of candidate filling, often failing to remove any candidates rendered impossible by the new digits completed. Most of the participants filled in the candidates in some sections and ignored them elsewhere.

---

[4] Participant IDs start at 7, 1-6 were used in the initial development stages

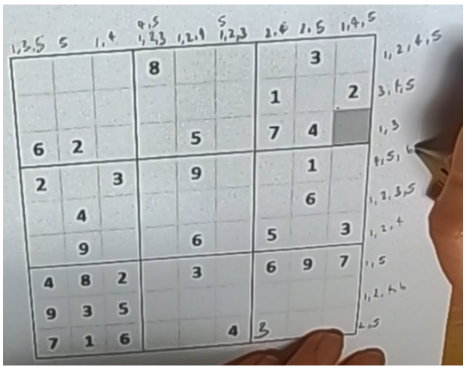


Figure 3.9: Example of Dimension Candidate Notation. The participant is writing the candidates missing from each row and column, noting all occurrences of each digit in order

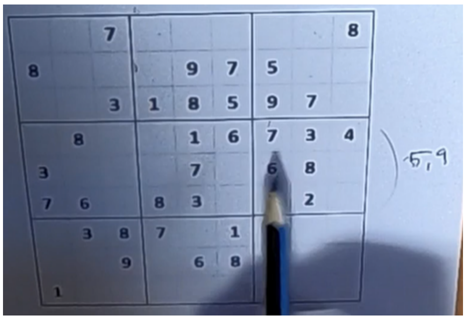


Figure 3.10: Example of unusual notation used by one participant to indicate Naked Pairs.

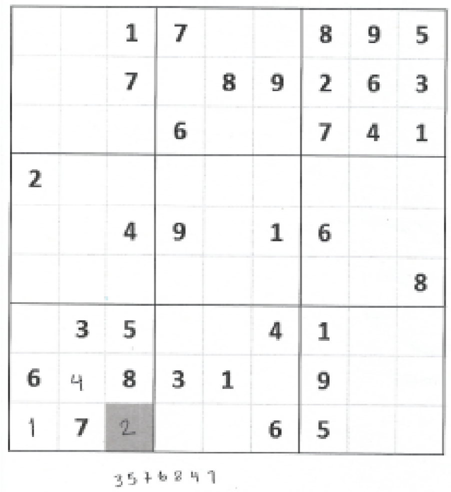


Figure 3.11: Example of Grey Box notation

The participants used different levels of annotation on different puzzles. The need for annotation appeared to affect the solving experience; participant 14 stated "I find it less fun if I have to write down the smaller numbers".

Participant 18 was the only participant to not use annotations in either session. They successfully solved puzzles requiring naked pairs in the first session, though they considered the naked pair puzzle in session 2 impossible to solve without a 'leap of faith'. Other participants also managed to solve naked pair puzzles without the use of annotation - notably of the 6 participants who attempted puzzle C without the use of annotation, 4 successfully solved the puzzle.

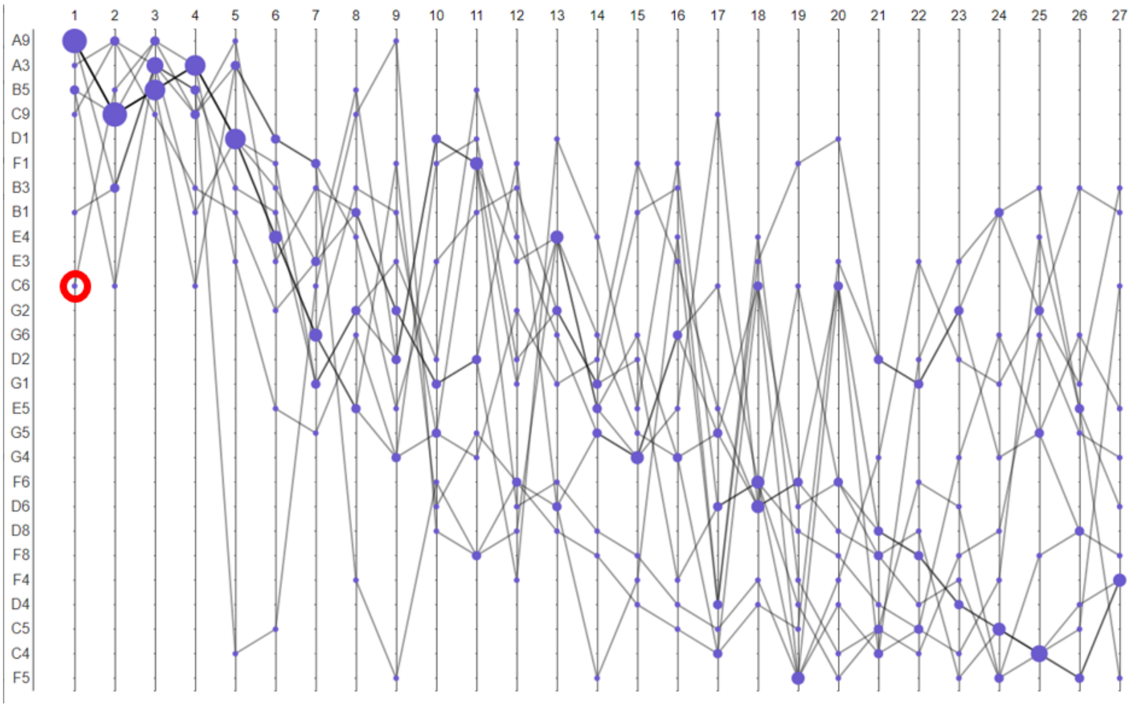


Figure 3.12: Path taken by participants through M1 puzzle, circles sized by the number of players that filled in the cell at that step. Y-axis is the cell filled in (ordered by the average step they were filled), X-axis is the step. Excludes participants that made a mistake and backtracked. The expected entry point is circled in red, showing that only one participant started the puzzle from the expected entry point. The ordering of the y-axis further shows that the expected entry point was, on average, the 11th cell filled in.

3.3.2 Order of play

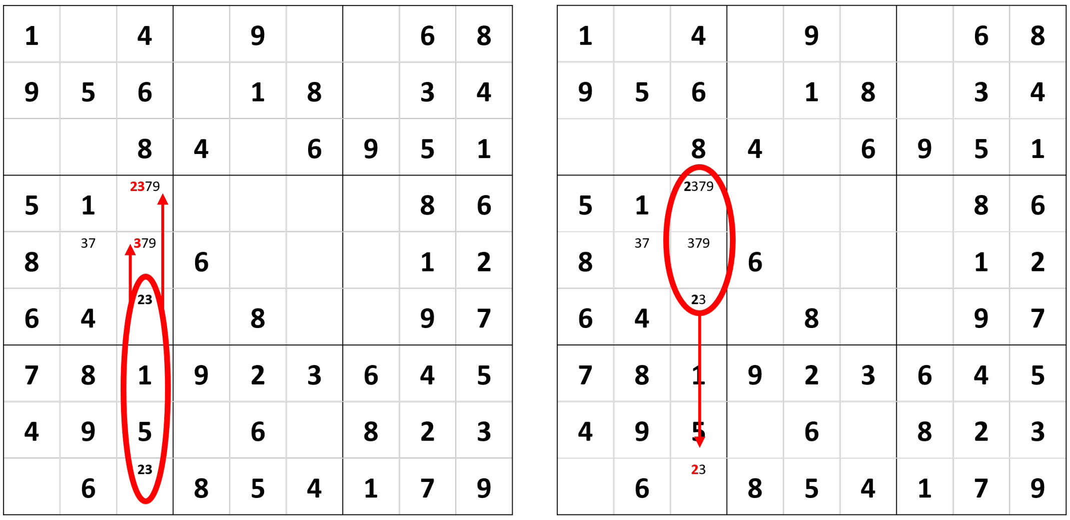
The order in which players complete the cells in a Sudoku is somewhat dictated by the design of the puzzle. The number of routes player can take through a puzzle is partly determined by how challenging particular techniques are to identify. However, even if a consistent difficulty ordering is used for the techniques, and the easiest available technique is applied at each step, there may be multiple places where that technique can be applied, allowing multiple routes through the puzzle. Other puzzles may only have one easy step available at each stage, allowing only one 'easiest' route through the puzzle. However, we noticed that even in Sudokus that we expected to have very limited entry points [5] participants picked a variety of entry points and followed varied routes through the puzzle (see Figure 3.12).

This was particularly noticeable in the M puzzles, which we expected to have a single entry point, based on the Sudoku Wiki solver [142, 141]. However, of

---

[5] Entry point meaning the first cell that a player can complete

the 21 participants that attempted M1, only one participant started the puzzle with the expected cell (including the eight participants who made mistakes). The paths the participants took through the puzzle can be seen in Figure 3.12. The entry via A9 and C9 demonstrates that participants performed a pointing pair before the expected naked pair (as the naked pair would remove the pointing pair). The two approaches are shown in Figure 3.13 Puzzle R1 had several possible entry points, one of which was the first column, which only had one empty cell (r8c1[6]) see Figure 3.14. We expected it to be the most common entry point, as a single missing cell is considered the easiest to spot [140]. However, of the 18 participants (including the 4 who made mistakes) that completed puzzle R1, only one completed this cell first. This may have been due to numerical approaches - where participants started by searching the grid for all the occurrences of 1, then for 2 etc. 10 of the participants completed r7c3 with 1 as their first step. Only six participants had completed r8c1 by their fifth step. R1 only required naked singles, and it is interesting that there does not appear to be a difference between the discovery of the last naked single in a row/column and one that requires more careful checking of the overlapping dimensions.



1		4		9			6	8
9	5	6		1	8		3	4
		8	4		6	9	5	1
5	1	2379					8	6
8	37	379	6				1	2
6	4	23		8			9	7
7	8	1	9	2	3	6	4	5
4	9	5		6		8	2	3
	6	23	8	5	4	1	7	9

(a) The M1 puzzle with naked pair shown

1		4		9			6	8
9	5	6		1	8		3	4
		8	4		6	9	5	1
5	1	2379					8	6
8	37	379	6				1	2
6	4	23		8			9	7
7	8	1	9	2	3	6	4	5
4	9	5		6		8	2	3
	6	23	8	5	4	1	7	9

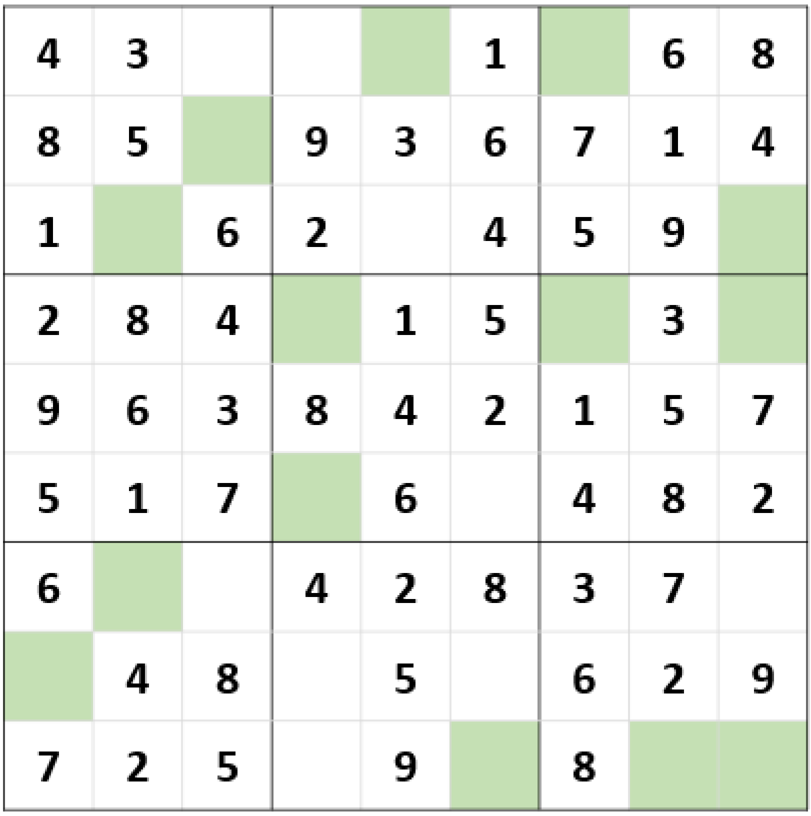
(b) The M1 puzzle with pointing pair shown

Figure 3.13: M1 puzzle with naked pair (left) and pointing pair (right) shown

Q1 had three expected entry points; however, 12 of the 13 participants who attempted the puzzle started in the same cell (r7c2) and 11 of those 12 completed r2c3 as their second step. Both participants who differed from this pattern made

---

[6] rX indicates the row in the Sudoku, numbered 1-9, cX indicates the column, numbered 1-9



4	3			🟩	1	🟩	6	8
8	5	🟩	9	3	6	7	1	4
1	🟩	6	2		4	5	9	🟩
2	8	4	🟩	1	5	🟩	3	🟩
9	6	3	8	4	2	1	5	7
5	1	7	🟩	6		4	8	2
6	🟩		4	2	8	3	7	
🟩	4	8		5		6	2	9
7	2	5		9	🟩	8	🟩	🟩

Figure 3.14: Puzzle R1 is shown, with entry points highlighted in pale green.

errors while completing the puzzle. Interestingly, in the puzzle Q2, which was an isomorphic version of Q1 and shared the pattern of a bottom left box with a single empty cell in r7c2, 10 of the 14 participants that attempted the puzzle started in r7c2. It is possible that a box with only one empty cell remaining is easier to spot than a column or row with only one empty cell remaining; however, the sample size is too small to be conclusive.

The interviews and questionnaires provided some insight into the order in which the participants approached the puzzles. While some participants described looking for easy approaches and then applying harder ones. "I try the easy tactics and if they do not work, I have to try some more difficult tactics" - participant 7[7]. 9 out of the 31 participants mentioned that they look for dimensions with the fewest empty cells as a starting point, participant 15 specifically noted that they prefer to look for the most filled 3x3 box before looking at the rows and columns. 17 out of 31 participants explained that they approach the Sudoku in numerical order, either checking all the ones, then twos, etc. until the nines or in reverse.

---

[7] Participant IDs start at 7, 1-6 were used in the initial development stages

Participant 34 described it as “I’d try and do all the ones and then try and do all the twos and then try and do all the threes etc”. Those who worked through the easiest to hardest techniques described different difficulty orderings.

3.3.3 Mistakes

20 of the 31 participants made at least one mistake during session 1. The types of mistakes made are discussed below. The causes of these mistakes can be simplified to either marking candidates as impossible that were actually viable or failing to exclude impossible candidates, of which the latter was more common. The errors in both sessions propagated from the cell where the error occurred to the rest of the puzzle.

Digit already present in a dimension. A common error occurred when candidates completed digits in a dimension that already contained a particular digit. In some cases, participants even completed digits directly adjacent to the digit that indicated its impossibility. An example is shown in Figure 3.16.

Impossible candidate not excluded. This error is in some ways a super-set of the ‘Digit already present in a dimension error’. In this case, we refer to errors that result from missing a deduction that would eliminate a candidate. This results in the player making deductions that include possible digits that should have been marked as impossible. For example, in Figure 3.17, 2 has not been excluded from r9c3, leaving 3 as the only candidate, which makes it clear that 3 cannot be placed in r9c1.

Incorrect Candidate exclusion. This is the reverse of not excluding an impossible candidate; in this case, the error occurred when a participant incorrectly excluded a candidate from a cell, leading to the appearance of a single remaining value. In this example (Figure 3.15, right) the participant incorrectly excluded the digit 8 from the grey box cell, leaving 9 as the only possible candidate. The completed digit should have been 8, as 9 should be placed in cell r1c7, as that is the only candidate that can go into that cell.

Incorrect Guess. 19 participants stated that they guessed when they could not make any further deductions. For example, on the right hand side of Figure 3.17, the participant stated that they were unable to make any further progress, so guessed (incorrectly) between the two allowed values (2 and 3) for r9c1.

Error propagation. When participants made errors, if they did not notice them immediately, the error propagated through the puzzle. For example, shown in Figure 3.17, the first error was entering 6 in cell r8c3, which seems to result from the incorrect exclusion of 5 as a candidate. Completion of r8c3 as 6 leaves r9c3 as the only place in the bottom left box where 5 can be completed. This is also incorrect, despite the deduction that leads to it being correct. Completion of r8c3 also leaves cell r7c9 as the only valid cell for 5 in the bottom right box. The error does not propagate up the column because the 5 and 6 still occupy the same set of cells that they would occupy if they had been done correctly. Filling them in allows 4 to be completed in cell r6c3 (as the only candidate remaining), leading to 9 being completed in cell r1c3 (also the only candidate remaining). It is not clear how the participant deduced that 4 should go into cell r1c1 (though it is correct). Five participants explicitly commented (via the post-puzzle questionnaire) that the propagation of the errors caused frustration as it was hard to backtrack to the source of an error, and they often simply started again.

3.3.4 Perceived Difficulties

The session 1 puzzles provided overarching data about the challenge a whole Sudoku posed. Participants indicated (via the post-puzzle questionnaire) that they found Q and F more challenging than the other puzzles that only required naked singles. F is a larger puzzle than the others [It had the most empty cells] which could explain the discrepancy, but Q was the smallest puzzle [It has the least empty cells]. When discussing the Q puzzles, both through the questionnaire and the interview, the participants stated that having an entirely empty 3×3 box increased the challenge. "The empty box in the middle made it trickier" - Participant 31, post puzzle questionnaire. They also expressed that having so much of the grid already filled in added to the challenge, it "Took a minute to identify where to start as so much was already completed" (participant 16) and, similarly, "I prefer an emptier grid to start with and having to fill in more" - participant 28. In contrast, participant 30 mentioned that they liked the empty 3×3 box.

Although many of the techniques that the participants described using were recognisable from the literature, the difficulty they associated with them was not

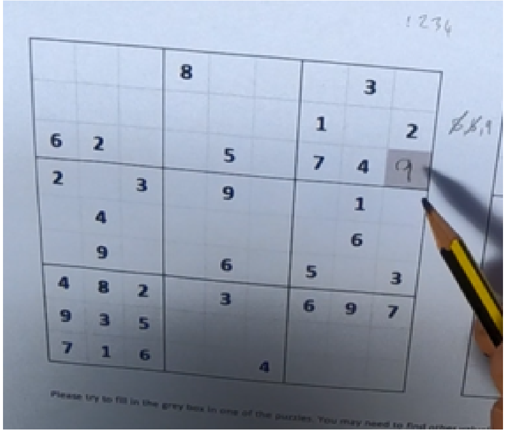


Figure 3.15: Example of incorrect candidate exclusion in puzzle B2, r3c9 should be 8 not 9

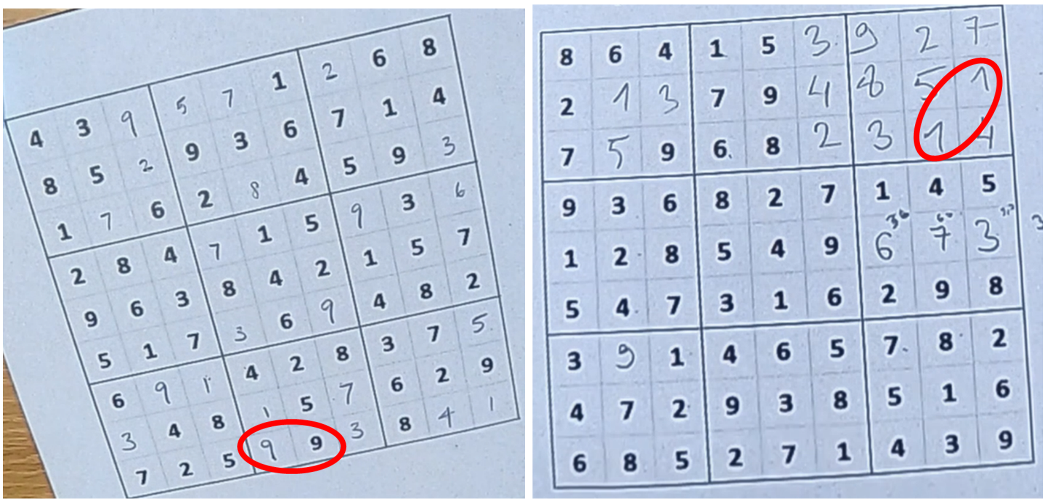


Figure 3.16: Left: Example of a 9 incorrectly completed in puzzle R1, directly adjacent to the clue that excluded it, Right: Example of two 1s incorrectly completed in the same box in puzzle Q1

consistent with the literature. Many participants described hidden singles and did not consider naked singles as an option, or described them as the step after looking for hidden singles. Participant 14 described hidden singles as 'any obvious numbers'. The difficulty of naked pairs varied between participants, ranging from considered equivalent to hidden and naked singles, to being considered more challenging than hidden and pointing pairs. Participant 11 used them so readily that they described it as "I kind of use it so much I don't think of it as technique" and "I do that without realising I'm doing it. I don't think". Some participants found pointing pairs the easiest of the techniques beyond naked and hidden singles, with participant 15 describing them as "It's the one that's easiest for me

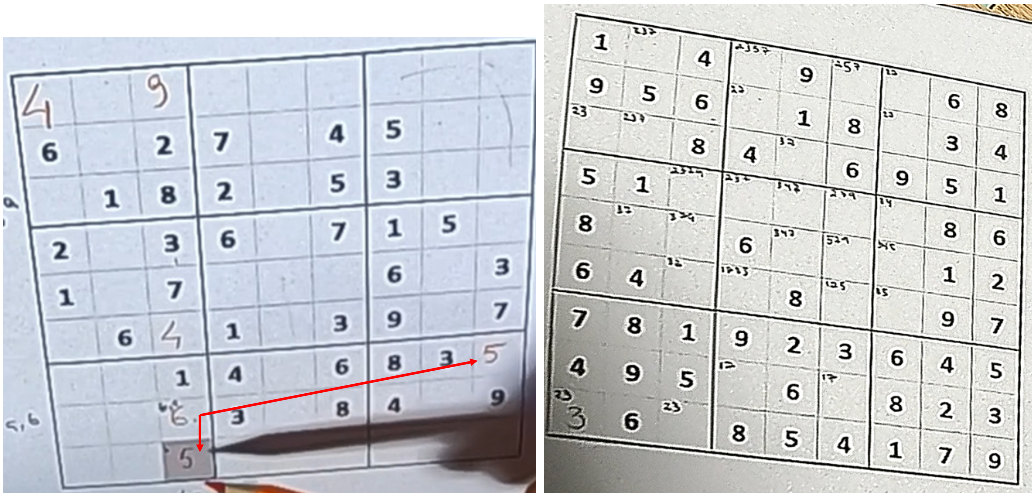


Figure 3.17: Left: Example of error propagation in puzzle O1, a mistake in excluding 5 from cell r8c3 leads to errors in r9c3 and r7c9, Right: Example of an incorrect guess (as described by participant) in r9c1

to approach”. This is supported by the preference the participants showed for employing a pointing pair rather than a naked pair when starting puzzle M.

The number of entry points into a puzzle had an impact on perceived difficulty and enjoyment, participant 15 stated “...if you can fill in a couple of firsts at the beginning and then get stuck, [false start], one of them was like oh, there is more multiple choices, I have to think about it now, I quite enjoy that. Whereas, if I can’t fill one out from the beginning it’s just a little bit frustrating”. This is consistent with the literature on the interaction between difficulty and enjoyment [72, 7, 73].

The variation in missing digits also affected both the perception of difficulty and enjoyment. 3 participants explicitly mentioned that they found puzzles where all the instances of a given number were missing both harder and less enjoyable, participant 16 stated “I’d rather there was more numbers missing but more variety of numbers”. 6 participants reported that they found the most frustrating puzzle to also be the most rewarding. These participants successfully completed the puzzle they viewed as most frustrating.

The session 2 puzzles focused on particular techniques. The challenge ratings provided by the participants indicated that, in session 2, the U puzzles were found to be the least challenging, despite them requiring Hidden Singles instead of Naked Singles. This is supported by the time taken to complete the puzzle, where U puzzles had the lowest average time to complete. This was further supported

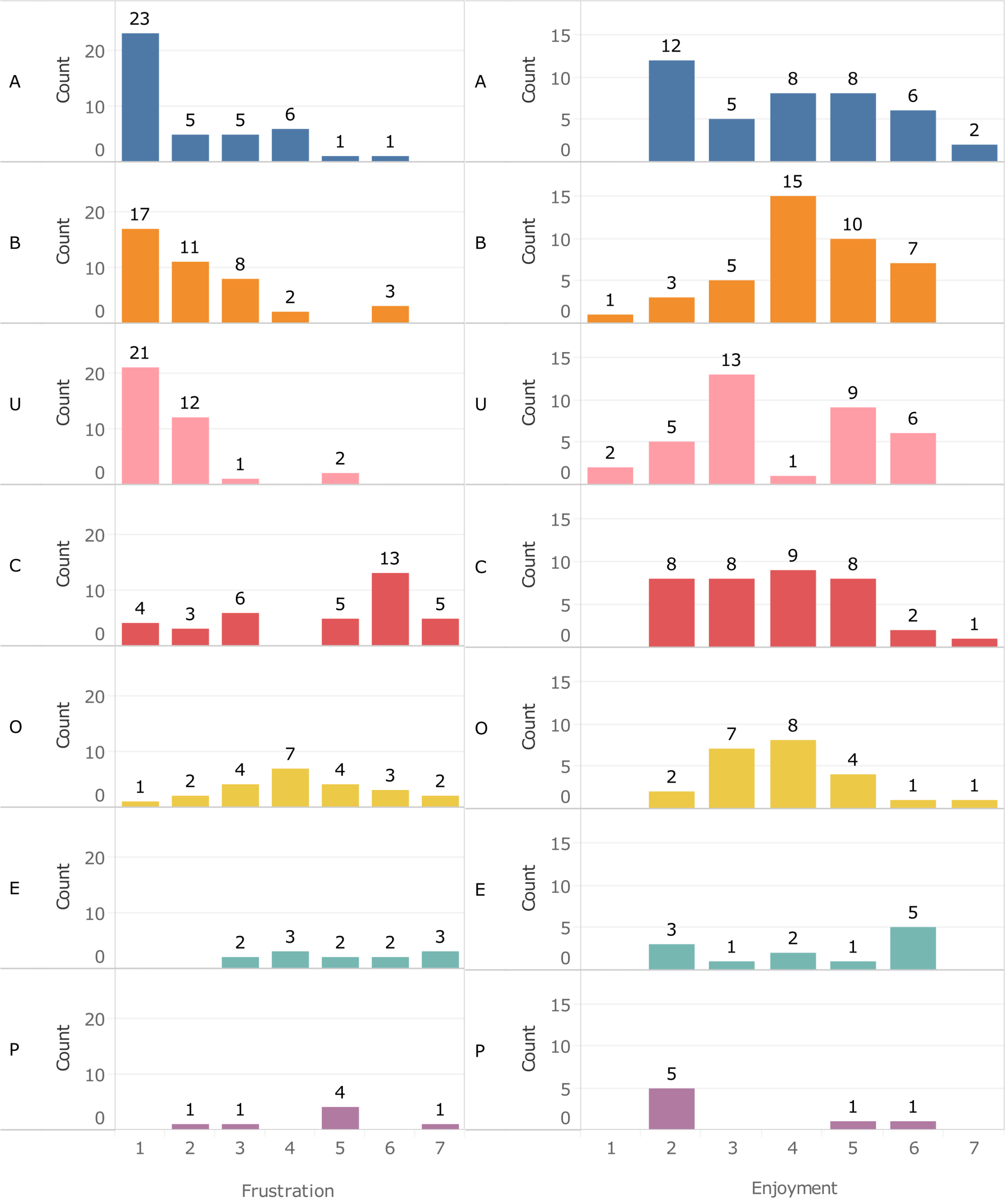


Figure 3.18: The left hand side shows the frustration Likert ratings (from 1 (no frustration) to 7 (extreme frustration)) for the puzzles in the Second Session. The right hand side shows the enjoyment Likert ratings (from 1 (no enjoyment) to 7 (extreme enjoyment)) for the puzzles in the Second Session.

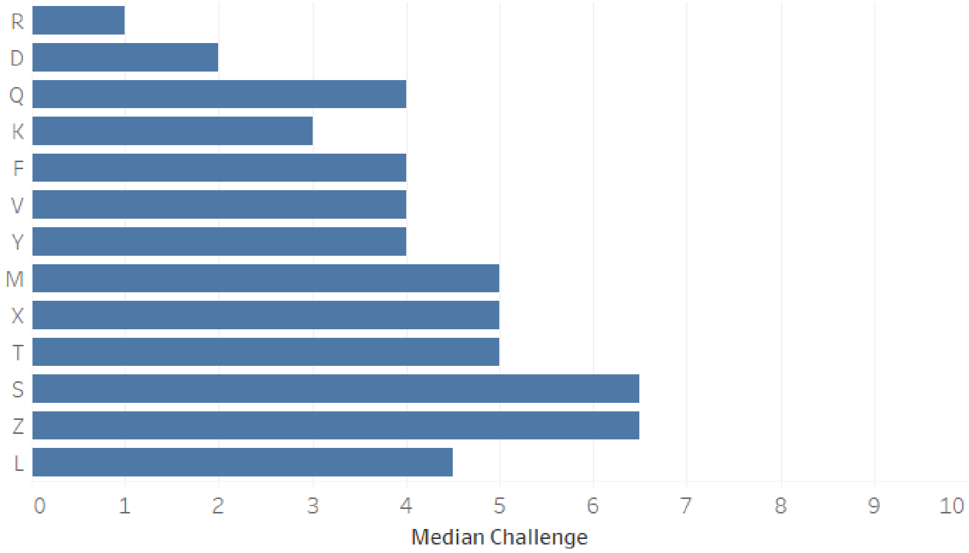


Figure 3.19: Median challenge ratings for Session 1 based on the post puzzle questionnaires, grouped by puzzle type

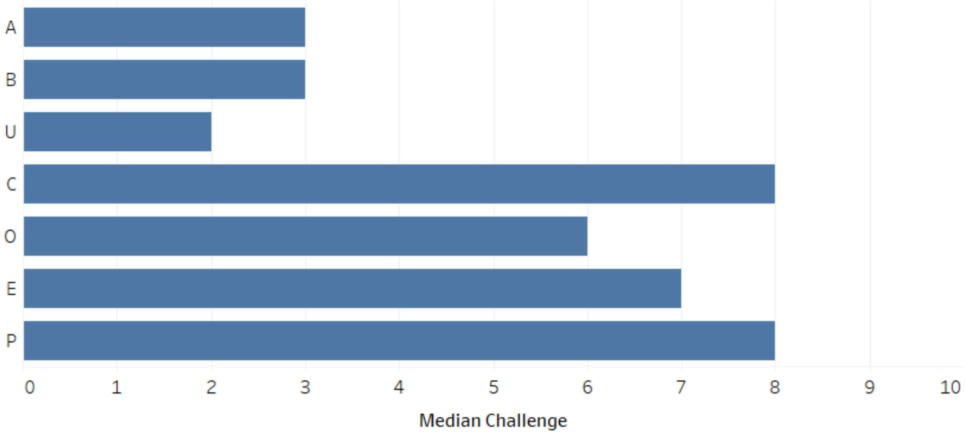


Figure 3.20: Median challenge ratings for session 2 based on post-puzzle questionnaires, grouped by puzzle type

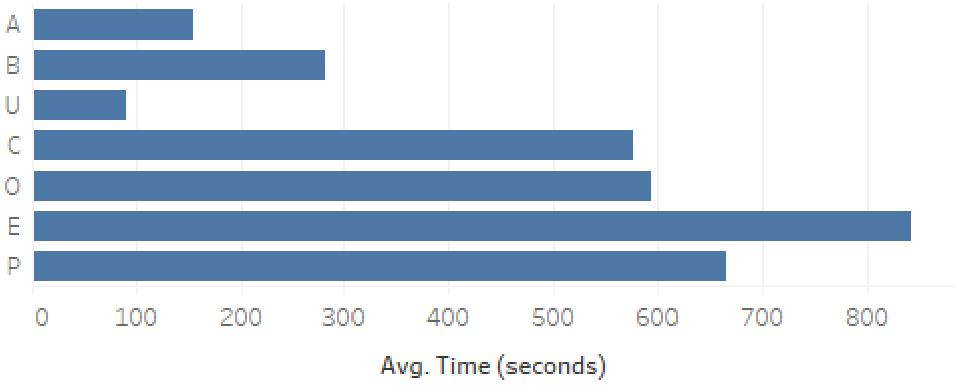
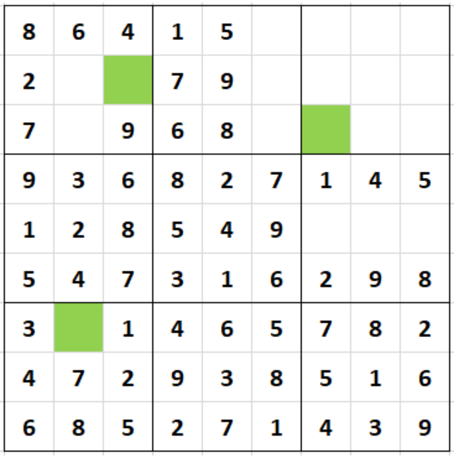


Figure 3.21: Mean time taken for each puzzle type in session 2



8	6	4	1	5				
2		🟩	7	9				
7		9	6	8		🟩		
9	3	6	8	2	7	1	4	5
1	2	8	5	4	9			
5	4	7	3	1	6	2	9	8
3	🟩	1	4	6	5	7	8	2
4	7	2	9	3	8	5	1	6
6	8	5	2	7	1	4	3	9

Figure 3.22: Puzzle Q with the three entry points highlighted

by the interviews, where five participants described hidden singles as the easiest technique. Although 5 other participants described approaching the puzzles by checking all the possible values for each cell without necessarily noting them down and felt this was the easiest technique. This is consistent with the findings of session 1, in which participants attempted to apply hidden singles before naked singles.

There was a notable increase in perceived challenge and time from the puzzles A, B & U (Naked and Hidden Singles) to puzzle C which includes a Naked Pair. Intermediate and advanced participants rated the Naked Pair as more challenging than the Hidden Pair, while beginner players rated it less challenging. (Although due to the small sample of beginner participants, it is not clear how representative that is.) This remained the case when participants who did not attempt the Hidden Pair were excluded from analysis. Participants readily resorted to chain methods in puzzle C despite it being solvable using a Naked Pair, theoretically a much easier technique than the chain methods.

3.4 Sudoku Solving Study Discussion

Our findings allow us to explore the validity of the assumptions used by existing computational models; in particular, the impact of candidate filling, the order of deductions, the experience of difficulty, and the impact of error. In each case, our findings demonstrate that the assumptions the existing computational models make do not align with the players.

3.4.1 Notation

Most guides (and techniques) assume that players systematically fill in all possiblecandidates for each cell and systematically eliminate impossible candidates as they progress through the puzzle. Computational models of Sudoku players share this assumption. However, our findings showed extensive variation in the way players take notes and how they use those notes in their solving process.

The systematic approach discussed above rarely occurred in practise. The approaches used by many participants meant that some techniques, such as naked singles, were harder to employ. This suggests that future guides and computational models should avoid the assumption that players will behave systematically and consistently.

In particular, many online systems assume players are performing systematic candidate filling and mark other notations as “wrong”', which may confuse or upset players using an alternative notation scheme. It is true that other notation approaches may make some of the standard techniques harder to employ, but for simple Sudokus these techniques are not necessary, and better supporting the player's chosen notation style could make for a more rewarding experience. Alternatively, online systems could coach players towards using systematic candidate filling.

The variety of notation that occurs would require a very free-form interface to support all options.

3.4.2 Relative Difficulty of Techniques

Based on the existing assumptions in the literature [140, 107], we expected:

- A linear increase in difficulty between different techniques.
- Techniques would be attempted in the established order of difficulty.
- The steps in which techniques were required would not have an impact on perceived challenge.

We found that all of these assumptions are flawed.

The accepted order of technique difficulty did not apply. Many participants found Hidden Singles easier than Naked Singles; this could be related to annotations.

Finding a naked single is easier when using systematic candidate filling, as a cell with only onecandidate stands out. However, without systematic candidate filling checking all the places in a row/column/box that a digit can go (Hidden Single), could be easier than checking all the numbers that can go in a particular cell to see if only one is possible (Naked Single). This is supported by the minimal use of annotation when solving Puzzle U, which contained only Hidden Singles. This finding could also suggest that hidden singles are easier to discover independently.

Similarly, participants finding Hidden Pairs easier than Naked Pairs could be related to non-standard notation making hidden singles easier to detect. Participants using pointing pairs before naked pairs could also be related to non-standard notation. In fact, participant 15 described pointing pairs as the easiest technique. Which differs from the literature which generally considers pointing pairs to be the most challenging of the pair techniques [140, 36]; although Sudoku Dragon by Silurian Software, has pointing pairs as the easiest technique and naked pairs as the hardest (albeit using slightly different names) [144].

The substantial jump in challenge from Naked/Hidden Singles to Naked/Hidden Pairs suggests that the impact more advanced techniques have on difficulty is greater than previously assumed. This suggests that when transitioning to Sudokus of higher difficulty levels, the environment (book, magazine, app, etc.) should provide players with support in solving using the more advanced techniques (for example, by including several examples, all of which can be used to solve the puzzle).

The ready use of chain techniques by participants suggests that they may be more intuitive than previously accepted.

It is clear that existing assumptions about techniques' relative challenge and their impact on a puzzle's difficulty are flawed. The finding that participants did not apply techniques in the established order of difficulty suggests that the ordering needs to be re-evaluated. However, the participants also disagreed on the difficulty of techniques, partially based on their notation approach. This suggests that digital tools could adapt their grading of puzzles based on an individual's notation approaches. Digital tools could also assess how challenging an individual player finds a particular technique and adapt their difficulty grading accordingly.

Overall, it is clear that a variety of factors must be considered when assessing difficulty. Computational models could consider designing different 'players' with varying difficulty ordering and notation approaches to test against, or test against real players. These findings have implications for teaching and help systems. The systems may need to be more flexible in the direction they guide players in - as a player could find being assisted only towards an 'easy' cell that they don't find that easy frustrating. An alternative could be a system that guided players towards the conventionally easiest square, but still provided guidance towards cells that required the second 'easiest' technique; this would mean players that did not find the 'easiest' technique that easy would still be given some useful guidance. Help systems could also consider that players may resort to chain techniques unnecessarily and provide guidance if they detect a chain being used unnecessarily.

3.4.3 Spatial layout

The findings that Q and F were considered more challenging than the other very easy puzzles were unexpected. We conclude that the spatial arrangement of the puzzle can impact the challenge a player experiences and therefore should be considered by designers, whether human or AI. However, further research is needed to fully formulate the impact it has on player experience.

3.4.4 Order of Play

Figure 3.12 demonstrates that players can consistently perform a deduction that differs from that suggested by current models. This affects both the difficulty grading and the help systems. This suggests the use a data-driven approach, where the moves taken by a majority of players are used for difficulty modelling and suggested to future players. It could also be used to develop a more accurate difficulty ordering and to assess the impact the spatial arrangement of the Sudoku has on the perception of difficulty and the player's actions. A full analysis using a data-driven approach would require a much larger data set; therefore, it is left for future work.

3.4.5 Impact of Error

Frequency of and recovery from errors contribute to the experience of a player in most games and is normally an important consideration for designers. However, in games like Sudoku, where each move should result from a logical deduction, player error is rarely considered. Our findings clearly show that participants made errors that were often missed – Figure 3.16 shows two final submissions with adjacent repeated digits.

Looking at participants' errors, they appear to come from both flawed visual searches and flawed logical deductions. The common error of completing a digit already present in one of the dimensions overlapping the cell often results from a flawed visual search. This error occurred even when the participant had completed the other occurrence of the digit themselves, strongly implying that the players have a limited ability to store the state of the puzzle in their head.

Errors, once made, propagate, as the grid is now in an incorrect state, so information deduced from it is flawed. The propagation of errors is a key contributing factor to making error recovery challenging. Participants reported increasing frustration when they made an error, as it often resulted in them restarting because recovery was too difficult.

Overall, our findings demonstrate that player error is not unusual and raises questions about the impact mistakes have on players' experience and how puzzle or interface design could be used to mitigate it. Players will sometimes correct mistakes almost immediately, therefore immediately correcting errors as soon as they appear would reduce player agency. However, unnoticed errors reduce player enjoyment, as they are often unrecoverable without restarting. Better systems to "fix" errors may improve player's enjoyment. For example, allowing the deductions that propagated from the error to be tracked and reverted would allow error recovery without restarting the puzzle or losing valid deductions made alongside erroneous ones.

3.4.6 Sudoku Solving Study Limitations

Despite 31 participants being a large cohort for this style of qualitative assessment, the study is primarily exploratory. It demonstrated that many previously published assumptions need to be reconsidered and provides direction in which they can be further investigated; however, it may not be sufficient to provide in-depth guidance to future model designs.

It was impossible to know exactly what people were thinking when they made specific notations. Recordings of their actions and subsequent interviews are not sufficient to fully understand the underlying thought processes.

Most of the participants in this study are university-educated and may not be representative of the general population.

Although this study provides rich qualitative information, it is an exploratory study and some areas would benefit from further large-scale quantitative studies. We leave this to future work.

3.5 Conclusion

Many of the assumptions that existing guides, designers, and AI models make about Sudoku players are flawed. The extensive variation, both in notation and in logical approaches, strongly suggests that Sudoku design or models based on rigid assumptions regarding player approaches are unlikely to produce puzzles of predictable challenge and reward. Designers (whether human or AI) should attempt to consider the different approaches players use when solving Sudokus, including the different methods of annotation, logical deductions, and mistakes. It is also important to explore the narrative that players could take through the puzzle, both the points of challenge and the number of deductions required to complete it. Treating players as automata, who always perform the easiest available technique and pick randomly if there are several options at the same level, does not reflect the player's behaviour. Furthermore, it should be considered that the most recent steps taken impact the player's current deductions and focus. It seems likely that in order to produce rewarding puzzles of predictable difficulty, the different paths which players may take through the puzzle need to be considered.

Our findings also have implications for tutoring systems and scaffolding. Puzzle games are used extensively throughout education, therefore providing better support systems has the potential to increase student attention and engagement. However, it is clear from this research that tutoring systems would need a way to interpret the notation style used, as assuming that a student is systematically noting down all possible options has been shown to be flawed. They may be noting down subsets or noting down values that have been excluded from possibility. Furthermore, it is important when providing assistance to decide if the goal is to teach students to apply methods in a chosen order or guide them towards the technique they would find easiest to apply.

Overall, the findings of this study have demonstrated that the existing, highly structured assumptions about puzzle game players are flawed. Players exhibit idiosyncratic solving techniques, are unsystematic, and make errors more often than previously assumed. In the following chapter, we will present a novel approach to providing guidance, ‘hints’, to players, the design of which incorporates insights gained from this study.

CHAPTER FOUR

DESIGN OF A NOVEL HINT SYSTEM

In this chapter we discuss the process of designing our novel hint system and the final design. The studies discussed in the previous chapter provided the motivation for much of the work and discussed in this chapter. The study of how people solve Sudoku discussed in the previous chapter demonstrated that players are erratic, error-prone and inconsistent. This means that any assistive system needed to be highly flexible - what works for one person may not work for another. We discuss this towards the end of Section 4.1. The variation in notation styles, discussed in Section 3.3.1.1 resulted in our work focusing on binary Progressive Pen & Paper Puzzle Game (PPPP)s, discussed in Section 4.2.3.1. The high rates of errors made by players during the study also demonstrated that we needed an error handling approach, or it would significantly compromise our results, we discuss the approach we chose in Section 4.2.3.2.

4.1 Survey of Existing Hint Systems

To evaluate the existing commercial approach to hint systems in PPPPs, we examined hints systems from the first 23 PPPP apps listed on the Android App Store, in 2020. We categorised the hint systems into the four categories described in Table 4.2.

Of the systems examined, only one provided information that guided the user towards the next deduction, rather than completing a cell. This system indicated to the user which cell could be filled in next; however, the criteria by which this

PPPP	Description
Sudoku	16
Tents and Trees	5
Other	2

Table 4.1: Types of PPPPs in surveyed apps

Hint System	Description	Number of apps
Textual step through	Fills in the next easiest square but explains why, often has a tap through where it talks through the explanation.	8
Fill in random cell	Fills in the value of a random cell	5
Fill in selected cell	Fills in the correct value into a selected cell.	6
Highlight the next easiest cell	Adds an indication to a cell that should be solvable. It does not indicate why though, and it was sometimes hard to work out why that cell was easier than an alternative.	1

Table 4.2: Summary of the types of hint systems found when surveying 23 apps in the Android app store. 3 had no hint system.

was defined were not clear and we could not always identify why the chosen cell would be the next step, instead finding an alternative cell to fill in which caused the suggested cell to change.

Hint systems provide an interesting problem: They should reduce frustration and help guide the player without detracting from the learning goals. In that sense two of the hint systems discussed in Table 4.2 are not applicable - ‘Fill in random cell’ and ‘Fill in selected cell’ do not provide the user with increased understanding or an increased ability to understand and engage with the task at hand. They do allow the player to progress with the puzzle, which may reduce frustration. The ’Textual step through’ works better; it talks the user through the process by which they can make the deduction, which may allow them to make the deduction independently in future, although the current step has been done for them. However, they sometimes tell the player things that they already understand but had not noticed due to poor visual search.

The 'highlight the next easiest cell' approach is interesting. It avoids the issue of telling the player the answer; instead, it guides the player towards a cell that they should be able to make a deduction about. However, if the player cannot make that deduction, it is probable that they will experience an increase in frustration, and there was no option to be led towards an alternative cell or to provide information on the deduction that is expected. In the second session of our Sudoku Solving Study, see Chapter 3, we indicated which cell would be the next easiest to solve (based on Stuart's difficulty ordering [140]). Our results indicated that if the next step was too challenging, the players became very frustrated (see Figure 3.18).

Our survey showed that there was scope for the development of a novel hint system that focused on guiding the player towards easier deductions, while providing options for the user to move on to alternative cells if they are unable to solve the current cell.

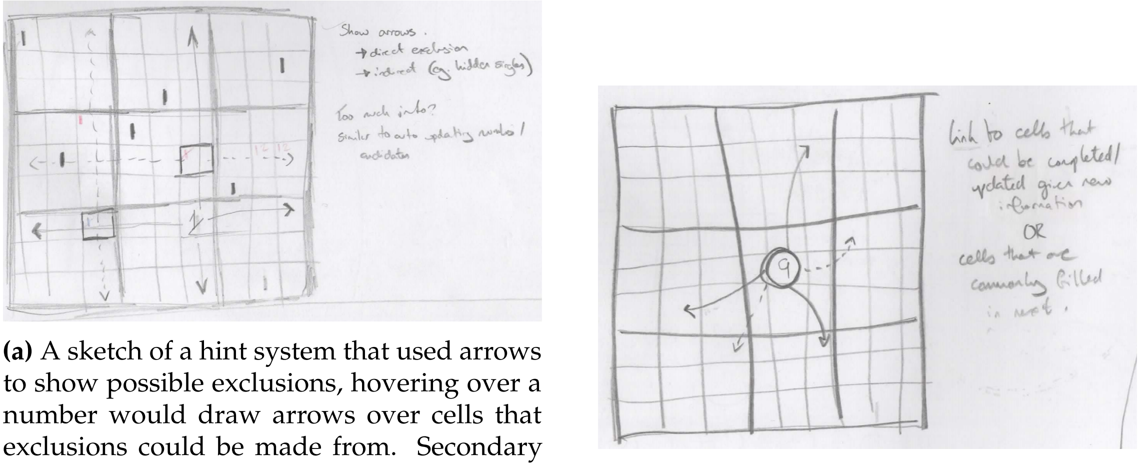
4.2 Designing the interface

We considered a range of different possible approaches for a novel hint system. The goal of the system was to guide the player towards a deduction rather than giving them the next step. Our design principle was based on the belief that providing a player with the next action, even with an associated explanation, will generally be found to be less rewarding and less conducive to learning than finding the solution themselves. We also aimed, as discussed in Section 2.4.2.1, to avoid a sense of resentment and 'cheating' in the players. Finally, the hint system should provide hints only on demand; Wauck & Fu found that players disliked systems where they were provided with hints automatically or adaptively. In both cases, the players felt that they had received more hints than necessary [155]. We aimed to provide procedural hints - a type of hint that helps a player progress to a new state that is closer to a solution -instead of providing remedial hint [81]. Finally, it needed to be generalisable to most PPPPs.

4.2.1 Considered Designs

We considered a range of designs prior to settling on the design we implemented (and then iteratively refined).

We considered guiding the players by highlighting the relationships between a cell (or candidate) and other elements of the puzzle. An early idea was to show


Show arrows.
→ direct exclusion
→ indirect (eg. hidden singles)
Too much info?
Similar to auto updating numbers / candidates
link to cells that could be completed/ updated given new information
OR
cells that are commonly filled in next.
9

(a) A sketch of a hint system that used arrows to show possible exclusions, hovering over a number would draw arrows over cells that exclusions could be made from. Secondary arrows would highlight the deductions that could be made as a result of the initial deduction.

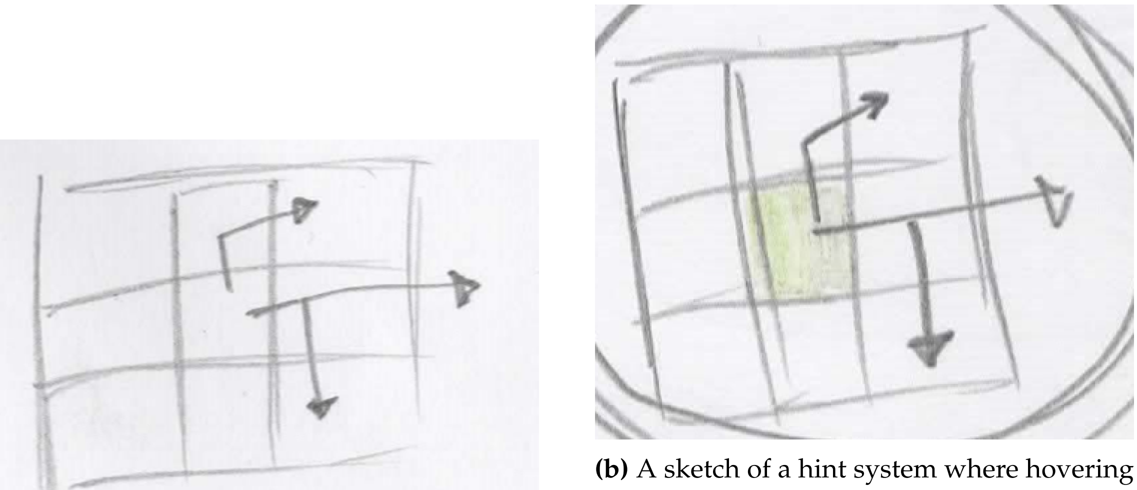
(b) A hint system that links to cells that the highlighted cell impacts, or cells that are commonly updated after the selected cell

Figure 4.1: Two sketches of possible hint systems using links

the eliminations that could be made using a completed digit. The example in Figure 4.1a would draw arrows over all cells that could have candidates eliminated from them using the completed digit; it would also highlight secondary deductions that could be made once the elimination was done. However, this was similar to existing systems that automatically removed candidates from overlapping dimensions when a completion was done; this type of system was normally considered an interface choice, rather than a hint system. It also would not provide much support to players in using or learning more complex deduction techniques. This could be improved by linking only to cells that could be in some way updated given a piece of information (whether cell completion or candidate) or cells that were commonly updated after the chosen cell was updated. An example is shown in Figure 4.1b, and while it is an improvement over only highlighting impacted dimensions, it may still give the user too much information for simple deductions.

Alternatively, the puzzle elements required for the deduction could be highlighted, as shown in Figure 4.3c. This approach allows the player to guide which cell they want to deduce while the hint system indicates where they should look for a deduction. This could allow more advanced techniques to be employed or learnt, such as an X-Wing in a Sudoku puzzle. However, it may not translate as well to other puzzles, such as tents and trees, where dimension highlighting may not be useful.

Providing the player with guidance towards the most useful, rather than the



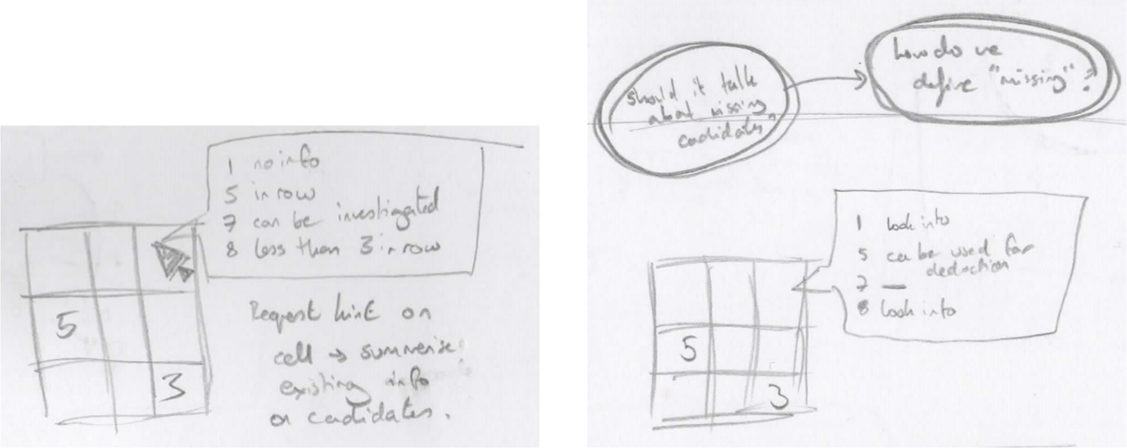
(a) A sketch of a hint system where hovering over a cell draws paths to the cells which will become easier to solve if the intial cell is solved.

(b) A sketch of a hint system where hovering over a cell draws paths to the cells which will become easier to solve if the initial cell is solved, combined with highlighting of ′easier′ to solve cells.

Figure 4.2: Two sketches of possible hint systems using lines to indicate which cells become easier to solve.

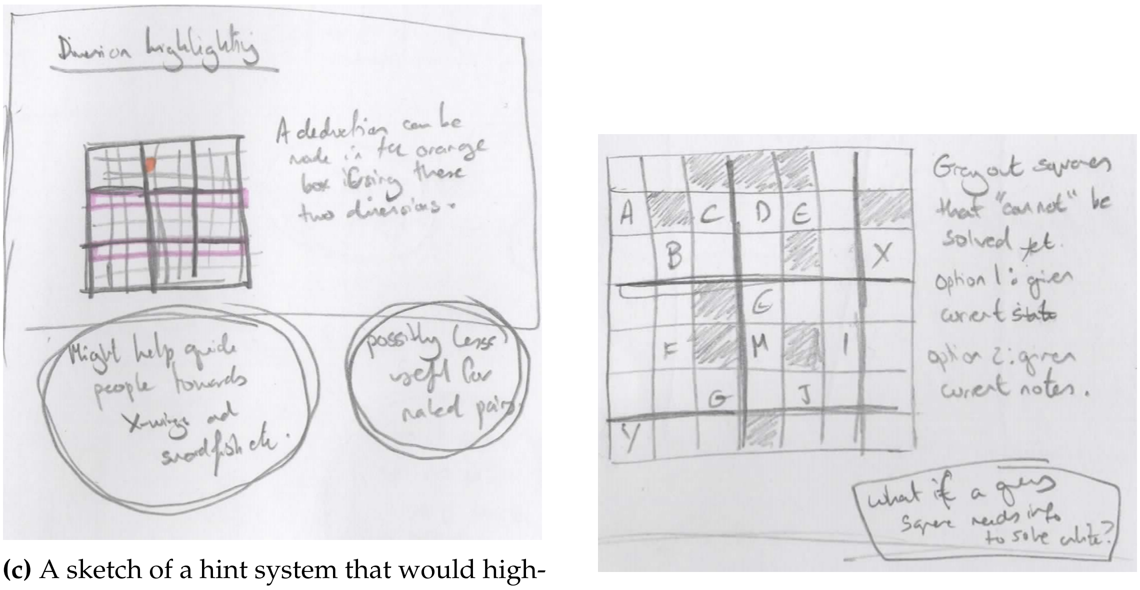
easiest, deduction was considered. Figure 4.2a shows a hint system where selecting an empty cell produced an indication of which cells would become easier to solve if it, the selected cell, were solved. This could help players focus on the cells which, if solved, would facilitate a large progression in the puzzle. However, the player has no idea how hard the deduction will be to make, and impossible deductions might facilitate the solving of many cells, resulting in the player wasting time attempting to solve impossible cells. This could be avoided by only providing the hints on ′solvable′ cells or highlighting ′solvable′ or ′easy′ cells and then providing the indication of which cells it will render solvable, shown in Figure 4.2b.

Focusing on helping the player′s process, we considered a system that summarised all the information about candidates in a cell. For example, when a user hovered over the cell it would tell the user if a candidate was in a dimension already, if there were less than a threshold number of that candidate in a dimension, and similar information. An example is shown in Figure 4.3a. However, this type of summary might encourage the player down unproductive paths - it is common for players to focus on candidates with only a few instances left in a dimension. We observed this behaviour in the study discussed in Chapter 3. Players often employed subset notation, a notation exclusively used to note candidates when there were only 2 or 3 possible places left in a dimension. However, limited numbers of possible placements in a dimension does not always mean there is an easy solving


1 no info
5 in row
7 can be investigated
8 less than 3 in row
5
3
Request hint on cell → summarise existing info on candidates.
Should it talk about missing candidates
How do we define "missing"?
1 look into
5 can be used for deduction
7 —
8 look into
5
3

(a) A sketch of a hint system that summarises information about the candidates in a selected cell

(b) A sketch of a hint system that indicates which candidates in a cell are sensible to focus the player's attention on


Dimension highlighting
A deduction can be made in the orange box using these two dimensions.
Might help guide people towards X-wings and swordfish etc.
possibly less useful for naked pairs
A
C
D
E
B
X
G
F
M
I
G
J
Y
Grey out squares that "can not" be solved yet.
option 1: given current state
option 2: given current notes.
what if a grey square needs info to solve white?

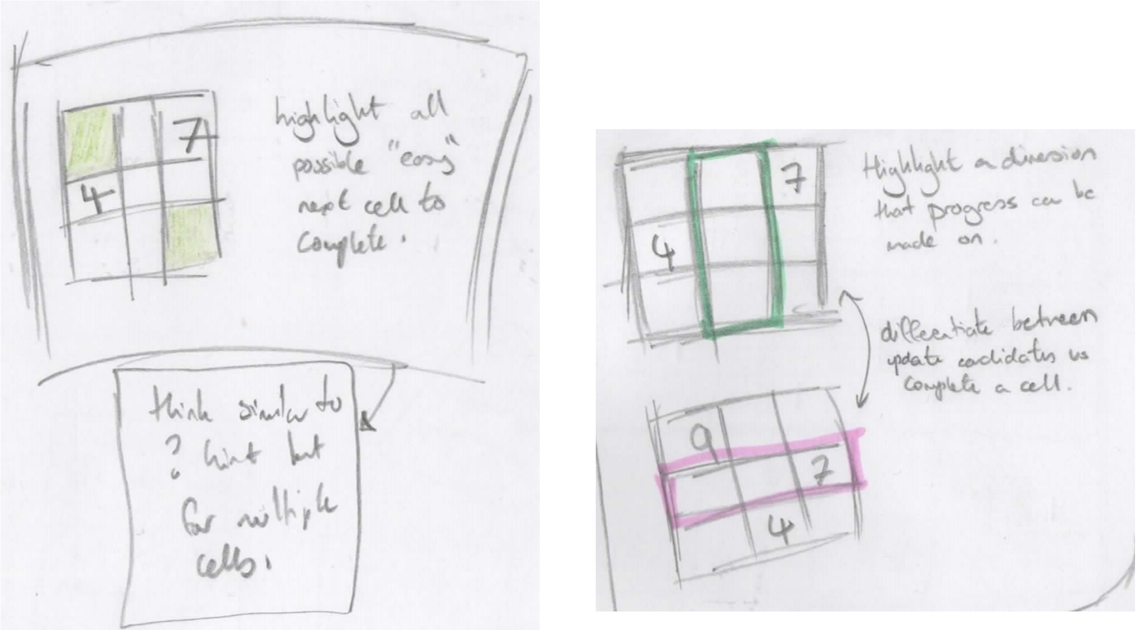
(c) A sketch of a hint system that would highlight the dimensions of the puzzle required to make a particular deduction.

(d) A sketch of a hint system that marked out 'unsolvable' cells in dark grey

Figure 4.3: Four sketches of possible hint system designs

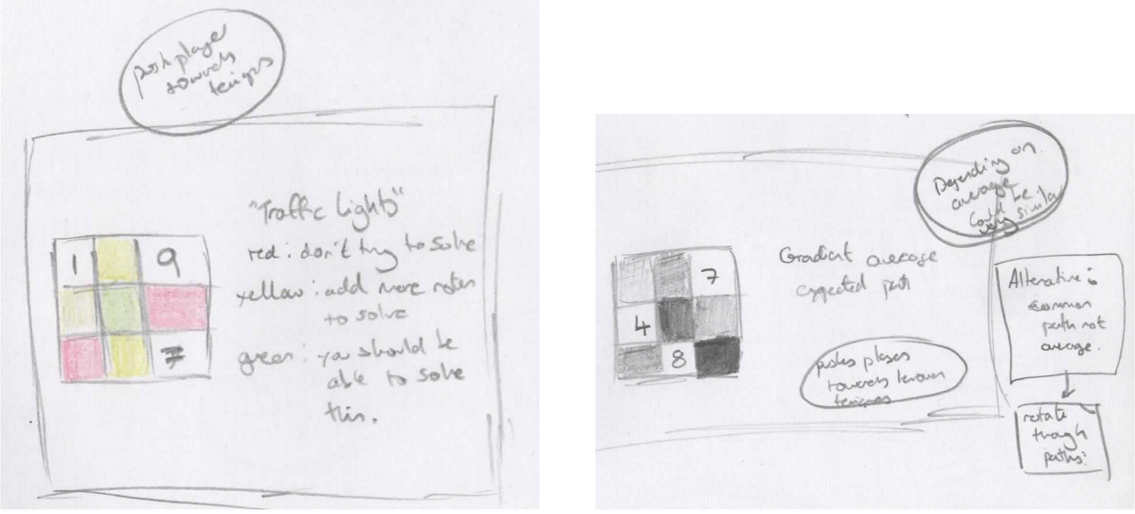
step available. Players attach more weight to information provided to them and therefore are more likely to fall into this trap [35, 55]. A variation of this approach, shown in Figure 4.3b, would be to provide the player with information as to how easy it would be to make a deduction regarding a given candidate. It could even be combined with the approaches discussed in the previous paragraph and provide information about which candidates could (if eliminated or completed) lead to most secondary deductions. All of these approaches depend on the puzzle having candidates - binary PPPPs, where only one of two values can be placed in a cell, do not require candidates meaning these approaches would not be generically applicable across PPPPs.

Guiding the player towards places in the puzzle where progress could most easily be made would be flexible, generic, and useful to the player. This is what the 'highlight the next easiest cell' (see Table 4.2) is aiming to do. However, as we found in session 2 of the Sudoku study discussed in Chapter 3, participants found the Complete Particular Cell (CPC) approach frustrating when they couldn't solve the cell. Rather than point players towards a specific cell we wanted to guide them towards a range of 'easy' cells, allowing them to move on from a particular cell without being left without guidance. We considered highlighting a dimension where progress could be made, shown in Figure 4.3c. This had the limitation that there might still only be one cell in the dimension that the player could make progress on - which would produce the same issues with frustration as a single box. A possible approach was to guide players away from impossible options, towards cells in a puzzle that were possible to be solved. Therefore, we developed early designs of a system that greyed-out squares that were unsolvable, as shown in Figure 4.3d. This approach directed players away from 'unsolvable' cells but didn't necessarily guide them towards 'easier' cells. We could, instead of directing players away from unsolvable cells, direct the towards the easier cells using a similar highlighting system, shown in Figure 4.4a. However, if there were very few easy cells, a player might still end up directing attention at unsolvable cells. Therefore, a better system would be to highlight both the easy and unsolvable cells in different ways; Figures 4.4c and 4.4d show two possible approaches. This approach is advantageous because it can be applied to the majority of PPPPs, and it provides assistance to the player while still requiring them to make their own deductions and progress. It could also, in future work, be combined with other approaches discussed above. However, if a player is stuck and cannot make any progress it will not help, which may render it ineffective as a hint system. It also


7
4
highlight all possible "easy" next cell to complete.
think similar to ? hint but for multiple cells.
7
4
Highlight a dimension that progress can be made on.
differentiate between update candidates vs complete a cell.
9
7
4

(a) A sketch of a hint system that would highlight the easiest cells in light green

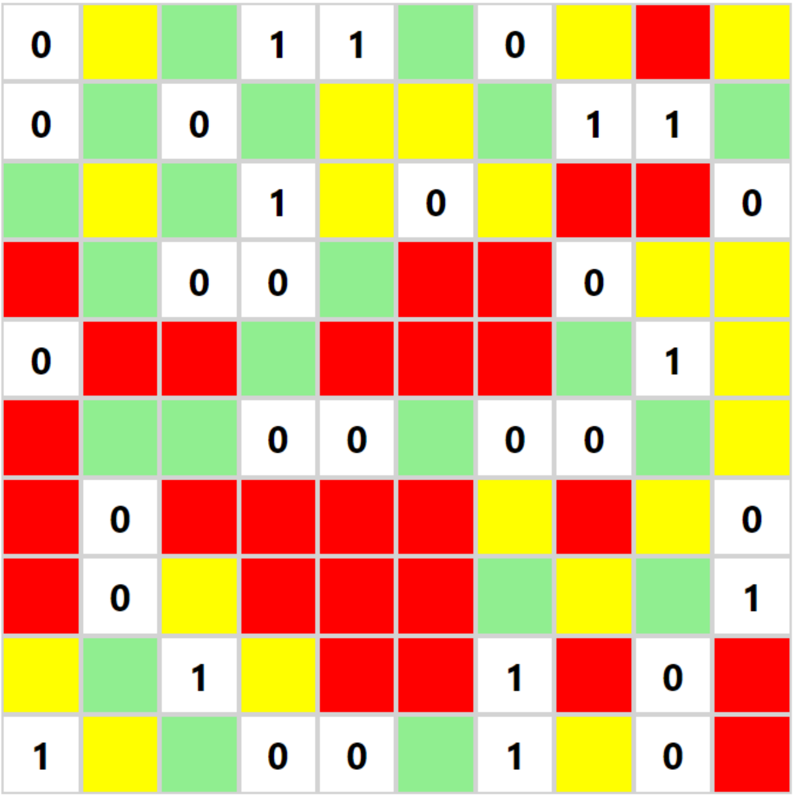
(b) A sketch of a hint system that summarises information about the candidates in a selected cell


push player towards techniques
"Traffic lights"
red: don't try to solve
yellow: add more notes to solve
green: you should be able to solve this.
1
9
7
Depending on average be could similar
Gradient average expected path
7
4
8
pushes players towards known techniques
Alternative is common path not average.
rotate through paths?

(c) A sketch of a hint system that would highlight unsolvable cells in red, easy cells in green and all other cells in yellow.

(d) A sketch of a hint system that uses saturation to indicate how easy or hard a cell is to solve

Figure 4.4: Four sketches of hint system designs that guide players towards the 'easiest' cells



0	🟨	🟩	1	1	🟩	0	🟨	🟥	🟨
0	🟩	0	🟩	🟨	🟨	🟩	1	1	🟩
🟩	🟨	🟩	1	🟨	0	🟨	🟥	🟥	0
🟥	🟩	0	0	🟩	🟥	🟥	0	🟨	🟨
0	🟥	🟥	🟩	🟥	🟥	🟥	🟩	1	🟨
🟥	🟩	🟩	0	0	🟩	0	0	🟩	🟨
🟥	0	🟥	🟥	🟥	🟥	🟨	🟥	🟨	0
🟥	0	🟨	🟥	🟥	🟥	🟩	🟨	🟩	1
🟨	🟩	1	🟨	🟥	🟥	1	🟥	0	🟥
1	🟨	🟩	0	0	🟩	1	🟨	0	🟥

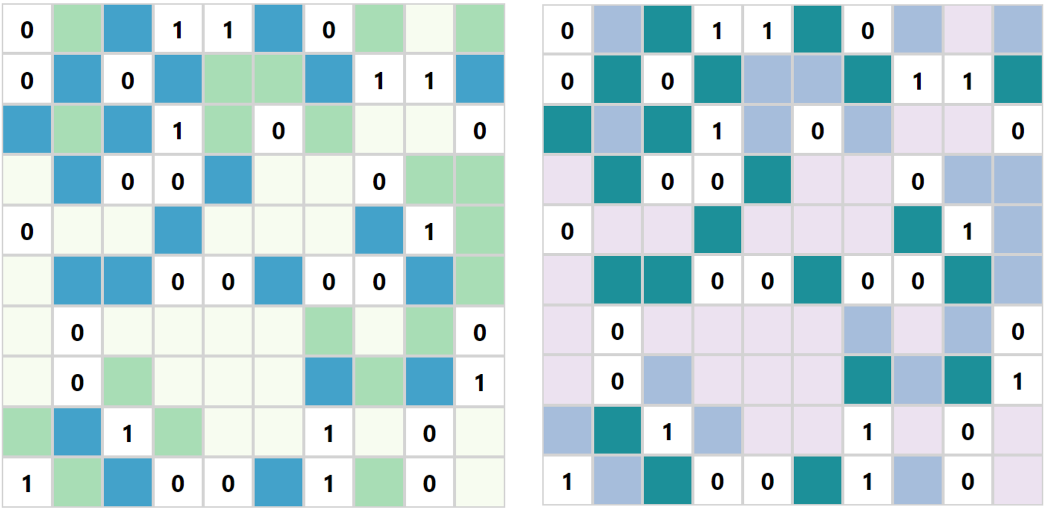
Figure 4.5: A test of the hint system with a traffic light colour scheme.

does not provide the player with guidance on how to make the deduction. While these are serious concerns, the advantages of a flexible generic system that guided players outweighed them.

4.2.2 Our Novel Hint System Design

We designed a novel hints system based on the design discussed at the end of Section 4.2.1.

The hint system would indicate how easy a cell was to solve, based on a difficulty rating. We considered a variety of approaches to indicate how hard a cell was likely to be. We decided against textures, as they might interact visually with the puzzle cells and do not have an intrinsic ordering. A "traffic light" system was considered, with red for hard, green for easy, and yellow for everything in between (Figure 4.5). However, this type of colour scheme is not universal, it depends on cultural norms that can vary, and is often not colour blind friendly [37]. We considered a gradient of colour with a highlight in a different colour showing the easiest cell, Figure 4.6 . However, early pilot studies within the research group



0			1	1		0			
0		0					1	1	
			1		0				0
		0	0				0		
0								1	
			0	0		0	0		
	0								0
	0								1
		1				1		0	
1			0	0		1		0	

0			1	1		0			
0		0					1	1	
			1		0				0
		0	0				0		
0								1	
			0	0		0	0		
	0								0
	0								1
		1				1		0	
1			0	0		1		0	

Figure 4.6: Two colour schemes with an alternate colour highlight for the easiest squares. Both shown on a Binairo puzzle.

made it clear that picking a highlight colour that was not misinterpreted by a high percentage of users was an area that needed further research. It was also challenging to find a highlight colour that was colour-blind friendly.

We chose saturation as the channel to convey the challenge, as it was both colour-blind friendly and inherently ordered [99]. We chose higher saturation to indicate easier cells and lower saturation to indicate harder cells. We considered an opposite mapping (lower saturation indicated easier cells and higher saturation indicated harder cells). However, as the overall page was white the eye was drawn by the highly saturated cells [99], and as we wanted to direct attention away from the harder cells towards the easier ones we chose to map high saturation to easier cells. Both approaches are shown in Figure 4.7 . The chosen colour scheme (Figure 4.7, right) was confirmed to be colour-blind friendly using the online tool Coblis [92]. Future investigation comparing the effectiveness of different colour schemes would allow a more optimal choice, but was outside of the scope of the initial research.

We decided to use DEMYSTIFY, the only available system to generate custom explanations for any PPPP (discussed in Section 2.2.3), to assess which cells were harder and which were easier. This is discussed below in Section 4.4.

We considered a continuous scale for mapping saturation to difficulty. The challenge of the harder cells covered a large range and therefore skewed the scale, making the easier cells very difficult to differentiate. This did not support

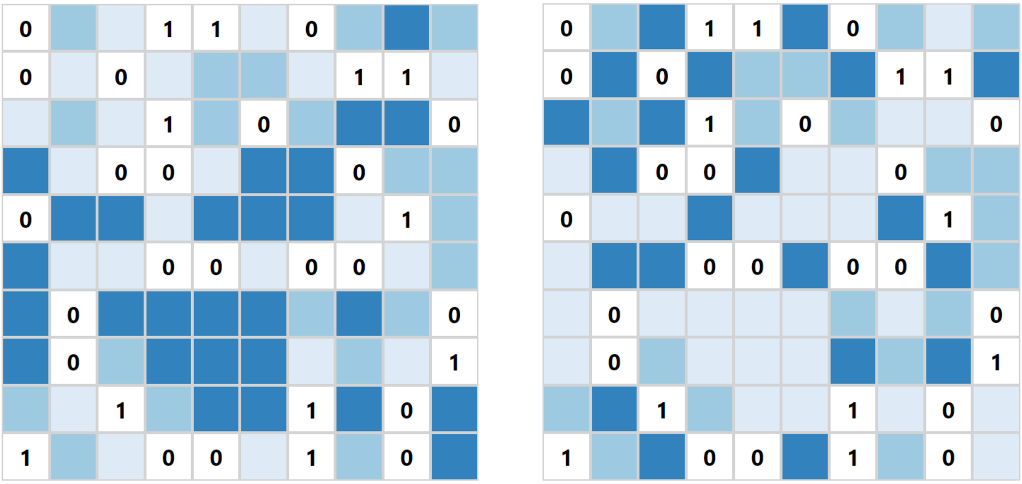


Figure 4.7: Example of grid hint using saturation as the key channel. Left: higher saturation indicates harder cells (rejected design). Right: lower saturation indicates harder cells (final design). Both shown on a Binairo puzzle.

the goal of directing players towards easy cells, and if they could not do the simplest cells, the directing them towards the next simplest cells. It was more important to direct the players towards the easier cells than to differentiate the harder cells. Therefore, we decided instead on grouping the difficulties into 3 bins, the simplest cells for the current puzzle state, the next simplest cells for the current puzzle state, and the hard cells for the current puzzle state.

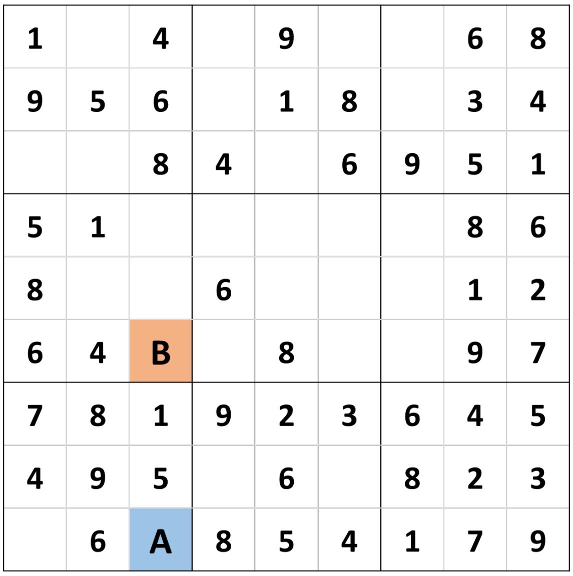
To conclude, we designed a hint system that shaded every not-completed cell in the puzzle grid. Each cell would be shaded with one of three saturation levels, the darkest level indicating that the cell is among the easiest to solve, the next level indicating that it is among the next easiest to solve, and the lightest level indicating that it is very challenging to solve.

4.2.3 Prototype Design Considerations

This section discusses the design decisions that were made in order to build a system that implements the concepts described in Section 4.2.2 while supporting a study evaluating the impact of the novel hint system design on the player experience compared to a traditional (fill in the next cell) system.

4.2.3.1 Issues with Candidates and Notation

The design process raised the question of whether we should be basing the state of the puzzle on the completed cells or the player's notations. We concluded that at



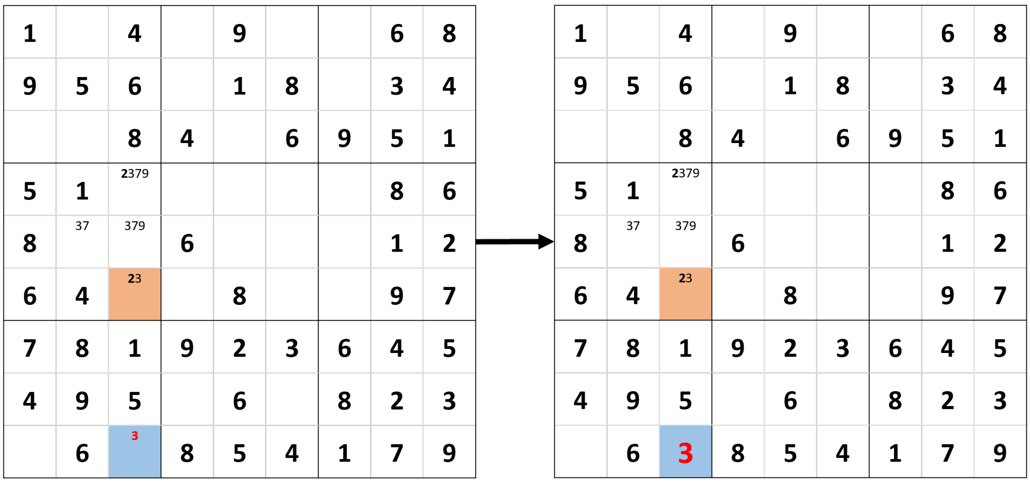
1		4		9			6	8
9	5	6		1	8		3	4
		8	4		6	9	5	1
5	1						8	6
8			6				1	2
6	4	B		8			9	7
7	8	1	9	2	3	6	4	5
4	9	5		6		8	2	3
	6	A	8	5	4	1	7	9

Figure 4.8: A Sudoku with two options, A & B, that could be the next easiest cell depending on the state of the candidates

this stage it should depend on the completed cells and therefore the study should use binary PPPPs. We explain our reasoning below.

Many techniques for solving PPPPs rely on the gradual elimination of candidates from cells until the player can complete a cell. Some of these reasoning chains require the elimination of tens of candidates before a completion can be made. It would be appealing to base the hint system on player notation, as it would best reflect the state of the puzzle; and if the hint system relies on only the completed cells, it will not reflect the state of the puzzle from the player's perspective. The next easiest moves will depend on the progress the player has made so far, and therefore if the state of the player's progress is not understood, the hint system will not provide useful information to the player.

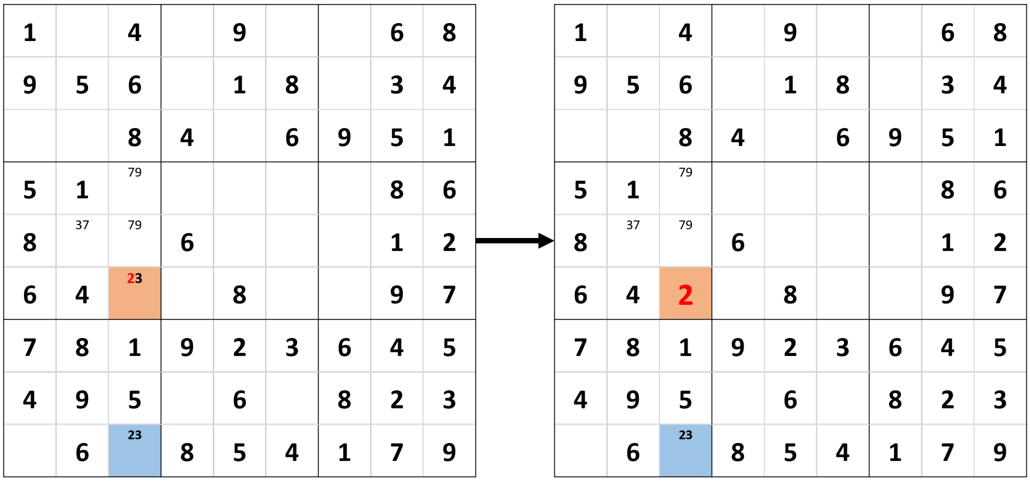
For example, Figure 4.8 shows a simple instance where there are two possible 'next easiest' cells depending on the candidates the player has eliminated (whether mentally or physically), shown in Figure 4.9. The completion of cell B results from a Naked Pair and cell A results from a Pointing Pair, shown in Figure 4.10. Stuart considers Pointing Pairs more challenging than Naked Pairs [140], therefore, if ignoring player notation and using Stuart's ratings, cell B would be highlighted to the player. However, if the player had successfully performed the pointing pair first (which is likely as only 1/21 participants in the study discussed in Chapter 3 did the naked pair in this puzzle first) the hint system would guide them away



1		4		9			6	8
9	5	6		1	8		3	4
		8	4		6	9	5	1
5	1	2379					8	6
8	37	379	6				1	2
6	4	23		8			9	7
7	8	1	9	2	3	6	4	5
4	9	5		6		8	2	3
	6	3	8	5	4	1	7	9

1		4		9			6	8
9	5	6		1	8		3	4
		8	4		6	9	5	1
5	1	2379					8	6
8	37	379	6				1	2
6	4	23		8			9	7
7	8	1	9	2	3	6	4	5
4	9	5		6		8	2	3
	6	3	8	5	4	1	7	9

(a) The state of the candidates in the puzzle produce a naked single in r9c3, resulting in its completion with 3



1		4		9			6	8
9	5	6		1	8		3	4
		8	4		6	9	5	1
5	1	79					8	6
8	37	79	6				1	2
6	4	23		8			9	7
7	8	1	9	2	3	6	4	5
4	9	5		6		8	2	3
	6	23	8	5	4	1	7	9

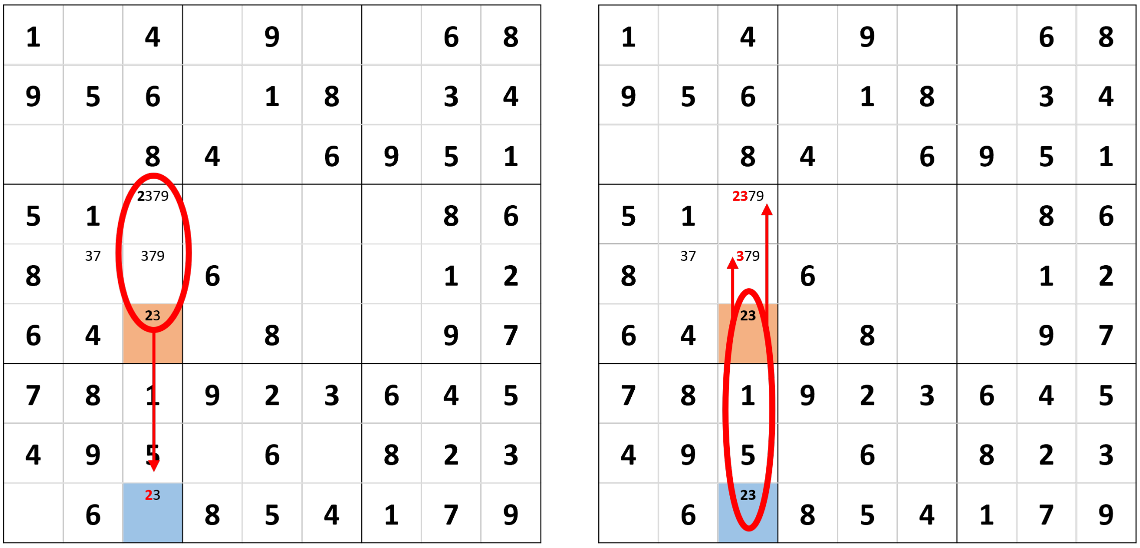
1		4		9			6	8
9	5	6		1	8		3	4
		8	4		6	9	5	1
5	1	79					8	6
8	37	79	6				1	2
6	4	2		8			9	7
7	8	1	9	2	3	6	4	5
4	9	5		6		8	2	3
	6	23	8	5	4	1	7	9

(b) The state of the candidates in the puzzle produce a hidden single in r6c3, resulting in its completion with 2

Figure 4.9: Two different next easiest steps, depending on the candidates eliminated by the player, for the Sudoku shown in Figure 4.8

from the square containing the naked single, shown in Figure 4.9b, and instead direct their attention towards cell B. In order to indicate the next easiest cell, you need to know which of these the player did. This example shows a short chain of eliminations; most deductions come at the end of a much longer chain of eliminations, which allow a wide variety of player states and make it increasingly likely that the hint system will not usefully guide the player, unless it uses the player's notation.

Unfortunately, our first study, described in Chapter 3, demonstrated that player's notation was varied, unsystematic and inconsistent. Unless a player's notation



(a) The deduction leading to the completion of cell A in Figure 4.8, shown in Figure 4.9a

(b) The deduction leading to the completion of cell B in Figure 4.8, shown in Figure 4.9b

Figure 4.10: Two different deductions for the Sudoku shown in Figure 4.8

is understood, the state of the puzzle, from the player's perspective, is unknown. The players may track candidates mentally or on paper, even if we prevent them being used in the interface; the inability to accurately reflect the player's current deductions and resultant puzzle state would be a serious confounding factor in any assessment of the efficaciousness of the novel hint system. It would be hard to differentiate feedback on unhelpful hints from feedback on the hint system itself.

As discussed in Section 2.1.1 there is no benefit to notation in a binary PPPP; as there are only two values, once one has been eliminated the other can be completed. Therefore, we decided to focus on Binary PPPPs for the development and testing of the system. We leave to future work the interpretation of player's notation.

4.2.3.2 Error Handling

In this section we discuss the need for error handling in a system used to assess the novel hint approach, discuss various approaches and conclude that the appropriate error handling system is automatic error highlighting after a short delay.

Producing meaningful hints when the puzzle is in an error state requires remedial hints. Meaningful procedural hints cannot be provided for a PPPP in an error state. Telling the player what the next easiest deduction is for a PPPP in an erroneous state, would be at best unhelpful and at worst would actively mislead the player down an incorrect path. Hints that mislead the player are, unsurprisingly, resented by players and considered worse than no assistance. This means the hint system would not provide any hints when the puzzle was in an erroneous state, which could be used by the player as an indicator of error. Players requesting hints as a way to check for error, would be a serious confounding factor in the data (as we couldn't interpret whether a hint was requested in order to have a hint or in order to detect errors), therefore, explicitly indicating error to the player is preferable.

Errors could be highlighted immediately to a player, after a short delay, or only on request. Providing error highlighting only on request would be consistent with Wauck *et al*'s work on hints [155], which indicated that players prefer to control when they are provided with assistance. However, once the puzzle is in an error state the player makes deductions based on the erroneous state, therefore this can lead to a mixture of correct (by-chance) and erroneous cell completion. If the erroneous cells are removed, it may be hard for the user to recover 'flow' and continue with the puzzle[94]; the puzzle will be in a strange state where information has been built without method, and the player's mental model of the puzzle will need to be updated. We found, in the study discussed in Chapter 3, that players often preferred to start again rather than recover from error. Resetting the puzzle to the state prior to the first error, would avoid the mixture of cells found by chance during the erroneous path, but would not avoid the player needing to rebuild their mental model. More importantly, the player could still use the hint system as an alternative error detection system, and might choose to in order to attempt to find the errors themselves, which would, as stated above, be a serious confounding factor in the data.

Immediately highlighting errors might frustrate players who had incorrectly interacted with the interface, either by misclicking or mistyping. We expect players would feel 'punished' and discouraged if they are told they have made a mistake that they know is an error and is a brief physical rather than mental error.

We decided on highlighting erroneous squares after a brief delay to allow the player to correct physical errors themselves. Highlighting the erroneous squares is the best compromise between the needs of the player and the experimental goals. We leave the best method of error handling for future work.

4.3 Modelling problems for DEMYSTIFY

Our previous study, discussed in Chapter 3, demonstrated that players do not always make the easiest move expected by guides. Therefore, we use DEMYSTIFY [38] to classify the challenge presented by all possible cell completions; allowing us to indicate to the user which cell completions the guides would assess as similar in challenge to the next easiest move. We use the size of the explanation sets returned by the DEMYSTIFY Python library [38] to provide this assessment.

As briefly discussed in Section 2.2.3, the design of the models used in DEMYSTIFY impacts the explanation sets returned. For example, if you use the standard Sudoku model, shown in Figure 2.16, the resulting explanations are not a good match. The constraints are too good, Naked and Hidden Singles/Pairs/Triples/Quads are all the same difficulty which no Sudoku guide agrees with [39]. Optimising models for explanations is likely to be an interesting future research area, as optimising constraint models for speed and efficiency has been for many years [39].

DEMYSTIFY expanded ESSENCE to facilitate human-readable explanations. This was achieved by adding to the language a set of 'annotations': [39]:

`$#VAR` : A variable (or matrix of variables) that must be assigned a value (by the player) for the puzzle to be completed [39].

`$#CON` : A constraint of the problem. Each constraint is given an English description, which is displayed when the constraint is required while solving the puzzle.

For example, in the Sudoku DEMYSTIFY model (Figure 4.11), the `col_contains` constraint, Lines 20 and 21, attached the explanation "Column {a[0]} must contain a {a[1]}" where {a[0]} will be replaced with the column number, and {a[1]} will be replaced with the missing digit. The matrix of boolean variables `col_contains` contains a variable for every column and digit combination. The ESSENCE constraint involving the variable matrix, Lines 47 and 48, checks each column for each digit and assigns the result (true if the digit is present, false if it is not) to the corresponding variable in the matrix.

`col_alldiff`, Lines 8 and 9, ensures that digits are not repeated in the same column. It has the attached explanation of "cells ({a[0]},{a[1]}) and ({a[0]},{a[2]}) cannot both be {a[3]} as they are in the same column", where {a[0]} is replaced with the column number, {a[1]} is replaced with the row number of the first cell the digit appears in, {a[2]} is replaced with the row number of the second cell the digit appears in. {a[3]}

```
letting D be domain int(1..9)
letting C be domain int(0..2)
given fixed : matrix indexed by [D,D] of int(0..9)

$#VAR grid
find grid : matrix indexed by [D,D] of D

$#CON col_alldiff "cells ({a[0]},{a[1]}) and ({a[0]},{a[2]}) cannot both be {a[3]} as they are in the same column"
find col_alldiff: matrix indexed by [D,D,D,D] of bool

$#CON row_alldiff "cells ({a[0]},{a[1]}) and ({a[0]},{a[2]}) cannot both be {a[3]} as they are in the same row"
find row_alldiff: matrix indexed by [D,D,D,D] of bool

$#CON box_alldiff "cells ({3*int(a[0])+int(a[2])},{3*int(a[1])+int(a[3])}) and ({3*int(a[0])+int(a[4])},{3*int(a[1])+int(a[5])}) cannot both be {a[6]} as they are in the same box"
find box_alldiff: matrix indexed by [C,C,C,C,C,C,D] of bool

$#CON row_contains "Row {a[0]} must contain a {a[1]}"
find row_contains: matrix indexed by [D,D] of bool

$#CON col_contains "Column {a[0]} must contain a {a[1]}"
find col_contains: matrix indexed by [D,D] of bool

$#CON box_contains "The box starting at ({1+3*int(a[0])},{1+3*int(a[1])}) must contain a {a[2]}"
find box_contains: matrix indexed by [C,C,D] of bool

such that
    forAll i,j: D.
        fixed[i,j] != 0 -> grid[i,j]=fixed[i,j],

    forAll i:D.
        forAll j1,j2:D. j1 < j2 ->
            forAll d:D. col_alldiff[i,j1,j2,d] -> !(grid[i,j1] = d /\ grid[i,j2]=d),

    forAll j:D.
        forAll i1,i2:D. i1 < i2 ->
            forAll d:D. row_alldiff[j,i1,i2,d] -> !(grid[i1,j] = d /\ grid[i2,j]=d),

    forAll a,b,i1,j1,i2,j2:C. (i1*3+j1) < (i2*3+j2) ->
        (
        forAll d:D. box_alldiff[a,b,i1,j1,i2,j2,d] ->
            !(grid[1+3*a+i1,1+3*b+j1] = d /\ grid[1+3*a+i2,1+3*b+j2] = d)
        ),

    forAll i,d:D.
         row_contains[i,d] -> or([grid[i,j]=d | j:D]),

    forAll i,d:D.
         col_contains[i,d] -> or([grid[j,i]=d | j:D]),

    forAll a,b:C. forAll d:D.
        (
            box_contains[a,b,d] ->
               or([grid[1+3*a+i,1+3*b+j] = d | i : C, j : C])
        )

```

Figure 4.11: A Sudoku model in DEMYSTIFY [38]

is replaced with a digit. The ESSENCE constraint involving the variable matrix, Lines 30 to 32 checks if the digit appears twice in a column and assigns the result (false if the digit repeats, true if it does not) to the corresponding variable in the matrix.

The same approach is used in Figure 4.11 to ensure rows and boxes are correctly constrained.

`$#AUX` is not used in any of our models. It indicates that a variable should not be included in explanations visible to the player. Further information can be found in the paper by Espasa *et al* [39].

All the annotations are applied to ESSENCE variables. DEMYSTIFY models include a greater number of variables than their equivalent ESSENCE model, in order to facilitate better explanations and a better match to human understanding.

4.3.1 DEMYSTIFY Sudoku Modelling Case Study

Figure 4.11 shows the DEMYSTIFY model, discussed in Section 4.3, of a standard Sudoku puzzle, as presented in [39].

The techniques described in below are introduced Section 2.1.2.

4.3.1.1 Naked Single in DEMYSTIFY.

The Naked Single is the observation that a cell which can only take a single value must take that value. This is a Minimal Unsatisfiable Set (MUS) of size 0 (once the critical constraint is removed) because in constraints it is implicit that once a variable can only take one value they must be assigned to that value.

4.3.1.2 Hidden Single in DEMYSTIFY.

The Hidden Single that can be used to solve the highlighted cell in Figure 4.13 has a MUS consisting of a single constraint: one of the constraints defined on Lines 50 to 54 in Figure 4.11. The human readable explanation is defined on Line 23 and for the example shown in Figure 4.13 would be "The box starting at (1,1) must contain a 1".

4.3.1.3 Naked Pair in DEMYSTIFY.

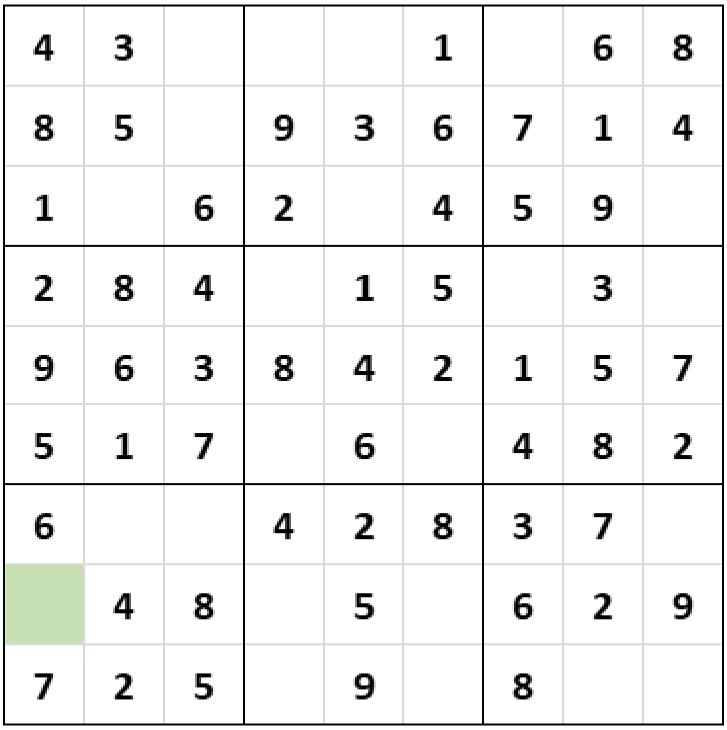
The Naked Pair shown in r5c9 and r6c9 in Figure 4.14 allows the elimination of three candidate 9s. Each elimination requires an individual step in DEMYSTIFY. We will use the candidate in r9c9 as an example. DEMYSTIFY will create an unsolvable problem by adding the constraint $r9c9 = 9$. Then it will look for MUSes and find MUS: $\{r9c9 = 9, !(r5c9 = 9 \land r9c9 = 9), !(r6c9 = 9 \land r9c9 = 9), !(r5c9 = 5 \land r6c9 = 5)\}$. $!(r5c9 = 9 \land r6c9 = 9)$ is not needed in the MUS, and its addition would result in the set not being minimal, as it could be removed without rendering the set satisfiable.

The current domain of r5c9 and r6c9 is {5,9}. The MUS is unsatisfiable as $r9c9 = 9$ combined with $!(r5c9 = 9 \land r9c9 = 9)$ means that r5c9 cannot be 9. $r9c9 = 9$ combined with $!(r6c9 = 9 \land r9c9 = 9)$ means that r6c9 cannot be 9. $!(r5c9 = 5 \land r6c9 = 5)$ means that both cannot be 5, which would leave r5c9 or r6c9 without a possible assignment.

The explanations associated with the constraints in the MUS, excluding the one introduced to make the problem unsatisfiable ($r9c9 = 9$), are returned. All the constraints used in the MUS are setup on Lines 34 to 36 in Figure 4.11, which uses *row_alldiff*, defined on Line 12. Resulting in a set of explanations: {"cells(5,9) and (9,9) cannot both be 9 as they are in the same row","cells(6,9) and (9,9) cannot both be 9 as they are in the same row","cells(5,9) and (6,9) cannot both be 5 as they are in the same row"}.

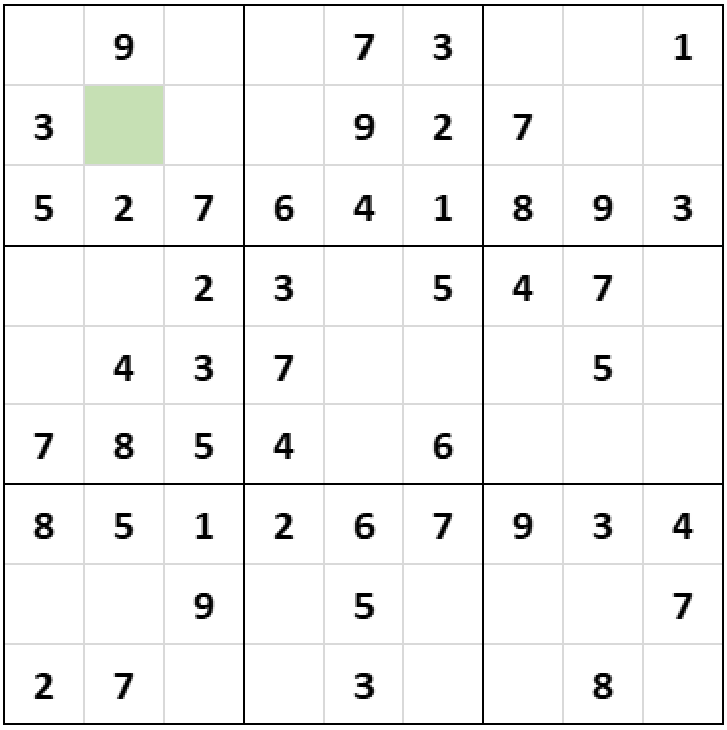
4.3.2 Further Models For DEMYSTIFY

We used the Binairo model (Figure 4.15) developed for the DEMYSTIFY paper by Espasa *et al* [38]. The Aquarium model, Figure 4.16, we developed for the study. It went through several iterations before it produced hints that approximated player moves. There was a significant challenge in ensuring the flow of water along a row was calculated as appropriate MUS difficulties. At the moment the design of models for DEMYSTIFY is more of an art than a science, therefore in this section a best attempt was made to produce a model which followed the intuition of the researchers and Aquarium guides [39]. An interesting area of future work is to automate the process of producing high quality models.



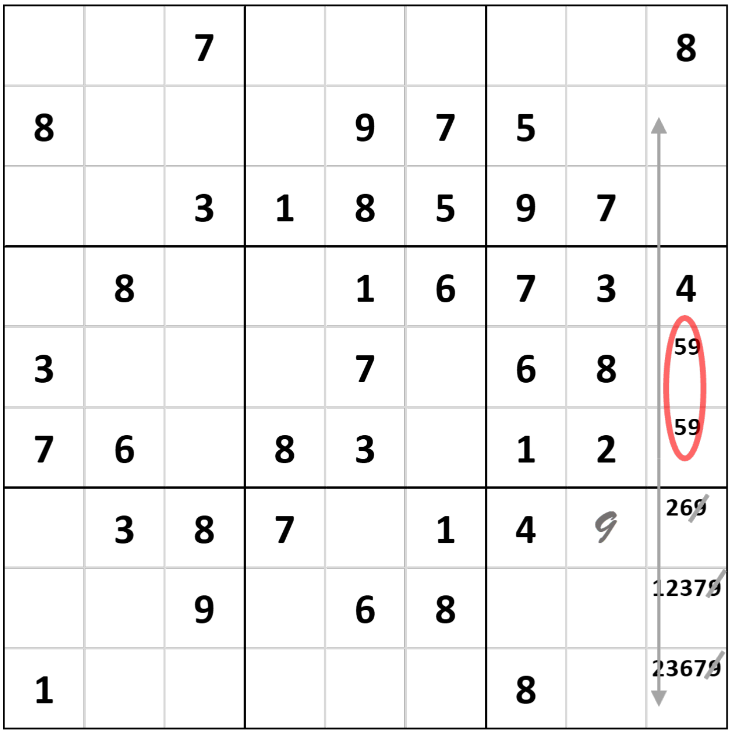
4	3				1		6	8
8	5		9	3	6	7	1	4
1		6	2		4	5	9	
2	8	4		1	5		3	
9	6	3	8	4	2	1	5	7
5	1	7		6		4	8	2
6			4	2	8	3	7	
🟩	4	8		5		6	2	9
7	2	5		9		8		

Figure 4.12: Sudoku puzzle, the highlighted cell can be solved using a Naked Single



	9			7	3			1
3	🟩			9	2	7		
5	2	7	6	4	1	8	9	3
		2	3		5	4	7	
	4	3	7				5	
7	8	5	4		6			
8	5	1	2	6	7	9	3	4
		9		5				7
2	7			3			8	

Figure 4.13: Sudoku puzzle, the highlighted cell can be solved using a Hidden Single



		7						8
8				9	7	5		
		3	1	8	5	9	7	
	8			1	6	7	3	4
3				7		6	8	59
7	6		8	3		1	2	59
	3	8	7		1	4	9	26~~9~~
		9		6	8			1237~~9~~
1						8		2367~~9~~

Figure 4.14: Sudoku Puzzle, showing a Naked Pair in r5c9 and r6c9, the resulting elimination of the candidate nines in c9, revealing a hidden single in r7c8.

4.3.2.1 Aquarium DEMYSTIFY model

The Aquarium model is shown in Figure 4.16 . `given` indicates a variable that changes for each instance of the problem (i.e. a level of a PPPP), these are passed to the ESSENCE model as a parameter file. Lines 1 to 4 of Figure 4.16 show the parameters of the Aquarium model. Aquariums vary in size, therefore unlike Sudoku, the grid size is provided as a parameter, Line 1. The number of separate aquarium regions is provided, Line 2, and the arrangement of the aquariums within the grid, Line 3. The regions are indicated by numbers in a matrix, as shown in Figure 4.17.

`griddim`, Line 6, is used for convenience to represent the numbers 1 to the size of the grid later in the model.

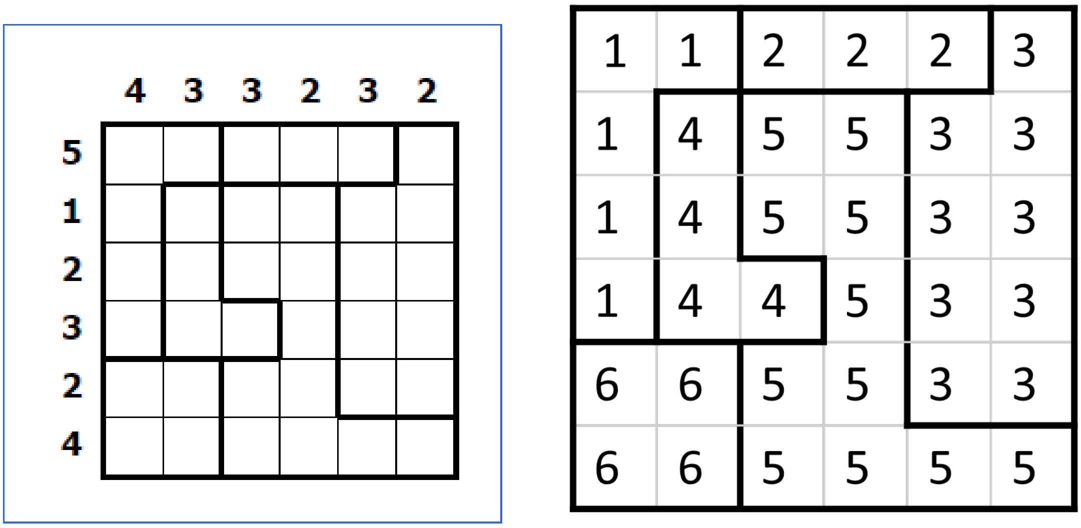
`find` is used to indicate variables with unknown values that must be assigned to find a solution. The DEMYSTIFY model has more `find` statements than an equivalent plain ESSENCE model would. Lines 13, 21, 28 and 31 are introduced to allow the generation of explanations. DEMYSTIFY interprets them as constraints on the problem, as indicated by `$#CON` on Lines 11, 12, 19, 20, 27 and 30. `rowup` and

```
given n: int
letting half = n/2
letting ndim be int(1..n)
letting ndim2 be int(1..n-2)
given initial: matrix indexed by [ndim, ndim] of int(0,1,2)
$ 0: black, 1: white, 2=empty

$#VAR grid
find grid: matrix indexed by [ndim, ndim] of bool

$#CON rowwhite "row {a[0]} must be at least half white"
find rowwhite: matrix indexed by [ndim] of bool
$#CON rowblack "row {a[0]} must be at least half black"
find rowblack: matrix indexed by [ndim] of bool
$#CON colwhite "col {a[0]} must be at least half white"
find colwhite: matrix indexed by [ndim] of bool
$#CON colblack "col {a[0]} must be at least half black"
find colblack: matrix indexed by [ndim] of bool

$#CON rowmatchwhite "row {a[0]} cannot have three white starting at {a[1]}"
find rowmatchwhite: matrix indexed by [ndim, ndim2] of bool
$#CON rowmatchblack "row {a[0]} cannot have three black starting at {a[1]}"
find rowmatchblack: matrix indexed by [ndim, ndim2] of bool
$#CON colmatchwhite "col {a[0]} cannot have three white starting at {a[1]}"
find colmatchwhite: matrix indexed by [ndim, ndim2] of bool
$#CON colmatchblack "col {a[0]} cannot have three black starting at {a[1]}"
find colmatchblack: matrix indexed by [ndim, ndim2] of bool

$#CON alldiffrow "rows {a[0]} and {a[1]} must be different"
find alldiffrow: matrix indexed by [ndim, ndim] of bool
$#CON alldiffcol "cols {a[0]} and {a[1]} must be different"
find alldiffcol: matrix indexed by [ndim, ndim] of bool

such that
forAll i,j: ndim.
    ((initial[i,j] = 1 -> grid[i,j])/\(initial[i,j] = 0 ->!grid[i,j])),

forAll i: ndim.
    rowwhite[i] -> sum([toInt(grid[i,j]) | j : ndim]) >= half,
forAll i: ndim.
    rowblack[i] -> sum([toInt(!grid[i,j]) | j : ndim]) >= half,
forAll i: ndim.
    colwhite[i] -> sum([toInt(grid[j,i]) | j : ndim]) >= half,
forAll i: ndim.
    colblack[i] -> sum([toInt(!grid[j,i]) | j : ndim]) >= half,


forAll i:ndim. forAll j: ndim2.
    (rowmatchwhite[i,j] -> !(grid[i,j] /\ grid[i,j+1] /\ grid[i,j+2])),
forAll i:ndim. forAll j: ndim2.
    (rowmatchblack[i,j] -> !(!grid[i,j] /\ !grid[i,j+1] /\ !grid[i,j+2])),
forAll i:ndim. forAll j: ndim2.
    (colmatchwhite[i,j] -> !(grid[j,i] /\ grid[j+1,i] /\ grid[j+2,i])),
forAll i:ndim. forAll j: ndim2.
    (colmatchblack[i,j] -> !(!grid[j,i] /\ !grid[j+1,i] /\ !grid[j+2,i])),

forAll i,j: ndim. (i!=j) ->
    (alldiffrow[i,j] -> exists k : ndim. grid[i,k] != grid[j,k]),
forAll i,j: ndim. (i!=j) ->
    (alldiffcol[i,j] -> exists k : ndim. grid[k,i] != grid[k,j])
```

Figure 4.15: A Binairo model in Demystify [38]

```
given grid: int
given numaquariums: int
given aquariums: matrix indexed by [griddim, griddim] of int(1..numaquariums)
given rowsums, colsums: matrix indexed by [griddim] of int(1..grid)

letting griddim be domain int(1..grid)

$#VAR water
find water: matrix indexed by [griddim, griddim] of bool

$#CON rowup "at least {params[’rowsums’][a[0]]} water in row ({a[0]})"
$#CON rowdown "at most {params[’rowsums’][a[0]]} water in row ({a[0]})"
find rowup, rowdown: matrix indexed by [griddim] of bool
such that
    forAll i: griddim.
        rowup[i] -> (sum([toInt(water[i,j]) | j : griddim]) >= rowsums[i]),
    forAll i: griddim.
        rowdown[i] -> (sum([toInt(water[i,j]) | j : griddim]) <= rowsums[i])
$#CON colup "at least {params[’colsums’][a[0]]} water in col ({a[0]})"
$#CON coldown "at most {params[’colsums’][a[0]]} water in col ({a[0]})"
find colup, coldown: matrix indexed by [griddim] of bool
such that
    forAll i: griddim.
        colup[i] -> (sum([toInt(water[j,i]) | j : griddim]) >= colsums[i]),
    forAll i: griddim.
        coldown[i] -> (sum([toInt(water[j,i]) | j : griddim]) <= colsums[i])
$#CON water_flood "If there is water in cell {a}, then there is water inside
     this region, there is water in this row and below this cell in the column"
find water_flood: matrix indexed by [griddim, griddim] of bool

$#CON air_flood "If there is air in cell {a}, then there is air inside this
     region, there is air in this row and above this cell in the column"
find air_flood: matrix indexed by [griddim, griddim] of bool

such that
    forAll i,j: griddim.
        water_flood[i,j] -> (
            (water[i,j]) -> (
            forAll row,col: griddim. ( (aquariums[i,j] = aquariums[row,col] /\
     i <= row /\ (i = row \/ j = col)) -> water[row,col] ))),

    forAll i,j: griddim.
        air_flood[i,j] -> (
            (!water[i,j]) -> (
            forAll row,col: griddim. ( (aquariums[i,j] = aquariums[row,col] /\
     i >= row /\ (i = row \/ j = col) ) -> !water[row,col] )))
```

Figure 4.16: An Aquarium model in DEMYSTIFY developed as part of this thesis



1	1	2	2	2	3
1	4	5	5	3	3
1	4	5	5	3	3
1	4	4	5	3	3
6	6	5	5	3	3
6	6	5	5	5	5

Figure 4.17: An example of the model representation of the aquarium regions. The puzzle is shown on the left, the regions are shown on the right. The borders have been included in the matrix on the right for easy of comparison. They are not represented in the parameter file.

`rowdown` ensure that the total number cells containing water in the row matches the number at the end of the row. They have been separated into two constraints, one indicating insufficient water in the row, the other indicating excessive water in the row. This helps to facilitate better explanations because sometimes only one of these is required. This pattern of splitting equality constraints into two inequalities is a common approach when designing DEMYSTIFY models [39]. Similar reasoning applies to `colup` and `coldown`. Lines 15 and 16 assign the value of the variables in `rowup` to true or false based on whether the sum of the water in a given row is greater than or equal to that row's value in `rowsums`. The equivalent (less than or equal) for `rowdown` occurs on Lines 17 and 18.

`water_flood` and `air_flood`, Lines 27 and 30, ensure that water and air fill the Aquarium correctly. Both fill the row of the Aquarium region they are in, water fills all cells below it in the Aquarium region it is in, and, as water fills *all* cells below it, once a cell contains air all cells above it must be filled with air. The constraint on Lines 34 to 37 ensures that the variables in `water_flood` evaluate to true only if all cells directly below and in the same row (within the same aquarium) are also water. The most difficult part of modelling Aquarium is the water and air constraints. If we model that once one cell is water everything at the same level or below is water, the MUSes did not distinguish between moves of different difficulty. We also attempted a model where if a cell contains water the directly adjacent cells to the left, right and below also had to contain water. However, this resulted in too much discrimination - a cell 3 columns to the left was seen as much harder than the cell directly next to the cell containing water; this was also inappropriate. The final edition where cells in the same row and same column were included in the constraint was consistent with guides for other PPPPs which focus on rows and columns. It was not considered reasonable to run an experiment comparing the models within the scope of this project, therefore the researchers had to pick the model that best matched the existing guides and their own expertise.

Comparing different versions of the DEMYSTIFY models is an interesting area of future research.

4.4 Generating the Hints

A key consideration in the hint generation process was that it needed to be possible to generate all the hints in response to user interaction with the puzzle. Generating hints for every possible valid puzzle state would be unfeasible. Even a small puzzle such as the Binairo shown in Figure 4.18, with only 25 empty cells, has 2^{25} possible valid puzzle states[1]. Pre-generating likely states a user would enter would be useful, and being able to generate hints when a user hits a novel state would be essential.

4.4.1 Hint Grid

In order to generate a grid of hints, we used DEMYSTIFY to calculate the MUSes required to fill a value in each square. For candidate PPPPs, we would calculate the MUSes required to eliminate each candidate and to fill in the final value and take the minimum. However, given the restriction to binary puzzles discussed in Section 4.2.3.1, calculating only the MUSes required to complete the value was a viable optimisation, as eliminating a candidate immediately leads to completing a cell.

---

[1] The cells can be either empty or contain the correct value. The number of possible (including invalid) puzzle states is much larger

	0			0	
0			0		
1			0		1
	1			0	
		0			
1					

Figure 4.18: Example Binairo

4.4.2 Next Cell Hints

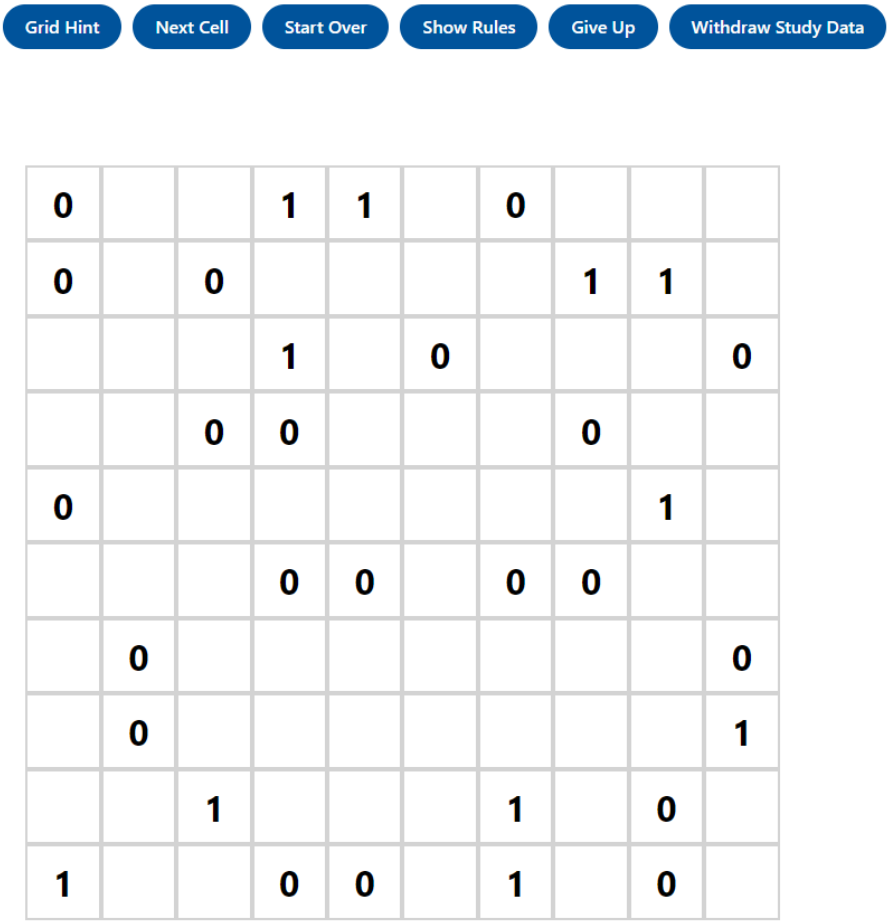
In order to compare our novel hint system with existing hint systems, we needed an equivalent system. We chose to fill in the next 'easiest' square. The definition of 'easiest' needed to match the hint grid system to avoid introducing a confounding factor where one hint system better matched the player's mental model. Therefore, we used DEMYSTIFY to provide the next step - it was asked what the next step would be, and if multiple possible next steps were returned, we selected one at random. This required minor updates to DEMYSTIFY to add an option to prevent the puzzle state from being updated when an explanation was requested.

4.5 Implementation

The system was implemented using Flask 2.0.2 and Python 3.9 for the back-end server, the front-end was implemented using React 17.0.1.

4.5.1 Puzzle Interface

The puzzle interface was built using React 17.0.1 and is shown in Figure 4.19. Users could click through possible values in a given cell. For example, in a Binairo



0			1	1		0			
0		0					1	1	
			1		0				0
		0	0				0		
0								1	
			0	0		0	0		
	0								0
	0								1
		1				1		0	
1			0	0		1		0	

Figure 4.19: The puzzle interface, showing a Binairo puzzle.

puzzle, one click would result in a 0, a second click would result in a 1, a third click would render it blank again etc.

The puzzle grid was reactive, adjusting in size to the display. However, it was not optimised for mobile, which participants were warned about in the information screen. The puzzle had a button for each hint system available to the participant (eg. if they could only use the grid hint system they would only see the grid hint button and not the next cell button). They had the option to start over or give up via buttons at the top of the screen. There was a button to show/hide the game's rules which was available throughout. Finally, there was a button to withdraw their data from the study, this is discussed further in Chapter 5.

4.5.2 Providing hints

The generation of a grid of hints initially took an unacceptably long time. A 2004 study found that users were unwilling to wait longer than 2s [101]. Since then, a 2012 article speaking to several industry experts and found that response times of a quarter of a second reduced user engagement [83]. The hint generation time was initially around 10s, sometimes substantially longer, which was likely to negatively impact participant engagement with the hint systems.

This was partially model-dependent - the Aquarium model consistently took longer than the Binairo model, even its final form.

This section discusses the steps we took both to improve generation speeds and to avoid avoid users having to wait while hints are generated.

4.5.2.1 Serialisation optimisation

We established that setting up a puzzle instance in DEMYSTIFY was a major contributor to the time required to generate hints. Setting up a puzzle instance in demystify requires running the entire ESSENCE toolchain [43] to generate the structures which demystify then uses to calculate MUSes. Serialisation of the resulting DEMYSTIFY puzzle state seemed to be a plausible solution. We serialised it using the Python pickle library [121]. We tested it on 50 Binairo instances and found an improvement of between 6 and 8 seconds. We compared the default and highest protocol, but found no significant difference, therefore we used the highest protocol.

4.5.2.2 Maximum MUS optimisation

Some of the cells in the grid would require very large deductions in order to be able to solve them from the current puzzle state, these took a long time for DEMYSTIFY to find a solution and explanation size. However, as the hard cells were grouped by the hint system, there was no differentiation between cells that had explanation sets one or two larger than the somewhat easy cells, and cells that had explanation sets fifty or more larger than the somewhat easy cells. Therefore, we added a configurable parameter, MUSGIVEUP, to DEMYSTIFY that told it to stop trying to find a MUS if its size was greater than MUSGIVEUP. We then adjusted the hint generation algorithm, so it updated MUSGIVEUP to the smallest explanation size in the ‘hard’ group. This meant processing time wasn’t wasted trying to find

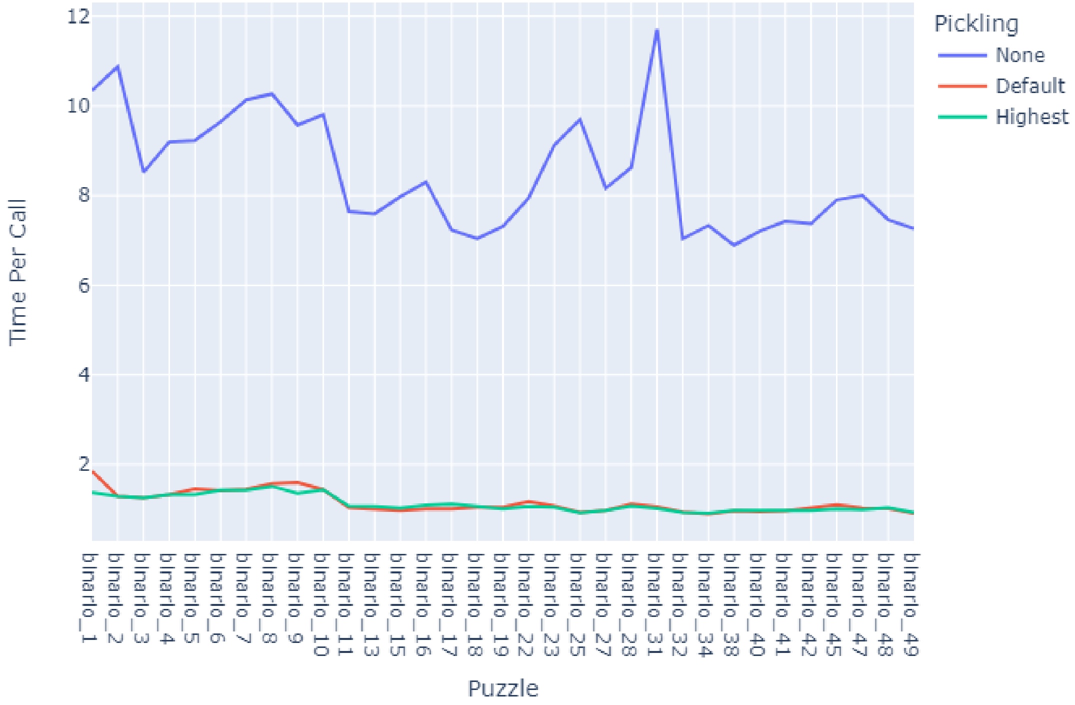


Figure 4.20: Graph comparing the performance of hint generation, without pickling the DEMYSTIFY puzzle object, pickling it with the default protocol, and pickling it with the highest protocol. Tested on 50 Binairo puzzles.

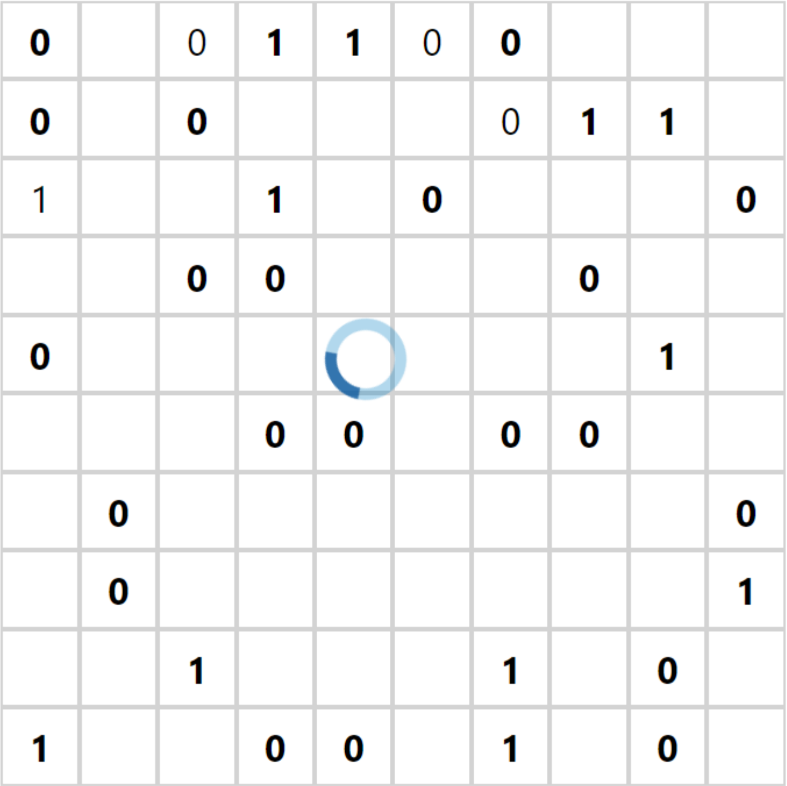
explanations for cells that would be marked as 'hard' either way.

4.5.2.3 Pre-Generating and Caching Hints

In order to reduce the user waiting time, we pre-generated a large number of hints for puzzle states. It was not feasible to generate hints for every possible puzzle state, because as mentioned above, even a relatively small puzzle will have tens of millions of potential puzzle states.

To generate the hints we took the starting puzzle state, generated hints, then filled a value in and repeated the process of generating hints then filling in a value until the puzzle was complete.

Two different methods were used to fill a value in: either a randomly selected cell was filled in or choosing (at random) one of the cells of next easiest difficulty was filled in.



0		0	1	1	0	0			
0		0				0	1	1	
1			1		0				0
		0	0				0		
0								1	
			0	0		0	0		
	0								0
	0								1
		1				1		0	
1			0	0		1		0	

Figure 4.21: Loading Wheel Example

To further reduce waiting times for users, when a user enters a puzzle state where hints have not been generated, hints are generated and stored (whether or not the user requests a hint).

4.6 Changes Post Pilot Study

The design was adjusted and improved following the feedback from initial pilot studies within the research group. The key changes were to do with players finding novel states and requiring new hints to be generated. This caused a noticeable delay, and we found that players pressed the hint buttons repeatedly while waiting for a new hint. We adapted the server to prevent it from spawning additional processes while already generating a set of hints. More significantly, we adapted the front end, adding a loading wheel which reassured the player that it was an expected behaviour, see Figure 5.1. We also disabled the hint buttons while the player was waiting for hints. These changes significantly improved the feedback from the pilot sessions, the players stopped reporting that the interface was broken, and the general feedback became much more positive.

4.7 Summary

In summary, we used the study described in Chapter 3 to provide the motivation and structure of our design of a novel hint system. In this chapter we have discussed the iterative design process by which we arrived at the final design: a light-touch guidance based approach which pointed players towards a range of possible cells which they should find easier to solve than others in the puzzle. We hoped this approach would both enhance the player experience and supported the wide range of player approaches we discovered. We chose to implement the system using DEMYSTIFY as it was the only tool which provided a flexible and generalisable difficulty metric. In the following chapter we will discuss the studies we conducted to compare our novel hint system to an existing approach.

CHAPTER FIVE

ASSESSMENT OF OUR NOVEL HINT SYSTEM

In this chapter, we discuss four experiments carried out to assess whether our novel hint system, discussed in Chapter 4, improves player experience and engagement.

Our key hypotheses were that the novel hint system would, when compared to traditional hint system:

- Feel less like cheating to the player.
- Reduce frustration in the player.
- Improve player experience.

Therefore, the studies were designed to compare the impact on player experience of the novel hint system to a traditional ‘fill-in next square’ hint system.

We also conducted informal pilot studies within the research group prior to the launch of the first major pilot study. The impact of the pilot studies on the design of the interface was discussed in Section 4.6. The impact of the pilot studies on the study design is discussed in part of Section 5.3.

5.1 Selection of Progressive Pen & Paper Puzzle Games (PPPPs) for the Novel Hint System Studies

There are a range of binary PPPPs, many of which are discussed in Section 2.1.1. We wanted to evaluate the hint system using at least two different binary PPPPs which were qualitatively different. We selected Binairo and Aquarium (see Section 2.1.1) and explain their selection in this section. We chose to use binary PPPPs rather than candidate PPPPs, such as Sudoku, due to the challenges with interpreting notation, as discussed in Section 4.2.3.1. It is not possible to provide meaningful assistance to players without an accurate understanding of the state of the puzzle.

We wanted to assess how the novel hint system performs when directing players towards both trivial and complex deductions. Binairo and Aquarium provided a mix of these types of deductions.

Some PPPPs tend to have more local deductions, while others require players to include entire regions of the grid in their reasoning. It was important to assess the hint system with a puzzle that tended toward local deductions and a puzzle that tended toward global deductions. Binairo tends towards local deductions, while Aquarium tends towards more global deductions (with the exception of the flow of air and water). For example, many of Binairo's deductions rely on local information - spotting patterns of 0s and 1s in a column. In contrast, many of Aquarium's deductions rely on the player looking at an entire row or entire column.

We therefore chose Binairo and Aquarium, as they represented a fairly well known binary PPPP and a more obscure one, respectively, while also fulfilling the above criteria.

5.2 Puzzle instances

This section discusses the generation and grading of the instances (levels) of the puzzles used in the assessment studies of the hint system. Binairo instances were generated using the algorithm discussed in Section 5.2.2. The Aquarium puzzles were provided by puzzles-aquarium.com [115]. The expected challenge of the puzzles was graded using DEMYSTIFY, as described below in Section 5.2.1.

5.2.1 Grading the puzzle instances

The probable challenge a puzzle would present was assessed using DEMYSTIFY, Section 2.2.3 and further evaluated using informal pilot studies within the research group. Given the established challenge of computationally grading a puzzle's difficulty, as discussed in Section 2.3.1, the DEMYSTIFY grading was assumed to be a rough estimate.

An accurate difficulty assessment was not essential for the study, as the initial goal was simply to provide puzzles of a difficulty that would encourage players to use a hint without frustrating them so much that they left the study. Evaluation of the efficacy of the difficulty grading was not a goal of these studies.

The puzzles were graded as part of the generation process discussed in Section 5.2.2. They were graded using DEMYSTIFY, shown in Algorithm 1. DEMYSTIFY generates a solving path through the puzzle; it aims for the minimum solving path (every step is as easy as possible), although due to the random nature of the Minimal Unsatisfiable Set (MUS) finding algorithm, it may not find a perfect path.

The deduction for each step has a MUS associated with it. The DEMYSTIFY deduction size is the size of the MUS minus the constraint that rendered it unsatisfiable. Deductions with a size of 1 are classed by DEMYSTIFY as *"simpleDeductions"*. The difficulty assessment of the puzzle counted the total number of steps that required a simple deduction. For more complex deductions, it outputs a list of the sizes of all complex deductions. This allowed the total number of steps, the number of 'simple' steps, and the number and size of 'complex' deductions, to all be used when assessing the probable challenge a puzzle would present. The grading tables for each study show the total number of simple deductions, the total number of complex deductions, and the size of the largest complex deduction. The full list of complex deductions is excluded for ease of interpretation.

The expectation was that more complex deductions, with a higher maximum deduction size would be more challenging; this is the approach used by the current best grading techniques[39, 142, 108].

5.2.2 Building the puzzle instances

This approach for building puzzle instances is designed for puzzles where the set of values that the player completes are the same type as the clues provided in the

Algorithm 1 Puzzle Grading Algorithm

```
procedure CALCULATEDIFFICULTY(puzzleGrid)
    explainer = DEMYSTIFY.Explainer(puzzleGrid)
    solvingSteps = explainer.explain_steps()
    simpleCount = 0
    complexList = [ ]
    for step in solvingSteps do
        simpleCount += count(step['simpleDeductions'])
        complexList.append(step['deductions'])
    return simpleCount, complexList
```

starting state of the puzzle. To use an example, the ‘clues’ in Sudoku are numbers in the grid, and the players fill in the missing numbers. In contrast, in Tents and Trees, the starting state consists of numbers around the edge and tree positions in the grid, while the player fills in grass and tents.

This means that removing or adding values to the puzzle can still (as long as there is still a unique solution) result in a valid starting state. This allows an approach to generating puzzles where a completed version of the puzzle (where every variable is assigned a value) is used as a starting point. A completed version of the puzzle can be generated by providing the puzzle rules and asking for one of the possible solutions to a blank starting state.

To generate instances of the puzzle, we used an algorithm, Algorithm 2, which takes a completed puzzle grid and removes values from cells. As more cells are unassigned, the puzzle will become more challenging; however, it may also gain an extra solution and become invalid. The challenge is determining which subset of cells to unassign in order to produce a puzzle of the difficulty we want, while maintaining a single solution. For each starting puzzle we produced a large number of instances and graded their difficulty - see Section 5.2.1.

Algorithm 2 takes the completed puzzle grid and a configuration parameter (containing the values listed in Table 5.1). It repeats the following process a number of times as defined by the configured variable, *variants*. First, a random cell is removed, as shown on Line 4. The algorithm then checks if the number of empty cells is less than the configured minimum number of empty cells (*minEmptyCells*), and if it is, it recursively calls GENERATEPUZZLES with the new puzzle grid (the grid with the removed cell). The minimum number of empty cells is configured to avoid grading a large number of puzzles with very few values for the player to fill in. A Binairo with only three values available to be completed would be unlikely to engage a player. The algorithm then checks that the grid is not completely empty[1] and exits if it is. Next, it checks whether the current puzzle grid has already been assessed, if it has, it moves on to the next variant.

Otherwise, the algorithm checks that the new puzzle grid still has a unique solution; it will always have a solution, but removing cells might result in multiple solutions. If it has multiple solutions, it moves on to the next variant. Otherwise, it calculates the difficulty, see Section 5.2.1, and checks that the most complex deduction is less than the configured maximum difficulty. The difficulty will always increase when a cell is removed because the puzzles are progressive; which means that each completed cell adds information to the puzzle. If the most complex deduction in the solve path is less than the maximum difficulty, the puzzle grid and difficulty are recorded and added to the global puzzle store. GENERATEPUZZLES is then recursively called with the new puzzle grid (with the removed cell). If the maximum difficulty has been exceeded, the algorithm continues to the next variant.

Algorithm 2 Puzzle Generating Algorithm

```
procedure GENERATEPUZZLES(startingPuzzleGrid, config)
    totalCellCount=calculateTotalCellsInGrid(startingPuzzleGrid)
    for $i = 0; i < config.variants; i++$ do
        puzzleGrid = removeRandomCell(startingPuzzleGrid)
        emptyCells = calculateEmptyCells(puzzleGrid)
        if $emptyCells < config.minEmptyCells$ then
            GeneratePuzzles(puzzleGrid, config)
        else if $emptyCells == totalCellCount$ then return
        else if $puzzleGrid\ in\ config.puzzleStore$ then
            continue
        else
            solution = checkUniqueSolution(puzzleGrid)
            if $solution$ then
                difficulty = CalculateDifficulty(puzzleGrid)
                if $maxDeductionSize(difficulty) <= config.maxDifficulty$ then
                    recordPuzzleAndDifficulty(puzzle_grid, difficulty)
                    GeneratePuzzles(puzzleGrid, config)
```

---

[1] In most puzzles this check will not trigger as a completely empty grid will not be a valid puzzle with a unique solution; but it is included to ensure the algorithm terminates for all puzzles.

variants	The number of different paths to explore at each removal step. High numbers produce a much wider variety of instances but take longer to run.
minEmptyCells	The minimum number of cells that needed to be empty prior to starting to calculate and record the puzzle's difficulty
maxDifficulty	The maximum acceptable size for a complex deduction in the puzzle, see Section 5.2.1 for further explanation
puzzleStore	The global store of tested puzzles, used to avoid grading puzzle instances that had already been recorded.

Table 5.1: Definitions of the values contained in the $config$ parameter of Algorithm 2

Table 5.2

5.3 Study Procedure

This section describes the general procedure for the four studies described later in this chapter. The specifics of each study's design are described in the individual study's design section.

The studies were conducted as unsupervised online experiments. The study design was moved to unsupervised online experiments as a result of the 2020 Covid-19 pandemic and associated restrictions. This change provided the benefits of: the quantitative experiment allowed for larger participant samples, and participating online allowed the possibility of a more diverse group of participants.

Participants participated via a website and their data were collected entirely anonymously. Anonymity was important both for data retention and for encouraging participation. Any free text boxes with identifying information will be removed before the data is made available in a repository.

Consent was indicated via button click. An example of the information and consent page is included in Appendix B. The website was built as an embedded application and later views[2] could not be accessed directly, ensuring that participants entered through the information and consent page. The participants had the option to withdraw their data at any point, which was done via a button in the puzzle interface. The information page at the start of the study explained (prior to consent) that this was the only way to withdraw their data and partially completed entries

---

[2] Different screens within the application

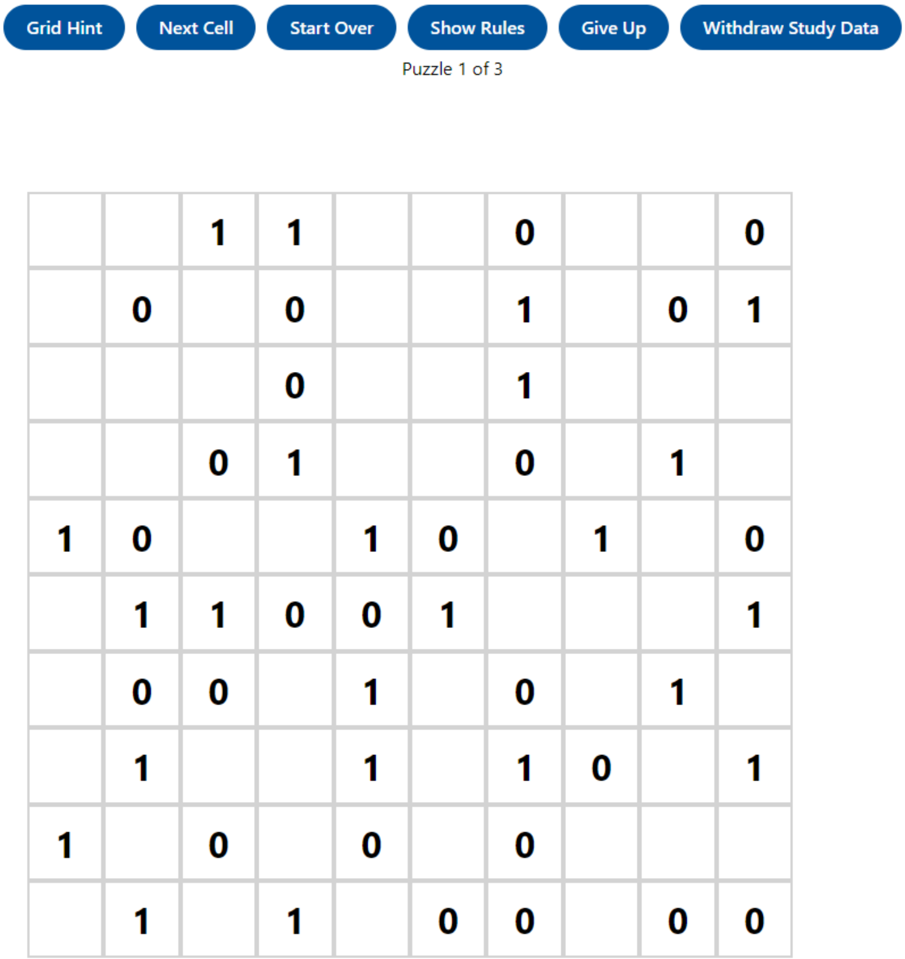
would be retained and analysed, unless explicitly withdrawn.

Simple demographic data was collected at the start of the study, and participants were asked if they had participated in the study already (and, in later studies, if they had participated in previous studies). As the data was entirely anonymous, asking if players had already participated was the simplest approach to allow the exclusion of repeat participants. Participants could of course lie, but all other approaches we considered either deanonymised the data or were equally unreliable. The pre-study questionnaires are available in Appendix A.3.1.

The players were then presented with a number of puzzles (the number varied between studies). Puzzles were presented using the interface discussed in Section 4.5.1. All data from all the puzzles was processed and analysed, as described in the information and consent page. The first puzzle was labelled 'Practice Puzzle', the remainder were labelled *X* of *Y* where *Y* was the total number of puzzles, and *X* the number of the puzzle, for example, puzzle 2 of 3, see Figure 5.1.

After each puzzle, they were presented with a questionnaire that assessed their experience with the puzzle. This questionnaire was heavily redesigned between the first Binairo pilot study (discussed in Section 5.5) and subsequent studies. The reason for the redesign is discussed in detail in Section 5.5.5. In summary, which hint systems participants used was found to be more important than which hint system they had access to; therefore we moved from a between group study to a within group study design, where all participants had access to both hint systems. Furthermore, the questionnaires were redesigned to be explicit about the hint system rather than indirectly assessing them via the player's experience of the puzzle. The initial experience assessment element of the questionnaire was based on the Game Experience Questionnaire [63]. We assessed competence using GEQ-Core 2 and 15, flow using GEQ-Core 5, annoyance using GEQ-Core 22 and 29, challenge using GEQ-Core 26, negative affect using GEQ-Core 8, and positive affect using GEQ-Core 20. The complete questionnaire is included in Appendix A.3.2.1. This was redesigned after the first study, as discussed in Sections 5.5.5 and 5.6.1.

During the puzzle solving stage, participants' interactions were recorded. All participant clicks on puzzle cells were recorded, all requests for hints were recorded, every time the interface highlighted an error, it was recorded, and



		1	1			0			0
	0		0			1		0	1
			0			1			
		0	1			0		1	
1	0			1	0		1		0
	1	1	0	0	1				1
	0	0		1		0		1	
	1			1		1	0		1
1		0		0		0			
	1		1		0	0		0	0

Figure 5.1: An example of the final interface with the practice puzzle labelled.

if the participant gave up on the puzzle it was recorded. We did not record if they opened the puzzle rules.

Participants were provided with access to one or both hint systems (in all except the first study, Section 5.5, participants had access to both).

5.4 Participant Recruitment

Participants were recruited online via social media. They were encouraged to share the link with other people in the hope of snowballing recruitment.

Participants were not offered a reward for participating in the study. This allowed the participation to be completely anonymous. This also eliminated any confounding factors that might be introduced by external reward motivation.

5.5 Binairo Pilot Study

This section discusses the first study assessing the novel hint system. The study used Binairo puzzles to test the hint system.

5.5.1 Binairo Pilot Study Design

Our generic study design is described in Section 5.3; in this section we discuss the specific design decisions for this study. This study was designed as a between-group comparison of the novel hint system, the traditional hint system, and the use of both systems. Participants were randomly assigned to one of the three conditions.

Participants were asked to solve six Binairo instances of varying sizes and different predicted challenge levels. After the last puzzle (and its associated questionnaire), participants were shown the debrief screen, and were offered the option to continue solving puzzles. There were eleven puzzles available after the debrief screen; if participants completed all of them, they were shown a final "End" screen. The additional puzzles were included due to the difficulty in accurately assessing the challenge a puzzle presents and participants’ variation in skill; the additional puzzles were intended to allow further data to be collected from skilled participants without pressuring other participants into participating beyond the expected study time.

The grading of the puzzles is discussed in Section 5.2.1.

5.5.2 Participant Demographics of Binairo Pilot Study

Participants were asked their age and gender at the start of the experiment. 269 participants started the study and did not withdraw their data. Of those 269 participants, 223 started the first puzzle and 137 completed the six puzzles that appeared before the debrief screen. The age distribution of the 223 participants who started the puzzles is shown in Figure 5.6 and their gender distribution is shown in Figure 5.7.

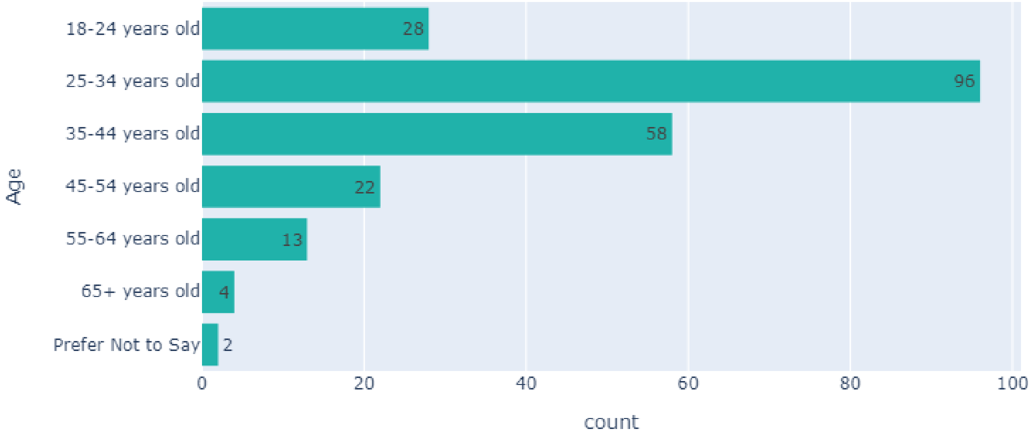


Figure 5.2: The Age distribution of participants in the Binairo Pilot Study. Blank responses have been combined with Prefer not to Say.

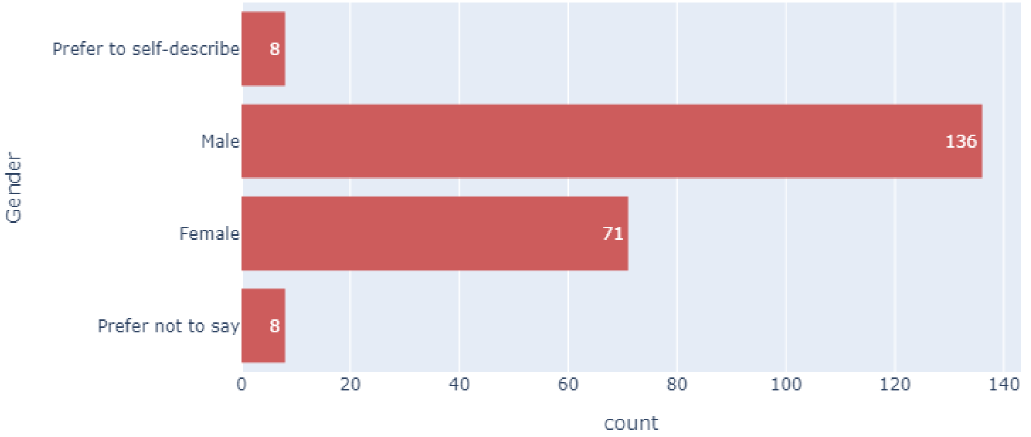


Figure 5.3: The Gender distribution of participants in the Binairo Pilot Study. Blank responses have been combined with Prefer not to Say.

Puzzle	Size	Total Simple Deductions	Total Complex Deductions	Max Size of Complex Deduction	ID
Practice	6 by 6	43	0	N/A	206
1	8 by 8	78	1	2	203
2	6 by 6	40	5	2	208
3	10 by 10	76	1	2	212
4	8 by 8	61	4	4	201
5	8 by 8	53	10	2	205
Debrief Screen					
a	10 by 10	109	5	2	211
b	12 by 12	86	1	2	217
c	8 by 8	49	9	4	202
d	8 by 8	50	0	N/A	204
e	6 by 6	41	1	2	207
f	6 by 6	30	5	4	209
g	10 by 10	97	7	3	210
h	10 by 10	104	0	N/A	213
i	12 by 12	113	9	3	214
j	12 by 12	128	0	N/A	215
k	12 by 12	135	7	2	216

Table 5.3: The puzzles used in the Binairo pilot study, in the order they were presented to the participants. See Section 5.2.1 for the explanation of simple and complex deductions.

5.5.3 Results of Binairo Pilot Study

We ran a linear mixed-effects model for the experience assessment statements (which asked participants to rate their agreement), using the 'statsmodel' package in python [133], with *condition* and *puzzle* as fixed effects and a random effect of *participant*. We mapped responses to numeric values, as shown in Table 5.4.

For most of the experience statements, puzzle ID had a strong effect on responses (the full results are shown in Appendix C.1), while the effect of the study condition was nonsignificant. All models are presented without interaction effects for ease of interpretation.

Text	Numeric Value
Strongly disagree	0
Somewhat disagree	1
Neither agree nor disagree	2
Somewhat agree	3
Strongly agree	4

Table 5.4: Mapping of 5-point Likert scale showing agreement to numeric values.

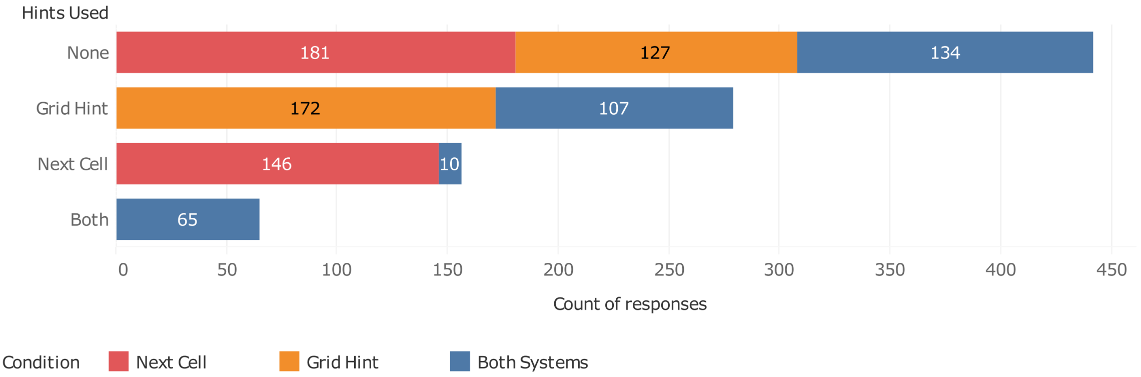


Figure 5.4: The number of responses that used either no hint system, just the next cell system, the novel hint system or both hint systems on a puzzle. The coloured sections indicate which study condition the participant giving the response was under and therefore which hint systems they had access to.

We also ran a linear mixed-effects model for the experience assessment statements, with the R package 'lme4' [10], with *puzzle* as a random effect shown in Appendix C.2. This did not have an impact on the significance of the effect of the study condition.

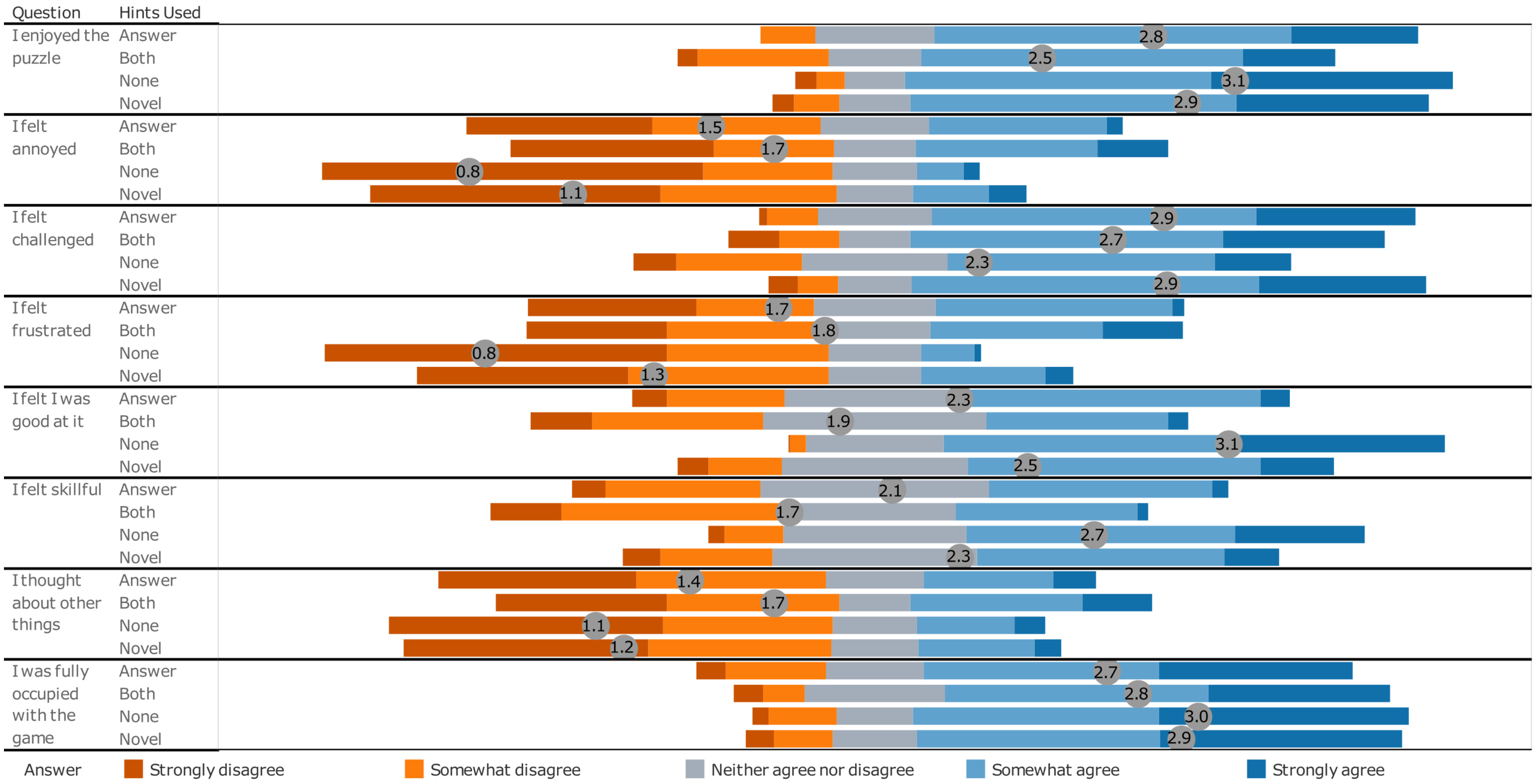


Figure 5.5: Likert visualisation of the responses to the experience assessment matrix for the Binairo pilot Study: A diverging stacked bar chart, centred around the centre of the neutral (Neither Agree nor Disagree) responses. There is a bar for each statement for each hint system that participants used (not the study condition they were under). Each sub-section of the bar matches a possible level of agreement with the statement on the left (see legend below the chart). The size of the subsections corresponds to the percentage of responses that gave that answer for a given statement and condition.

5.5.3.1 Analysis of Hint Systems Used

Finally, we ran a linear mixed-effects model of the experience assessment statements, using the same R package, with the hint system the participants used rather than their study condition. The number of participants that used each hint system(s) is shown in Figure 5.4. The % agreement ratings, and their mean value (calculated from numeric mappings of the statements of agreement, as shown in Table 5.4) are shown in Figure 5.5. *Participant* and *puzzle* were treated as random effects. This analysis showed significant effects; summarised in Table 5.5. There were significant differences between the traditional and novel systems for the statements "I felt frustrated", "I felt annoyed", and "I felt skillful". The responses expressed less agreement with the statement "I felt frustrated" when they had used only the novel hint system than when they had only used the traditional (next cell) system. Responses also expressed less agreement with the statement "I felt annoyed" when they had used only the novel hint system than when they had only used the traditional (next cell) system. The responses expressed greater agreement with the statement "I felt skillful" when they had used only the novel hint system than when they had only used the traditional (next cell) system. There were also some significant differences between responses that used the traditional system and those that used neither system. The responses indicated greater agreement with the statements "I felt I was good at it", "I enjoyed the puzzle", "I felt skillful", and "I was fully occupied with the game" when neither system had been used on the puzzle than when the traditional hint system had been used. The responses indicated less agreement with the statements "I felt frustrated" and "I felt annoyed" when neither system had been used on the puzzle than when the traditional hint system had been used. The responses indicated less agreement with the statement "I felt I was good at it" when they used both systems than when they used only the traditional hint system.

5.5.4 Discussion of Binairo Pilot Study

No significant differences were found in player experience between the three study conditions: access to both hint systems, access to only the traditional hint system, and access to only the novel hint system. However, when the analysis was run again, looking at the effect of which hint system(s) were reported to be *actually used* when solving the puzzle[3], several significant differences were observed, which

---

[3] As opposed to the systems they had access to

Statement	Hint System Used	β	Std. Err	t-value
I felt frustrated	Both	0.021	0.190	0.109
	Neither System	-0.642	0.121	-5.303
	Only Novel Hint System	-0.386	0.141	-2.744
I enjoyed the puzzle	Both	-0.240	0.139	-1.729
	Neither System	0.471	0.088	5.330
	Only Novel Hint System	0.171	0.103	1.664
I felt I was good at it	Both	-0.339	0.141	-2.405
	Neither System	0.841	0.089	9.413
	Only Novel Hint System	0.193	0.104	1.854
I felt annoyed	Both	0.234	0.187	1.256
	Neither System	-0.579	0.119	-4.880
	Only Novel Hint System	-0.500	0.139	-3.594
I felt skillful	Both	-0.288	0.155	-1.857
	Neither System	0.749	0.098	7.669
	Only Novel Hint System	0.264	0.114	2.314
I felt challenged	Both	-0.121	0.169	-0.716
	Neither System	-0.139	0.109	-1.276
	Only Novel Hint System	0.034	0.123	0.276
I was fully occupied with the game	Both	-0.059	0.152	-0.384
	Neither System	0.317	0.096	3.309
	Only Novel Hint System	0.021	0.114	0.185
I thought about other things	Both	0.256	0.188	1.356
	Neither System	-0.040	0.119	-0.335
	Only Novel Hint System	-0.040	0.141	-0.285

Table 5.5: The results of the Linear Mixed-Effects Model, with the experience ratings of participants that used only the traditional hint system as the reference category, and puzzle and participant as random effects. Showing the effects of participants only using the novel hint system, using neither hint system, and using both systems. T values less than -1.96 or greater than 1.96 are highlighted and considered significant.

are discussed below. However, participants were restricted by the study condition under which they participated. Participants under the study condition of both could use both systems, just one, or neither; whereas participants under the other conditions could only use the hint system to which they had access, or no hint system.

The statements "I felt frustrated" and "I felt annoyed" showed significant differences between responses that used the novel grid system and those that used the next cell system. There was more disagreement with the statements for the novel grid system than there was for the next cell system. However, there was the least frustration and annoyance expressed for neither hint system being used. This may mean that using either of the hint systems increases frustration and annoyance; however, as participants choose when to use the hint systems, it may mean that participants don’t use the hint systems on puzzles that do not frustrate or annoy them.

The statement "I enjoyed this puzzle" only showed a significant difference between using the next cell system and using neither system. Participants indicated that they enjoyed puzzles most when they used neither system. This could, as discussed above, have less to do with the impact of the hint systems and more to do with participants enjoying the puzzles that do not make them feel like they need a hint.

The statement "I felt I was good at it" showed significant differences between the next cell system and using either both systems or no systems. Responses indicated more agreement when using neither system, which, again, may be more due to not needing a hint than any reflection on the hint systems themselves. There was slightly less agreement when responses had to use both hint systems, which is consistent with the idea that having to use a hint system reduces how good at a puzzle a player felt they were.

The statement "I felt skillful" showed significant differences between the next cell system and using either just the novel hint system or neither system. Using just the novel hint system or none of the systems showed greater agreement with the statement, suggesting that the novel hint system impacted feelings of skilfulness less than the next cell system. The significant difference between the next cell system and using neither system suggests that players feel more skillful when they solve a puzzle without using hints.

There were no significant differences between the ratings of "I felt challenged" for the system(s) used. It is unexpected that there was not significantly less challenge experienced on puzzles where the participants did not feel the need for hints.

There was no difference found between the systems used for the agreement rating of the statement "I thought about other things". The statement "I was fully occupied with the game" showed a difference between the next cell system and using neither system. This suggests that needing hints may have broken the player's flow somewhat. Alternatively, players may have been more likely to use hints on puzzles where they were not fully focused on the game.

Overall, while no concrete conclusions can be drawn due to the different study conditions, it is clear that there are some differences between the impact of the two systems. It is also clear that not needing to use hints has the most significant impact on the participant's experience.

5.5.5 Impact of Binairo Pilot Study on Main Binairo Study Design

The analysis of this study showed that which hint system participants *used* was more important than which they *had access to*, and that the participants were resistant to using a hint system. Therefore, we moved from a between-group study design to a within-group study design where all participants had access to both hint systems for all puzzles. This would allow participants that made use of the hint systems to be compared without further confounding factors.

Furthermore, we redesigned the post-puzzle questionnaire to explicitly refer to the hint system and directly asked participants to evaluate their experience of the hint system, instead of relying on a comparison of their general experience with the puzzle. We removed questions related to how occupied players were with the game, as this study suggested that the hint systems have limited impact on occupation. We focussed the questions on our key hypotheses.

There was limited engagement with the puzzles after the debrief screen; therefore, we removed them for the following study, as they were not adding value to our data and were using some participants' time.

5.6 Main Binairo Study

The main Binairo study was conducted with a completely redesigned post-puzzle questionnaire and easier puzzles than the previous Binairo Study. This section discusses the changes to the study design and its results.

5.6.1 Main Binairo Study Design

The pre-study questionnaire was adjusted to ask if participants had done the previous Binairo study and to add non-binary with optional text box as one of the responses to the "How do you describe yourself?" question.

The questionnaires presented to participants after each puzzle were adapted following the results of the pilot study, as discussed in Section 5.5.5. The new questionnaire design focused on asking participants about their experience with the hint systems directly. They were still asked to rate the difficulty of the puzzle presented; however, this was moved to a seven-point Likert scale to improve the granularity of the results [67]. They were asked which hint systems they used and, if they used both, which they preferred. For each hint system (if they used it), they were asked to rate their agreement with a set of statements, on a seven-point Likert scale of agreement. The statements were as follows:

- I found it helpful
- It gave me the type of help I wanted
- I found these hints reduced my enjoyment
- I found it felt like cheating
- I found it enhanced my experience

The full questionnaire is included in Appendix A.3.2.2.

These were generated on a face-validity basis by the research team. The first two statements tie back to the question of which system was the most helpful in the Binairo pilot study. The latter three are based on the existing literature on hints.

In the previous study, few participants engaged with the optional extra puzzles and their data were excluded from most of the analysis; therefore, we did not provide the option in this study.

5.6.1.1 Puzzle Selection for Main Binairo Study

The puzzles selected for this study were expected to be significantly easier than the puzzles selected for the Binairo pilot study. Table 5.6 lists the puzzles selected. This adjustment was made due to the feedback of the pilot study, as discussed in Section 5.5.5.

Puzzle	Size	Total Simple Deductions	Total Complex Deductions	Max Size of Complex Deduction
Practice	6 by 6	38	0	N/A
1	10 by 10	95	0	N/A
2	12 by 12	173	0	N/A
3	10 by 10	111	0	N/A

Table 5.6: The puzzles used in the main Binairo study, in the order they were presented to the participants. See Section 5.2.1 for the explanation of simple and complex deductions.

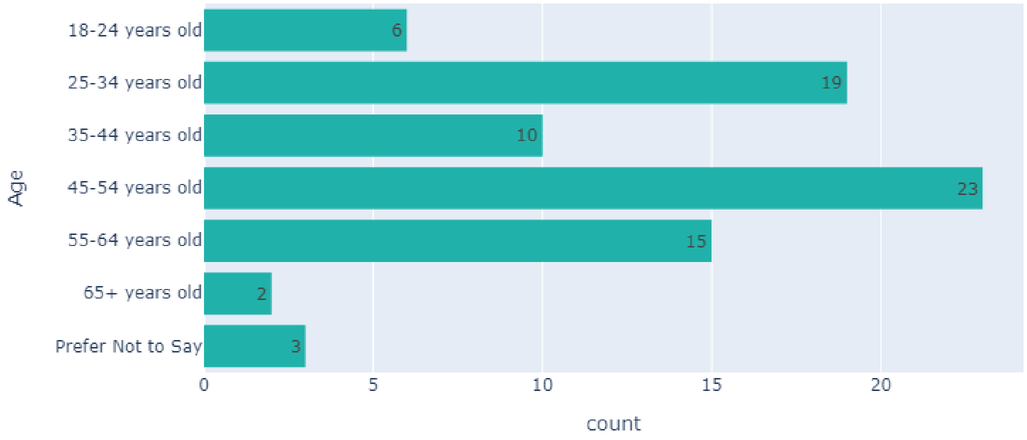


Figure 5.6: The Age distribution of participants in the Main Binairo Study. Blank responses have been combined with Prefer not to Say.

5.6.2 Participant Demographics of Main Binairo Study

Participants were asked their age and gender at the start of the experiment. 100 participants started the study and did not withdraw their data. Of those 100 participants 78 started the first puzzle, and 63 of those attempted all 4 puzzles. The age distribution of the 79 that started the puzzles is shown in Figure 5.6 and their gender distribution is shown in Figure 5.7.

5.6.3 Results of Main Binairo Study

There were a total of 274 responses to the post puzzle questionnaire from 76 participants. 41% (113) of responses indicated that they made use of at least one of

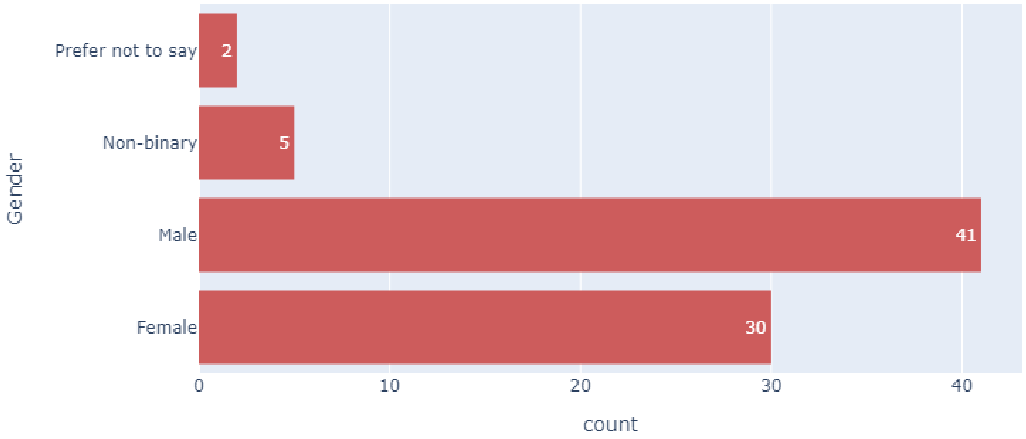


Figure 5.7: The Gender distribution of participants in the Main Binairo Study. Blank responses have been combined with Prefer not to Say. There were no responses of "Prefer to self-describe".

the two hints systems, see Figure 5.8 for exact numbers.

It became clear from some of the free text boxes that some participants had mistaken the questions about the coloured grid hint system for questions about the error handling system. For example, "I was just a little slow to click and the colored grid hint activated! How do I turn it off? I don't want hints at all!" The actual grid hint system was only available via a button click. Responses in which the free text box made it clear that they were referring to the error handling system were excluded from the analysis. However, it is not clear how many of the remaining ratings of the novel grid hint system were actually attempting to give ratings of the error handling system. Given the evidence in the free text boxes of how disliked the error handling system was, the most likely impact of this mistake would be to create a bias in favour of the traditional (next cell) hint system. Therefore, where we find evidence that the novel system is preferred, we can be fairly confident in those results.

Participants were asked which hint systems they used after each puzzle, shown in Figure 5.8, and their experience with the hint system(s) they used, by rating their agreement with a set of statements. There were 78 ratings of the novel grid hint system and 35 ratings of the next cell hint system[4] , all using the set

---

[4] Of the 62 responses that indicated that they used the novel grid hint system, 1 response

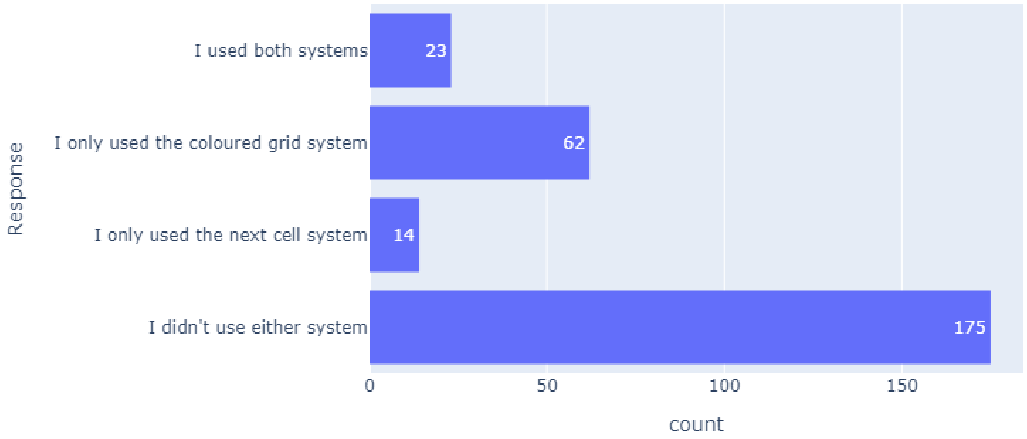


Figure 5.8: The Main Binairo study responses to the question "Which help system(s) did you use during this puzzle?". Responses are for each puzzle each participant attempted. The responses from participants that indicated in the free text boxes that they had confused the grid hint system with the error handling system were removed.

Text	Numeric Value
Strongly disagree	0
Disagree	1
Somewhat disagree	2
Neither agree nor disagree	3
Somewhat agree	4
Agree	5
Strongly agree	6

Table 5.7: Mapping of 7-point Likert scale showing agreement to numeric values.

of statements discussed in Section 5.6.1. The full questionnaire is included in Appendix A.3.1.2. The % agreement ratings, and their mean value (calculated from numeric mappings of the statements of agreement, as shown in Table 5.7) are shown in Figure 5.10.

---

indicated in the free text box that they actually used neither hint system and provided no ratings, 2 responses didn't provide any ratings, 3 responses indicated in the free text box that they were actually rating the error handling system and their answers were excluded. Of the 23 that indicated that they used both systems, 1 indicated in the free text box for the novel hint system that they were actually rating the error handling system, and their ratings of the novel hint system were excluded. Of the 14 that only used the next cell system, 2 provided no ratings.

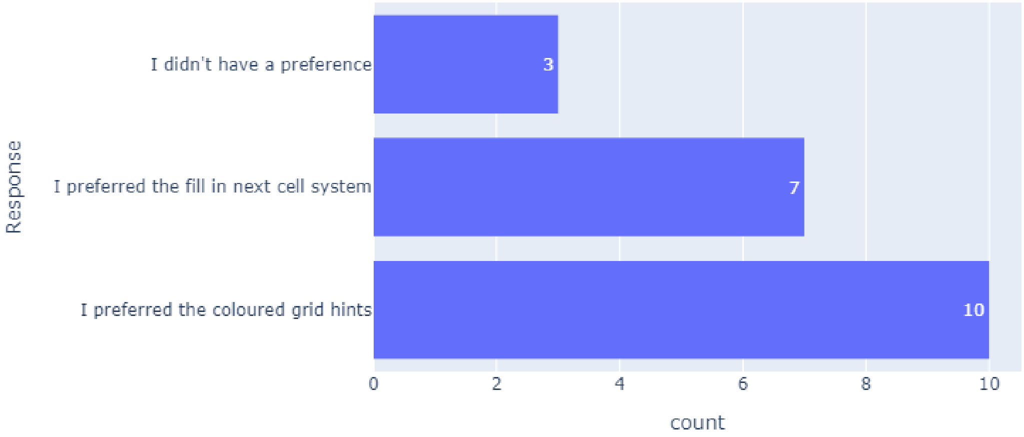


Figure 5.9: The Main Binairo study responses to the question "Which help system did you prefer during this puzzle?". Responses are for each puzzle each participant attempted. Participants were only asked for their preference if they used both hint systems when attempting the puzzle.

Question	β	Std. Err.	t value
I found it enhanced my experience	0.5084	0.2459	2.068
I found it helpful	-0.2714	0.3049	-0.89
I found it felt like cheating	-0.8694	0.2997	-2.901
I found these hints reduced my enjoyment	-1.3600	0.2847	-4.777
It gave me the type of help I wanted	0.6180	0.3283	1.883

Table 5.8: The results of the Linear Mixed-Effects Model, with the next cell rating as reference parameter, and puzzle and participant as random effects. The results shown are for the parameter Grid Hint ratings. T-values of less than -1.96 or greater than 1.96 are highlighted and considered significant.

We ran a linear mixed-effects model for agreement with the experience statements using the R package lme4 [10]. The hint system being rated (either the novel hint system or the traditional hint system) was a fixed effect and the model had random effects of *participant* and *puzzle*. The traditional (next cell) hint system was used as the reference category. The effect of using the novel hint system as opposed to the traditional one is shown in Table 5.8. A t-value of ±1.96 is equivalent to a p-value of 0.05, which is the chosen significance limit for this set of studies [44].

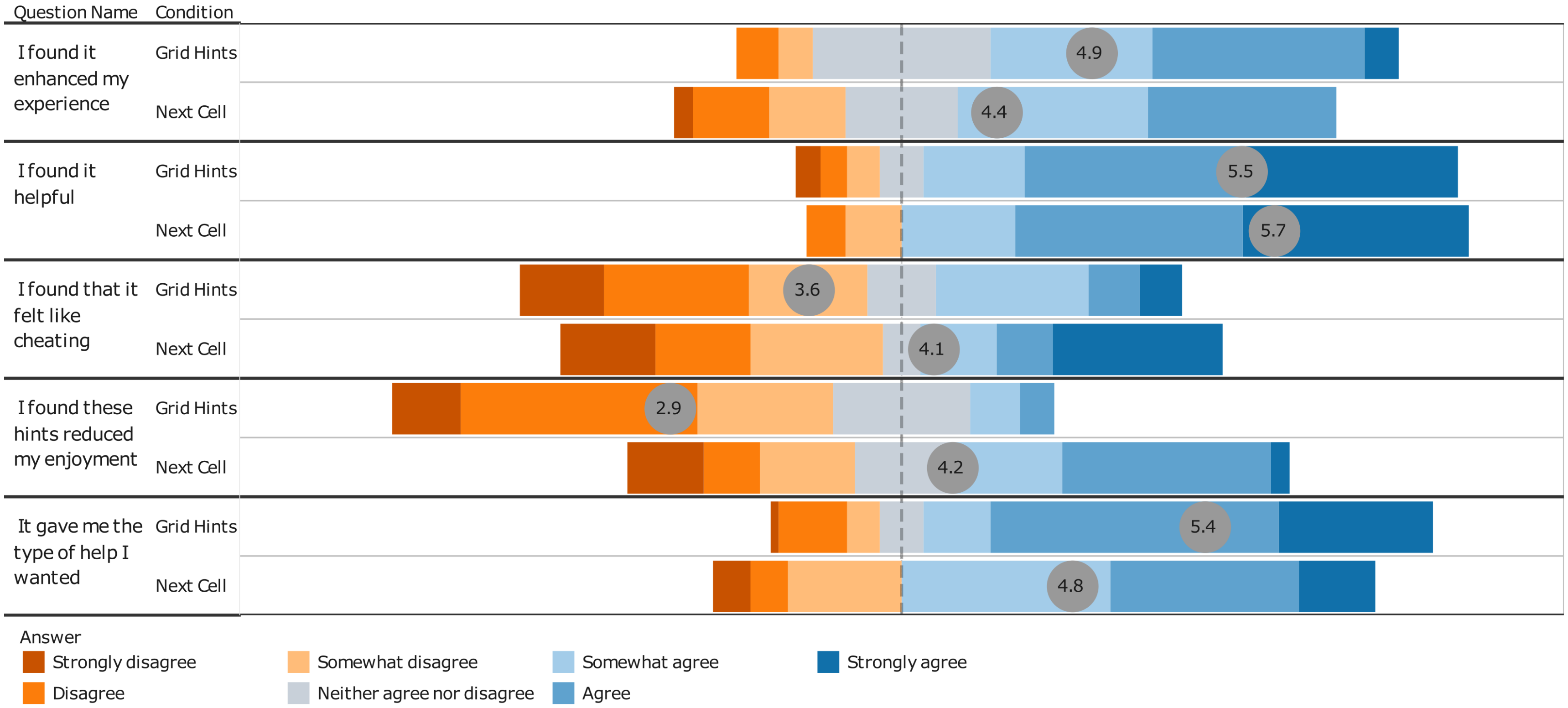


Figure 5.10: Likert visualisation of hint system assessment matrix for the main Binairo Study: A diverging stacked bar chart, centred around the centre of the neutral (Neither Agree nor Disagree) responses. There is a bar for each statement for each hint system. Each sub-section of the bar matches a possible level of agreement with the statement on the left (see legend below the chart). The size of the subsections corresponds to the percentage of responses that gave that answer for a given statement and condition. Answers where the free-text box made it clear the participant had provided responses for the error handling system rather than the grid hint system were excluded.

There was no significant difference between the two systems for the statements "I found it helpful" and "It gave me the type of help I wanted". The responses for "I found it helpful", shown in Figure 5.10, are nearly identical, suggesting that both hint systems are equally helpful to players.

There were significant differences between the two systems for the three remaining statements, as shown in Table 5.8. Figure 5.10 and Table 5.8 show that, while participants agreed that both hint systems enhanced their experience, they agreed more strongly that the novel hint system enhanced their experience. For, "I found it felt like cheating" Figure 5.10 and Table 5.8 show that participants disagreed that the novel hint system felt like cheating, but agreed that the traditional (next cell) system felt like cheating. The strongest difference was seen for agreement with the statement "I found these hints reduced my enjoyment", where participants disagreed that the novel hint system reduced their enjoyment, while agreeing that the traditional next cell system reduced their enjoyment.

5.6.3.1 Free text box discussion of the Error Handling System

Assessment of the error handling system was not a goal of the study and was therefore not assessed directly using the questionnaire. However, several participants used the free text boxes to comment on the error handling system, shown in Table 5.9. Only eight responses expressed an opinion on the error handling system, therefore no strong conclusions can be drawn; all but one agree with the expectation that the error system would be unpopular with players. However, several of the responses indicated that participants wanted to guess and see if it led to a contradiction (a chain technique, similar to Sudoku chain techniques, discussed in Section 2.1.2.5). However, the reasons for the delayed error feedback discussed in Section 4.2.3.2 still applied, so the error handling system was not adjusted for subsequent studies. The responses suggested that the choice to have a slight delay rather than an instant response was beneficial.

5.6.4 Discussion of Main Binairo Study

The results of this study demonstrated that there were significant differences between the participants' experience of the two hint systems, in general their experience of the novel hint system was more favourable. In addition to rating it more favourably, participants were more likely to use the novel hint system in the

it felt annoying that the hints were an incorrect/correct. i wanted to be able to work out why a cell is one way or the other and the coloured cells did not help with this
i found it irritating as it got in the way of how i wanted to play - i wanted to run some 1s and 0s as guesses and see what result it was and undo if the guesses turned out to be wrong, however with the coloured grids appearing and telling me straight away i couldn't do this option
I was just a little slow to click and the colored grid hint activated! How do I turn it off? I don't want hints at all! **&** Is there no way to deactivate the colo(u)red hints?
The coloured grid hints appeared automatically, and I didn't want them. I wanted them to appear when I chose, to see if I could work it out for myself.
In this puzzle I had to do some read-ahead, and in other puzzles when I do this I would write down a possible path, but here because it warns you if you are wrong after a short wait, I couldn't do that on the puzzle screen itself, which would have been useful
Need another tiny bit of time before I get the red box. I'm an impulsive placer and sometimes I was still thinking about it when I could see wrong answer pop up
The red square when you make a mistake was a way of checking what I was doing in a very definite way. I like 'that doesn't work' hints in story puzzles, not so much in this type of puzzle, where making and error and having to unpick to where you got it wrong can be an enjoyable part of the thinking process.
Error highlighting helped

Table 5.9: Free text boxes expressing an opinion on the error handling boxes (some refer to the error handling system as the coloured grid hint system, but it is clear from context and later comments that they meant the error handling system). Comments that stated they were surprised by the error handling system but expressed no further opinion were excluded.

first place. 85 responses indicated that they chose to use the novel system, while 37 responses indicated that they chose to use the next cell system[5].

There was no significant difference between the two hint systems in how helpful participants found them; this is consistent with the results of the previous study. The next cell system should always be helpful, as it increases the information in the puzzle. We expected that the novel grid hint could be considered less helpful than the next cell system; if the player cannot work out how to solve the 'easier' cell, they cannot make progress. It is interesting that, despite this, both systems are considered equally helpful.

The difference between participant ratings of "It gave me the type of help I wanted"

---

[5] These number are the combination of responses that used just a given hint system with the responses that used both, separate numbers are shown in Figure 5.8

for the next cell system and the grid hint system were not statistically significant, although it was very close to the borderline. The average response for both was positive (showing that participants mostly agreed with the statement), suggesting that both types of hint system provide a type of help that is desirable to some players. However, there is a potential bias in the statement ratings, as we did not force participants to use a hint system and we described the systems ahead of time. This meant that we only have ratings for each system from participants who were willing to use that system. Participants are likely to mainly use the system that provides the type of help they want; this may explain the positivity of the ratings. Therefore, although that comparison is not statistically significant, it is indicative that more than twice as many responses used the grid hint system as used the next cell system; therefore, the grid hint system may be closer to the type of help participants wanted.

Participant ratings of "I found these hints reduced my enjoyment" showed significantly less agreement for the novel grid hint system than they did for the next cell system. This was consistent with our hypothesis that a hint system that guided the player towards making deductions themselves, rather than telling them the answer, would have less negative influence on the player's experience. Only 12% (9/77) ratings indicated any agreement ("Somewhat Agree" or "Agree", there were no ratings of "Strongly Agree") with the statement when rating the grid hint system. In contrast, 51% (18/35) of the responses regarding the next cell system expressed agreement ("Somewhat Agree" or "Agree", or "Strongly Agree") with the statement. The high level of agreement that the next cell system reduced enjoyment is consistent with the existing literature showing that hints and feedback that simply tell the player what to do next negatively affect player experience. The low level of agreement regarding the novel grid hint system suggests that its visual guidance-based approach mitigates the negative impact hints can have on player experience.

The ratings of the statement "I found it enhanced my experience" indicated that the novel grid hint system enhanced players' experience more than the next cell system did. The average response to both was towards agreement with the statement, suggesting that access to a hint system when needed enhances player's experience. However, we only have ratings for participants that clicked on a hint system, therefore we expect ratings for this statement to be positive. There is no reason to think that this would bias one system more heavily than the other.

Similarly, the statement "I found it felt like cheating" showed significantly less agreement from participants for the novel grid hint system than for the next cell system. The mean response for the next cell system showed slight agreement with the statement, while the mean response for our novel hint system showed disagreement with the statement. This was consistent with our hypothesis, which was that a system that guided players rather than gave them the answer would feel less like cheating.

Overall, the results of this study show that our novel hint system is just as helpful as a more traditional approach, but our system enhances player experience compared to the traditional hint system and does not feel as much like cheating.

5.6.5 Impact of Main Binairo Study on Aquarium Study Design

It was clear from the free text box answers to "Do you have any comments on the colouring of the grid hints?" that a few participants had become confused between the coloured grid hint system and the error highlighting. The questionnaire was adjusted to add a clarification: "These are the hints provided by the Grid Hint button, not the error highlighting". No other changes were made to the questionnaires as a result of this study.

The number of responses that made use of the hint button was less than we hoped, therefore, the expected challenge of the puzzles was slightly increased for the following Aquarium Study.

5.7 Aquarium Pilot Study

The Aquarium study was conducted to assess the hint system against a second PPPP. The choice of Aquarium is explained in Section 5.1.

5.7.1 Aquarium Pilot Study Design

The pre-study questionnaire was slightly adjusted to ask if the participants had done either of the previous Binairo studies. The post-puzzle questionnaire was slightly adjusted to make it clearer that the ’coloured grid hint system’ meant the novel grid hint system and not the error highlighting. This followed the confusion in the previous study, as discussed in Section 5.6.5.

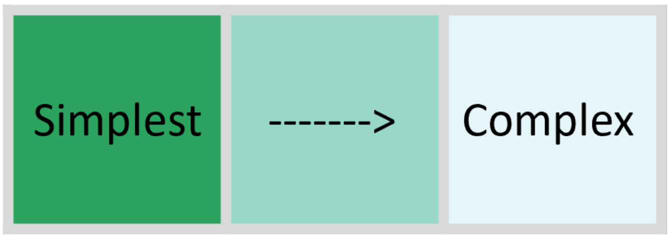


Figure 5.11: The legend used with the green variation of the hint grid colouring.

The puzzles were selected with the intention of offering a slightly increased challenge to the participants compared to the previous study, Section 5.6. Puzzles were provided by puzzles-aquarium.com [115]. The puzzles chosen are shown in Table 5.10.

Puzzle	Size	Total Simple Deductions	Total Complex Deductions	Max Size of Complex Deduction	ID
Practice	6 by 6	44	4	2	3
1	6 by 6	40	7	3	16
2	6 by 6	39	9	2	24
3	10 by 10	128	12	3	33

Table 5.10: The puzzles used in the second Binairo study, in the order they were presented to the participants. See Section 5.2.1 for the explanation of simple and complex deductions.

Participants were not presented with optional extra puzzles following the main ones, as in the previous study.

The colour of the novel hint system was changed for this study, as the blue was hard to separate from the water in the aquarium puzzles; an example is shown in Figure 5.12.

5.7.2 Participant Demographics of Aquarium Pilot Study

Participants were asked their age and gender at the start of the experiment. 58 participants started the study and did not withdraw their data. Of those 58 participants, 45 started the first puzzle, and 34 of those attempted all 4 puzzles. The age distribution of the 45 that started the puzzles is shown in Figure 5.13 and their gender distribution is shown in Figure 5.14.

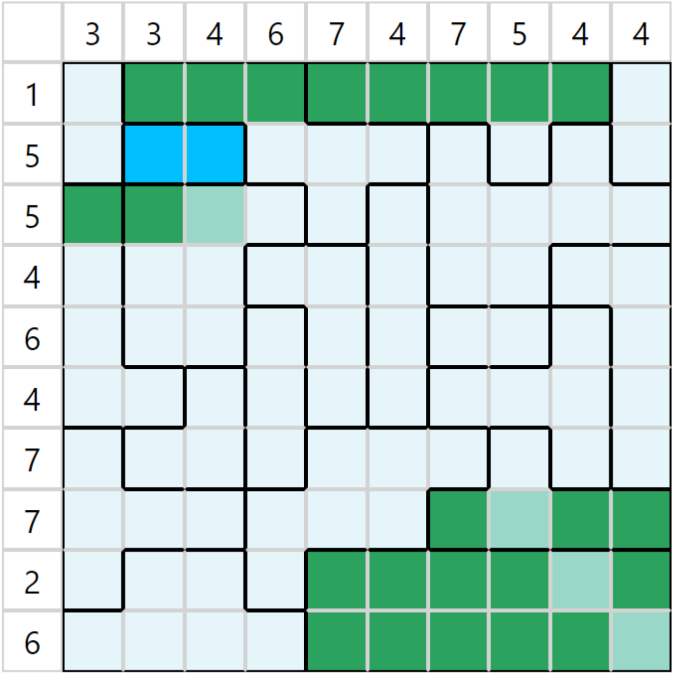


Figure 5.12: An example of Aquarium with the green hint system variant.

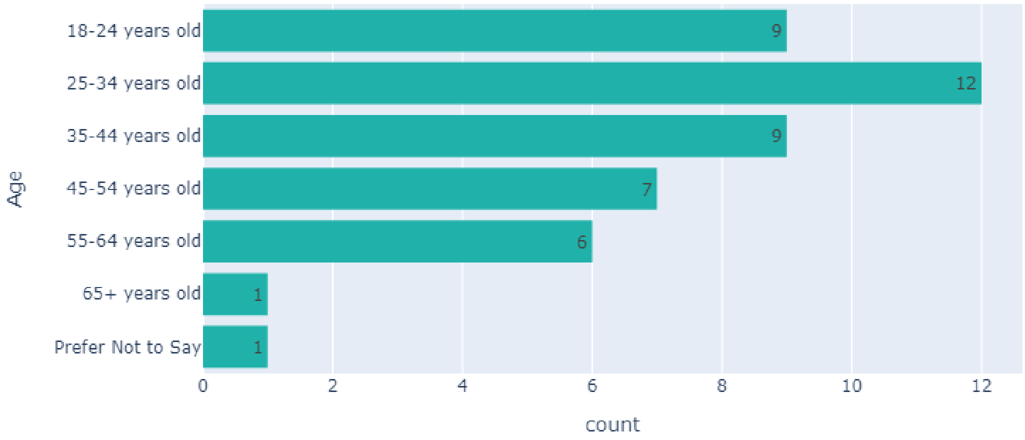


Figure 5.13: The Age distribution of participants in the Aquarium Pilot Study. Blank responses have been combined with Prefer not to Say.

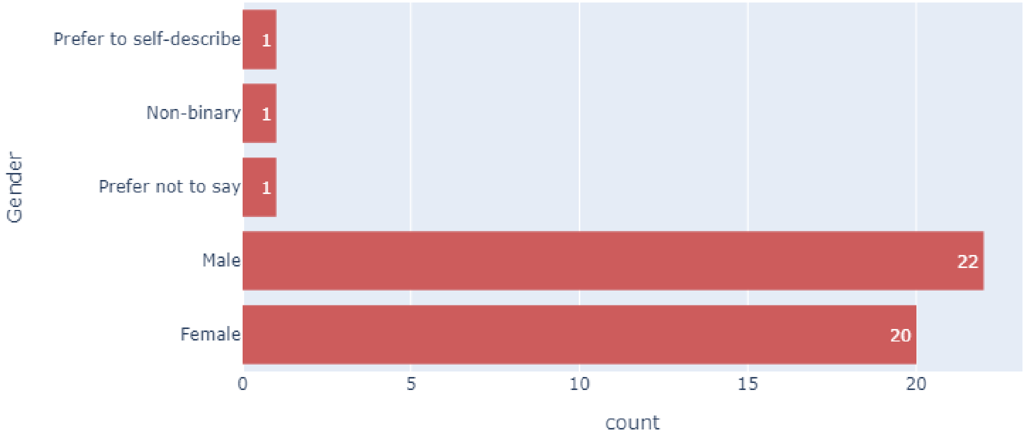


Figure 5.14: The Gender distribution of participants in the Aquarium Pilot Study. Blank responses have been combined with Prefer not to Say.

5.7.3 Results of Aquarium Pilot Study

The study was ended early, when it became clear that the puzzles were not prompting enough interaction with the hint system. As a result of the early finish and the limited interaction with the hint system (12% of responses), the sample sizes are very small.

There were a total of 148 responses to the post-puzzle questionnaire (one additional one was started but all questions were unanswered) from 42 participants.

Participants were asked which hint systems they used after each puzzle, shown in Figure 5.15, and their experience with the hint system(s) they used, by rating their agreement with a set of statements. There were 15 ratings of the grid hint system and 4 ratings of the next cell system [6]. These were provided by 14 unique participants (the ratings were for each puzzle, not each participant). The % value of each rating is shown in Figure 5.16.

Only the 3 responses that marked that they used both systems were asked which system they preferred. 1 marked that they did not have a preference, 1 marked

---

[6] 3 participants marked that they used both systems, but one did not provide ratings and stated in the free text box that they did not actually use either system. 15 participants marked that they used only the coloured grid hint system, of those 1 stated in the free text box that they did not use any hints and their responses (all neutral) were excluded from the analysis, and 2 provided no ratings. Both participants that marked that they only used the next cell hint system provided ratings.

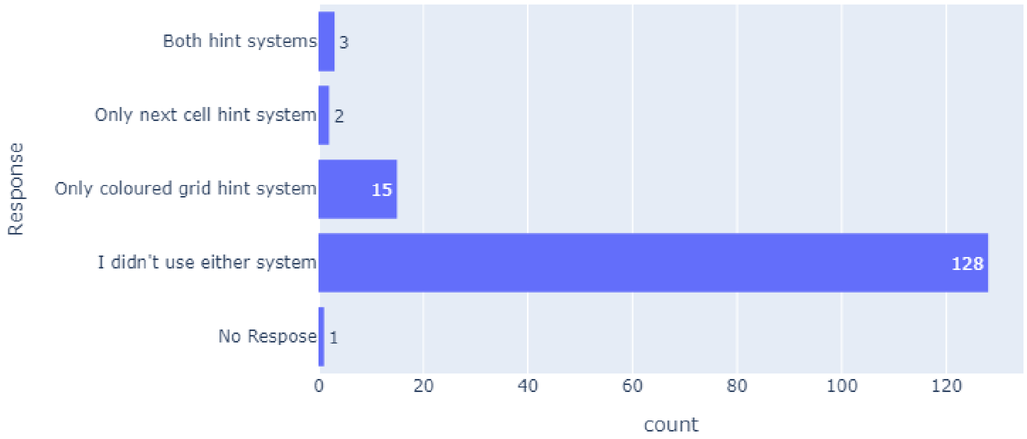


Figure 5.15: The Aquarium pilot study responses to the question "Which help system(s) did you use during this puzzle?". Responses are for each puzzle each participant attempted.

that they preferred the next cell system, and 1 marked that they preferred the novel grid hint system. It was too small a sample size to draw conclusions.

An attempt was made to run a linear mixed-effects model for agreement with the experience statements using the R package, lme4 [10], however, due to the small sample size of the next cell ratings (4), the model would not converge and therefore is not reported.

The study did produce a data set of 45 people playing Aquarium. This data set could still be of use for future research and will be provided in the data repository.

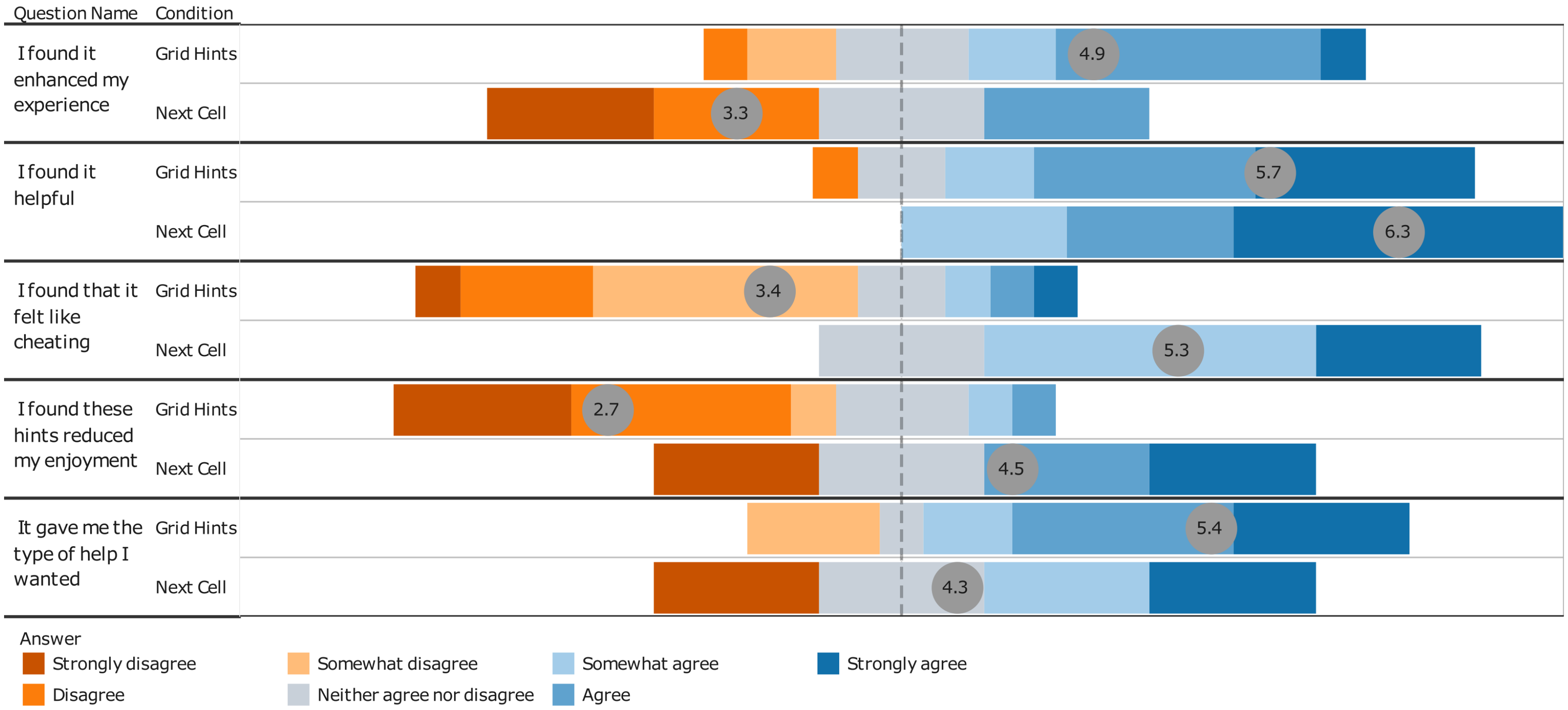


Figure 5.16: Likert visualisation of hint system assessment matrix for the Aquarium Pilot Study: A diverging stacked bar chart, centred around the centre of the neutral (Neither Agree nor Disagree) responses. There is a bar for each statement for each hint system. Each sub-section of the bar matches a possible level of agreement with the statement on the left (see legend below the chart). The size of the sub-sections corresponds to the percentage of responses that gave that answer for a given statement and condition, see Footnote 6.

Figure 5.16 does show that the mean values for the ratings of the novel hint system are consistent with the previous study, with at most a difference of 0.2. The mean values for "I found it enhanced my experience" and "It gave me the type of help I wanted" are identical. The mean values for the next cell ratings showed substantial differences from the previous study, however, this is unsurprisingly, as the small sample size gives each of the 4 responses a significant impact on the results. The ratings for the next cell system for this study are not considered meaningful due to the small sample size.

5.7.4 Discussion of Aquarium Pilot Study

Unfortunately the low level of engagement with the hint systems (4 responses used the next cell, and 15 used the novel grid hint) in this study made it difficult to come to any meaningful conclusions. The experience ratings for the novel hint system are consistent with the previous study.

The lack of ratings for the next cell system make it impossible to draw meaningful conclusions about the player experience with the next cell system.

5.7.5 Impact of Aquarium Pilot Study on Main Aquarium Study Design

Only 12% of responses stated that they used either of the hint systems; this was substantially less than the 41% that stated that they used them in the previous Binairo Study. Clearly, the challenge presented by the chosen puzzles was not sufficient to prompt participants to need a hint. Therefore, we conducted a main Aquarium study, discussed below in Section 5.8, where we increased the expected challenge of the puzzles. We reintroduced the extra puzzles following the debrief screen for skilled participants. These puzzles were expected to be more challenging than those preceding the debrief screen. The goal of the extra puzzles was to induce skilled participants to engage with the hint systems and to avoid them feeling cheated or distressed by the puzzles being too easy[7].

There was no indication that the confusion between the novel hint system and the error handling system had occurred, the clarification seemed to have resolved the

---

[7] A participant reached out to the researchers directly to express their distress at the puzzles being too simple, this was supported by some of the free text box responses also complaining the puzzles were too easy

issue. Although the small sample size could also explain why the issue was not seen.

5.8 Main Aquarium Study

This Aquarium study was conducted, as the previous Aquarium study did not provide meaningful results due to lack of engagement with the hint system.

5.8.1 Main Aquarium Study Design

The pre-study questionnaire was slightly adjusted to also ask if participants had done the previous Aquarium Study.

The puzzles selected for this study were intended to be more challenging than the puzzles in the previous study. We also reintroduced the post debrief screen puzzles. The puzzles are listed in Table 5.11.

Puzzle	Size	Total Simple Deductions	Total Complex Deductions	Max Size of Complex Deduction	ID
Practice	6 by 6	53	5	2	15
1	6 by 6	26	16	7	29
2	10 by 10	106	19	4	47
3	10 by 10	99	24	6	57
Debrief Screen					
4	10 by 10	119	16	2	43
5	10 by 10	91	30	6	53
6	10 by 10	106	22	7	56
7	10 by 10	87	32	6	60
8	10 by 10	73	33	17	55

Table 5.11: The puzzles used in the main Aquarium study, in the order they were presented to the participants. See Section 5.2.1 for the explanation of simple and complex deductions.

5.8.2 Participant Demographics of Main Aquarium Study

Participants were asked their age and gender at the start of the experiment. 1,059 participants started the study. Of those 1,059 participants, 319 started the first

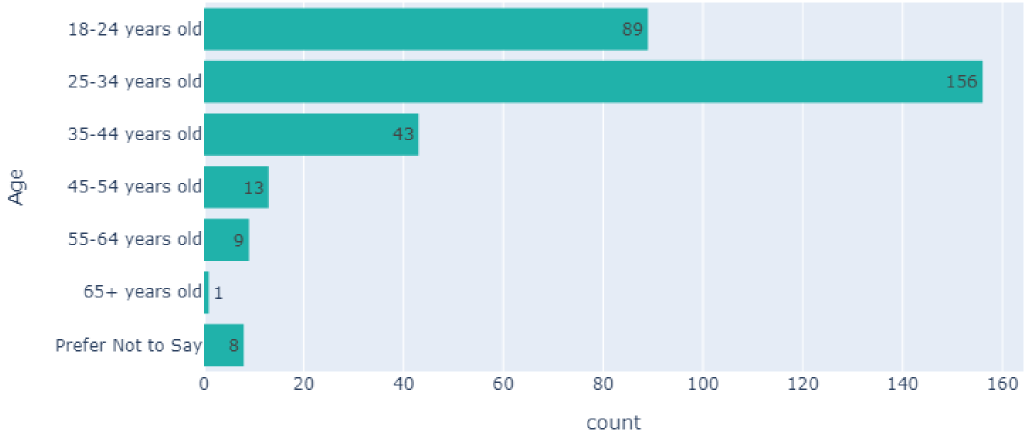


Figure 5.17: The Age distribution of participants in the Main Aquarium Study. Blank responses have been combined with Prefer not to Say.

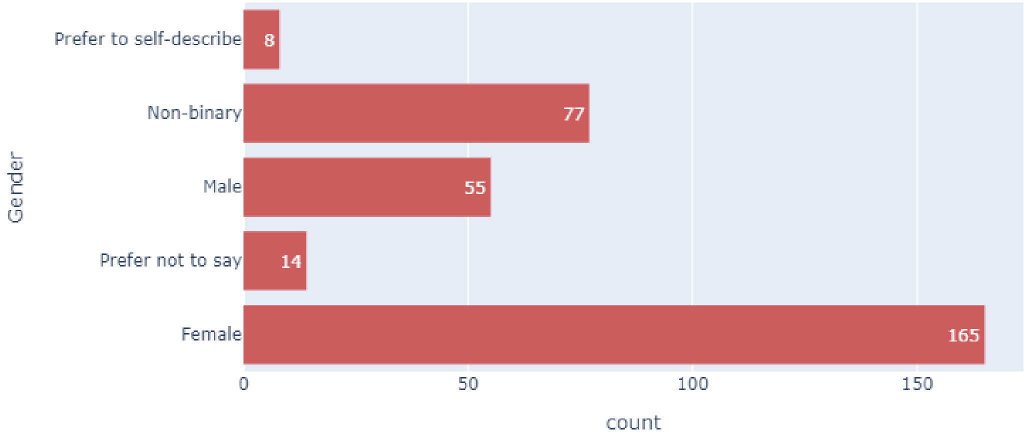


Figure 5.18: The Gender distribution of participants in the Main Aquarium Study. Blank responses have been combined with Prefer not to Say.

puzzle, and 84 of those attempted all the puzzles prior to the debrief screen. The age distribution of the 319 participants that started the puzzles is shown in Figure 5.17 and their gender distribution is shown in Figure 5.18.

5.8.3 Results of Main Aquarium Study

This study was impacted by a sudden spike of popularity, which resulted in over 1000 attempts at participation in 7 hours. Our system (which had previously seen peaks of at most 50 participants over the course of a day) struggled and many participants found the system slow and buggy. Many of them were unable to start the puzzles and many that started struggled to complete multiple puzzles. 14 responses rated the systems but stated they weren't working. Their ratings were excluded from the analysis. 3 response indicated that the coloured grid hint system was working but the next cell system was not (and rated both). Their ratings for both and their preference between them was excluded from the analysis.

Some responses indicated that they had confused the coloured grid system with the error handling system, for example "i would like the color hint delay to be longer so i can color it in and look at the whole puzzle before it hints". This was despite the changes made following the main Binairo study, attempting to make it very clear which system was meant. As in the previous study, it is not clear how many of the remaining ratings of the novel grid hint system were actually trying to give ratings of the error handling system. However, again, given the evidence in the free text boxes of how disliked the error handling system was, the most likely impact of this mistake would be to create a bias in favour of the traditional (next cell) hint system. Therefore, where we find evidence that the novel system is preferred, we can be fairly confident in those results.

1 participant seemed to assume across all their responses that they were doing a different puzzle, for example "The dark lines do not seem to mean anything...". However, they did not use the hint systems (They stated: "Maybe I'm a paranoid bastard, but since your instructions made no sense with those dark lines, I don't trust that your "hints" will be any more useful. The striped mistake indicator is enough hint for me - so far!") and therefore have not impacted the analysis. It is not clear whether other participants had the same confusion.

Participants were asked which hint systems they used after each puzzle, shown in Figure 5.20, and their experience with the hint system(s) they used, by rating their agreement with a set of statements. There were 200 ratings of the novel grid hint system and 115 ratings of the next cell hint system[8] , all using the set

---

[8] Of the 231 responses that indicated that they used the novel grid hint system, 17 ratings were excluded due to technical difficulties (discussed earlier) indicated in the free text boxes, 7 ratings were excluded as the free text boxes indicated they were actually rating the error handling system,

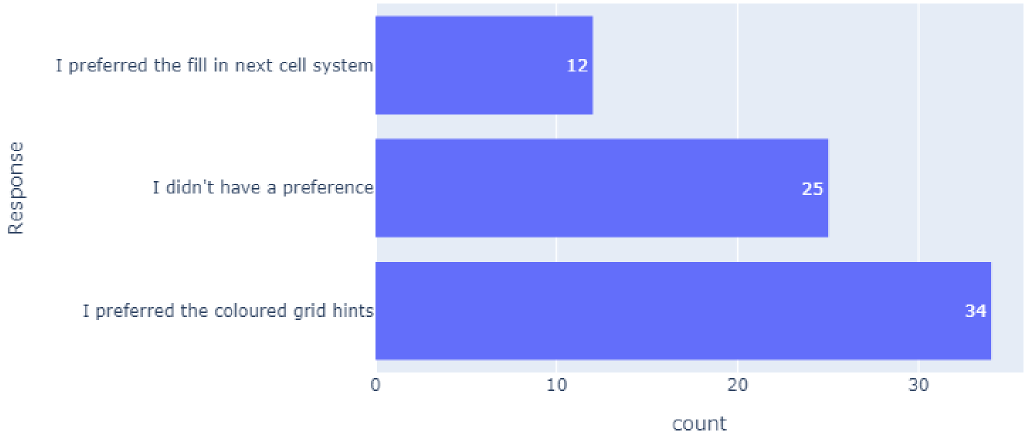


Figure 5.19: The Main Aquarium study responses to the question "Which help system did you prefer during this puzzle?". Responses are for each puzzle each participant attempted. Participants were only asked for their preference if they used both hint systems when attempting the puzzle. 3 response were excluded as participants indicated that the next cell system wasn't working properly.

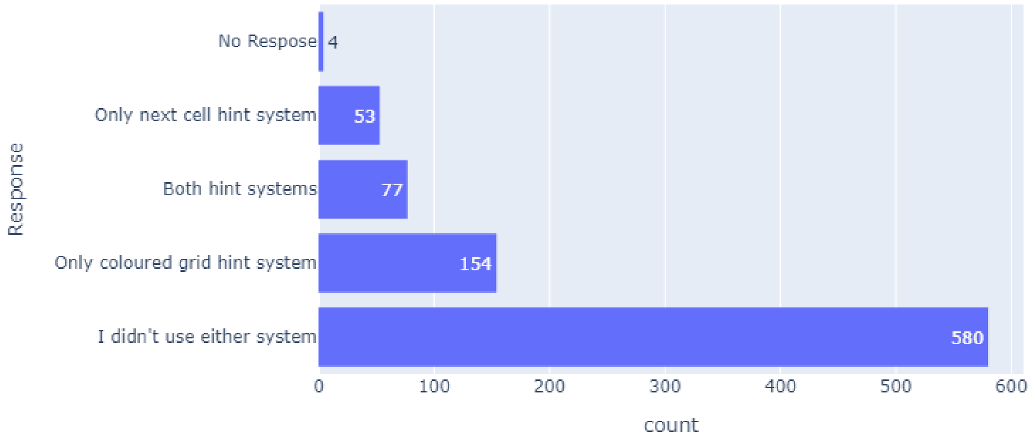


Figure 5.20: The Main Aquarium study responses to the question "Which help system(s) did you use during this puzzle?". Responses are for each puzzle each participant attempted.

of statements discussed in Section 5.6.1, the full questionnaire is included in Appendix A.3.2.4. The % agreement ratings, and their mean value (calculated from numeric mappings of the statements of agreement, as shown in Table 5.7) are shown in Figure 5.21.

---

1 response was excluded as the responses indicated that they hadn't actually used the hint systems, 1 response was excluded as they seemed to have the two hint systems confused, and 5 responses provided no ratings. Of the 130 responses rating the next cell system, 11 were excluded due to technical difficulties (discussed earlier), 1 was excluded because they seemed to have confused the two hint systems, and 3 provided no ratings

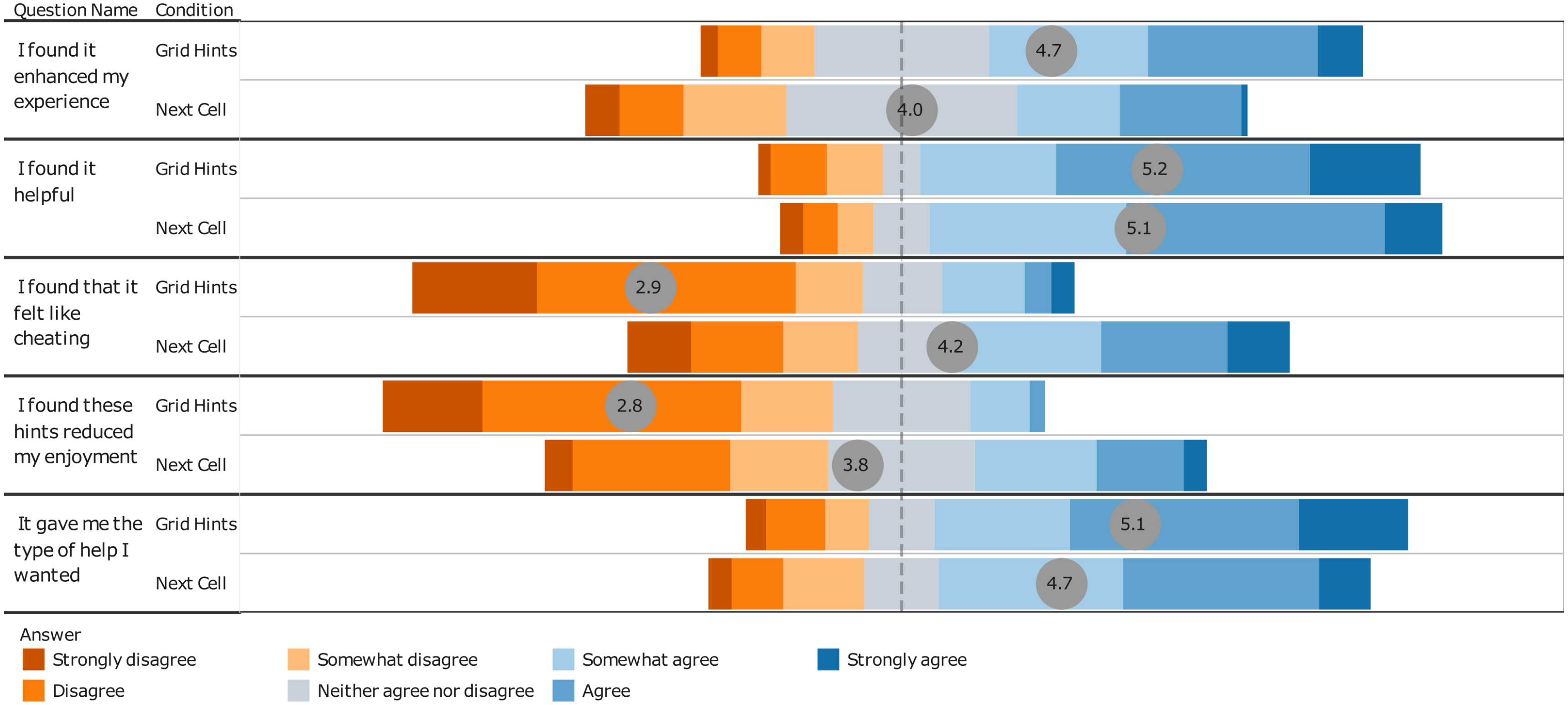


Figure 5.21: Likert visualisation of hint system assessment matrix for the main Aquarium Study: A diverging stacked bar chart, centred around the centre of the neutral (Neither Agree nor Disagree) responses. There is a bar for each statement for each hint system. Each sub-section of the bar matches a possible level of agreement with the statement on the left (see legend below the chart). The size of the subsections corresponds to the percentage of responses that gave that answer for a given statement and hint system, see Footnote 8 for exclusions.

Question	β	Std. Err.	t value
I found it enhanced my experience	0.5690	0.1543	3.687
I found it helpful	-0.0310	0.1758	-0.176
I found it felt like cheating	-1.2685	0.1712	-7.41
I found these hints reduced my enjoyment	-0.9814	0.1532	-6.408
It gave me the type of help I wanted	0.2568	0.1808	1.421

Table 5.12: The results of the Linear Mixed-Effects Model for the Main Aquarium Study, with the next cell rating as reference parameter, and puzzle and participant as random effects. The results shown are for the parameter Grid Hint ratings. T-values of less than -1.96 or greater than 1.96 are highlighted and considered significant.

We ran a linear mixed-effects model for agreement with the experience statements using the R package, lme4 [10]. The hint system being rated (either the novel hint system or the traditional hint system) was a fixed effect and the model had a random effect of participant and puzzle. The traditional (next cell) hint system was used as the reference category. The effect of using the novel hint system as opposed to the traditional one is shown in Table 5.12. t-values of ±1.96 is equivalent to a p-value of 0.05, which is the chosen significance limit for this set of studies [44].

The results of the analysis were consistent with the results of the same analysis on the Main Binairo Study (and, to the extent it could be assessed, the Aquarium pilot study). There was again no significant difference between the two systems for the statements "I found it helpful" and "It gave me the type of help I wanted". The responses to "I found it helpful", shown in Figure 5.10, were again nearly identical, suggesting that both hint systems are equally helpful to players.

There were significant differences between the two systems for the three remaining statements, as shown in Table 5.12, this is consistent with the analysis results of the Main Binairo Study. Figure 5.21 and Table 5.12 show that participants were, on average, neutral for "I found it enhanced my experience" when rating the traditional next cell system, but expressed agreement with the statement when rating the novel hint system. The agreement ratings for this statement of the next cell system were more negative (expressed more disagreement) than in the main Binairo study.

For, "I found it felt like cheating" Figure 5.21 and Table 5.12 show that participants disagreed that the novel hint system felt like cheating, but agreed that the traditional (next cell) system felt like cheating. The disagreement that the novel hint system felt like cheating was stronger than it was in the Main Binairo Study (this study had a mean rating of 2.9, while the Main Binairo study had a mean rating of 3.6), while the agreement that the next cell system felt like cheating was very similar to the Main Binairo Study (a mean of 4.2 in this study and 4.1 in the Main Binairo Study).

The ratings of the statement "I found these hints reduced my enjoyment", compared to the Main Binairo Study, showed that while participants still disagreed that the novel hint system reduced their enjoyment, the mean rating of the next cell system had moved to slight disagreement (a mean of 4.2 in the Main Binairo Study, to a mean of 3.8 in this study). However, the level of disagreement was significantly higher for the novel hint system (2.8), indicating that the novel hint system reduced enjoyment less than the next cell system.

5.8.4 Discussion of Main Aquarium Study

The results of this study were consistent with the results of the main Binairo study. They are also consistent with the novel hint system results from the Aquarium Pilot Study; there were too few responses for the next cell system in the Aquarium pilot study to draw meaningful results. There were significant differences in player experience when rating agreement with the statements "I found it enhanced my experience", "I found it felt like cheating", and "I found these hints reduced my enjoyment".

There is no reason to believe that the technological difficulties faced by participants as a result of the unexpected popularity of this study would have impacted either hint system more than the other, and therefore, while it could have had an overall negative effect on the general player experience, it should not have impacted the relative experience of the two hint systems. The consistency of the results of this study and the results of the main Binairo study support this expectation.

We can conclude from the similarities between the results of the main Binairo study and this study that the system is generalisable.

5.9 Limitations

There were specific, different, challenges faced by each study which are discussed in their sections. This section discusses the limitations common to all the studies.

Due to the nature of a large-scale unsupervised, online experiment, a balance had to be struck between asking participants about their experience and avoiding long questionnaires that could result in abandonment of the study. Whenever we wanted to add new questions, other questions had to be removed. For example, when we added explicit questions about the hint systems, we had to remove the general questions about player experience.

A second challenge found with a large-scale, unsupervised, online experiment was that participants could not ask for clarifications if they misunderstood. Therefore, there were some instances of misinterpretation of the rules of the game, how the hint systems worked, and what the questionnaires asked. There do not seem to have been very many of these and where the free text boxes indicated misunderstanding, they were excluded. These exclusions are listed in each study section.

5.10 Discussion

We conducted two studies using the Binairo puzzle and two studies using the Aquarium puzzle assessing the performance of the novel hint system against a traditional next cell system. The four studies produced consistent results that show a strong player preference for our novel hint system over the more traditional next cell system. This was shown both in the ratings of player experience and in the preference for using our novel hint system over the traditional next cell one.

In the Binairo pilot study, we assigned participants to different study conditions. They were allowed access to either the novel hint system, the next cell system or both. The Binairo pilot study did not demonstrate any difference between the study conditions, but showed significant differences (in favour of the novel hint system) in player experience ratings between the use of the next cell system and the use of the novel hint system. However, since not all players had access to both systems, it is hard to draw concrete conclusions from this data.

The remaining three studies were very similar in design; however, we corrected the problem seen in the pilot Binairo by moving to a within group design, ensuring that all participants had equal access to the two hint systems. The ratings of the novel hint system were consistent across all three studies. The pilot aquarium study had too few users of the next cell system to allow analysis; the ratings of the next cell system were consistent between the main Binairo study and the main

Aquarium study.

The player experience ratings showed that the participants felt the novel hint system did not feel like cheating, while the next cell system did. Furthermore, the novel hint system improved player experience compared to the next cell. It did not reduce their enjoyment and enhanced their experience, when compared to the next cell system.

Both hint systems were found to be helpful and to provide the type of help that the player wanted. As discussed previously, the second result is unsurprising, as the participant knew what each hint system would do and chose which one to use. We expected the novel hint system to be considered slightly less helpful than the next cell system, as, if the participant could not work out the solution when directed, the system would not be 'helpful'; whereas the next cell system would always provide information. However, there was no significant difference between the two and the means of the ratings of the novel hint system were slightly higher across all three studies that used this questionnaire.

Participants indicated that they used the novel hint system more than the next cell system. In the Main Binairo study, the Pilot Aquarium study, and the Main Aquarium study, participants used the novel hint system alone on more puzzles than they used both systems, which was again more often than they used only the next cell system. The results for the Binairo Pilot study are impacted by the different study conditions. In all studies participants tended to use no hints rather than either of the hint systems. This is consistent with the existing literature that found that players prefer not to use hints if they can avoid it. The free text boxes further supported this, with comments such as "It'll be a cold day in hell before I use hints". This combined with the indication that participants did not feel that the novel hint system "felt like cheating" suggests that the approach of guiding players would be a productive direction to take future research on providing player assistance.

The free text box responses for the error handling system, discussed in more detail in the Main Binairo study (Section 5.6.3) but seen throughout the studies, are consistent with Wauck *et al*'s finding that players dislike automatically provided 'hints' or corrections. This was an expected result and the error handling system was not under evaluation during these studies.

The error handling system may have contributed to reducing engagement with both hint systems. Several free text boxes indicated that they had not used the hint systems because the error handling system rendered it unnecessary. This should not have impacted one hint system more than the other, as using the error handling system this way rendered hints unnecessary.

Overall, it seems likely that a hint system which guides players towards the next easiest square (and reduces the need for visual search) is an effective hint system, which players are more willing to use, which feels less like cheating, and which enhances their experience. We have demonstrated that it is an improvement over the traditional next cell system. The data generated from these studies will be placed in a data repository to facilitate further research.

CHAPTER SIX

FUTURE WORK AND CONCLUSIONS

In this thesis, we have made the following contributions. We have demonstrated serious flaws in existing assumptions about how people play Progressive Pen & Paper Puzzle Games (PPPPs) through a qualitative study of people playing Sudoku (Contribution 1), discussed in Chapter 3. We have designed a novel approach to providing players with hints, based on the findings of our in-person study and existing literature on hints, discussed in Chapter 4. We have implemented that system and demonstrated via a series of large-scale online experiments (compared against a standard existing approach), discussed in Chapter 5, that the hint system we designed succeeded in its goals of improving player experience and allowing players to use it without feeling like they were cheating (Contribution 2).

We set out to explore two research questions: "How do people solve PPPPs?" and "How do we provide assistance to players of PPPPs in a way that improves player experience compared to current systems?".

To investigate the first question, "How do people solve PPPPs?", we conducted an in person study of how people solve Sudoku, discussed in Chapter 3. This study found that none of the participants behaved in the manner described in existing literature. This does not completely answer the question, but it does demonstrate significant weaknesses in the literature. This study provided both motivation for the second stage of the research and guidance on how to structure an assistance system. It would have been very challenging to build an assistance system without the better understanding of how people play PPPPs this work provided.

To investigate the second question, "How do we provide assistance to players of PPPPs in a way that improves player experience compared to current systems?", we explored various hint system designs and went through an iterative process to develop our final design, discussed in Chapter 4. Once we had developed a novel hint system we conducted a serious of studies with two PPPPs to assess the efficacious of the novel hint system. These studies and their results were discussed in Chapter 5. The findings of the studies were that our novel hint system design significantly improved both the player experience and the players' willingness to engage with a hint system when compared to existing systems. The work in this thesis provides a light-touch, guidance based approach for providing assistance to players of PPPPs that improves player experience when compared to the existing standards for hint systems.

6.1 Key findings

This work provides better understanding of a task millions of people do everyday. The current assumptions made by both guides and models about how people play PPPPs are deeply flawed. From the way players make notes, to the high number and frequency of mistakes, to the variation in how hard they find a given technique; players of PPPPs are more unsystematic, idiosyncratic and error prone than has been previously acknowledged. This was found in the study discussed in Chapter 3. This thesis therefore corrects a long standing error in the literature and this research will ensure that future research in this area can start from a better understanding of player behaviour.

The high frequency of error, even in 'finished' puzzles, found in the study discussed in Chapter 3, was very unexpected. Players can check puzzles for correctness, and the assumption in existing work was that players would make either no errors or very few errors. The discovery that errors occur frequently when playing PPPPs motivates research in how to help players recover from erroneous puzzle states. It is essential, particularly if using a PPPP as a serious game (see Section 2.1.3), for players to be able to recover from errors.

It is widely acknowledged that players resist using hint systems [12]. Our findings were consistent with this, however the degree to which players avoided engaging with any hint systems, during the studies discussed in Chapter 5, was unexpected. Most participants avoided using the hint systems whenever possible, and many left comments expressing their desire to never use a hint system. This highlights both a challenge for future research on hint systems and an interesting area of further study. Players need to use the hint system for researchers to be able to assess it's impact.

The findings of this thesis provide a new paradigm for providing assistance in PPPPs; the priority should be guiding rather than telling players what to do next. This approach is gaining traction in games outside of PPPPs and has long been key in pedagogical approaches; however existing PPPP systems and research focus on telling the player what to do next (sometimes with an accompanying explanation) rather than providing assistance on finding the next step themselves. Our novel hint system, discussed in Chapter 4, provided very limited assistance - it directed players towards cells they should find easier to solve. Yet, it was found to be considered just as helpful as a system that told the player what to fill in the next cell with. It also enhanced player experience more than filling in the next cell did and did not feel like cheating. The latter element may also explain why players showed a strong preference for using our novel hint system over the fill in cell system.

Overall, the findings of this thesis have demonstrated that the current state of understanding of players of PPPPs is inaccurate and inadequate. More research is needed on players of PPPPs, particularly if they continue to be used as serious games. We have provided a new approach to providing assistance to PPPP players and demonstrated that players prefer it over existing approaches. We have demonstrated that this approach is generalisable across binary PPPPs. Refining and expanding this approach and confirming that it is generalisable across all PPPPs by testing against more PPPPs is the next step in working towards meaningful assistance and balance in PPPPs.

6.2 Future Work

Further study of players would allow the building of a more accurate model of the various player behaviours. Building a data set of players' use of notation could allow the building of a system which could understand the most likely meaning of players' notation and respond accordingly. This would allow understanding of the state of the puzzle for candidate PPPPs, allowing the proposed hint system to be employed. It would also allow better support for players that prefer not to use the recommended notation styles, possibly increasing player engagement.

The work in this thesis motivates further study into player error and error recovery. We found that players made frequent and unexpected errors. This was at odds with existing literature which expected there to be little to no error in PPPPs. As PPPPs and similar games are increasingly employed in serious contexts it is important to understand and be able to support players in recovering from error without significant set back or loss of engagement. The work in this thesis demonstrates that players make frequent errors and it is essential to provide support for player's to recover from errors. Generalising this work to other games would therefore be very valuable.

There is a direction of research suggested by this work into why and when people make errors when they have access to all the information required to avoid them. In PPPPs players have access to all the information required to check their move is correct. We believe this is largely the reason the literature assumed that players would never or very rarely make an error. Despite this, players made frequent errors. The assumption that if all the information is there people will make use of it correctly and avoid error is seen throughout gaming, education, and system design. Further research into why players make errors in PPPPs may lead to a better understanding of how to reduce these errors, both in serious games, general education and system design.

This work highlights a key issue with studying player assistance - players actively avoid using assistance systems. As serious games continue to grow and player assistance becomes increasingly important, research into the study design for this type of experiment becomes essential. It could be optimising the parameters of studies to encourage hint usage or compensating for the negative impact of forcing hints on participants or some combination of the two. Regardless of the form, improvements in study design will facilitate further improvement in assistance systems, both within games and other contexts, for example, assistive technology for the elderly.

An important area of future work will be exploring the applications of the assistive approaches designed within this thesis within pedagogical contexts. As serious games are used increasingly in educational contexts providing good quality assistance that maintains both player engagement and the goals of the game is essential. This research provides a new paradigm for providing assistance that can be explored both within serious PPPPs and applied to other games.

The design of the Aquarium model made it clear that there is compelling area of research into comparing different DEMYSTIFY models of the same PPPP with players' experience of difficulty. This research should allow better support and understanding of players' experience of PPPPs and other puzzles. It may also lead to a mapping of constraints to mental models. Better DEMYSTIFY models could facilitate a more rewarding and engaging experience for players. This would be of benefit both in a purely entertainment context but also in a serious games context.

Our hint system design demonstrates that an extremely 'light touch' approach which guides players while adding very little additional information improves player experience. This approach could be applied within intelligent tutoring systems, expanded to other serious games or used within system design to better support users. It could significantly improve outcomes when using games outside a purely entertainment context.

Overall, the research presented in this thesis provides a new paradigm for player assistance which motivates further research within serious games, education and system design contexts.

REFERENCES

[1] Osama Abdel Raouf, Ibrahim El Henawy, and Mohamed Abdel Baset. “A Novel Hybrid Flower Pollination Algorithm with Chaotic Harmony Search for Solving Sudoku Puzzles”. In: *International Journal of Modern Education and Computer Science* 6.3 (Mar. 8, 2014), pp. 38–44. ISSN: 20750161, 2075017X. DOI: `10.5815/ijmecs.2014.03.05`. URL: `http://www.mecs-press.org/ijmecs/ijmecs-v6-n3/v6n3-5.html` (visited on 06/07/2023).

[2] Clark C. Abt. *Serious Games*. 4501 Forbes Blvd., Suite 200,Lanham,MD,20706: University Press of America, 1970. 200 pp. ISBN: 978-0-8191-6148-2.

[3] Aenigmatis. *Color Sudoku for Adults: No Numbers Just Colors - 3x3 Box Format*. N/A: Independently Published, July 16, 2018. 46 pp. ISBN: 978-1-71777-816-1. Google Books: `0lvIxQEACAAJ`.

[4] Justin T Alexander, John Sear, and Andreas Oikonomou. “An Investigation of the Effects of Game Difficulty on Player Enjoyment”. In: *Entertainment Computing* 4 (2013), pp. 53–62. DOI: `10.1016/j.entcom.2012.09.001`. URL: `http://dx.doi.org/10.1016/j.entcom.2012.09.001`.

[5] Alzheimer Society. *Challenging Your Brain*. Alzheimer Society of Canada. URL: `http://alzheimer.ca/en/help-support/im-living-dementia/living-well-dementia/challenging-your-brain` (visited on 06/05/2023).

[6] Dennis Ang and Alex Mitchell. “Comparing Effects of Dynamic Difficulty Adjustment Systems on Video Game Experience”. In: *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. CHI Play. CHI PLAY ’17. New York, NY, USA: Association for Computing Machinery, Oct. 15, 2017, pp. 317–327. ISBN: 978-1-4503-4898-0. DOI: `10.1145/3116595.3116623`. URL: `https://doi.org/10.1145/3116595.3116623` (visited on 01/06/2021).

[7] Maria-Virginia Aponte, Guillaume Levieux, and Stéphane Natkin. "Scaling the Level of Difficulty in Single Player Video Games". In: *Entertainment Computing – ICEC 2009*. Ed. by Stéphane Natkin and Jérôme Dupire. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2009, pp. 24–35. ISBN: 978-3-642-04052-8. DOI: `10.1007/978-3-642-04052-8_3`.

[8] Astraware Limited. *Sudoku Difficulty*. Sudoku of the Day. URL: `https://www.sudokuoftheday.com/about/difficulty/` (visited on 02/16/2021).

[9] André F. S. Barbosa, Pedro N. M. Pereira, João A. F. F. Dias, and Frutuoso G. M. Silva. "A New Methodology of Design and Development of Serious Games". In: *International Journal of Computer Games Technology* 2014 (2014), pp. 1–8. ISSN: 1687-7047, 1687-7055. DOI: `10.1155/2014/817167`. URL: `http://www.hindawi.com/journals/ijcgt/2014/817167/` (visited on 06/05/2023).

[10] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software* 67.1 (2015), pp. 1–48. DOI: `10.18637/jss.v067.i01`.

[11] Bernier-Lucien, Sherri. *Think! Teaching Logic in the Elementary or Middle Grades Classroom*. Charlotte Teachers Institute. URL: `https://charlotteteachers.org/2012/05/think-teaching-logic-in-the-elementary-or-middle-grades-classroom/` (visited on 07/01/2021).

[12] Big Fish Games. *Drawn: The Painted Tower | Drawn Games Official Fan Site*. drawngame. URL: `http://www.drawngame.com/games/the-painted-tower` (visited on 06/05/2023).

[13] Thiago Bittar, Luanna Lobato, Rodrigo Brum, D Peres, A Cintra, and Elson Longo. "Development of Two Educational Web Games: Chemical Sudoku and Nanotechnology Puzzle". In: (May 11, 2023).

[14] Carlton Books. *Sumoji: More Than 100 Emoji Sudoku*. Carlton Books, Limited, Sept. 6, 2018. 156 pp. ISBN: 978-1-78739-168-0. Google Books: `pk7UtwEACAAJ`.

[15] Daniel Bor. *The Ravenous Brain: How the New Science of Consciousness Explains Our Insatiable Search for Meaning*. Illustrated edition. New York: Basic Books, Sept. 13, 2012. 352 pp. ISBN: 978-0-465-02047-8.

[16] Jennifer L Branch. "Investigating the Information-Seeking Processes of Adolescents: The Value of Using Think Alouds and Think Afters". In: *Library & Information Science Research* 22.4 (2000), pp. 371–392. ISSN: 0740-8188.

[17] Richard Buday, Tom Baranowski, and Debbe Thompson. "Fun and Games and Boredom". In: *Games for Health Journal* 1.4 (Aug. 2012), pp. 257–261. ISSN: 2161-783X. DOI: `10.1089/g4h.2012.0026`. pmid: `24761316`. URL: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3833369/` (visited on 06/05/2023).

[18] Tristan Cazenave. "A Search Based Sudoku Solver". In: ().

[19] Jared E. Cechanowicz, Carl Gutwin, Scott Bateman, Regan Mandryk, and Ian Stavness. "Improving Player Balancing in Racing Games". In: *Proceedings of the First ACM SIGCHI Annual Symposium on Computer-human Interaction in Play*. CHI PLAY '14: The Annual Symposium on Computer-Human Interaction in Play. Toronto Ontario Canada: ACM, Oct. 19, 2014, pp. 47–56. ISBN: 978-1-4503-3014-5. DOI: `10.1145/2658537.2658701`. URL: `https://dl.acm.org/doi/10.1145/2658537.2658701` (visited on 06/07/2023).

[20] Seth B Chadwick, Rachel Krieg, and Christopher Granade. "Ease and Toil: Analyzing Sudoku". In: University of Alaska Fairbanks, 2007, pp. 363–380. URL: `http://eaton.math.rpi.edu/faculty/kramer/mcm/2008mcmsolutions.pdf#page=182`.

[21] Harrison Chapman and Malcolm E Rupert. "A Group-theoretic Approach to Human Solving Strategies in Sudoku". In: (2012), p. 17.

[22] Meng-Tzu Cheng, Jhih-Hao Chen, Sheng-Ju Chu, and Shin-Yen Chen. "The Use of Serious Games in Science Education: A Review of Selected Empirical Research from 2002 to 2013". In: *Journal of Computers in Education* 2.3 (Sept. 1, 2015), pp. 353–375. ISSN: 2197-9995. DOI: `10.1007/s40692-015-0039-9`. URL: `https://doi.org/10.1007/s40692-015-0039-9` (visited on 05/11/2023).

[23] Shira Chess. "A Time for Play: Interstitial Time, Invest/Express Games, and Feminine Leisure Style". In: *New Media & Society* 20.1 (Jan. 1, 2018), pp. 105–121. ISSN: 1461-4448. DOI: `10.1177/1461444816660729`. URL: `https://doi.org/10.1177/1461444816660729` (visited on 05/11/2023).

[24] Shira Chess and Christopher A. Paul. “The End of Casual: Long Live Casual”. In: *Games and Culture* 14.2 (Mar. 1, 2019), pp. 107–118. ISSN: 1555-4120. DOI: 10.1177/1555412018786652. URL: https://doi.org/10.1177/1555412018786652 (visited on 05/11/2023).

[25] Chris Campbell. *Casual Meets Core for a Drink: Developing Drawn*. Gamasutra. 2010. URL: https://www.gamedeveloper.com/design/casual-meets-core-for-a-drink-developing-i-drawn-i- (visited on 04/26/2022).

[26] Thomas M. Connolly, Elizabeth A. Boyle, Ewan MacArthur, Thomas Hainey, and James M. Boyle. “A Systematic Literature Review of Empirical Evidence on Computer Games and Serious Games”. In: *Computers & Education* 59.2 (Sept. 1, 2012), pp. 661–686. ISSN: 0360-1315. DOI: 10.1016/j.compedu.2012.03.004. URL: https://www.sciencedirect.com/science/article/pii/S0360131512000619 (visited on 06/05/2023).

[27] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, and Foldit Players. “Predicting Protein Structures with a Multiplayer Online Game”. In: *Nature* 466.7307 (7307 Aug. 2010), pp. 756–760. ISSN: 1476-4687. DOI: 10.1038/nature09304. URL: https://www.nature.com/articles/nature09304 (visited on 05/11/2023).

[28] *Core and Casual: What's the Difference?* VentureBeat. Apr. 30, 2011. URL: https://venturebeat.com/gbunfiltered/core-and-casual-whats-the-difference/ (visited on 05/11/2023).

[29] Amanda C Cote. “Casual Resistance: A Longitudinal Case Study of Video Gaming's Gendered Construction and Related Audience Perceptions”. In: *Journal of Communication* 70.6 (Dec. 17, 2020), pp. 819–841. ISSN: 0021-9916. DOI: 10.1093/joc/jqaa028. URL: https://doi.org/10.1093/joc/jqaa028 (visited on 05/11/2023).

[30] Nelson Cowan. “The Magical Mystery Four: How Is Working Memory Capacity Limited, and Why?” In: *Current directions in psychological science* 19.1 (2010), pp. 51–57.

[31] Thomas D Crute and Stephanie A Myers. *Chemistry for Everyone Sudoku Puzzles as Chemistry Learning Tools W*. 4. 2007. URL: www.cambridgesoft.com/software/ChemDraw/.

[32] Tom Davis. *Kenken for Teachers*. 2010. URL: http://geometer.org/mathcircles/kenken.pdf.

[33] Jean-Paul Delahaye. “The Science behind Sudoku”. In: *Scientific American* 294.6 (2006), pp. 80–87. JSTOR: `pdf/26061494.pdf`. URL: `https://www.jstor.org/stable/pdf/26061494.pdf`.

[34] Xiu Qin Deng and Yong Da Li. “A Novel Hybrid Genetic Algorithm for Solving Sudoku Puzzles”. In: *Optimization Letters* 7 (2013), pp. 241–257.

[35] Diane Dufort, Federico Tajariol, Ioan Roxin, and Université de Franche-Comté. “Bridging the Gap between Game Designers and Cultural Institutions: A Typology to Analyse and Classify Cultural Pervasive Games”. In: *Pervasive Games* (2016).

[36] Easybrain. *Sudoku Rules - Strategies, Solving Techniques and Tricks*. Sudoku.com. URL: `https://sudoku.com/sudoku-rules/` (visited on 04/02/2023).

[37] Rune Verpe Engeset, Gerit Pfuhl, Camilla Orten, Jordy Hendrikx, and Audun Hetland. “Colours and Maps for Communicating Natural Hazards to Users with and without Colour Vision Deficiency”. In: *International Journal of Disaster Risk Reduction* 76 (June 15, 2022), p. 103034. ISSN: 2212-4209. DOI: `10.1016/j.ijdrr.2022.103034`. URL: `https://www.sciencedirect.com/science/article/pii/S2212420922002539` (visited on 06/14/2023).

[38] Joan Espasa Arxer, Ian P Gent, Ruth Hoffmann, Christopher Jefferson, Matthew J McIlree, and Alice M Lynch. “Towards Generic Explanations for Pen and Paper Puzzles with MUSes”. In: *Proceedings of the SICSA eXplainable Artifical Intelligence Workshop 2021*. Aberdeen: SICSA, 2021, pp. 1–8.

[39] Joan Espasa Arxer, Ian P. Gent, Ruth Hoffmann, Christopher Jefferson, and Alice M. Lynch. “Using Small MUSes to Explain How to Solve Pen and Paper Puzzles”. In: *arXiv preprint* arXiv:2104.15040 (Apr. 30, 2021), pp. 1–24. URL: `https://research-repository.st-andrews.ac.uk/handle/10023/24086` (visited on 10/14/2021).

[40] Bahare Fatemi, Seyed Mehran Kazemi, and Nazanin Mehrasa. “Rating and Generating Sudoku Puzzles Based On Constraint Satisfaction Problems”. In: 8.10 (2014).

[41] Nicola Ferreira, Adrian Owen, Anita Mohan, Anne Corbett, and Clive Ballard. “Associations between Cognitively Stimulating Leisure Activities, Cognitive Function and Age-Related Cognitive Decline”. In: *International*

Journal of Geriatric Psychiatry 30.4 (2015), pp. 422–430. ISSN: 1099-1166. DOI: 10.1002/gps.4155. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/gps.4155 (visited on 06/05/2023).

[42] M. Fitzsimmons and H. Kunze. "Combining Hopfield Neural Networks, with Applications to Grid-Based Mathematics Puzzles". In: *Neural Networks* 118 (Oct. 1, 2019), pp. 81–89. ISSN: 0893-6080. DOI: 10.1016/j.neunet.2019.06.005. URL: https://www.sciencedirect.com/science/article/pii/S0893608019301789 (visited on 06/07/2023).

[43] Alan M. Frisch, Warwick Harvey, Chris Jefferson, Bernadette Martínez-Hernández, and Ian Miguel. "Essence: A Constraint Language for Specifying Combinatorial Problems". In: *Constraints* 13.3 (Sept. 1, 2008), pp. 268–306. ISSN: 1572-9354. DOI: 10.1007/s10601-008-9047-y. URL: https://doi.org/10.1007/s10601-008-9047-y (visited on 06/06/2023).

[44] Andrzej Gałecki and Tomasz Burzykowski. *Linear Mixed-Effects Models Using R: A Step-by-Step Approach*. New York, NY, USA: Springer Science & Business Media, Feb. 5, 2013. 558 pp. ISBN: 978-1-4614-3900-4. Google Books: rbk_AAAAQBAJ.

[45] Zong Woo Geem. "Harmony Search Algorithm for Solving Sudoku". In: *Knowledge-Based Intelligent Information and Engineering Systems*. Ed. by Bruno Apolloni, Robert J. Howlett, and Lakhmi Jain. Vol. 4692. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 371–378. ISBN: 978-3-540-74817-5. DOI: 10.1007/978-3-540-74819-9_46. URL: http://link.springer.com/10.1007/978-3-540-74819-9_46 (visited on 06/07/2023).

[46] Ian P. Gent, Ian Miguel, and Peter Nightingale. "Generalised Arc Consistency for the AllDifferent Constraint: An Empirical Survey". In: *Artificial Intelligence* 172.18 (Dec. 2008), pp. 1973–2000. ISSN: 00043702. DOI: 10.1016/j.artint.2008.10.006. URL: https://linkinghub.elsevier.com/retrieve/pii/S0004370208001410 (visited on 06/06/2023).

[47] Gergana Georgieva, Sylvester Arnab, Margarida Romero, and Sara de Freitas. "Transposing Freemium Business Model from Casual Games to Serious Games". In: *Entertainment Computing* 9–10 (June 1, 2015), pp. 29–41. ISSN: 1875-9521. DOI: 10.1016/j.entcom.2015.07.003. URL: https://www.sciencedirect.com/science/article/pii/S1875952115000099 (visited on 05/11/2023).

[48] Samuel J. Gershman, Eric J. Horvitz, and Joshua B. Tenenbaum. “Computational Rationality: A Converging Paradigm for Intelligence in Brains, Minds, and Machines”. In: *Science* 349.6245 (July 17, 2015), pp. 273–278. ISSN: 0036-8075, 1095-9203. DOI: `10.1126/science.aac6076`. pmid: `26185246`. URL: `https://science.sciencemag.org/content/349/6245/273` (visited on 02/15/2021).

[49] Jacob Goldberger. “Solving Sudoku Using Combined Message Passing Algorithms”. School of Engineering, Bar-Ilan University.

[50] Fiorella Grossi and Wayne Gould. *The Addict’s Guide to Everything Sudoku*. Fair Winds, 2007. ISBN: 978-1-61059-517-9. Google Books: `2MLsVxlmFAsC`.

[51] Stefan Gudmundsson, Philipp Eisen, Erik Poromaa, Alex Nodet, Sami Purmonen, Bartlomiej Kozakowski, Richard Meurling, and Lele Cao. “Human-Like Playtesting with Deep Learning”. In: Aug. 1, 2018, pp. 1–8. DOI: `10.1109/CIG.2018.8490442`.

[52] Sönke Hartmann. “Puzzle—More Logic Puzzle Apps Solved by Mathematical Programming”. In: *INFORMS Transactions on Education* 20.1 (Sept. 2019), pp. 49–55. ISSN: 1532-0545, 1532-0545. DOI: `10.1287/ited.2019.0212`. URL: `https://pubsonline.informs.org/doi/10.1287/ited.2019.0212` (visited on 06/07/2023).

[53] Hao He, Yulin Zhu, and Wei Cai. “An Adaptive Hint System for Puzzle Games: A Multimodal-based Approach”. In: *2022 IEEE Games, Entertainment, Media Conference (GEM)*. 2022 IEEE Games, Entertainment, Media Conference (GEM). Nov. 2022, pp. 1–6. DOI: `10.1109/GEM56474.2022.10017301`.

[54] C. Heath, J. Hindmarsh, and P. Luff. *Video in Qualitative Research*. Introducing Qualitative Methods Series. SAGE Publications, 2010. ISBN: 978-1-4462-4669-6.

[55] Cecilia Heyes. “Born Pupils? Natural Pedagogy and Cultural Pedagogy”. In: *Perspectives on Psychological Science* 11.2 (Mar. 1, 2016), pp. 280–295. ISSN: 1745-6916. DOI: `10.1177/1745691615621276`. URL: `https://doi.org/10.1177/1745691615621276` (visited on 06/14/2023).

[56] Eilean Hooper-Greenhill. *The Educational Role of the Museum*. Psychology Press, 1999. 366 pp. ISBN: 978-0-415-19826-4.

[57] Talke Klara Hoppmann. “Examining the ‘Point of Frustration’. The Think-Aloud Method Applied to Online Search Tasks”. In: *Quality & Quantity* 43.2 (2009), pp. 211–224. ISSN: 1573-7845.

[58] *How to Solve Star Battle Puzzles*. URL: http://www.clarity-media.co.uk/puzzleblog/how-solve-star-battle-strategy (visited on 06/05/2023).

[59] Kenton Taylor Howard. “Free-to-Play or Pay-to-Win? Casual, Hardcore, and Hearthstone”. In: *Transactions of the Digital Games Research Association* 4.3 (Oct. 11, 2019). ISSN: 2328-9422, 2328-9414. DOI: 10.26503/todigra.v4i3.103. URL: http://todigra.org/index.php/todigra/article/view/103 (visited on 05/11/2023).

[60] Sen Huang, Jin-cai Huang, Qing Chen, Sheng-yun Liu, and Yan-jun Liu. “The Research of Efficient Algorithm to Generate Sudoku Puzzle”. In: *Proceedings of the 2012 International Conference on Information Technology and Software Engineering*. Ed. by Wei Lu, Guoqiang Cai, Weibin Liu, and Weiwei Xing. Lecture Notes in Electrical Engineering. Berlin, Heidelberg: Springer, 2013, pp. 721–730. ISBN: 978-3-642-34522-7. DOI: 10.1007/978-3-642-34522-7_77.

[61] Martin Hunt, Christopher Pong, and George Tucker. “Difficulty-Driven Sudoku Puzzle Generation”. In: *The UMAP Journal* 29.3 (2007), pp. 343–362.

[62] IGN Staff. *The Death of the Manual - IGN*. IGN. 2010. URL: https://www.ign.com/articles/2010/04/22/the-death-of-the-manual (visited on 05/10/2023).

[63] Wijnand A IJsselsteijn, Yvonne AW De Kort, and Karolien Poels. “The Game Experience Questionnaire”. In: (2013).

[64] Petr Jarušek and Radek Pelánek. “Difficulty Rating of Sokoban Puzzle”. In: *Proceedings of the 2010 Conference on STAIRS 2010: Proceedings of the Fifth Starting AI Researchers’ Symposium*. NLD: IOS Press, Aug. 11, 2010, pp. 140–150. ISBN: 978-1-60750-675-1.

[65] Dhanya Job and Varghese Paul. “Recursive Backtracking for Solving 9*9 Sudoku Puzzle”. In: *Bonfring International Journal of Data Mining* 6.1 (Jan. 31, 2016), pp. 07–09. ISSN: 2250107X, 22775048. DOI: 10.9756/BIJDM.8128. URL: http://journal.bonfring.org/abstract.php?id=2&archiveid=484 (visited on 06/05/2023).

[66] Mark R Johnson. “Casual Games Before Casual Games: Historicizing Paper Puzzle Games in an Era of Digital Play”. In: *Games and Culture* 14.2 (2019), pp. 119–138. DOI: `10.1177/1555412018790423`. URL: `https://journals.sagepub.com/doi/pdf/10.1177/1555412018790423`.

[67] Ankur Joshi, Saket Kale, Satish Chandel, and D. Pal. “Likert Scale: Explored and Explained”. In: *British Journal of Applied Science & Technology* 7.4 (Jan. 10, 2015), pp. 396–403. ISSN: 22310843. DOI: `10.9734/BJAST/2015/14975`. URL: `https://journalcjast.com/index.php/CJAST/article/view/381` (visited on 06/08/2023).

[68] Narendra Jussien. *A-Z of Sudoku*. ISTE Ltd., 2007.

[69] Yasmin B. Kafai. “Constructionist Visions: Hard Fun with Serious Games”. In: *International Journal of Child-Computer Interaction* 18 (Nov. 1, 2018), pp. 19–21. ISSN: 2212-8689. DOI: `10.1016/j.ijcci.2018.04.002`. URL: `https://www.sciencedirect.com/science/article/pii/S2212868917300478` (visited on 06/05/2023).

[70] Nobuhiko Kanamoto. *Why Hand Made?* Nikoli. URL: `https://www.nikoli.co.jp/en/puzzles/why_hand_made.html` (visited on 02/13/2021).

[71] Setargew Kenaw. “Hubert L. Dreyfus’s Critique of Classical AI and Its Rationalist Assumptions”. In: *Minds and Machines* 18.2 (June 1, 2008), pp. 227–238. ISSN: 0924-6495. DOI: `10.1007/s11023-008-9093-7`. URL: `https://doi.org/10.1007/s11023-008-9093-7` (visited on 09/13/2020).

[72] Madison Klarkowski, Daniel Johnson, Peta Wyeth, Mitchell McEwan, Cody Phillips, and Simon Smith. “Operationalising and Evaluating Sub-Optimal and Optimal Play Experiences through Challenge-Skill Manipulation”. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI ’16. New York, NY, USA: Association for Computing Machinery, May 7, 2016, pp. 5583–5594. ISBN: 978-1-4503-3362-7. DOI: `10.1145/2858036.2858563`. URL: `https://dl.acm.org/doi/10.1145/2858036.2858563` (visited on 05/09/2023).

[73] Raph Koster. *Theory of Fun for Game Design*. 2nd ed. O’Reilly Media, Inc., 2013. ISBN: 1-4493-6321-0.

[74] Niklas Kühl, Marc Goutier, Lucas Baier, Clemens Wolff, and Dominik Martin. *Human vs. Supervised Machine Learning: Who Learns Patterns Faster?* Nov. 30, 2020. arXiv: `2012.03661` [cs]. URL: `http://arxiv.org/abs/2012.03661` (visited on 06/05/2023). preprint.

[75] Gihwon Kwon and Himanshu Jain. "Optimized CNF Encoding for Sudoku Puzzles". In: *In Proc. 13th International Conference on Logic for Programming Artificial Intelligence and Reasoning* LPAR2006 (2006), pp. 1–5.

[76] Susan M. Landau, Shawn M. Marks, Elizabeth C. Mormino, Gil D. Rabinovici, Hwamee Oh, James P. O'Neil, Robert S. Wilson, and William J. Jagust. "Association of Lifetime Cognitive Engagement and Low β-Amyloid Deposition". In: *Archives of Neurology* 69.5 (May 1, 2012), pp. 623–629. ISSN: 0003-9942. DOI: `10.1001/archneurol.2011.2748`. URL: `https://doi.org/10.1001/archneurol.2011.2748` (visited on 06/05/2023).

[77] Frédéric Lardeux, Eric Monfroy, Frédéric Saubion, Broderick Crawford, and Carlos Castro. "Overlapping Alldierent Constraints and the Sudoku Puzzle". In: (June 7, 2023).

[78] Chiapello Laureline. "Formalizing Casual Games: A Study Based on Game Designers' Professional Knowledge". In: (2014). URL: `http://www.digra.org/wp-content/uploads/digital-library/paper_168.pdf`.

[79] Richard L. Lewis, Andrew Howes, and Satinder Singh. "Computational Rationality: Linking Mechanism and Behavior Through Bounded Utility Maximization". In: *Topics in Cognitive Science* 6.2 (Apr. 2014), pp. 279–311. ISSN: 17568757. DOI: `10.1111/tops.12086`. URL: `http://doi.wiley.com/10.1111/tops.12086` (visited on 02/15/2021).

[80] Conor Linehan, George Bellord, Ben Kirman, Zachary H. Morford, and Bryan Roche. "Learning Curves: Analysing Pace and Challenge in Four Successful Puzzle Games". In: *Proceedings of the First ACM SIGCHI Annual Symposium on Computer-human Interaction in Play*. CHI PLAY '14: The Annual Symposium on Computer-Human Interaction in Play. Toronto Ontario Canada: ACM, Oct. 19, 2014, pp. 181–190. ISBN: 978-1-4503-3014-5. DOI: `10.1145/2658537.2658695`. URL: `https://dl.acm.org/doi/10.1145/2658537.2658695` (visited on 02/12/2021).

[81] Zhongxiu Liu. "Data-Driven Hint Generation from Peer Debugging Solutions". In: *International Educational Data Mining Society* (2015), p. 3.

[82] Huw Lloyd and Martyn Amos. "Solving Sudoku With Ant Colony Optimization". In: *IEEE Transactions on Games* 12.3 (Sept. 2020), pp. 302–311. ISSN: 2475-1510. DOI: 10.1109/TG.2019.2942773.

[83] Steve Lohr. "For Impatient Web Users, an Eye Blink Is Just Too Long to Wait". In: *New York Times* (February 29 2012).

[84] Inês Lynce and Joël Ouaknine. "Sudoku as a SAT Problem." In: *AI&M*. 2006.

[85] Lynch Alice, Jefferson Chris, and Hinrichs Uta. "Considering the Person in the Puzzle: Challenging Common Assumptions about Sudoku Player Strategies". In: *DiGRA &rsquo22 &ndash Proceedings of the 2022 DiGRA International Conference: Bringing Worlds Together*. 2022. URL: http://www.digra.org/wp-content/uploads/digital-library/DiGRA_2022_paper_124.pdf.

[86] Arnab Kumar Maji, Sunanda Jana, and Rajat Kumar Pal. "A Comprehensive Sudoku Instance Generator". In: *Advanced Computing and Systems for Security: Volume 2*. Ed. by Rituparna Chaki, Agostino Cortesi, Khalid Saeed, and Nabendu Chaki. New Delhi: Springer India, 2016, pp. 215–233. ISBN: 978-81-322-2653-6. DOI: 10.1007/978-81-322-2653-6_15. URL: https://doi.org/10.1007/978-81-322-2653-6_15.

[87] Arnab Kumar Maji and Rajat Kumar Pal. "Sudoku Solver Using Minigrid Based Backtracking". In: *2014 IEEE International Advance Computing Conference (IACC)*. 2014 IEEE International Advance Computing Conference (IACC). Feb. 2014, pp. 36–44. DOI: 10.1109/IAdCC.2014.6779291.

[88] Pavlos Malakonakis, Miltiadis Smerdis, Euripides Sotiriades, and Apostolos Dollas. "An FPGA-based Sudoku Solver Based on Simulated Annealing Methods". In: *2009 International Conference on Field-Programmable Technology*. 2009 International Conference on Field-Programmable Technology. Dec. 2009, pp. 522–525. DOI: 10.1109/FPT.2009.5377608.

[89] Timo Mantere. "Improved Ant Colony Genetic Algorithm Hybrid for Sudoku Solving". In: *2013 Third World Congress on Information and Communication Technologies (WICT 2013)*. 2013 Third World Congress on Information and Communication Technologies (WICT 2013). Dec. 2013, pp. 274–279. DOI: 10.1109/WICT.2013.7113148.

[90] Timo Mantere and Janne Koljonen. "Solving, Rating and Generating Sudoku Puzzles with GA". In: *2007 IEEE Congress on Evolutionary Computation*. 2007 IEEE Congress on Evolutionary Computation. Sept. 2007, pp. 1382–1389. DOI: 10.1109/CEC.2007.4424632.

[91] Catherine Marshall and Gretchen Rossman. *Designing Qualitative Research - Catherine Marshall, Gretchen B. Rossman - Google Books*. 6th ed. SAGE Publications, 2014.

[92] Matthew Wickline. *Coblis — Color Blindness Simulator – Colblindor*. Colblindor. 2001. URL: https://www.color-blindness.com/coblis-color-blindness-simulator/ (visited on 06/14/2023).

[93] Melvin. *Discover How Grading Sudoku Puzzles And Games Is Done*. Sudoku Essentials. 2021. URL: https://www.sudokuessentials.com/grading-sudoku-puzzles (visited on 02/13/2021).

[94] Mihaly Csikszentmihalyi. *FLOW: The Psychology of Optimal Experience*. HaperPerennial, 1990. URL: https://static1.squarespace.com/static/547fd964e4b082d85a18eab7/t/54e2b1b7e4b0902efd66fe58/1424142775852/flow_the_psychology_of_optimal_experience-2-2.pdf.

[95] Tetsuya Miyamoto. *The Times: KenKen Book 1: The New Brain-Training Puzzle Phenomenon*. HarperCollins Publishers Limited, May 2008. 176 pp. ISBN: 978-0-00-728824-3. Google Books: CwgdNQAACAAJ.

[96] Atasi Mohanty, Rahul Sarkar, and Siddhartha Chaudhury. "Design Engineering Design and Development of Digital Game- Based Learning Software for Incorporation into School Syllabus and Curriculum Transaction". In: Jan. 23, 2023, pp. 4864–4900.

[97] Todd K. Moon and Jacob H. Gunther. "Multiple Constraint Satisfaction by Belief Propagation: An Example Using Sudoku". In: *2006 IEEE Mountain Workshop on Adaptive and Learning Systems*. 2006 IEEE Mountain Workshop on Adaptive and Learning Systems. July 2006, pp. 122–126. DOI: 10.1109/SMCALS.2006.250702.

[98] Alaa Morsy, Hegazy Sabry, and Naglaa Ragaa. "A BLACK WIDOW OPTIMIZATION ALGORITHM FOR SOLVING KENKEN PROBLEM". In: *Seybold Report* 17 (Oct. 16, 2022), p. 1232. DOI: 10.5281/zenodo.7106602.

[99] Tamara Munzner. “Visualization”. In: *Fundamentals of Computer Graphics*. AK Peters/CRC Press, 2018, pp. 665–699.

[100] Jaclyn M Murawska. “KenKen Puzzles for Developing Number Sense and Positive Mathematics Identify in Elementary School”. In: *Success in High-Need Schools Journal* 14.1 (2018), p. 21.

[101] Fiona Fui-Hoon Nah. “A Study on Tolerable Waiting Time: How Long Are Web Users Willing to Wait?” In: *Behaviour & Information Technology* 23.3 (May 2004), pp. 153–163. ISSN: 0144-929X, 1362-3001. DOI: 10.1080/01449290410001669914. URL: http://www.tandfonline.com/doi/abs/10.1080/01449290410001669914 (visited on 06/06/2023).

[102] Tiia Ngandu, Jenni Lehtisalo, Alina Solomon, Esko Levälahti, Satu Ahtiluoto, Riitta Antikainen, Lars Bäckman, Tuomo Hänninen, Antti Jula, Tiina Laatikainen, Jaana Lindström, Francesca Mangialasche, Teemu Paajanen, Satu Pajala, Markku Peltonen, Rainer Rauramaa, Anna Stigsdotter-Neely, Timo Strandberg, Jaakko Tuomilehto, Hilkka Soininen, and Miia Kivipelto. “A 2 Year Multidomain Intervention of Diet, Exercise, Cognitive Training, and Vascular Risk Monitoring versus Control to Prevent Cognitive Decline in at-Risk Elderly People (FINGER): A Randomised Controlled Trial”. In: *The Lancet* 385.9984 (June 6, 2015), pp. 2255–2263. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(15)60461-5. pmid: 25771249. URL: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(15)60461-5/fulltext (visited on 06/05/2023).

[103] Peter Nightingale, Özgür Akgün, Ian P. Gent, Christopher Jefferson, Ian Miguel, and Patrick Spracklen. “Automatically Improving Constraint Models in Savile Row”. In: *Artificial Intelligence* 251 (Oct. 1, 2017), pp. 35–61. ISSN: 0004-3702. DOI: 10.1016/j.artint.2017.07.001. URL: https://www.sciencedirect.com/science/article/pii/S0004370217300747 (visited on 06/07/2023).

[104] Gali Noti. “Do Humans Play Equilibrium?: Modeling Human Behavior in Computational Strategic Systems”. In: *XRDS: Crossroads, The ACM Magazine for Students* 24.1 (Sept. 14, 2017), pp. 29–33. ISSN: 15284972. DOI: 10.1145/3123740. URL: http://dl.acm.org/citation.cfm?doid=3140569.3123740 (visited on 01/24/2020).

[105] Jaysonne A. Pacurib, Glaiza Mae M. Seno, and John Paul T. Yusiong. "Solving Sudoku Puzzles Using Improved Artificial Bee Colony Algorithm". In: *2009 Fourth International Conference on Innovative Computing, Information and Control (ICICIC)*. 2009 Fourth International Conference on Innovative Computing, Information and Control (ICICIC). Dec. 2009, pp. 885–888. DOI: `10.1109/ICICIC.2009.334`.

[106] *Pattern Recognition in Machine Learning*. June 1, 2023. URL: `https://labelyourdata.com/articles/pattern-recognition-in-machine-learning` (visited on 06/05/2023).

[107] Radek Pelánek. "Difficulty Rating of Sudoku Puzzles by a Computational Model". In: *Twenty-Fourth International FLAIRS Conference* (2011). URL: `www.aaai.org`.

[108] Radek Pelánek. *Difficulty Rating of Sudoku Puzzles: An Overview and Evaluation*. 2014. URL: `https://arxiv.org/pdf/1403.7373.pdf`.

[109] Radek Pelánek. *Human Problem Solving: Sudoku Case Study Human Problem Solving: Sudoku Case Study*. 2011. URL: `http://www.fi.muni.cz/reports/`.

[110] Alice L. Perez and G. Lamoureux. "Sudoku Puzzles for First-Year Organic Chemistry Students". In: *Journal of Chemical Education* 84.4 (Apr. 2007), p. 614. ISSN: 0021-9584, 1938-1328. DOI: `10.1021/ed084p614`. URL: `https://pubs.acs.org/doi/abs/10.1021/ed084p614` (visited on 05/11/2023).

[111] Michael B. Pitt, Emily C. Borman-Shoap, and Walter J. Eppich. "Twelve Tips for Maximizing the Effectiveness of Game-Based Learning". In: *Medical Teacher* 37.11 (Nov. 2, 2015), pp. 1013–1017. ISSN: 0142-159X, 1466-187X. DOI: `10.3109/0142159X.2015.1020289`. URL: `http://www.tandfonline.com/doi/full/10.3109/0142159X.2015.1020289` (visited on 06/14/2023).

[112] Lev Poretski and Anthony Tang. "Press A to Jump: Design Strategies for Video Game Learnability". In: *CHI Conference on Human Factors in Computing Systems*. CHI '22: CHI Conference on Human Factors in Computing Systems. New Orleans LA USA: ACM, Apr. 27, 2022, pp. 1–26. ISBN: 978-1-4503-9157-3. DOI: `10.1145/3491102.3517685`. URL: `https://dl.acm.org/doi/10.1145/3491102.3517685` (visited on 06/08/2023).

[113] Shammis Press. *Brain Training Sudoku Puzzle: Ideal for the Sudoku Solver Looking for Variety and Challenge, Our Books Are Available in Various Solving Levels*. Independently Published, July 11, 2020. 88 pp. Google Books: `2sOrzQEACAAJ`.

[114] Puzzle Magazine. *Binary Puzzle Solving Methods*. Puzzle Magazine. URL: `https://www.puzzle-magazine.com/solving-binary-puzzle.php` (visited on 06/05/2023).

[115] PuzzleTeam. *Aquarium - Online Puzzle Game*. Aquarium. 2020. URL: `https://web.archive.org/web/20230531084348/https://www.puzzle-aquarium.com/` (visited on 05/31/2023).

[116] PuzzleTeam. *Binairo - Online Puzzle Game*. 2020. URL: `https://www.puzzle-binairo.com/` (visited on 06/08/2023).

[117] PuzzleTeam. *Skyscrapers - Online Puzzle Game*. 2020. URL: `https://www.puzzle-skyscrapers.com/` (visited on 06/08/2023).

[118] PuzzleTeam. *Star Battle - Online Puzzle Game*. Star Battle. 2020. URL: `https://www.puzzle-star-battle.com/` (visited on 06/08/2023).

[119] PuzzleTeam. *Sudoku - Online Puzzle Game*. 2020. URL: `https://www.puzzle-sudoku.com/` (visited on 06/08/2023).

[120] PuzzleTeam. *Tents - Online Puzzle Game*. 2020. URL: `https://www.puzzle-tents.com/` (visited on 06/08/2023).

[121] Python Software Foundation. *Pickle — Python Object Serialization*. The Python Standard Library. 2021. URL: `https://docs.python.org/3/library/pickle.html` (visited on 05/06/2023).

[122] Stephen K. Reed. *Psychological Processes in Pattern Recognition*. Academic Press, Sept. 11, 2013. 261 pp. ISBN: 978-1-4832-6334-2. Google Books: `6INGBQAAQBAJ`.

[123] Christopher G Reeson, Kai-chen Huang, Kenneth M Bayer, and Berthe Y Choueiry. "An Interactive Constraint-Based Approach to Sudoku". In: (2007), pp. 1976–1977.

[124] Harold B Reiter, John Thornton, and G Patrick Vennebush. "Using KenKen to Build Reasoning Skills". In: *The Mathematics Teacher* 107.5 (2013), pp. 341–347.

[125] Shaghayegh Roohi, Asko Relas, Jari Takatalo, Henri Heiskanen, and Perttu Hämäläinen. "Predicting Game Difficulty and Churn Without Players". In: *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. CHI PLAY '20. New York, NY, USA: Association for Computing Machinery, Nov. 2, 2020, pp. 585–593. ISBN: 978-1-4503-8074-4. DOI: `10.1145/3410404.3414235`. URL: `https://doi.org/10.1145/3410404.3414235` (visited on 01/06/2021).

[126] Francesca Rossi, Peter van Beek, and Toby Walsh, eds. *Handbook of Constraint Programming*. Elsevier, 2006.

[127] Julian Runge, Peng Gao, Florent Garcin, and Boi Faltings. "Churn Prediction for High-Value Players in Casual Social Games". In: *2014 IEEE Conference on Computational Intelligence and Games*. 2014 IEEE Conference on Computational Intelligence and Games. Aug. 2014, pp. 1–8. DOI: `10.1109/CIG.2014.6932875`.

[128] Ed Russell and Frazer Jarvis. "Mathematics of Sudoku II". In: *Mathematical Spectrum* 39.2 (2006), pp. 54–58.

[129] Johnny Saldaña. *The Coding Manual for Qualitative Researchers*. Sage, 2015.

[130] Denis Savenkov and Eugene Agichtein. "To Hint or Not: Exploring the Effectiveness of Search Hints for Complex Informational Tasks". In: *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*. SIGIR '14: The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. Gold Coast Queensland Australia: ACM, July 3, 2014, pp. 1115–1118. ISBN: 978-1-4503-2257-7. DOI: `10.1145/2600428.2609523`. URL: `https://dl.acm.org/doi/10.1145/2600428.2609523` (visited on 05/08/2023).

[131] Ben Sawyer and Peter Smith. *Serious Games Taxonomy*. 2008. URL: `https://web.archive.org/web/20101117054600/http://www.seriousgames.org/presentations/serious-games-taxonomy-2008_web.pdf` (visited on 05/11/2023).

[132] Jerry Schnepp and Christian Rogers. "Just Give Me a Hint! An Alternative Testing Approach for Simultaneous Assessment and Learning". In: *Smart Education and Smart E-Learning*. Ed. by Vladimir L. Uskov, Robert J. Howlett, and Lakhmi C. Jain. Smart Innovation, Systems and Technologies. Cham: Springer International Publishing, 2015, pp. 141–150. ISBN: 978-3-319-19875-0. DOI: `10.1007/978-3-319-19875-0_13`.

[133] Skipper Seabold and Josef Perktold. “Statsmodels: Econometric and Statistical Modeling with Python”. In: *9th Python in Science Conference*. 2010.

[134] M Selinker and T Snyder. *Puzzle Craft: The Ultimate Guide on How to Construct Every Kind of Puzzle*. Sterling Publishing Company, Incorporated, 2013. ISBN: 978-1-4027-7924-4.

[135] *Serious Games Market Share, Size, Industry Forecast 2023-2028*. URL: `https://www.imarcgroup.com/serious-games-market` (visited on 06/05/2023).

[136] Simon Tatham. *KenKen*. Online Math Learning Interactive Area. URL: `https://www.interactive.onlinemathlearning.com/fun_kenken.php` (visited on 06/08/2023).

[137] Helmut Simonis. “Sudoku as a Constraint Problem”. In: CP Workshop on Modeling and Reformulating Constraint Satisfaction Problems. Vol. 12. Citeseer, 2005, pp. 13–27.

[138] Jason J. Sroka and Louis D. Braida. “Human and Machine Consonant Recognition”. In: *Speech Communication* 45.4 (Apr. 1, 2005), pp. 401–423. ISSN: 0167-6393. DOI: `10.1016/j.specom.2004.11.009`. URL: `https://www.sciencedirect.com/science/article/pii/S0167639304001499` (visited on 06/05/2023).

[139] Roger T. Staff, Michael J. Hogan, Daniel S. Williams, and L. J. Whalley. “Intellectual Engagement and Cognitive Ability in Later Life (the “Use It or Lose It” Conjecture): Longitudinal, Prospective Study”. In: *BMJ* 363 (Dec. 10, 2018), k4925. ISSN: 0959-8138, 1756-1833. DOI: `10.1136/bmj.k4925`. pmid: `30530522`. URL: `https://www.bmj.com/content/363/bmj.k4925` (visited on 06/05/2023).

[140] A Stuart. *The Logic of Sudoku*. Michael Mepham, 2007. ISBN: 978-0-9554841-0-0.

[141] Andrew Stuart. *SudokuWiki.Org*. 2008. URL: `http://www.sudokuwiki.org/` (visited on 11/12/2018).

[142] Andrew C Stuart. “Sudoku Creation and Grading”. In: *Mathematica* 39.6 (2007), pp. 126–142.

[143] Andrew C Stuart. *SudokuWiki.Org - The Relative Incidence of Sudoku Strategies*. URL: `https://www.sudokuwiki.org/The_Relative_Incidence_of_Sudoku_Strategies` (visited on 02/13/2021).

[144] *Sudoku Dragon: Sudoku Puzzle Solving Strategies*. URL: http://sudokudragon.com/sudokustrategy.htm (visited on 11/12/2018).

[145] *Sudoku Snake - Solving Techniques*. URL: http://www.sudokusnake.com/techniques.php (visited on 11/12/2018).

[146] Chuen-Tsai Sun, Dai-Yi Wang, and Hui-Ling Chan. "How Digital Scaffolds in Games Direct Problem-Solving Behaviors". In: *Computers & Education* 57 (2011), pp. 2118–2125. DOI: 10.1016/j.compedu.2011.05.022. URL: https://ac.els-cdn.com/S036013151100128X/1-s2.0-S036013151100128X-main.pdf?_tid=bf4bfaac-8f5c-4eba-a60c-37a26f9ab257&acdnat=1539614255_111ddba793d2156f7109f17b7026b2cc.

[147] Tarja Susi, Mikael Johannesson, and Per Backlund. "Serious Games: An Overview". In: (2007).

[148] Aaquib Tabrez, Shivendra Agrawal, and Bradley Hayes. "Explanation-Based Reward Coaching to Improve Human Performance via Reinforcement Learning". In: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI). Daegu, Korea (South): IEEE, Mar. 2019, pp. 249–257. ISBN: 978-1-5386-8555-6. DOI: 10.1109/HRI.2019.8673104. URL: https://ieeexplore.ieee.org/document/8673104/ (visited on 01/24/2020).

[149] SENIOR TASSE. *16x16 Sudoku Puzzles for Adults Easy to Hard: Handmade Puzzles: Attractive and Pleasing to the Eye | 16x16 Size: Best for Memory and Focus | Large ... per Page | Sudoku Puzzle Book for Adults*. Independently published, Mar. 3, 2023. 140 pp.

[150] Loren G. Terveen. "Overview of Human-Computer Collaboration". In: *Knowledge-Based Systems*. Human-Computer Collaboration 8.2 (Apr. 1, 1995), pp. 67–81. ISSN: 0950-7051. DOI: 10.1016/0950-7051(95)98369-H. URL: http://www.sciencedirect.com/science/article/pii/095070519598369H (visited on 09/16/2020).

[151] Y L Teresa Ting. "Learning to Hypothesize with Confidence through Sudoku Game Play–". In: *$Number$* 1 (2009), p. 5.

[152] Valve. *Portal on Steam*. URL: https://store.steampowered.com/app/400/Portal/ (visited on 05/10/2023).

[153] MW Van Someren, YF Barnard, and JAC Sandberg. "The Think Aloud Method: A Practical Approach to Modelling Cognitive". In: *London: AcademicPress* (1994).

[154] Daniel J. Ware. *The Slow and Silent Death of the Video Game Manual*. SUPERJUMP. Feb. 26, 2020. URL: https://medium.com/super-jump/the-slow-and-silent-death-of-the-video-game-manual-cb22eb3167bf (visited on 05/10/2023).

[155] Helen Wauck and Wai-Tat Fu. "A Data-Driven, Multidimensional Approach to Hint Design in Video Games". In: *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. IUI'17: 22nd International Conference on Intelligent User Interfaces. Limassol Cyprus: ACM, Mar. 7, 2017, pp. 137–147. ISBN: 978-1-4503-4348-0. DOI: 10.1145/3025171.3025224. URL: https://dl.acm.org/doi/10.1145/3025171.3025224 (visited on 10/19/2020).

[156] Matthew M White. *Learn to Play: Designing Tutorials for Video Games*. CRC Press, 2014.

[157] Markus Wiemker, Errol Elumir, and Adam Clare. "Escape Room Games". In: *Game based learning* 55 (2015), pp. 55–75.

[158] Alan Williams-Key. *Ultimate Sudoku Secrets: Challenging Hand-Crafted Sudoku Puzzles and How to Solve Them*. 1st edition. CreateSpace Independent Publishing Platform, Dec. 17, 2014. 312 pp. ISBN: 978-1-5055-3244-9.

[159] Donghee Yvette Wohn. "Gender and Race Representation in Casual Games". In: *Sex Roles* 65.3 (Aug. 1, 2011), pp. 198–207. ISSN: 1573-2762. DOI: 10.1007/s11199-011-0007-4. URL: https://doi.org/10.1007/s11199-011-0007-4 (visited on 05/11/2023).

[160] Chungen Xu and Weng Xu. "The Model and Algorithm to Estimate the Difficulty Levels of Sudoku Puzzles". In: *Journal of Mathematics Research* 1.2 (2009), pp. 43–46. URL: https://pdfs.semanticscholar.org/87cc/6591845d4023aeeec8121aa20f72dc4d32c7.pdf.

PART I

APPENDIX

APPENDIX A

STUDY QUESTIONNAIRES AND SURVEYS

This appendix contains all questionnaires and surveys used during the work described in this thesis.

A.1 Sudoku Survey Questions

Sudoku Background Survey

Start of Block: Information and Consent

Q23 **Studying Strategies and Preferences when Solving Puzzles** **Researchers** *Alice Lynch, Uta Hinrichs, Chris Jefferson* The aim of this survey is to collect research on people's approaches to solving puzzles, in particular Sudoku and what makes them enjoyable. All data that you provide will be anonymized prior to any presentation or publication. You can abort the survey at any point and without providing an explanation. All information you have provided until this point will be discarded. We will NOT ask you for any identifying information (e.g. name) as part of this survey. However, at the end of the survey, there is an option for you to provide your name and email address if you would like to participate in a follow up study. The secondary study will take place in person in St Andrews, Fife, UK. Its goal is to look at solving techniques in more detail. Again, this is completely optional and does not represent agreement to the follow up study – just that you are happy to be contacted regarding it. All data you provide through the survey tool will be stored in a secure database which is only accessible by the researchers listed above. Data will be stored for 3 years before being destroyed. Results of this research may be published or presented in anonymized form at academic venues. We will not present or publish any identifying information about survey participants. There are no known risks associated with this research that exceed that of everyday situations. If you have further questions or concerns, please contact us. What should I do if I have concerns about this study? A full outline of the procedures governed by the University Teaching and Research Ethical Committee is available at http://www.st-andrews.ac.uk/utrec/guidelinespolicies/complaints/

Q24 I agree to participate in the study given the above information

○ Yes (1)

○ No (2)

Skip To: End of Survey If I agree to participate in the study given the above information = No

End of Block: Information and Consent

Start of Block: About You

Q4 Gender

○ Male (1)

○ Female (2)

○ Other (3)

Q5 Age Group

○ 18 - 24 (1)

○ 25 - 34 (2)

○ 35 - 44 (3)

○ 45 - 54 (4)

○ 55 - 64 (5)

○ 65 - 74 (6)

○ 75 - 84 (7)

○ 85 or older (8)

Q18 Highest Educational Level

○ Further Education beyond undergraduate (1)

○ Undergraduate Degree (2)

○ A-level (or equivalent) (3)

○ GCSE (or equivalent) (4)

Display This Question:

If Highest Educational Level = Undergraduate Degree

Or Highest Educational Level = Further Education beyond undergraduate

Q26 Please specify your undergraduate degree topic

____________________*single line free text box*_______________________

Display This Question:

If Highest Educational Level = Further Education beyond undergraduate

Q27 Please specify the level(s) and topic(s) of your further education

__________________*single line free text box*_____________________

Q3 How would you describe your primary area of work or study?

__________________*single line free text box*_____________________

Q2 Rate your knowledge of the following topics

	I don't know what it is (1)	I don't know anything about it (2)	I know a little about it (3)	I have informally studied it (4)	I have/am studying it as part of a qualification (5)	I have a specialized qualification in it (6)	I am a specialist in this area (7)
Formal Logic (1)	○	○	○	○	○	○	○
Artificial Intelligence (2)	○	○	○	○	○	○	○
Constraints Programming (3)	○	○	○	○	○	○	○
Programming (4)	○	○	○	○	○	○	○
Complex project scheduling (5)	○	○	○	○	○	○	○
Physical Sciences (6)	○	○	○	○	○	○	○
Advanced Mathematics (7)	○	○	○	○	○	○	○

End of Block: About You

Start of Block: Background (in Sudoku?)

Q1 How often do you solve Sudoku?

○ I have never played Sudoku (1)

○ I play once a year or less (3)

○ I play on average every six months (9)

○ I play on average once a month (4)

○ I play on a weekly basis (5)

○ I play on a daily basis (6)

○ I play more than once a day (7)

Skip To: End of Block If How often do you solve Sudoku? = I have never played Sudoku

Display This Question:

If How often do you solve Sudoku? != I have never played Sudoku

Q6 How would you rate your Sudoku expertise?

○ Complete Novice (1)

○ Intermediate Player (2)

○ Master solver (3)

End of Block: Background (in Sudoku?)

Start of Block: What's fun?

Q7 How important is completing the Sudoku to your enjoyment?

○ I don't enjoy doing Sudoku if I can't finish it. (1)

○ It significantly reduces my enjoyment if I can't finish the Sudoku (2)

○ Being able to finish the Sudoku has no impact on my enjoyment (3)

○ I prefer Sudokus that are too hard for me to finish (4)

○ I only enjoy doing Sudoku if I can't finish it. (5)

Q8 Do you ever get frustrated when solving Sudoku? What causes that frustration?

○ Yes (1) ____

○ No (2)

Q25 Where do you play Sudoku? (select all that apply)

☐ Newspapers/Magazines (1)

☐ Books (2)

☐ Websites (3)

☐ Apps (4)

☐ Other (5) ____

Q19 Do you look up the solutions to Sudokus?

○ Never (1)

○ Only once I've solved the puzzle (2)

○ If I've been stuck for a long time (3)

○ Once it starts being frustrating I check the solution (4)

○ I always check the solution (5)

Q20 Have you investigated solving techniques?

○ Yes (1)

○ No (2)

Display This Question:

If Have you investigated solving techniques? = Yes

Q25
Where do you investigate solving techniques? (Please select all that apply)

☐ I read books (1)

☐ I look at websites (2)

☐ I discuss solving approaches with friends (3)

☐ I took a course that included Sudoku solving techniques (4)

☐ I took a course on how to solve a Sudoku (5)

☐ Other (6) ____

Q9 How long do you think a fun (ie. not too easy, not too hard) Sudoku should take to solve?

________________________*single line free text box*________________________

Q10 Do you play against other people?

○ Yes (1)

○ No (2)

Display This Question:

If Do you play against other people? = Yes

Q11 When you play against other people, how do you compete (e.g. for time, for amount of notation, and/or hardest difficulty solved)?

________________________*single line free text box*________________________

Q12 What do you find fun about Sudoku? (or what makes it lack fun?)

__________________________*multiline free text box*__________________________

End of Block: What's fun?

Start of Block: Other Puzzles

Q14 Do you play any other puzzle games?

○ Yes (1)

○ No (2)

Display This Question:

If Do you play any other puzzle games? = Yes

Q15 Can you list your favorite puzzle games (other than Sudoku)? For each if you can, specify any similarities to Sudoku.

______________________*multiline free text box*___________________________

Q16 Please describe what makes a good puzzle game for you?

______________________*multiline free text box*___________________________

Q17 Please describe what makes a bad puzzle game for you?

______________________*multiline free text box*___________________________

End of Block: Other Puzzles

Start of Block: Block 5

Q29
Thank you taking part in this survey and contributing to my research.

If you wish to take part in the second stage of this research (even if you have never played Sudoku) which will require you to be physically present in St Andrews, Scotland and has a reward of a £10 book voucher please leave your email below. Your email will only be used for this purpose.

Q28 Please enter your email if you wish to be contacted regarding the follow up study

____________________*single line free text box*_______________________

End of Block: Block 5

A.2 Sudoku Study Questionnaires

A.2.1 Pre Session 1 Questionnaire

Pre Study questionnaire

Partcipant ID:

Would you describe yourself as…

☐ Completely new to Sudoku
☐ A beginner
☐ An intermediate player
☐ An advanced player
☐ A very advanced player

How experienced are you at solving Sudoku?

☐ I have never played a Sudoku before
☐ I have played less than 10 Sudokus
☐ I have played approximately between 10 and 150 Sudokus
☐ I have played between 150 and 500 Sudokus
☐ I have played over 500 Sudokus

How recently have you solved a Sudoku?

☐ Today
☐ In the last week
☐ In the last fortnight
☐ In the last month
☐ In the last six months
☐ In the last year
☐ Over a year ago

Have you prepared for this study? If yes please briefly describe how.

☐ Yes
☐ No

A.2.2 Pre Session 2 Questionnaire

Pre Study questionnaire Session 2

Participant ID:

Since the first session of the study have you played…

- ☐ No Sudoku
- ☐ Less than 10 Sudoku
- ☐ Approximately 10-30 Sudoku
- ☐ Approximately 30-50 Sudoku
- ☐ Over 50 Sudoku

Have you researched Sudoku or Sudoku solving techniques since the first session?

- ☐ Yes
- ☐ No

If yes, could you briefly note where? (eg. Books, websites etc)

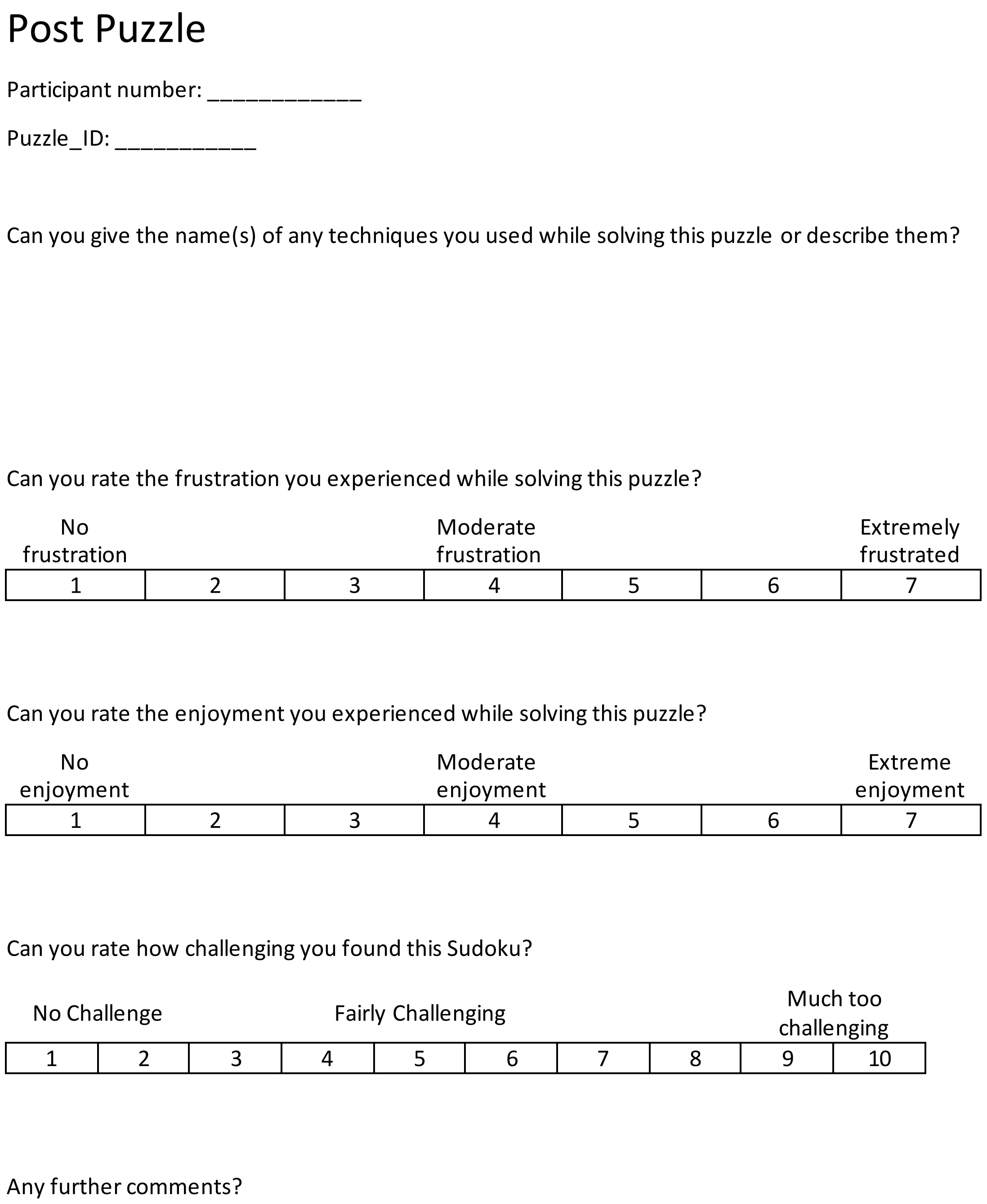
Would you describe yourself as…

- ☐ Completely new to Sudoku
- ☐ A beginner
- ☐ An intermediate player
- ☐ An advanced player
- ☐ A very advanced player

How recently have you solved a Sudoku?

- ☐ Today
- ☐ Yesterday
- ☐ 3 days ago
- ☐ More than 3 days ago

A.2.3 Post Puzzle Questionnaire



Post Puzzle

Participant number: ____________

Puzzle_ID: ___________

Can you give the name(s) of any techniques you used while solving this puzzle or describe them?

Can you rate the frustration you experienced while solving this puzzle?

No frustration			Moderate frustration			Extremely frustrated
1	2	3	4	5	6	7

Can you rate the enjoyment you experienced while solving this puzzle?

No enjoyment			Moderate enjoyment			Extreme enjoyment
1	2	3	4	5	6	7

Can you rate how challenging you found this Sudoku?

No Challenge				Fairly Challenging				Much too challenging	
1	2	3	4	5	6	7	8	9	10

Any further comments?

A.2.4 Interview Questions

The interview used the following guidelines for the semi-structured interviews conducted at the end of each session of the Sudoku solving study discussed in Chapter 3, Section 3.2.

Post Problem Solving Study interviews

These semi-structured interviews will be run with participants following the problem-solving study.

They will initially be asked for general thoughts on the puzzles and then focus will be drawn to the below topics. The question format may be subject to change but the topics will remain consistent.

Questions

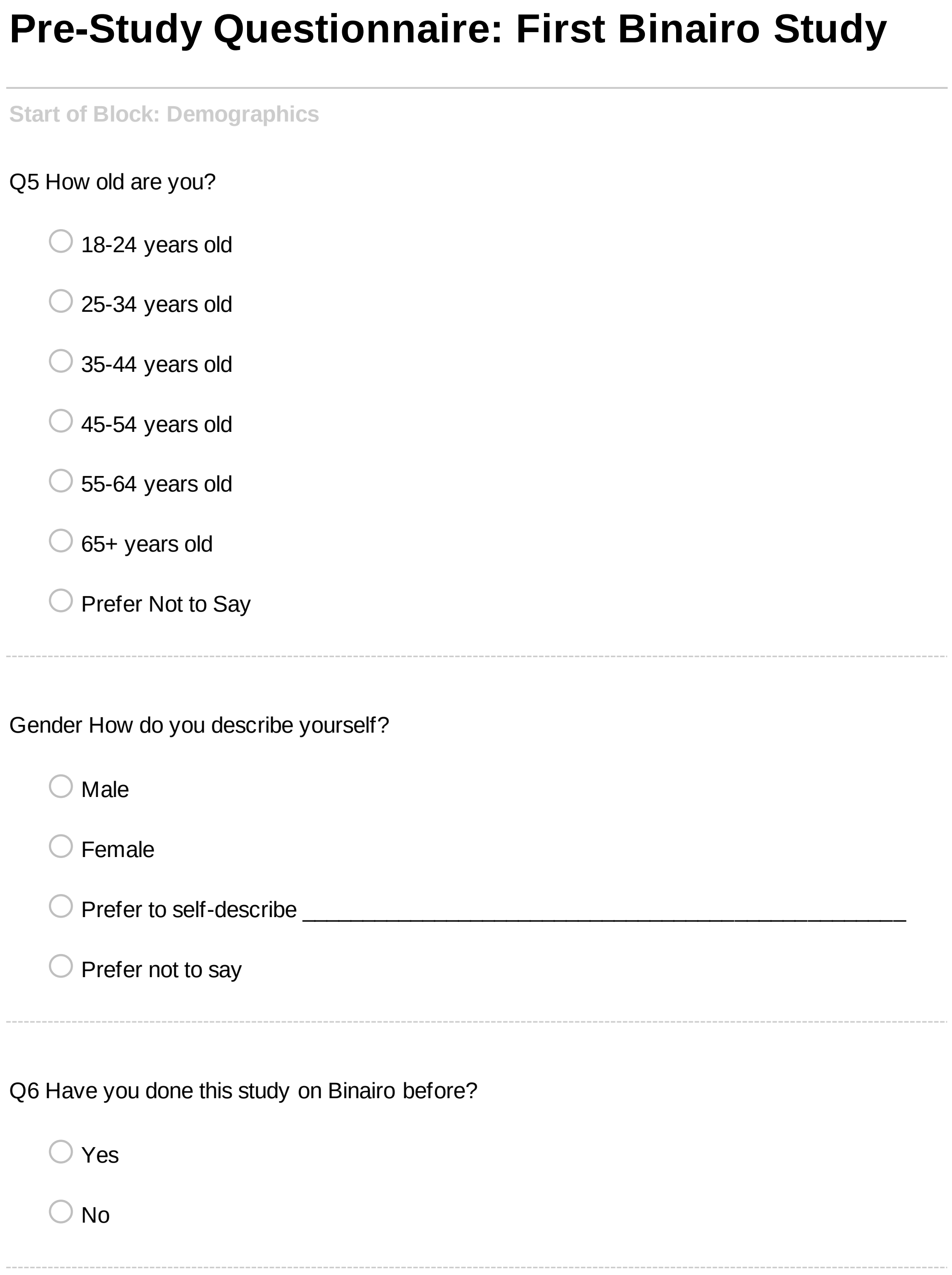
- Can you talk me through your approach to solving these problems?
- Did you enjoy any problems especially and why?
- Did you find any problems particularly frustrating and why?
- Can you explain how your notations works? (If participant used any form of notation)
- Did you come up with any approaches you haven’t used before?

A.3 Novel Hint System Study Questionnaires

This section contains the Pre-Study and Post-Puzzle Questionnaires from the studies discussed in Chapter 5.

A.3.1 Pre-Study Questionnaires

A.3.1.1 First Binairo Study Pre-Study Questionnaire



Pre-Study Questionnaire: First Binairo Study

Start of Block: Demographics

Q5 How old are you?

○ 18-24 years old

○ 25-34 years old

○ 35-44 years old

○ 45-54 years old

○ 55-64 years old

○ 65+ years old

○ Prefer Not to Say

Gender How do you describe yourself?

○ Male

○ Female

○ Prefer to self-describe ____

○ Prefer not to say

Q6 Have you done this study on Binairo before?

○ Yes

○ No

Display This Question:

If Have you done this study on Binairo before? = Yes

Q8 Which hint systems have you seen before?

☐ The coloured grid hint system

☐ The fill in next cell hint system

☐ I can't remember which I saw

End of Block: Demographics

Start of Block: Experience

Q7 Please describe your experience with traditional pen and paper puzzle games such as Sudoku, Starbattle, and Binairo

____

____

____

____

____

End of Block: Experience

A.3.1.2 Second Binairo Study Pre-Study Questionnaire

Pre-Study Questionnaire: Second Binairo Study

Start of Block: Demographics

Q5 How old are you?

- ○ 18-24 years old
- ○ 25-34 years old
- ○ 35-44 years old
- ○ 45-54 years old
- ○ 55-64 years old
- ○ 65+ years old
- ○ Prefer Not to Say

Gender How do you describe yourself?

- ○ Female
- ○ Male
- ○ Non-binary ____
- ○ Prefer to self-describe ____
- ○ Prefer not to say

Q6 Have you done this study on Binairo in 2023 before?

○ Yes

○ No

End of Block: Demographics

Start of Block: Experience

Q7 Please describe your experience with traditional pen and paper puzzle games such as Sudoku, Starbattle, and Binairo

____

____

____

____

____

End of Block: Experience

A.3.1.3 First Aquarium Study Pre-Study Questionnaire

Pre-Study Questionnaire: First Aquarium Study

Start of Block: Demographics

Q5 How old are you?

○ 18-24 years old

○ 25-34 years old

○ 35-44 years old

○ 45-54 years old

○ 55-64 years old

○ 65+ years old

○ Prefer Not to Say

Gender How do you describe yourself?

○ Female

○ Male

○ Non-binary ____

○ Prefer to self-describe ____

○ Prefer not to say

Q6 Have you done either of the previous studies on Binairo?

○ Yes

○ No

Q9 Have you done this study on Aquarium before?

○ Yes

○ No

End of Block: Demographics

Start of Block: Experience

Q7 Please describe your experience with traditional pen and paper puzzle games such as Sudoku, Starbattle, and Binairo

____

____

____

____

____

End of Block: Experience

A.3.1.4 First Aquarium Study Pre-Study Questionnaire

Pre-Study Questionnaire: Second Aquarium Study

Start of Block: Default Question Block

Q5 How old are you?

- ○ 18-24 years old (2)
- ○ 25-34 years old (3)
- ○ 35-44 years old (4)
- ○ 45-54 years old (5)
- ○ 55-64 years old (6)
- ○ 65+ years old (7)
- ○ Prefer Not to Say (8)

Gender How do you describe yourself?

- ○ Female (2)
- ○ Male (1)
- ○ Non-binary (7) ____
- ○ Prefer to self-describe (4) ____
- ○ Prefer not to say (5)

Q6 Have you done either of the previous studies on Binairo?

○ Yes (1)

○ No (2)

Q10 Have you done the previous study on Aquarium?

○ Yes (1)

○ No (2)

Q9 Have you done this study on Aquarium before?

○ Yes (1)

○ No (2)

End of Block: Default Question Block

Start of Block: Block 2

Q7 Please describe your experience with traditional pen and paper puzzle games such as Sudoku, Starbattle, and Binairo

____

____

____

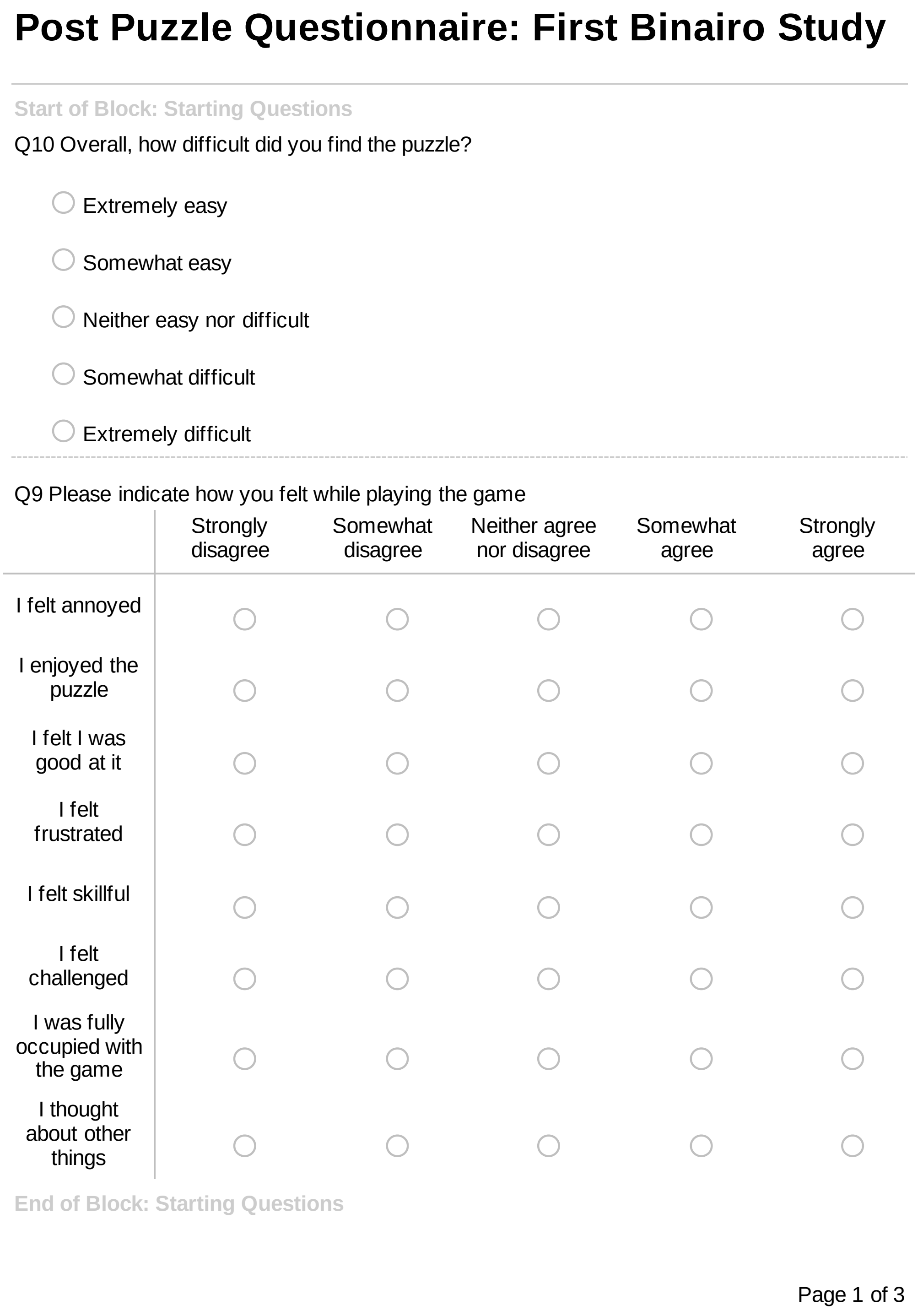
____

____

End of Block: Block 2

A.3.2 Post Puzzle Questionnaires

A.3.2.1 First Binairo Study Post Puzzle Questionnaire



Post Puzzle Questionnaire: First Binairo Study

Start of Block: Starting Questions

Q10 Overall, how difficult did you find the puzzle?

○ Extremely easy

○ Somewhat easy

○ Neither easy nor difficult

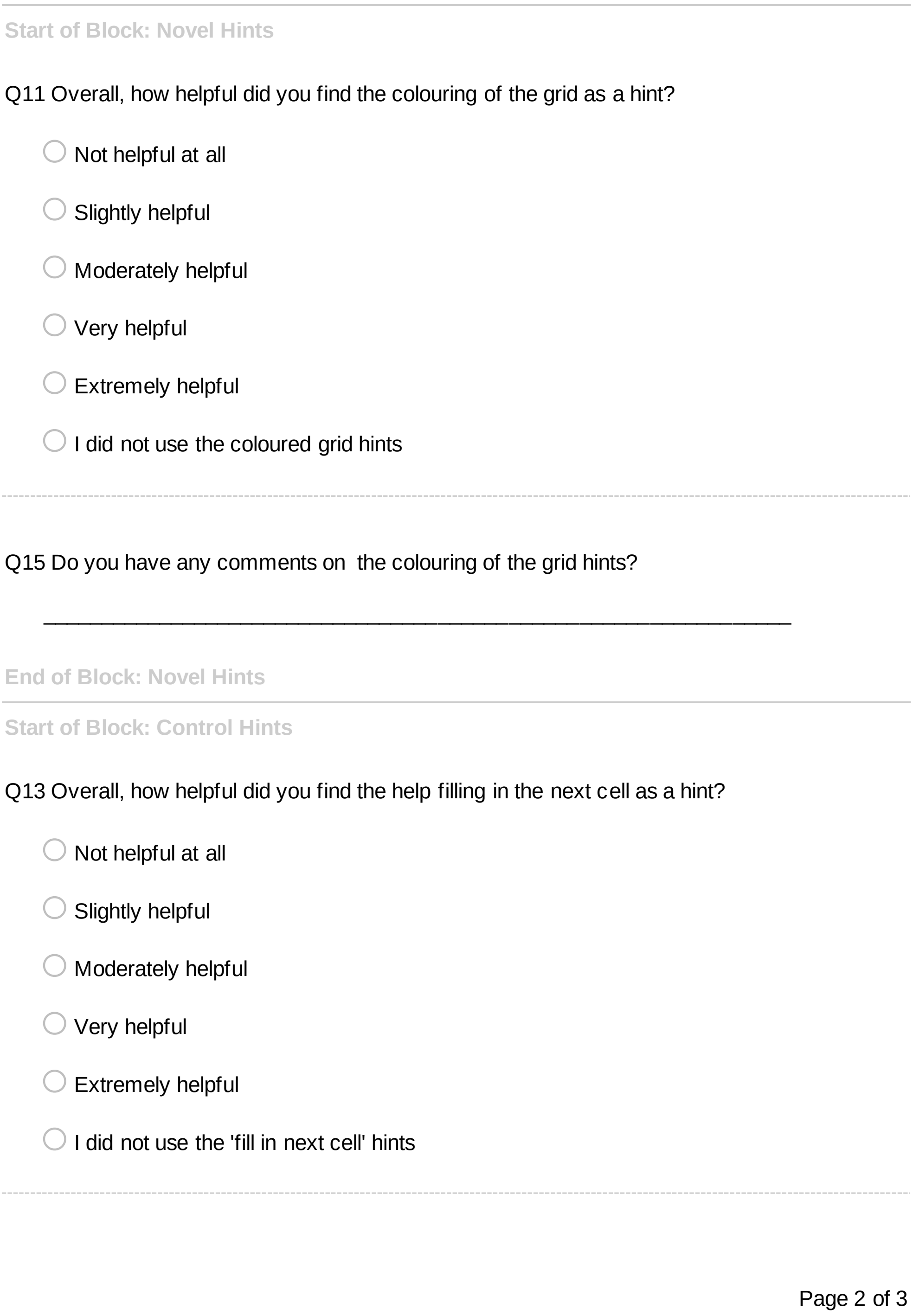
○ Somewhat difficult

○ Extremely difficult

Q9 Please indicate how you felt while playing the game

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
I felt annoyed	○	○	○	○	○
I enjoyed the puzzle	○	○	○	○	○
I felt I was good at it	○	○	○	○	○
I felt frustrated	○	○	○	○	○
I felt skillful	○	○	○	○	○
I felt challenged	○	○	○	○	○
I was fully occupied with the game	○	○	○	○	○
I thought about other things	○	○	○	○	○

End of Block: Starting Questions



Start of Block: Novel Hints

Q11 Overall, how helpful did you find the colouring of the grid as a hint?

○ Not helpful at all

○ Slightly helpful

○ Moderately helpful

○ Very helpful

○ Extremely helpful

○ I did not use the coloured grid hints

Q15 Do you have any comments on the colouring of the grid hints?

____

End of Block: Novel Hints

Start of Block: Control Hints

Q13 Overall, how helpful did you find the help filling in the next cell as a hint?

○ Not helpful at all

○ Slightly helpful

○ Moderately helpful

○ Very helpful

○ Extremely helpful

○ I did not use the 'fill in next cell' hints

Page 2 of 3

Q16 Do you have any comments on the help filling in the next square?

____

End of Block: Control Hints

Start of Block: Final Questions

Q14 Do you have any other comments?

____

____

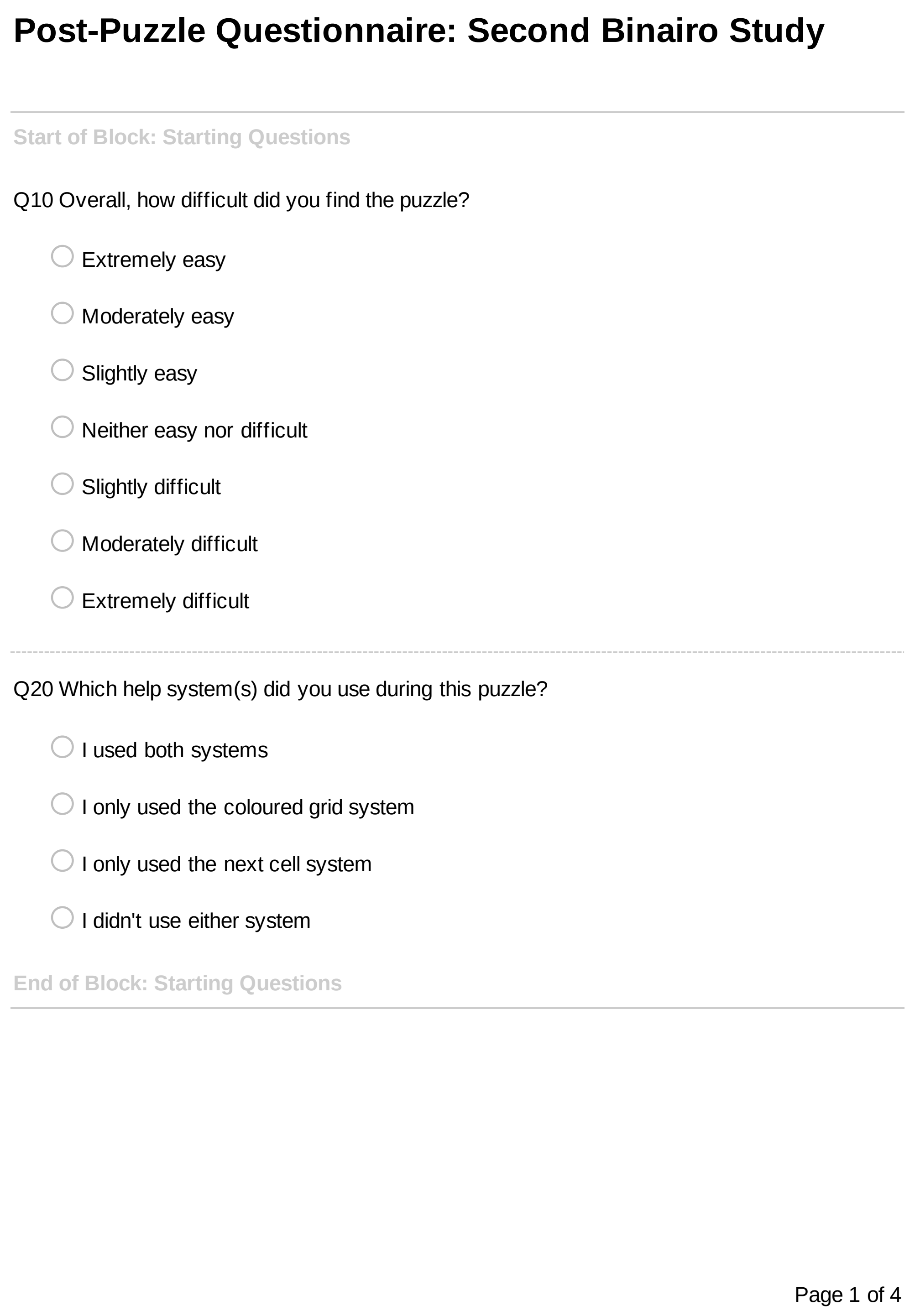
____

____

____

End of Block: Final Questions

A.3.2.2 Second Binairo Study Post Puzzle Questionnaire



Post-Puzzle Questionnaire: Second Binairo Study

Start of Block: Starting Questions

Q10 Overall, how difficult did you find the puzzle?

- ○ Extremely easy
- ○ Moderately easy
- ○ Slightly easy
- ○ Neither easy nor difficult
- ○ Slightly difficult
- ○ Moderately difficult
- ○ Extremely difficult

Q20 Which help system(s) did you use during this puzzle?

- ○ I used both systems
- ○ I only used the coloured grid system
- ○ I only used the next cell system
- ○ I didn't use either system

End of Block: Starting Questions

Start of Block: Preference Block

Q21 Which help system did you prefer during this puzzle?

○ I preferred the coloured grid hints

○ I preferred the fill in next cell system

○ I didn't have a preference

End of Block: Preference Block

Start of Block: Coloured Grid Hints

Q18 Please rate your agreement with the following statements regarding the **coloured grid hints**

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
I found it helpful	○	○	○	○	○	○	○
It gave me the type of help I wanted	○	○	○	○	○	○	○
I found these hints reduced my enjoyment	○	○	○	○	○	○	○
I found that it felt like cheating	○	○	○	○	○	○	○
I found it enhanced my experience	○	○	○	○	○	○	○

Q15 Do you have any comments on the colouring of the grid hints?

____

End of Block: Coloured Grid Hints

Page 2 of 4

Start of Block: Next Cell

Q19 Please rate your agreement with the following statements regarding the **next cell system**

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
I found it helpful	○	○	○	○	○	○	○
It gave me the type of help I wanted	○	○	○	○	○	○	○
I found these hints reduced my enjoyment	○	○	○	○	○	○	○
I found that it felt like cheating	○	○	○	○	○	○	○
I found it enhanced my experience	○	○	○	○	○	○	○

Q16 Do you have any comments on the help filling in the next square?

____

End of Block: Next Cell

Page 3 of 4

Start of Block: Final Questions

Q14 Do you have any other comments?

____

____

____

____

____

End of Block: Final Questions

A.3.2.3 First Aquarium Study Post Puzzle Questionnaire

Post-Puzzle Questionnaire: First Aquarium Study

Start of Block: Starting Questions

Q10 Overall, how difficult did you find the puzzle?

○ Extremely easy

○ Moderately easy

○ Slightly easy

○ Neither easy nor difficult

○ Slightly difficult

○ Moderately difficult

○ Extremely difficult

Q20 Which help system(s) did you use during this puzzle?

○ I used both below systems

○ I only used the coloured grid system (These are the hints provided by the Grid Hint button, not the error highlighting)

○ I only used the next cell system (via pressing the Next Cell button)

○ I didn't use either system

End of Block: Starting Questions

Start of Block: Preference Block

Q21 Which help system did you prefer during this puzzle?

○ I preferred the coloured grid hints (These are the hints provided by the Grid Hint button, not the error highlighting)

○ I preferred the fill in next cell system (via pressing the Next Cell button)

○ I didn't have a preference

End of Block: Preference Block

Start of Block: Coloured Grid Hints

Q18 Please rate your agreement with the following statements regarding the **coloured grid hints** (These are the hints provided by the Grid Hint button, not the error highlighting)

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
I found it helpful	○	○	○	○	○	○	○
It gave me the type of help I wanted	○	○	○	○	○	○	○
I found these hints reduced my enjoyment	○	○	○	○	○	○	○
I found that it felt like cheating	○	○	○	○	○	○	○
I found it enhanced my experience	○	○	○	○	○	○	○

Q15 Do you have any comments on the colouring of the grid hints?

____

End of Block: Coloured Grid Hints

Start of Block: Next Cell

Q19 Please rate your agreement with the following statements regarding the **next cell system**

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
I found it helpful	○	○	○	○	○	○	○
It gave me the type of help I wanted	○	○	○	○	○	○	○
I found these hints reduced my enjoyment	○	○	○	○	○	○	○
I found that it felt like cheating	○	○	○	○	○	○	○
I found it enhanced my experience	○	○	○	○	○	○	○

Q16 Do you have any comments on the help filling in the next square?

____

End of Block: Next Cell

Start of Block: Final Questions

Q14 Do you have any other comments?

____

____

____

____

____

End of Block: Final Questions

A.3.2.4 Second Aquarium Study Post Puzzle Questionnaire

Post-Puzzle Questionnaire: Second Aquarium Study

Start of Block: StartingQuestions

Q10 Overall, how difficult did you find the puzzle?

○ Extremely easy (36)

○ Moderately easy (37)

○ Slightly easy (38)

○ Neither easy nor difficult (39)

○ Slightly difficult (40)

○ Moderately difficult (41)

○ Extremely difficult (42)

Q20 Which help system(s) did you use during this puzzle?

○ I used both below systems (1)

○ I only used the coloured grid system (These are the hints provided by the Grid Hint button, not the error highlighting) (2)

○ I only used the next cell system (via pressing the Next Cell button) (3)

○ I didn't use either system (4)

End of Block: StartingQuestions

Start of Block: PreferenceBlock

Q21 Which help system did you prefer during this puzzle?

○ I preferred the coloured grid hints (These are the hints provided by the Grid Hint button, not the error highlighting) (1)

○ I preferred the fill in next cell system (via pressing the Next Cell button) (2)

○ I didn't have a preference (3)

End of Block: PreferenceBlock

Start of Block: ColouredGridHints

Q18 Please rate your agreement with the following statements regarding the **coloured grid hints** (These are the hints provided by the Grid Hint button, not the error highlighting)

	Strongly disagree (9)	Disagree (10)	Somewhat disagree (11)	Neither agree nor disagree (12)	Somewhat agree (13)	Agree (14)	Strongly agree (15)
I found it helpful (1)	○	○	○	○	○	○	○
It gave me the type of help I wanted (10)	○	○	○	○	○	○	○
I found these hints reduced my enjoyment (2)	○	○	○	○	○	○	○
I found that it felt like cheating (3)	○	○	○	○	○	○	○
I found it enhanced my experience (9)	○	○	○	○	○	○	○

Q15 Do you have any comments on the colouring of the grid hints?

____

End of Block: ColouredGridHints

Start of Block: NextCell

Q19 Please rate your agreement with the following statements regarding the **next cell system**

	Strongly disagree (9)	Disagree (10)	Somewhat disagree (11)	Neither agree nor disagree (12)	Somewhat agree (13)	Agree (14)	Strongly agree (15)
I found it helpful (1)	○	○	○	○	○	○	○
It gave me the type of help I wanted (10)	○	○	○	○	○	○	○
I found these hints reduced my enjoyment (2)	○	○	○	○	○	○	○
I found that it felt like cheating (3)	○	○	○	○	○	○	○
I found it enhanced my experience (9)	○	○	○	○	○	○	○

Q16 Do you have any comments on the help filling in the next square?

____

End of Block: NextCell

Start of Block: FinalQuestions

Q14 Do you have any other comments?

____

____

____

____

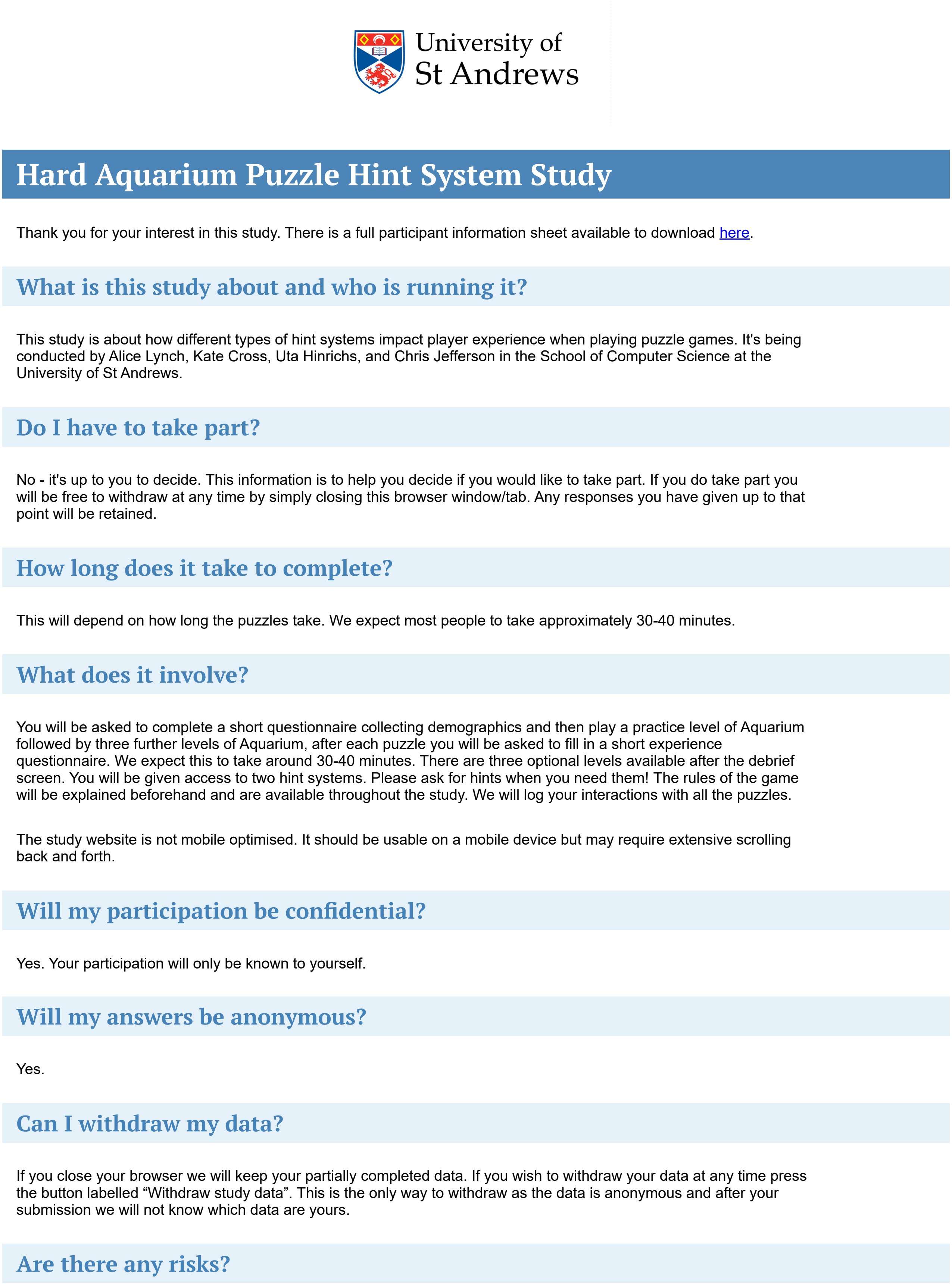
____

End of Block: FinalQuestions

APPENDIX B

INFORMATION AND CONSENT PAGE

This appendix contains an example of the information and consent webpage used for the studies discussed in Chapter 5



University of
St Andrews

Hard Aquarium Puzzle Hint System Study

Thank you for your interest in this study. There is a full participant information sheet available to download here.

What is this study about and who is running it?

This study is about how different types of hint systems impact player experience when playing puzzle games. It's being conducted by Alice Lynch, Kate Cross, Uta Hinrichs, and Chris Jefferson in the School of Computer Science at the University of St Andrews.

Do I have to take part?

No - it's up to you to decide. This information is to help you decide if you would like to take part. If you do take part you will be free to withdraw at any time by simply closing this browser window/tab. Any responses you have given up to that point will be retained.

How long does it take to complete?

This will depend on how long the puzzles take. We expect most people to take approximately 30-40 minutes.

What does it involve?

You will be asked to complete a short questionnaire collecting demographics and then play a practice level of Aquarium followed by three further levels of Aquarium, after each puzzle you will be asked to fill in a short experience questionnaire. We expect this to take around 30-40 minutes. There are three optional levels available after the debrief screen. You will be given access to two hint systems. Please ask for hints when you need them! The rules of the game will be explained beforehand and are available throughout the study. We will log your interactions with all the puzzles.

The study website is not mobile optimised. It should be usable on a mobile device but may require extensive scrolling back and forth.

Will my participation be confidential?

Yes. Your participation will only be known to yourself.

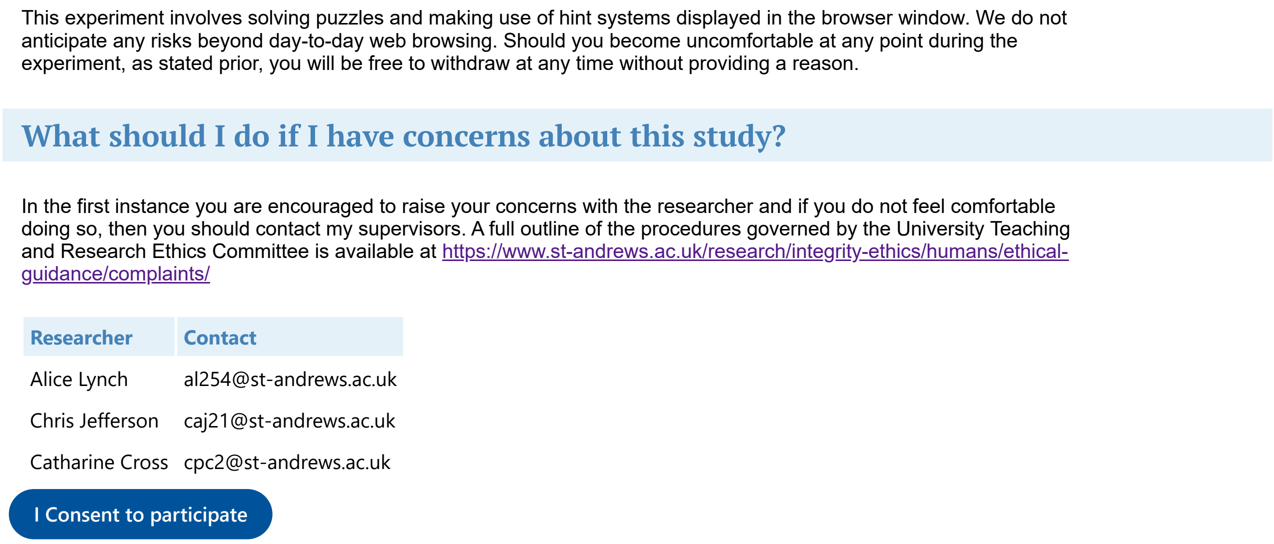
Will my answers be anonymous?

Yes.

Can I withdraw my data?

If you close your browser we will keep your partially completed data. If you wish to withdraw your data at any time press the button labelled “Withdraw study data”. This is the only way to withdraw as the data is anonymous and after your submission we will not know which data are yours.

Are there any risks?



This experiment involves solving puzzles and making use of hint systems displayed in the browser window. We do not anticipate any risks beyond day-to-day web browsing. Should you become uncomfortable at any point during the experiment, as stated prior, you will be free to withdraw at any time without providing a reason.

What should I do if I have concerns about this study?

In the first instance you are encouraged to raise your concerns with the researcher and if you do not feel comfortable doing so, then you should contact my supervisors. A full outline of the procedures governed by the University Teaching and Research Ethics Committee is available at https://www.st-andrews.ac.uk/research/integrity-ethics/humans/ethical-guidance/complaints/

Researcher	Contact
Alice Lynch	al254@st-andrews.ac.uk
Chris Jefferson	caj21@st-andrews.ac.uk
Catharine Cross	cpc2@st-andrews.ac.uk

APPENDIX C

LINEAR MIXED-EFFECTS MODEL OF THE QUESTIONNAIRE RESULTS OF THE FIRST BINAIRO STUDY

This appendix contains the linear mixed-effects model analysis results of the First Binairo Study Post Puzzle Questionnaire responses. First with puzzle as a fixed effect, and secondly with puzzle as a random effect.

The 5-point Likert scales were mapped to numeric values of 0 to 4. The mappings are shown in Tables C.1 to C.3.

Text	**Numeric Value**
Strongly disagree	0
Somewhat disagree	1
Neither agree nor disagree	2
Somewhat agree	3
Strongly agree	4

Table C.1: Mapping of 5-point Likert scale showing agreement to numeric values.

Text	**Numeric Value**
Not helpful at all	0
Slightly helpful	1
Neither agree nor disagree	2
Moderately helpful	3
Very helpful	4

Table C.2: Mapping of 5-point Likert scale showing perceived helpfulness to numeric values.

Text	**Numeric Value**
Extremely easy	0
Somewhat easy	1
Neither easy nor difficult	2
Somewhat difficult	3
Extremely difficult	4

Table C.3: Mapping of 5-point Likert scale showing agreement to numeric values.

C.1 Linear Mixed-Effects Model Analysis with puzzle as a fixed effect

The tables in this section shows the linear mixed-effects model for the experience assessment statements (asking participants to rate their agreement), using the smf package in python, (statsmodel.formula.api), with condition and puzzle as fixed effects and a random effect of participant.

The intercept row contains the values for the reference values, the p-value is significant in all the tables, but this only means that the average was not 0.

The β column (for all rows apart from the intercept) shows the deviation of the mean of the data for the parameter (indicated in the first column) from the mean of the reference. The Std. Err column shows the standard error for that parameter. The final column, p, shows the p-value, for these experiments a p-value of <=0.05 is considered significant.

Tables C.4 to C.11 asked the participants to rate their agreement with the statements, using the Likert scale shown in Table C.1. A negative β value indicates that participants agreed less with the statement heading the table for a given parameter, than they did when rating the statement for the reference condition. The reference parameters for all tables in this section was Binairo 206 (the practice puzzle) with the traditional (next cell) hint system. A more positive β value indicates that participants expressed greater agreement with the statement than they did when rating the reference parameters. Therefore, in Table C.4 there are significant differences (indicated by p-values<=0.05) between the reference puzzle parameter (Binairo 206) and all other puzzles, there is not a significant difference between the reference condition parameter, Next cell, and all other conditions. All the β values for the puzzle parameters are positive, indicating increased agreement with the statement "I felt frustrated" for every puzzle, when compared to the ratings of Binairo 206.

C.2 Linear Mixed-Effects Model Analysis with puzzle as a random effect

The tables in this section shows the linear mixed-effects model for the experience assessment statements (asking participants to rate their agreement), using the smf

"I felt frustrated"

Parameter	β	Std. Err.	p
Intercept	0.758	0.120	< 0.001
Puzzle = Binairo 201	1.068	0.111	<0.001
Puzzle = Binairo 203	0.481	0.102	<0.001
Puzzle = Binairo 205	0.814	0.112	<0.001
Puzzle = Binairo 208	0.601	0.105	<0.001
Puzzle = Binairo 212	0.273	0.107	0.011
Condition = Both Systems	0.064	0.146	0.661
Condition = Grid Hint	-0.156	0.148	0.292

Table C.4: Linear Mixed-Effects Model (using the next cell condition and puzzle Binairo 206 as the reference) results showing that all puzzles are considered more frustrating than puzzle 206, and that there is no effect of condition on frustration levels

"I felt annoyed"

Parameter	β	Std. Err.	p
Intercept	0.723	0.129	< 0.001
Puzzle = Binairo 201	0.917	0.106	<0.001
Puzzle = Binairo 203	0.211	0.098	0.031
Puzzle = Binairo 205	0.738	0.108	<0.001
Puzzle = Binairo 208	0.611	0.101	<0.001
Puzzle = Binairo 212	0.352	0.102	0.001
Condition = Both Systems	0.030	0.162	0.854
Condition = Grid Hint	-0.140	0.163	0.391

Table C.5: Linear Mixed-Effects Model (using the next cell condition and puzzle Binairo 206 as the reference) results showing that participants felt more annoyed with all puzzles than they did with puzzle 206, and that there is no effect of condition on participant annoyance.

"I enjoyed the puzzle"

Parameter	β	Std. Err.	p
Intercept	3.214	0.103	< 0.001
Puzzle = Binairo 201	-0.464	0.083	<0.001
Puzzle = Binairo 203	0.019	0.076	0.808
Puzzle = Binairo 205	-0.362	0.084	<0.001
Puzzle = Binairo 208	-0.310	0.078	<0.001
Puzzle = Binairo 212	-0.256	0.080	0.001
Condition = Both Systems	-0.153	0.130	0.238
Condition = Grid Hint	-0.084	0.131	0.523

Table C.6: Linear Mixed-Effects Model (using the next cell condition and puzzle Binairo 206 as the reference) results showing that participants felt less enjoyment with all puzzles, except Binairo 203, than they did with puzzle 206. There is no effect of condition on participant enjoyment.

"I felt I was good at it"

Parameter	β	Std. Err.	p
Intercept	2.882	0.103	< 0.001
Puzzle = Binairo 201	-0.656	0.088	<0.001
Puzzle = Binairo 203	-0.286	0.081	<0.001
Puzzle = Binairo 205	-0.466	0.089	<0.001
Puzzle = Binairo 208	-0.453	0.084	<0.001
Puzzle = Binairo 212	0.170	0.085	0.046
Condition = Both Systems	0.024	0.129	0.851
Condition = Grid Hint	0.062	0.130	0.634

Table C.7: Linear Mixed-Effects Model (using the next cell condition and puzzle Binairo 206 as the reference) results showing that participants felt less good at puzzles Binairo 201, Binairo 203, Binairo 205, and Binairo 208 than they did with puzzle 206. They felt slightly more good at Binairo 212 than they did with Binairo 206. There is no effect of condition on how good players felt they were at the puzzles.

"I felt skillful"

Parameter	β	Std. Err.	p
Intercept	2.399	0.108	< 0.001
Puzzle = Binairo 201	-0.195	0.094	0.039
Puzzle = Binairo 203	0.056	0.087	0.516
Puzzle = Binairo 205	-0.047	0.095	0.618
Puzzle = Binairo 208	-0.101	0.089	0.260
Puzzle = Binairo 212	0.279	0.091	0.002
Condition = Both Systems	-0.113	0.133	0.395
Condition = Grid Hint	-0.107	0.134	0.426

Table C.8: Linear Mixed-Effects Model (using the next cell condition and puzzle Binairo 206 as the reference) results showing that participants felt less skillful at Binairo 201, Binairo 203, Binairo 205, and Binairo 208 than they did with puzzle 206. They felt slightly more good at Binairo 212 than they did with Binairo 206. There is no effect of condition on how skillful players felt they were at the puzzles.

"I felt challenged"

Parameter	β	Std. Err.	p
Intercept	2.084	0.104	<0.001
Puzzle = Binairo 201	0.794	0.102	<0.001
Puzzle = Binairo 203	0.743	0.094	<0.001
Puzzle = Binairo 205	0.794	0.103	<0.001
Puzzle = Binairo 208	0.515	0.097	<0.001
Puzzle = Binairo 212	0.127	0.098	0.197
Condition = Both Systems	0.106	0.123	0.389
Condition = Grid Hint	0.088	0.124	0.479

Table C.9: Linear Mixed-Effects Model (using the next cell condition and puzzle Binairo 206 as the reference) results showing that participants felt more challenged with all puzzles, except Binairo 212, than they did with puzzle 206. There is no effect of condition on how challenged players felt.

"I was fully occupied with the game"

Parameter	β	Std. Err.	p
Intercept	2.805	0.122	<0.001
Puzzle = Binairo 201	-0.189	0.089	0.034
Puzzle = Binairo 203	0.120	0.082	0.143
Puzzle = Binairo 205	-0.274	0.090	0.002
Puzzle = Binairo 208	-0.099	0.084	0.239
Puzzle = Binairo 212	0.169	0.086	0.049
Condition = Both Systems	0.185	0.158	0.241
Condition = Grid Hint	0.054	0.159	0.735

Table C.10: Linear Mixed-Effects Model (using the next cell condition and puzzle Binairo 206 as the reference) results showing that participants felt slightly less fully occupied while playing Binairo 201 and Binairo 205 than they did with Binairo 206. They felt slightly more occupied with Binairo 212 than they did with Binairo 206. There is no effect of condition on how fully occupied players felt.

"I thought about other things"

Parameter	β	Std. Err.	p
Intercept	1.317	0.142	<0.001
Puzzle = Binairo 201	0.246	0.102	0.016
Puzzle = Binairo 203	-0.095	0.094	0.312
Puzzle = Binairo 205	0.306	0.103	0.003
Puzzle = Binairo 208	0.018	0.097	0.853
Puzzle = Binairo 212	-0.251	0.098	0.011
Condition = Both Systems	0.015	0.183	0.933
Condition = Grid Hint	-0.054	0.185	0.768

Table C.11: Linear Mixed-Effects Model (using the next cell condition and puzzle Binairo 206 as the reference) results showing that participants felt slightly more distracted while playing Binairo 201 and Binairo 205 than they did with Binairo 206. They felt slightly more occupied with Binairo 212 than they did with Binairo 206. There is no effect of condition on how fully occupied the players felt.

How difficult did you find the puzzle?

Parameter	β	Std. Err.	p
Intercept	1.033	0.107	<0.001
Puzzle = Binairo 201	1.597	0.105	<0.001
Puzzle = Binairo 203	1.071	0.097	<0.001
Puzzle = Binairo 205	1.486	0.106	<0.001
Puzzle = Binairo 208	1.032	0.100	<0.001
Puzzle = Binairo 212	0.431	0.101	0.011
Condition = Both Systems	0.016	0.126	0.900
Condition = Grid Hint	0.085	0.128	0.509

Table C.12: Linear Mixed-Effects Model (using the next cell condition and puzzle Binairo 206 as the reference) results showing that participants assessed all the puzzles as more challenging than Binairo 206. There is no effect of condition on how difficult participants perceived the puzzles to be.

package in python, (statsmodel.formula.api), with condition as a fixed effect and random effects of participant and puzzle.

The intercept row contains the values for the reference values, the t-value is significant in this row in all the tables, but this only means that the average for the reference category was not 0.

The β column (for all rows apart from the intercept) shows the deviation of the mean of the data for the parameter (indicated in the first column) from the mean of the reference. The Std. Err column shows the standard error for that parameter. The final column, p, shows the t-value, for these experiments a t-value less than -1.96 or greater than 1.96 are considered significant and are highlighted.

Tables C.14 to C.21 asked the participants to rate their agreement with the statements, using the Likert scale shown in Table C.1. A negative β value indicates that participants agreed less with the statement for a given parameter, than they did when rating the statement for the reference condition. The reference category for all tables in this section the traditional (next cell) hint system. A more positive β value indicates that participants expressed greater agreement with the statement than they did when rating the reference parameters.

How difficult did you find the puzzle?

Parameter	β	Std. Err.	t-value
Intercept	2.097	0.273	7.659
Condition = Both Systems	-0.0546	0.149	-0.363
Condition = Grid Hint	0.078	0.150	0.518

Table C.13: Linear Mixed-Effects Model (using the next cell condition as the reference) with puzzle as a random effect. The results show that there is no effect of condition on how difficult participants perceived the puzzles to be.

I felt annoyed

Parameter	β	Std. Err.	t-value
Intercept	1.3591	0.203	6.703
Condition = Both Systems	-0.0351	0.196	-0.179
Condition = Grid Hint	-0.0984	0.197	-0.499

Table C.14: Linear Mixed-Effects Model (using the next cell condition as the reference) with puzzle as a random effect. The results show that there is no effect of condition on how annoyed participants were.

I felt challenged

Parameter	β	Std. Err.	t-value
Intercept	2.6841	0.174	15.408
Condition = Both Systems	0.0740	0.135	0.548
Condition = Grid Hint	0.1073	0.136	0.789

Table C.15: Linear Mixed-Effects Model (using the next cell condition as the reference) with puzzle as a random effect. The results show that there is no effect of condition on how challenged participants were.

I thought about other things

Parameter	β	Std. Err.	t-value
Intercept	1.5065	0.171	8.815
Condition = Both Systems	-0.0143	0.207	-0.069
Condition = Grid Hint	-0.0222	0.208	-0.107

Table C.16: Linear Mixed-Effects Model (using the next cell condition as the reference) with puzzle as a random effect. The results show that there is no effect of condition on how much participants thought about other things.

I enjoyed the puzzle

Parameter	β	Std. Err.	t-value
Intercept	3.0803	0.119	25.802
Condition = Both Systems	-0.1806	0.138	-1.309
Condition = Grid Hint	-0.0675	0.139	-0.487

Table C.17: Linear Mixed-Effects Model (using the next cell condition as the reference) with puzzle as a random effect. The results show that there is no effect of condition on how much participants enjoyed solving the puzzle.

I felt frustrated

Parameter	β	Std. Err.	t-value
Intercept	1.4563	0.207	7.050
Condition = Both Systems	0.0222	0.183	0.121
Condition = Grid Hint	-0.1207	0.184	-0.655

Table C.18: Linear Mixed-Effects Model (using the next cell condition as the reference) with puzzle as a random effect. The results show that there is no effect of condition on how frustrated participants were.

I felt I was good at it

Parameter	β	Std. Err.	t-value
Intercept	2.7084	0.151	17.876
Condition = Both Systems	-0.0160	0.142	-0.113
Condition = Grid Hint	-0.0490	0.143	-0.343

Table C.19: Linear Mixed-Effects Model (using the next cell condition as the reference) with puzzle as a random effect. The results show that there is no effect of condition on how good at the puzzle participants felt they were.

I was fully occupied with the game

Parameter	β	Std. Err.	t-value
Intercept	2.8634	0.129	22.135
Condition = Both Systems	0.1504	0.165	0.914
Condition = Grid Hint	0.0521	0.165	0.316

Table C.20: Linear Mixed-Effects Model (using the next cell condition as the reference) with puzzle as a random effect. The results show that there is no effect of condition on how occupied participants were with the game.

I felt skillful

Parameter	β	Std. Err.	t-value
Intercept	2.5160	0.117	21.465
Condition = Both Systems	-0.1438	0.148	-0.969
Condition = Grid Hint	-0.0962	0.149	-0.644

Table C.21: Linear Mixed-Effects Model (using the next cell condition as the reference) with puzzle as a random effect. The results show that there is no effect of condition on how skillful participants felt they were.

APPENDIX D

LINEAR MIXED-EFFECTS MODEL RESULTS FOR SECOND BINAIRO AND AQUARIUM STUDIES

This section shows the full Linear Mixed-Effects Model results run with the R package lme4 [10] on the results of the second Binairo study, Table D.1, and the second Aquarium study, Table D.2.

Statement	**Parameter**	β	**Std. Err.**	**t-value**
I found it enhanced my experience	Intercept	3.3356	0.237	14.070
	Condition = Grid Hint	0.5084	0.246	2.068
I found it helpful	Intercept	4.7717	0.321	14.85
	Condition = Grid Hint	-0.2714	0.305	-0.89
I found it felt like cheating	Intercept	3.3884	0.337	10.071
	Condition = Grid Hint	-0.869	0.300	-2.901
I found these hints reduced my enjoyment	Intercept	3.2546	0.263	12.370
	Condition = Grid Hint	-1.360	0.285	-4.777
It gave me the type of help I wanted	Intercept	3.7183	0.296	12.550
	Condition = Grid Hint	0.6180	0.328	1.883

Table D.1: The Linear Mixed-Effects Model (using the next cell condition as the reference) with hint system as a fixed effect and puzzle and participant as random effects of the Second Binairo Study.

Statement	**Parameter**	β	**Std. Err.**	**t-value**
I found it enhanced my experience	Intercept	3.0298	0.143	21.255
	Condition = Grid Hint	0.5690	0.154	3.687
I found it helpful	Intercept	4.1265	0.151	27.331
	Condition = Grid Hint	-0.0310	0.176	-0.176
I found it felt like cheating	Intercept	3.1280	0.171	18.30
	Condition = Grid Hint	-1.2685	0.171	-7.41
I found these hints reduced my enjoyment	Intercept	2.7296	0.144	18.921
	Condition = Grid Hint	-0.9814	0.153	-6.408
It gave me the type of help I wanted	Intercept	3.7321	0.160	23.366
	Condition = Grid Hint	0.2568	0.181	1.421

Table D.2: The Linear Mixed-Effects Model (using the next cell condition as the reference) with hint system as a fixed effect and puzzle and participant as random effects of the Second Aquarium Study.

APPENDIX E

ETHICS APPROVAL LETTERS

This appendix contains the letters of ethical approval for the studies conducted in this thesis. First, approval for the study described in Chapter 3, followed by an approval for an amendment adding two researchers. Secondly approval for the series of studies, described in Chapter 5, assessing the efficacy of hint systems.

University of St Andrews | FOUNDED 1413

University Teaching and Research Ethics Committee

19 June 2023

Dear Alice,

Thank you for submitting your ethical application, which was considered by the School of Computer Science Ethics Committee on Wednesday 16th January, where the following documents were reviewed:

1. Ethical Application Form
2. Participant Information Sheet
3. Consent Form
4. Debriefing Form
5. Interview Questions

The School of Computer Science Ethics Committee has been delegated to act on behalf of the University Teaching and Research Ethics Committee (UTREC) and has granted this application ethical approval. The particulars relating to the approved project are as follows -

Approval Code:	CS14059	**Approved on:**	05.02.19	**Approval Expiry:**	05.02.2024
Project Title:	Studying Strategies and Preferences when Solving Puzzles				
Researcher(s):	Alice Lynch				
Supervisor(s):	Uta Hinrichs and Chris Jefferson				

Approval is awarded for five years. Projects which have not commenced within two years of approval must be re-submitted for review by your School Ethics Committee. If you are unable to complete your research within the five year approval period, you are required to write to your School Ethics Committee Convener to request a discretionary extension of no greater than 6 months or to re-apply if directed to do so, and you should inform your School Ethics Committee when your project reaches completion.

If you make any changes to the project outlined in your approved ethical application form, you should inform your supervisor and seek advice on the ethical implications of those changes from the School Ethics Convener who may advise you to complete and submit an ethical amendment form for review.

Any adverse incident which occurs during the course of conducting your research must be reported immediately to the School Ethics Committee who will advise you on the appropriate action to be taken.

Approval is given on the understanding that you conduct your research as outlined in your application and in compliance with UTREC Guidelines and Policies (http://www.st-andrews.ac.uk/utrec/guidelinespolicies/). You are also advised to ensure that you procure and handle your research data within the provisions of the Data Provision Act 1998 and in accordance with any conditions of funding incumbent upon you.

Yours sincerely

School Ethics Committee Administrator

ethics-cs@st-andrews.ac.uk

The University of St Andrews is a charity registered in Scotland: No SC013532

School of Computer Science Ethics Committee

03 September 2021

Dear Alice,

Thank you for submitting your ethical amendment application.

The School of Computer Science Ethics Committee has approved this ethical amendment application:

Original Approval Code:	CS14059	**Original Approval Date:**	05.02.2019
Amendment Approval Date:	03.09.2021	**Approval Expiry Date:**	05.02.2024
Project Title:	Studying Strategies and Preferences when Solving Puzzles		
Researcher(s):	Alice Lynch	**Supervisor/PI:**	Uta Hinrichs and Chris Jefferson
School/Unit:	Computer Science		

[Delete if not applicable] The following supporting documents are also acknowledged and approved:

1. Ethical Amendment Application Form

This approval does not extend the originally granted approval period. If you require an extension to the approval period, you can write to your School Ethics Committee who may grant a discretionary extension of no greater than 6 months. For longer extensions, or for any further changes, you must submit an additional ethical amendment application. For all extensions, you should inform the School Ethics Committee when your study is complete.

You must report any serious adverse events, or significant changes not covered by this approval, related to this study immediately to the School Ethics Committee.

Approval is given on the following conditions:

- that you conduct your research in line with:
    - the details provided in your ethical amendment application (and the original ethical application where still relevant)
    - the University's Principles of Good Research Conduct
    - the conditions of any funding associated with your work
- that you obtain all applicable additional documents (see the 'additional documents' webpage for guidance) before research commences.

You should retain this approval letter with your study paperwork.

Yours sincerely,

SEC Administrator

School of Computer Science Ethics Committee
Dr Juan Ye/Convenor, Jack Cole Building, North Haugh, St Andrews, Fife, KY16 9SX
T: 01334 463252 E: ethics-cs@st-andrews.ac.uk

School of Computer Science Ethics Committee

25 January 2022

Dear Alice,

Thank you for submitting your ethical application which was considered at the School Ethics Committee on Wednesday 8th December 2021.

The School of Computer Science Ethics Committee, acting on behalf of the University Teaching and Research Ethics Committee (UTREC), has approved this application:

Approval Code:	CS15921	**Approved on:**	25.01.2022	**Approval Expiry:**	25.01.2027
Project Title:	The impact of feedback approaches on player experience in grid based progressive logic puzzles				
Researcher(s):	Alice Lynch, Chris Jefferson and Catherine Cross				
Supervisor(s):	Chris Jefferson				

The following supporting documents are also acknowledged and approved:

1. Application Form
2. Participant Information Sheet
3. Participant Consent Form
4. Participant Debrief Form
5. Advertisement
6. Questionnaire

Approval is awarded for 5 years, see the approval expiry data above.

If your project has not commenced within 2 years of approval, you must submit a new and updated ethical application to your School Ethics Committee.

If you are unable to complete your research by the approval expiry date you must request an extension to the approval period. You can write to your School Ethics Committee who may grant a discretionary extension of up to 6 months. For longer extensions, or for any other changes, you must submit an ethical amendment application.

You must report any serious adverse events, or significant changes not covered by this approval, related to this study immediately to the School Ethics Committee.

Approval is given on the following conditions:

- that you conduct your research in line with:
    - the details provided in your ethical application
    - the University's Principles of Good Research Conduct
    - the conditions of any funding associated with your work
- that you obtain all applicable additional documents (see the 'additional documents' webpage for guidance) before research commences.

You should retain this approval letter with your study paperwork.

School of Computer Science Ethics Committee
Dr Juan Ye/Convenor, Jack Cole Building, North Haugh, St Andrews, Fife, KY16 9SX
Telephone: 01334 463252 Email: ethics-cs@st-andrews.ac.uk

Yours sincerely,

SEC Administrator